Title: POLYNUCLEOTIDES ENCODING PHENYLPROPANOID AND FLAVONOID BIOSYNTHETIC PATHWAY ENZYMES IN COFFEE

Abstract: Polynucleotides encoding polypeptides that comprise the biosynthetic pathway for phenylpropanoids and flavonoids in the coffee plant are disclosed. Also disclosed are methods for using these polynucleotides and polypeptides for the manipulation of flavor, aroma, and other features of coffee beans, as well as the manipulation resistance to pathogen, herbivore, and insect attack in the coffee plant.
POLYNUCLEOTIDES ENCODING PHENYLPROPANOID AND FLAVONOID BIOSYNTHETIC PATHWAY ENZYMES IN COFFEE

FIELD OF THE INVENTION

The present invention relates to the field of agricultural biotechnology. In particular, the invention features polynucleotides from coffee plants that encode enzymes responsible for flavonoid synthesis, as well as methods for using these polynucleotides and polypeptides for gene regulation and manipulation of flavor, aroma and other features of coffee beans.

BACKGROUND OF THE INVENTION

Various publications, including patents, published applications and scholarly articles, are cited throughout the specification. Each of these publications is incorporated by reference herein, in its entirety. Citations not fully set forth within the specification may be found at the end of the specification.

Coffee aroma and flavor are key components in consumer preference for coffee varieties and brands. The characteristic aroma and flavor of coffee stems from a complex series of chemical reactions involving flavor precursors (Maillard reactions) that occur during the roasting of the bean. Flavor precursors include chemical compounds and biomolecules present in the green coffee bean. To date, over 800 chemicals and biomolecules have been identified as contributing to coffee flavor and aroma. (Flament, L., 2002 “Coffee Flavor Chemistry” J.Wiley U.K.). Because coffee consumers are becoming increasingly sophisticated, it is desirable to produce coffee with improved aroma and flavor in order to meet consumer preferences. Both aroma and flavor may be artificially imparted into coffee products through chemical means. See, for example, U.S. Pat. No. 4,072,761 (aroma) and U.S. Pat. No. 3,962,321 (flavor).

However, to date, there is little information concerning the influence of natural coffee grain components such as polysaccharides, proteins, pigments, and lipids, on coffee aroma and flavor. One approach is to select varieties from the existing germplasm that have superior flavor characteristics. A disadvantage to this approach is that, frequently, the highest quality varieties also possess significant negative agronomics traits, such as poor yield and low resistance to diseases and environmental stresses. It is also possible to select new varieties from breeding trials in which varieties with different industrial and agronomic traits are crossed and their progeny are screened for both high quality and good agronomic performance. However, this
latter approach is very time consuming, with one crossing experiment and selection over three growing seasons taking a minimum of 7-8 years. Thus, an alternative approach to enhancing coffee quality would be to use techniques of molecular biology to enhance those elements responsible for the flavor and aroma that are naturally found in the coffee bean, or to add aroma and flavor-enhancing elements that do not naturally occur in coffee beans. Genetic engineering is particularly suited to achieve these ends. For example, coffee proteins from different coffee species may be swapped. In the alternative, the expression of genes encoding naturally occurring coffee proteins that positively contribute to coffee flavor may be enhanced. Conversely, the expression of genes encoding naturally occurring coffee proteins that negatively contribute to coffee flavor may be suppressed.

Coffees from different varieties and origins exhibit significant flavor and aroma quality variations when the green grain samples are roasted and processed in the same manner. The quality differences are a manifestation of chemical and physical variations within the grain samples that result mainly from differences in growing and processing conditions, and also from differences in the genetic background of both the maternal plant and the grain. At the level of chemical composition, at least part of the flavor quality can be associated with variations in the levels of small metabolites, such as sugars, acids, phenolics, and caffeine found associated with grain from different varieties. It is accepted that there are other less well characterized flavor and flavor-precursor molecules. In addition, it is likely that structural variations within the grain also contribute to differences in coffee quality. One approach to finding new components in the coffee grain linked to coffee quality is to study the genes and proteins differentially expressed during the maturation of grain samples in different varieties that possess different quality characteristics. Similarly, genes and proteins that participate in the biosynthesis of flavor and flavor-precursor molecules may be studied.

The flavonoids form a large group of ubiquitous plant secondary metabolites, with over 4000 molecules of this class identified to date (Bovy et al. (2002) and Yilmaz et al. (2004)). Flavonoids are derived from the condensation of p-coumaroyl-CoA, which is synthesized from phenylalanine via the early phenylpropanoid pathway, and three molecules of malonyl-CoA, which is generated by the TCA cycle. (Dixon et al. (1999); Winkel-Shirley (2002); and Dixon (2005)). The various flavonoid metabolites contribute in different ways to the normal functioning and survival of the plant. For example, the red, blue and purple anthocyanin pigments found in flowers participate in plant reproduction by their involvement in attracting insects for pollination. (Winkel-Shirley (2002)). Other flavonoids, the brown proanthocyanidin
pigments (also termed condensed tannins), are believed to have antimicrobial properties and thus have been proposed to contribute to microbial resistance. (Sivakumaran et al. (2004); and Cos et al. (2005)). Yet another group of flavonoids, the isoflavones which are synthesised primarily in leguminous plants, are involved in plant-microbe interactions. (Dixon (2005)). For example, isoflavones are continuously excreted from the roots of legumes and molecules such as daidzein have been shown to induce nodulation related genes in the nodulating Rhizobium bacteria (Kobayashi et al. (2004)).

In recent years, an increasing number of studies have focused on the relationship between plant-derived foods containing flavonoids and human health. Both academic and applied interest in this area is stimulated by the fact that some widely consumed plant foods are relatively rich in flavonoid/phenolic compounds and by the fact that people who consume higher quantities of these foods appear to have lower risks for certain significant heath problems, such as cardiovascular disease and cancer. (Bazzano et al. (2002); Clifford (2004); Cos et al. (2005); and Go et al. (2005)). The importance of flavonoids to human and animal health is supported by detailed experimental data, which indicate that flavonoids can have specific functional interactions within mammalian cells. For example, the antioxidant properties of flavonols, such as kaempferol and quercetin, have been broadly shown to give some protection against oxidative stress. (Sugihara et al. (1999); and Duthie et al. (2000)). Daily oral administration of the flavonoid quercetin has been shown to exhibit both anti-hypertensive and antioxidative effects in hypertensive rats. (Garcia-Saura et al. (2005)).

The flavonoids present in dark chocolate are currently being intensively studied, and a recent study showed that consumption of dark chocolate rich in flavonoids may lower blood pressure, presumably through the ability of one or more of the flavonoids to increase nitric oxide bioavailability. (Grassi et al. (2005)). Resveratrol is another flavonoid related molecule that is currently of interest. This phytoalexin is found in grapes and other foods, and has been found to be active as a cancer chemoprevention agent, (Jang et al. (1997)), and to have the potential to delay aging via its ability to activate Sir-2 like proteins (Sirtuins). (Wood et al. (2004)). Higher dietary intake of other flavonoids, like the isoflavonoids found in soy, has also been associated with reduced levels of cancer, (Setchell et al. (1999)), and dietary intake of the isoflavone genistein has been shown to reduce the susceptibility of rats to mammary cancer, (Lamartiniere et al. (2002), and helps prevent bone loss caused by estrogen deficiency in female mice. (Ishimi et al. (1999)).
The early steps of the plant phenylpropanoid pathway leading to the key flavonoid precursors p-coumaryl-CoA have been described in several plants (Figure 1A; Dixon et al. (1995); and Winkel-Shirley (2002). The first step in the phenylpropanoid pathway is the deamination of phenylalanine to cinnamic acid by L-phenylalanine ammonia lyase (PAL). Four different PAL genes have been characterized in Arabidopsis and these appear to fall into two different groups. (Raes et al. (2003)). As expected for a major branch-point between the plant primary and secondary metabolic pathways, the expression and activities of the different PAL isoforms are under complex regulatory control. (Dixon et al. (1995); and Rohde et al. (2004)). The next enzyme in the pathway is the protein trans cinnamate-4-hydroxylase (C4H; CYP73A5). Only one gene has been found for this P450-dependent mono-oxygenase in arabidopsis, while in some other plants, two or more C4H genes can been found that fall into two distinct classes. (Raes et al. (2003)). The next step, the production of p-coumaryl-CoA, is carried out by 4-coumarate:CoA ligase (4CL). In the arabidopsis genome, there are at least four 4CL genes and nine 4CL-like genes. (Raes et al. (2003)). In addition to forming p-coumaryl CoA, the 4CL proteins characterized from arabidopsis, as well as the characterized 4CL proteins from soybean we found to be capable of forming CoA esters with caffeic acid and ferulic acid at different efficiencies, and the At4CL4 protein of arabidopsis and the Gm4CL1 protein of soybean were also found to be capable of forming CoA esters with 5-hydroxyferulic acid and sinapic acid. (Hu et al. (1998); Lindermayr et al. (2002); Schneider et al. (2003); and Hamberger et al. (2004)).

A number of recent reviews on the core flavonoid synthesis pathway have been published (Winkel-Shirley (2002) and Dixon (2005), and the current understanding of this pathway is outlined schematically in Figure 1B. (Winkel-Shirley (2002), and Xie et al. (2004)). The first step of this pathway, which is catalyzed by chalcone synthase (CHS), is the condensation of p-coumaryl CoA with three molecules of malonyl CoA to form tetrahydroxylchalcone (naringenin chalcone). In some plants, particularly the leguminous plants, the enzyme chalcone reductase (CHR) can also be present (Figure 1B). This enzyme is thought to act on an intermediate of the CHS multistep reaction, and the CHS/CHR coupled reaction is proposed to yield chalcone (4,2',4',6'-tetrahydroxylchalcone) and deoxychalcone (4,2',4'-trihydroxyxchalcone). (Bomati et al. (2005)). These CHS/CHR products are then precursors for a group of phytoalexins that are often produced in response to herbivore and pathogen attacks, and for the synthesis of CHR derived products that are involved in symbiotic root nodulation by nitrogen fixing Rhizobium bacteria. (Dixon et al. (1999)).
In the core flavonoid pathway, the product of CHS (naringenin chalcone) is transformed into (2S)-5,7,4’-trihydroxyflavanone (naringenin) by chalcone isomerase (CHI). Two types of CHI have been found, with type I being ubiquitous in the plant kingdom, while the type II CHI, which has a broader substrate range, appears to be most frequently found in leguminous plants. (Ralston et al. (2005)). The next reaction in the pathway is the addition of a hydroxyl group at the C3 position of the C ring to form 2,3-dihydrokaempferol (DHK) and is catalyzed by F3betaH, a 2-oxoglutarate dependent dioxygenase. (Dixon et al. (1999); Wellman et al. (2004); and Dixon (2005)). As indicated in Figure 1B, DHK can be further hydroxylated at the 3’ and 5’ positions of the B ring by the P450 dependent enzymes F3’H and F3’5’H forming 2,3-dihydroquercetin (DHQ) and 2,3dihydromyricetin (DHM) respectively. Dihydroflavonol-4-reductase (DFR) catalyses the next reaction, the addition of a hydroxyl group at the 4 position of ring C of DHK, DHQ, and DHM (synthesized by the F3H family of enzymes) to yield leucopelargonidin, leucocyanidin or leucodelphinidin respectively. However, some plant DFR proteins do not accept the monohydroxylated DHK and thus these plants may not be able to make the associated downstream products. (Johnson et al. (2001)). Only one DFR gene has been found in plants such as Arabidopsis and tomato, and interestingly, in plants with several DFR genes, it appears that only one of these genes produces an active protein. (Xie et al. (2004)). The products of DFR are key precursors for the synthesis of the anthocyanins and condensed tannins and it has been noted that herbivore attack induces DFR expression in plants. Furthermore, this induction is associated with an increase in the synthesis of condensed tannins, a group of molecules that have been implicated in protecting plants from herbivores. (Peters et al. (2002)). Immediately downstream of DFR, the enzyme anthocyanidin synthase (ANS; leucoanthocyanidin dioxygenase) is capable of forming anthocyanidins from the different DFR products (i.e., from the leucoanthocyanins), and these ANS products can be subsequently glycosylated to form the anthocyanins. In addition, leucoanthocyanidin reductase (LAR) can also convert the leucoanthocyanidins to form 2,3-trans-flavan-3-ols (catechins). The related 2,3-cis-flavan-3-ols (epicatechins) are formed via the action of ANS and anthocyanidin reductase (ANR) which uses NADPH to reduce anthocyanidins. (Xie et al. (2003)). Finally, there is little currently known about the last step(s) involved in the formation of condensed tannins from trans- and cis-flavan-3-ol monomers.

It is well known that flavonoids, and related glycosylated derivatives, make significant flavor contributions to beverages produced from plant ingredients. For example, grapefruit citrus are known to have a bitter flavor, which is in part due to the presence of a flavanone
(flavanone-7-neohesperidosides), while oranges, in contrast, are generally less bitter, having only tasteless flavonone-7-rutinosides. (Frydman A et al. (2004). Likewise, it is well known that the flavonoids and related molecules in grapes contribute significantly to the astringency and bitterness characteristics of different wines (Monagas M et al. (2005)). Finally, fruit juices and other beverages such as tea contain anthocyanidins at levels that contribute significantly to the flavor and astringency of these beverages. (Dixon R et al. (2005b); and, Lesschaeve I et al. (2005)). The anthocyanidins are monomers, oligomers and polymers of molecules produced by the flavonoid pathway. Considering the observations above, it can thus be expected that by altering the levels, and/or the molecular profiles of the precursors, or by altering the polymerization levels and profiles of the final products in the starting plant material, it could be possible to alter the flavor and astringency profiles of beverages made from these raw materials.

There is currently little information published on the presence of flavonoids in the green or roasted coffee grain and whether these molecules or derivatives thereof contribute to the flavor of coffee. However, there is one recent report that suggests that flavonoids are present in roasted coffee. (Yen et al. (2005)). It is noted, however, that the method used by these investigators to determine flavonoid content is a generalized total flavonoid method and thus could provide an artificially inflated measurement of this broad class of molecules in roasted coffee. Accordingly, more detailed work is required to examine the flavonoids present in both the coffee grain and in the roasted product. One study has recently been carried out that begins to address the flavonoids present in the fruit part of the coffee cherry (pericarp). It has been found that ripe coffee arabica cherry fruit contains three major classes of flavonoids: the flavan-3-ols (monomers and procyanidins), flavonols, and anthocyanidins. (Ramirez-Coronel et al. (2004)). Given the known roles these compounds play in other plants, it can be presumed that the coffee flavonoids are also involved in protecting the fruit tissues from UV and oxidation related stresses, and in protecting the cherries from microbial and insect attack. Because of the health benefits of flavonoids in the human diet, and because these molecules also have agronomic benefits (herbivore/pathogen and stress resistance), it is of interest to examine the flavonoid pathway in coffee.

From the foregoing discussion, it will be appreciated that modulating flavonoid content in coffee grain by genetically modulating the production of the proteins responsible for early phenylpropanoid and flavonoid biosynthesis would be of great utility to enhance the aroma and flavor of coffee beverages and coffee products produced from such genetically engineered coffee beans. Enhanced flavonoid content and/or altered flavonoid profile in the coffee bean may also
positively contribute to the overall health and wellness of consumers of coffee beverages and products produced from such coffee beans. In addition, modulating flavonoid content in the coffee plant has implications for protecting the coffee fruit from ultraviolet, oxidative, microbial, or insect stress or damage. Accordingly, a need exists to identify, isolate and utilize genes and enzymes from coffee that are involved in the biosynthesis of early phenylpropanoids and flavonoids.

SUMMARY OF THE INVENTION

The invention described herein features genes encoding enzymes in the pathways that lead to flavonoid biosynthesis in coffee plants, their encoded polypeptides, and methods for using these polynucleotides and polypeptides for gene regulation and manipulation of flavor, aroma and other features of coffee beans.

One aspect of the invention features a nucleic acid molecule isolated from coffee (Coffeea spp.), having a coding sequence that encodes a flavonoid pathway enzyme. In one embodiment, the enzyme is a phenylalanine ammonia lyase that is at least 85.2% identical to SEQ ID NO:20. In another embodiment, the enzyme is a phenylalanine ammonia lyase that is at least 83.9% identical to SEQ ID NO:21. In another embodiment, the enzyme is a phenylalanine ammonia lyase that is at least 82.6% identical to SEQ ID NO:22. In another embodiment, the enzyme is a trans cinnamate-4-hydroxylase that is at least 89.9% identical to SEQ ID NO:23. In another embodiment, the enzyme is a 4-coumarate:CoA ligase that is at least 81.1% identical to SEQ ID NO:24. In another embodiment, the enzyme is a 4-coumarate:CoA ligase that is at least 81.9% identical to SEQ ID NO:25. In another embodiment, the enzyme is a 4-coumarate:CoA ligase that is at least 81% identical to SEQ ID NO:26. In another embodiment, the enzyme is a chalcone synthase that is at least 90.5% identical to SEQ ID NO:27. In another embodiment, the enzyme is a chalcone reductase that is at least 60.7% identical to SEQ ID NO:28. In another embodiment, the enzyme is a chalcone reductase that is at least 61.6% identical to SEQ ID NO:29. In another embodiment, the enzyme is a chalcone reductase that is at least 61.3% identical to SEQ ID NO:30. In another embodiment, the enzyme is a chalcone isomerase that is at least 63.4% identical to SEQ ID NO:31. In another embodiment, the enzyme is a chalcone isomerase that is at least 64.3% identical to SEQ ID NO:32. In another embodiment, the enzyme is a flavanone 3-hydroxylase that is at least 82.4% identical to SEQ ID NO:33. In another embodiment, the enzyme is a flavanoid 3',5'-hydroxylase that is at least 67.8% identical to SEQ ID NO:34. In another embodiment, the enzyme is a dihydroflavonol-4-reductase that is at least
67.6% identical to SEQ ID NO:35. In another embodiment, the enzyme is a leucoanthocyanidin dioxygenase that is at least 73.1% identical to SEQ ID NO:36. In another embodiment, the enzyme is a leucoanthocyanidin reductase that is at least 59.9% identical to SEQ ID NO:37. In another embodiment, the enzyme is an anthocyanidin reductase that is at least 77.4% identical to SEQ ID NO:38.

In certain embodiments, the nucleic acid molecule is a gene having an open reading frame that comprises the coding sequence. Alternatively, it may comprise an mRNA molecule produced by transcription of that gene, or a cDNA molecule produced by reverse transcription of the mRNA molecule. The invention also features an oligonucleotide between 8 and 100 bases in length, which is complementary to a segment of the aforementioned nucleic acid molecule.

Another aspect of the invention features a vector comprising the above-described flavonoid pathway enzyme-encoding nucleic acid molecules. In certain embodiments, the vector is an expression vector selected from the group of vectors consisting of plasmid, phagemid, cosmid, baculovirus, bacmid, bacterial, yeast and viral vectors. In certain embodiments, the vector contains the coding sequence of the nucleic acid molecule operably linked to a constitutive promoter. In other embodiments, the coding sequence is operably linked to an inducible promoter. In other embodiments, the coding sequence of the nucleic acid molecule is operably linked to a tissue specific promoter, such as a seed specific promoter, preferably a coffee seed specific promoter.

According to another aspect of the invention, a host cell transformed with the aforementioned vector is provided. The host cell may be a plant, bacterial, fungal, insect or mammalian cell. In certain embodiments, the host cell is a plant cell selected from any one of coffee, tobacco, Arabidopsis, maize, wheat, rice, soybean barley, rye, oats, sorghum, alfalfa, clover, canola, safflower, sunflower, peanut, cacao, tomato tomatillo, potato, pepper, eggplant, sugar beet, carrot, cucumber, lettuce, pea, aster, begonia, chrysanthemum, delphinium, zinnia, and turfgrasses. The invention also features a fertile transgenic plant produced by regenerating the transformed plant cell. In a specific embodiment, the fertile transgenic plant is a Coffea species.

Another aspect of the invention features a method to modulate flavor or aroma of coffee beans. The method comprises modulating production of one or more flavonoid pathway enzymes within coffee seeds. In some embodiments, the method comprises increasing production of the one or more flavonoid pathway enzymes, e.g., by increasing expression of one or more endogenous flavonoid pathway enzyme-encoding genes within the coffee seeds, or by
introducing a flavonoid pathway enzyme-encoding transgene into the plant. In other embodiments, the method comprises decreasing production of the one or more flavonoid pathway enzymes, e.g., by introducing a nucleic acid molecule into the coffee that inhibits the expression of one or more of the flavonoid pathway enzyme-encoding genes.

Other features and advantages of the invention will be understood by reference to the drawings, detailed description and examples that follow.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1. Schematic overview of early phenylpropanoid and the flavonoid pathways. A) General phenylpropanoid metabolism. The early part of this pathway is dependent on the activity of the three enzymes noted. PAL, phenylalanine lyase; C4H, cinnamate-4-hydroxylase; and 4CL, 4-coumarate:coenzyme A ligase. B) Plant flavonoid pathway (modified from Xie et al. (2004) CHS, chalcone synthase; CHI, chalcone isomerase; F3'H, flavanone 3-hydroxylase; F3',5'H, flavonoid 3',5'-hydroxylase; DFR, dihydroflavonol-4-reductase; ANS, anthocyanidin synthase (also known as LDOX for leucoanthocyanidin dioxygenase); GT, anthocyanidin glucosyl transferase; LAR, leucoanthocyanidin reductase. ANR, anthocyanidin reductase (also called BAN, i.e., BANYULS protein); CHR, chalcone reductase (CHS/CHR are believed to work together to generate 4,2’,4’-trihydroxychalcone, also called deoxychalcone, Bomati et al., 2005). Tetrahydroxychalcone (4, 2’, 4’, 6’- tetrahydroxychalcone) and trihydroxychalcone (4,2’,4’- trihydroxychalcone) are also known as chalcone and deoxychalcone, respectively. A to C on the naringenin structure indicate the standard nomenclature assigned to the three flavonoid rings. After 3-O-glucosylation of anthocyanidins to form anthocyanins by GT, anthocyanins may be further modified by additional glycosylation, methylation, and acylation.

Figure 2. Characterization of the complete ORF for CaPAL1 from Coffea arabica. A) The full protein-coding region is shown as a hatched bar and the 5’ untranslated region and 3’ untranslated region are shown as a thick black and a thin black bars, respectively. The translation start codon (ATG) and stop (TGA) codons are indicated. B) The insert sequence (SEQ ID NO: 1) of pML8 was aligned with the cDNA sequences pccwc22w18n3 (SEQ ID NO: 39) and Race1_CcPAL1 (SEQ ID NO: 40). The alignment was done using the CLUSTAL W program (Lasergene package, DNASTAR) and manually optimized. Nucleic acids marked in grey match the pML8 (CaPAL1) insert sequence (SEQ ID NO: 1).
Figure 3. Characterization of the partial ORF for CcPAL2 from Coffea canephora.
A) In-silico generation of the consensus sequence for C. canephora CcPAL2 (SEQ ID NO: 3). The partial protein-coding region is shown as a hatched bar, the 3’ untranslated region is shown as a thin black line, and the grey bar represents a genomic sequence believed to be an intron. B) The insert sequence of pcccpl9k7 (SEQ ID NO: 41) and the partial insert sequence for the clone ccl25c13 (SEQ ID NO: 42) were aligned with the cDNA sequence Race1_CcPAL2 (SEQ ID NO: 43) and the genomic sequence GW1_CcPAL2 (SEQ ID NO: 44) to generate the in-silico partial sequence for CcPAL2 (SEQ ID NO: 3). The alignment was done using the CLUSTAL W program and manually optimized.

Figure 4. Characterization of the complete ORF for CaPAL3 (SEQ ID NO: 2) from Coffea arabica. A) The full protein-coding region is shown as a hatched bar and the 5’ untranslated region and 3’ untranslated region are shown as a thick black and a thin black bars, respectively. B) The insert sequence (SEQ ID NO: 2) of pML14 aligned with the cDNA sequences pcccpl611 (SEQ ID NO: 45), Race1_CcPAL3 (SEQ ID NO: 46) and the Coffea canephora PAL1 cDNA sequence (GenBank Accession Number AF460203) (SEQ ID NO: 47). The alignment was done using the CLUSTAL W program and manually optimized. Nucleic acids marked in grey match the pML14 (CaPAL3) insert sequence (SEQ ID NO: 2).

Figure 5. Protein sequence alignment of new PAL sequences with PAL sequences in the public database. A) Alignment of the protein sequences CaPAL1 (pML8), CcPAL2 (in-silico assembly, partial) and CaPAL3 (pML14) (SEQ ID NOs.: 20, 21) with other Coffea PAL sequences. The alignment was performed using ClustalW method in the MegAlign software. Amino acids marked in grey are generally conserved in the majority of the sequences at this position. GenBank Accession Numbers are given in parentheses. CcPAL1 AAN32866 (Coffea canephora, AAN32866, full protein) (SEQ ID NO: 48), CcPAL2 AAN32867 CcPAL2 (Coffea canephora, AAN32867, full protein) (SEQ ID NO: 49). B) Protein sequence alignment of CaPAL1 (pML8) (SEQ ID NO: 20), CcPAL2 (in-silico assembly, partial) (SEQ ID NO: 22) and CaPAL3 (pML14) (SEQ ID NO: 21) with other PAL protein sequences. The alignment was done with CLUSTAL-W. Amino acids marked in grey are generally conserved in the majority of the sequences at this position. GenBank Accession Numbers are given in parentheses. ZmPAL1 (Zea mays, AAL40137) (SEQ ID NO: 50), AtPAL1 (Arabidopsis thaliana, AAP59438, Cochrane et al., 2004) (SEQ ID NO: 51), AtPAL2 (Arabidopsis thaliana, AAP59439, Cochrane et al., 2004) (SEQ ID NO: 52), AtPAL4 (Arabidopsis thaliana, AAP59440, Cochrane et al., 2004) (SEQ ID NO: 53), PcPAL1 (Petroselinum crispum, Parsley,
CAA68938) (SEQ ID NO: 54), PcPAL2 (\textit{Petroselinum crispum}, Parsley, CAA57056) (SEQ ID NO: 55) and PcPAL3 (\textit{Petroselinum crispum}, Parsley, CAA57057) (SEQ ID NO: 56).

\textbf{Figure 6. Protein sequence alignment of CcC4H with plant C4H protein sequences.} The alignment of protein sequence (SEQ ID NO: 23) encoded by \textit{CcC4H} (SEQ ID NO: 4) with other C4H proteins available in the NCBI database was done using CLUSTAL W. Amino acids marked in grey are generally conserved in the majority of the sequences at this position. GenBank Accession Numbers are given in parentheses. AtC4H (\textit{Arabidopsis thaliana}, BAA24355) (SEQ ID NO: 57), MsC4H (\textit{Medicago sativa}, P37114) (SEQ ID NO: 158), PbC4H (\textit{Populus balsamifera}, AAC50231) (SEQ ID NO: 58).

\textbf{Figure 7. Characterization of the partial ORF for Cc4CL1 from \textit{Coffea canephora}.} A) Generation of an \textit{in-silico}-assembled partial sequence for Cc4CL1 from \textit{C. canephora}. The partial protein-coding region is shown as a hatched bar and the 3' untranslated region is shown as a thin black line. B) The insert sequence of pcccp27d21 (SEQ ID NO: 59) was aligned with the cDNA sequence Race1\_Cc4CL1 (SEQ ID NO: 60) to generate the \textit{in-silico} partial sequence for Cc4CL1 (SEQ ID NO: 6). The alignment was done using the CLUSTAL W program, and manually optimized.

\textbf{Figure 8. Isolation and characterization of the complete ORF for Ca4CL2 (SEQ ID NO: 5) from \textit{Coffea arabica} and Ca4CL2 (SEQ ID NO: 7) from \textit{Coffea canephora}.} A) Isolation of two cDNA containing the complete ORF sequences of 4CL2 from both \textit{C. arabica} and \textit{C. canephora} (SEQ ID NOs: 5, 7). The full protein-coding region is shown as a hatched bar, the 3' untranslated region is shown as a thin black line, and the grey bar represents genomic sequences 5' to the ORFs of Cc4CL2 and Ca4CL2 (upstream of the translation start codon, ATG. The border between the promoter and 5'UTR are not yet defined. B) The cDNA sequence of pcccl24i21 and the genomic sequences GW1\_Ca4CL2 (SEQ ID NO: 61) and GW1\_Cc4CL2 (SEQ ID NO: 62) were aligned with the inserts of pGC1 (Ca4CL2) (SEQ ID NO: 5) and pGC3 (Cc4CL2) (SEQ ID NO: 7). The alignment was done using the CLUSTAL-W program and manually optimized.

\textbf{Figure 9. Protein sequence alignment of the coffee Cc4CL1 (partial), Cc4CL2 (pGC3) and Ca4CL2 (pGC1) protein sequences with other plant 4CL sequences.} Alignment of the coffee 4CL protein sequences (\textit{Cc4CL1}, \textit{Cc4CL2} (in pGC3); \textit{and Ca4CL2} (in pGC1) (SEQ ID NOs: 25, 26, 24) with other 4CL proteins available in the NCBI database was done using the CLUSTAL-W program. Amino acids marked in grey are generally conserved in the majority of the sequences at this position. GenBank Accession Numbers are given in
parentheses: At4CL1 (Arabidopsis thaliana, Q42524) (SEQ ID NO: 63), At4CL2 (Arabidopsis thaliana, NP_188761.1) (SEQ ID NO: 64), At4CL3 (Arabidopsis thaliana, Q9S777) (SEQ ID NO: 65), At4CL4 (Arabidopsis thaliana, AAM19949) (SEQ ID NO: 66), Nt4CL1 (Nicotiana tabacum, O24145) (SEQ ID NO: 67), Nt4CL2 (Nicotiana tabacum, T03789) (SEQ ID NO: 68). Boxed regions have been proposed and presented in an alignment by Schneider et al., (2003): Blue boxes indicate two conserved peptide motifs. Green boxes indicate the 12 amino acids residues that are proposed to function as the 4CL substrate specificity.

**Figure 10.** Protein sequence alignment of CeCHS with plant chalcone synthase (CHS) and stilbene synthase protein sequences. Alignment of the protein (SEQ ID NO: 27) encoded by the CeCHS (SEQ ID NO: 8) with other CHS proteins available in the NCBI database was done using the CLUSTAL W program. Amino acids marked in grey are generally conserved in the majority of the sequences at this position. GenBank Accession Numbers are given in parentheses. Chalcone synthases (CHS) sequences: HaCHS (Hypericum androsaemum, AAG30295) (SEQ ID NO: 69) and LeCHS (Lycopersicon esculentum, CAA38981) (SEQ ID NO: 70). Stilbene Synthases (SS or STS): VSS (Vitis, AAB19887) (SEQ ID NO: 71) and PsSTS (Pinus strobus, CAA87013) (SEQ ID NO: 72).

**Figure 11.** Protein sequence alignment of CcCHR1, CcCHR2A and CcCHR2B with other plant CHR protein sequences. Manually optimized alignment of putative proteins (SEQ ID NOs: 28, 29, 30) encoded by CcCHR1, CcCHR2A and CcCHR2B (SEQ ID NOs: 9, 10, 11) with other CHR proteins available in the NCBI database was done in the MegAlign software. Amino acids marked in grey are generally conserved in the majority of the sequences at this position. GenBank Accession Numbers are given in parentheses. SrCHR (Sesbania rostrata, CAA11226) (SEQ ID NO: 73), PiCHR (Pueraria Montana var. lobata, AAM12529) (SEQ ID NO: 74) and MsCHR (Medicago sativa, AAB41555) (SEQ ID NO: 75).

**Figure 12.** Protein sequence alignment of CcCHI with plant CHI protein sequences. A) Manually optimized alignment of putative protein (SEQ ID NO: 31) encoded by CcCHI (SEQ ID NO: 12) with other CHI proteins available in the NCBI database was done using CLUSTAL W program. Amino acids marked in grey are generally conserved in the majority of the sequences at this position. GenBank Accession Numbers are given in parentheses. PhCHI (Petunia x hybrida, PIR ID ISPJA1) (SEQ ID NO: 76), LjCHI-2 (Lotus corniculatus var. japonicus, BAC53984 Type I CHI) (SEQ ID NO: 77) and GmCHI-2 (Glycine max, AAT94360, Type I CHI) (SEQ ID NO: 78). B) Alignment of putative protein (SEQ ID NO: 32) encoded by CcCHI-like (SEQ ID NO: 13) with other CHI proteins available in the NCBI database was done.
using the CLUSTAL W program in the MegAlign software. Amino acids marked in grey are
generally conserved in the majority of the sequences at this position. GenBank Accession
Numbers are given in parentheses GmCHI-4A (Glycine max, AAT94362) (SEQ ID NO: 79) and
GmCHI-1A (Glycine max, AAT94358, Type II CHI) (SEQ ID NO: 80).

**Figure 13. Protein sequence alignment of CcF3H with plant F3H protein sequences.**
A) Alignment of the protein (SEQ ID NO: 33) encoded by CcF3H (SEQ ID NO: 14) with other
F3H proteins available in the NCBI database was done using the CLUSTALW program. Amino
acids marked in grey are generally conserved in the majority of the sequences at this position.
GenBank Accession Numbers are given in parentheses. AtF3H (Arabidopsis thaliana,
AAC68584) (SEQ ID NO: 81), GmF3H (Glycine max, AAT94365) (SEQ ID NO: 82). B)
Alignment of partial putative protein (SEQ ID NO: 34) encoded by the CcF3'5'H (SEQ ID NO:
15) with biochemically-characterized F3'5'H proteins available in the NCBI database over the
same region was done using the CLUSTAL W program. Amino acids marked in grey are
generally conserved in the majority of the sequences at this position. GenBank Accession
Numbers are given in parentheses. CrF3'5'H (Catharanthus roseus, CAA09850) (SEQ ID NO:
83) and GtF3'5'H (Gentiana triflora, Q96581) (SEQ ID NO: 84).

**Figure 14. Protein sequence alignment of CcDFR with plant DFR protein
sequences.** Manually optimized alignment of putative protein (SEQ ID NO: 35) encoded by the
CcDFR (SEQ ID NO: 16) with biochemically-characterized DFR proteins from Medicago
truncatula MtDFR1 (GenBank Accession Number AAR27014) (SEQ ID NO: 85) and MtDFR2
(GenBank Accession Number AAR27015) (SEQ ID NO: 86) available in the NCBI database
was done using the MegAlign software (Lasergene package, DNASTAR). Amino acids marked
in grey are generally conserved in the majority of the sequences at this position.

**Figure 15. Protein sequence alignments.** A) Alignment of the protein (SEQ ID NO:
36) encoded by the CcLOX (SEQ ID NO: 17) with other LOX (ANS) proteins available in
the NCBI database was done using the CLUSTAL W program. Amino acids marked in grey are
generally conserved in the majority of the sequences at this position. GenBank Accession
Numbers are given in parentheses. AtLOX (Arabidopsis thaliana, CAD91994) (SEQ ID NO:
87), PfANS (Perilla frutescens, O04274) (SEQ ID NO: 88), FiANS (Forsythia x intermedia,
CAA73094) (SEQ ID NO: 89), InANS (Ipomoea nil, BAB71810) (SEQ ID NO: 90). B)
Alignment of the protein (SEQ ID NO: 37) encoded by the CcLAR (SEQ ID NO: 18) with other
LAR proteins available in the NCBI database was done using the CLUSTAL W program.
Amino acids marked in grey are generally conserved in the majority of the sequences at this
position. GenBank Accession Numbers are given in parentheses. VvLAR (Vitis vinifera, CAI26309) (SEQ ID NO: 91), LuLAR (Lotus uliginosus, AAU45392) (SEQ ID NO: 92), DuLAR (Desmodium uncinatum, Q84V83) (SEQ ID NO: 93). C) Protein sequence alignment of CcANR (SEQ ID NO: 38) with plant anthocyanin and leucoanthocyanin reductase protein sequences. Alignment of putative protein (SEQ ID NO: 38) encoded by the CcANR (SEQ ID NO: 19) with other anthocyanin and leucoanthocyanin reductase proteins available in the NCBI database was done using the CLUSTAL W program. Amino acids marked in grey are generally conserved in the majority of the sequences at this position. GenBank Accession Numbers are given in parentheses. MtANR (Medicago truncatula, AAN77735) (SEQ ID NO: 94), AtANR (Arabidopsis thaliana, AAF23859 Xie et al. (2003)) (SEQ ID NO: 95) and CsLAR (Camellia sinensis, AAT68773) (SEQ ID NO: 96). Despite the fact that CsLAR is currently annotated as an LAR protein, the alignment of this apparently uncharacterized protein strongly suggests it has an ANR-like activity and is not an LAR protein.

Figure 16. RNA blot analysis of CHS, CHI, DFR and F3H expression in whole plant and during coffee bean maturation in Coffea Arabica T2308. About 5 µg of RNA from roots (R), stem (St), leaves (L), flowers (F) and from coffee beans harvested at four different maturation stages Small-Green (SG), Large-Green (LG), Yellow (Y) and Red (R) were separated on denaturing agarose gel. For each maturation stage, coffee cherries have been separated into pericarp (P) and grain (G). Membranes have been probed with either CcCHS (ccc85j10 clone) (SEQ ID NO: 8 or 97) or CcCHI (ccc22k18 clone) (SEQ ID NO: 12) or CcDFR (ccc5115 clone) (SEQ ID NO: 16) or CcF3H (ccc5120 clone) (SEQ ID NO: 14) probes as indicated on the left. A photo of one of the gels showing the ethidium bromide staining of the rRNA25S demonstrates that equivalent amounts of RNA were loaded in each lane. Similar results were obtained for all the blots (data not shown).

Figure 17. Comparison of transcript accumulation of CHS, CHI and CHI-like genes involved in flavonoid metabolism in roots, stems, and leaves and during coffee bean maturation in Coffea arabica T2308 by RT-PCR. cDNA were prepared from 1 µg of total RNA isolated roots (R), stem (St), leaves (L), seed (S) and pericarp (P) at four maturation stages: Small Green (SG), Large Green (LG), Yellow (Y) and Red (R). RT-PCR analysis was performed using gene-specific primers for CcCHS, CcCHI, CcCHI-like, and GOS2.

Figure 18. Tissue-specific expression profile of PAL1, PAL2, PAL3, C4H, 4CL1, and 4CL2 (SEQ ID Nos: 1-7) in C. canephora (robusta, BP409) and C. arabica (arabica, T2308) using quantitative RT-PCR. Total RNA was isolated from root, stem, flower, leaf and from
the grain (G) and pericarp (P) of the cherries harvested at four different maturation stages: Small-Green (SG), Large-Green (LG), Yellow (Y) and Red (R). The RQ value for each tissue sample was determined by normalizing the transcript level of this sample versus the *rpl39* transcript level in that tissue. Data shown represent mean values obtained from three amplification reactions and the error bars indicate standard deviation.

**Figure 19. Analysis of Ca4CL2 (pGC5) (SEQ ID NO: 24) and Ce4CL2 (pGC8) (SEQ ID NO: 26) expression (A) and purification (B and C) performed on SDS-PAGE 12% stained with coomassie blue.** A-Lanes: 1. Crude extract of non-induced Btl21 recombinant cells with pGC5, 2. Crude extract of induced Btl21 recombinant cells with pGC5, 3. Crude extract of non-induced Btl21 recombinant cells with pGC8, 4. Crude extract of induced Btl21 recombinant cells with pGC8. B-Lanes: 1. first wash of column purification, 2-5. fractions of successive elutions for Ca4CL2. C-Lanes: 1. first wash of column purification, 2-5. fractions of successive elutions for Ce4CL2. MW: Molecular weight marker (Unstained Precision Broad Range (Biorad #161-0362)).

DETAILED DESCRIPTION OF ILLUSTRATIVE EMBODIMENTS

**Definitions:**

Various terms relating to the biological molecules and other aspects of the present invention are used throughout the specification and claims.

The term "phenylpropanoid and flavonoid biosynthetic pathways" refers to polypeptides that participate in phenylpropanoid and flavonoid biosynthesis in plants, and more specifically, in coffee plants. This term encompasses the specific mechanism of action of each respective protein in the pathway, and includes the enzymatic generation of flavonoid precursors in the early steps of the phenylpropanoid pathway. The polypeptides include without limitation, phenylalanine ammonia lyase ("PAL"), cinnamate 4-hydroxylase ("C4H"), 4-coumarate:CoA ligase ("4CL"), chalcone synthase ("CHS"), chalcone isomerase ("CHI"), chalcone reductase ("CHR"), flavone 3-hydroxylase ("F3H"), dihydroflavonol 4-reductase ("DFR"), anthocyanidin reductase ("ANR"), and anthocyanidin synthase, also referred to as leucoanthocyanidin dioxygenase ("ANS"), as exemplified herein.

"Isolated" means altered "by the hand of man" from the natural state. If a composition or substance occurs in nature, it has been "isolated" if it has been changed or removed from its original environment, or both. For example, a polynucleotide or a polypeptide naturally present in a living plant or animal is not "isolated," but the same polynucleotide or polypeptide
separated from the coexisting materials of its natural state is "isolated", as the term is employed herein.

"Polynucleotide," also referred to as "nucleic acid molecule", generally refers to any polynucleotide or polydeoxribonucleotide, which may be unmodified RNA or DNA or modified RNA or DNA. "Polynucleotides" include, without limitation single- and double-stranded DNA, DNA that is a mixture of single- and double-stranded regions, single- and double-stranded RNA, and RNA that is mixture of single- and double-stranded regions, hybrid molecules comprising DNA and RNA that may be single-stranded or, more typically, double-stranded or a mixture of single- and double-stranded regions. In addition, “polynucleotide” refers to triple-stranded regions comprising RNA or DNA or both RNA and DNA. The term polynucleotide also includes DNAs or RNAs containing one or more modified bases and DNAs or RNAs with backbones modified for stability or for other reasons. "Modified" bases include, for example, tritylated bases and unusual bases such as inosine. A variety of modifications can be made to DNA and RNA; thus, “polynucleotide” embraces chemically, enzymatically or metabolically modified forms of polynucleotides as typically found in nature, as well as the chemical forms of DNA and RNA characteristic of viruses and cells. “Polynucleotide” also embraces relatively short polynucleotides, often referred to as oligonucleotides.

"Polypeptide" refers to any peptide or protein comprising two or more amino acids joined to each other by peptide bonds or modified peptide bonds, i.e., peptide isosteres.

"Polypeptide" refers to both short chains, commonly referred to as peptides, oligopeptides or oligomers, and to longer chains, generally referred to as proteins. Polypeptides may contain amino acids other than the 20 gene-encoded amino acids. “Polypeptides” include amino acid sequences modified either by natural processes, such as post-translational processing, or by chemical modification techniques which are well known in the art. Such modifications are well described in basic texts and in more detailed monographs, as well as in a voluminous research literature. Modifications can occur anywhere in a polypeptide, including the peptide backbone, the amino acid side-chains and the amino or carboxyl termini. It will be appreciated that the same type of modification may be present in the same or varying degrees at several sites in a given polypeptide. Also, a given polypeptide may contain many types of modifications. Polypeptides may be branched as a result of ubiquitination, and they may be cyclic, with or without branching. Cyclic, branched and branched cyclic polypeptides may result from natural posttranslational processes or may be made by synthetic methods. Modifications include acetylation, acylation, ADP-ribosylation, amidation, covalent attachment of flavin, covalent

"Variant" as the term is used herein, is a polynucleotide or polypeptide that differs from a reference polynucleotide or polypeptide respectively, but retains essential properties. A typical variant of a polynucleotide differs in nucleotide sequence from another, reference polynucleotide. Changes in the nucleotide sequence of the variant may or may not alter the amino acid sequence of a polypeptide encoded by the reference polynucleotide. Nucleotide changes may result in amino acid substitutions, additions, deletions, fusions and truncations in the polypeptide encoded by the reference sequence, as discussed below. A typical variant of a polypeptide differs in amino acid sequence from another, reference polypeptide. Generally, differences are limited so that the sequences of the reference polypeptide and the variant are closely similar overall and, in many regions, identical. A variant and reference polypeptide may differ in amino acid sequence by one or more substitutions, additions or deletions in any combination. A substituted or inserted amino acid residue may or may not be one encoded by the genetic code. A variant of a polynucleotide or polypeptide may be naturally occurring, such as an allelic variant, or it may be a variant that is not known to occur naturally. Non-naturally occurring variants of polynucleotides and polypeptides may be made by mutagenesis techniques or by direct synthesis.

In reference to mutant plants, the terms "null mutant" or "loss-of-function mutant" are used to designate an organism or genomic DNA sequence with a mutation that causes a gene product to be non-functional or largely absent. Such mutations may occur in the coding and/or
regulatory regions of the gene, and may be changes of individual residues, or insertions or deletions of regions of nucleic acids. These mutations may also occur in the coding and/or regulatory regions of other genes which may regulate or control a gene and/or encoded protein, so as to cause the protein to be non-functional or largely absent.

The term "substantially the same" refers to nucleic acid or amino acid sequences having sequence variations that do not materially affect the nature of the protein (i.e. the structure, stability characteristics, substrate specificity and/or biological activity of the protein). With particular reference to nucleic acid sequences, the term "substantially the same" is intended to refer to the coding region and to conserved sequences governing expression, and refers primarily to degenerate codons encoding the same amino acid, or alternate codons encoding conservative substitute amino acids in the encoded polypeptide. With reference to amino acid sequences, the term "substantially the same" refers generally to conservative substitutions and/or variations in regions of the polypeptide not involved in determination of structure or function.

The terms “percent identical” and “percent similar” are also used herein in comparisons among amino acid and nucleic acid sequences. When referring to amino acid sequences, “identity” or “percent identical” refers to the percent of the amino acids of the subject amino acid sequence that have been matched to identical amino acids in the compared amino acid sequence by a sequence analysis program. “Percent similar” refers to the percent of the amino acids of the subject amino acid sequence that have been matched to identical or conserved amino acids. Conserved amino acids are those which differ in structure but are similar in physical properties such that the exchange of one for another would not appreciably change the tertiary structure of the resulting protein. Conservative substitutions are defined in Taylor (1986, J. Theor. Biol. 119:205). When referring to nucleic acid molecules, “percent identical” refers to the percent of the nucleotides of the subject nucleic acid sequence that have been matched to identical nucleotides by a sequence analysis program.

“Identity” and “similarity” can be readily calculated by known methods. Nucleic acid sequences and amino acid sequences can be compared using computer programs that align the similar sequences of the nucleic or amino acids and thus define the differences. In preferred methodologies, the BLAST programs (NCBI) and parameters used therein are employed, and the DNAstar system (Madison, WI) is used to align sequence fragments of genomic DNA sequences. However, equivalent alignments and similarity/identity assessments can be obtained through the use of any standard alignment software. For instance, the GCG Wisconsin Package version 9.1, available from the Genetics Computer Group in Madison, Wisconsin, and the
default parameters used (gap creation penalty=12, gap extension penalty=4) by that program may also be used to compare sequence identity and similarity.

"Antibodies" as used herein includes polyclonal and monoclonal antibodies, chimeric, single chain, and humanized antibodies, as well as antibody fragments (e.g., Fab, Fab’, F(ab’)2 and Fv), including the products of a Fab or other immunoglobulin expression library. With respect to antibodies, the term, "immunologically specific" or "specific" refers to antibodies that bind to one or more epitopes of a protein of interest, but which do not substantially recognize and bind other molecules in a sample containing a mixed population of antigenic biological molecules. Screening assays to determine binding specificity of an antibody are well known and routinely practiced in the art. For a comprehensive discussion of such assays, see Harlow et al. (Eds.), ANTIBODIES A LABORATORY MANUAL; Cold Spring Harbor Laboratory; Cold Spring Harbor, NY (1988), Chapter 6.

The term “substantially pure” refers to a preparation comprising at least 50-60% by weight the compound of interest (e.g., nucleic acid, oligonucleotide, protein, etc.). More preferably, the preparation comprises at least 75% by weight, and most preferably 90-99% by weight, the compound of interest. Purity is measured by methods appropriate for the compound of interest (e.g., chromatographic methods, agarose or polyacrylamide gel electrophoresis, HPLC analysis, and the like).

With respect to single-stranded nucleic acid molecules, the term “specifically hybridizing” refers to the association between two single-stranded nucleic acid molecules of sufficiently complementary sequence to permit such hybridization under pre-determined conditions generally used in the art (sometimes termed “substantially complementary”). In particular, the term refers to hybridization of an oligonucleotide with a substantially complementary sequence contained within a single-stranded DNA or RNA molecule, to the substantial exclusion of hybridization of the oligonucleotide with single-stranded nucleic acids of non-complementary sequence.

A “coding sequence” or “coding region” refers to a nucleic acid molecule having sequence information necessary to produce a gene product, such as an amino acid or polypeptide, when the sequence is expressed. The coding sequence may comprise untranslated sequences (e.g., introns or 5' or 3' untranslated regions) within translated regions, or may lack such intervening untranslated sequences (e.g., as in cDNA).
“Intron” refers to polynucleotide sequences in a nucleic acid that do not code information related to protein synthesis. Such sequences are transcribed into mRNA, but are removed before translation of the mRNA into a protein.

The term “operably linked” or “operably inserted” means that the regulatory sequences necessary for expression of the coding sequence are placed in a nucleic acid molecule in the appropriate positions relative to the coding sequence so as to enable expression of the coding sequence. By way of example, a promoter is operably linked with a coding sequence when the promoter is capable of controlling the transcription or expression of that coding sequence. Coding sequences can be operably linked to promoters or regulatory sequences in a sense or antisense orientation. The term “operably linked” is sometimes applied to the arrangement of other transcription control elements (e.g., enhancers) in an expression vector.

Transcriptional and translational control sequences are DNA regulatory sequences, such as promoters, enhancers, polyadenylation signals, terminators, and the like, that provide for the expression of a coding sequence in a host cell.

The terms “promoter,” “promoter region,” or “promoter sequence” refer generally to transcriptional regulatory regions of a gene, which may be found at the 5' or 3' side of the coding region, or within the coding region, or within introns. Typically, a promoter is a DNA regulatory region capable of binding RNA polymerase in a cell and initiating transcription of a downstream (3' direction) coding sequence. The typical 5' promoter sequence is bounded at its 3' terminus by the transcription initiation site and extends upstream (5' direction) to include the minimum number of bases or elements necessary to initiate transcription at levels detectable above background. Within the promoter sequence is a transcription initiation site (conveniently defined by mapping with nuclease S1), as well as protein binding domains (consensus sequences) responsible for the binding of RNA polymerase.

A “vector” is a repllicon, such as plasmid, phage, cosmid, or virus to which another nucleic acid segment may be operably inserted so as to bring about the replication or expression of the segment.

The term “nucleic acid construct” or “DNA construct” is sometimes used to refer to a coding sequence or sequences operably linked to appropriate regulatory sequences and inserted into a vector for transforming a cell. This term may be used interchangeably with the term “transforming DNA” or “transgene.” Such a nucleic acid construct may contain a coding sequence for a gene product of interest, along with a selectable marker gene and/or a reporter gene.
A “marker gene” or “selectable marker gene” is a gene whose encoded gene product confers a feature that enables a cell containing the gene to be selected from among cells not containing the gene. Vectors used for genetic engineering typically contain one or more selectable marker genes. Types of selectable marker genes include (1) antibiotic resistance genes, (2) herbicide tolerance or resistance genes, and (3) metabolic or auxotrophic marker genes that enable transformed cells to synthesize an essential component, usually an amino acid, which the cells cannot otherwise produce.

A “reporter gene” is also a type of marker gene. It typically encodes a gene product that is assayable or detectable by standard laboratory means (e.g., enzymatic activity, fluorescence).

The term “express,” “expressed,” or “expression” of a gene refers to the biosynthesis of a gene product. The process involves transcription of the gene into mRNA and then translation of the mRNA into one or more polypeptides, and encompasses all naturally occurring post-translational modifications.

“Endogenous” refers to any constituent, for example, a gene or nucleic acid, or polypeptide, that can be found naturally within the specified organism.

A “heterologous” region of a nucleic acid construct is an identifiable segment (or segments) of the nucleic acid molecule within a larger molecule that is not found in association with the larger molecule in nature. Thus, when the heterologous region comprises a gene, the gene will usually be flanked by DNA that does not flank the genomic DNA in the genome of the source organism. In another example, a heterologous region is a construct where the coding sequence itself is not found in nature (e.g., a cDNA where the genomic coding sequence contains introns, or synthetic sequences having codons different than the native gene). Allelic variations or naturally-occurring mutational events do not give rise to a heterologous region of DNA as defined herein. The term “DNA construct”, as defined above, is also used to refer to a heterologous region, particularly one constructed for use in transformation of a cell.

A cell has been “transformed” or “transfected” by exogenous or heterologous DNA when such DNA has been introduced inside the cell. The transforming DNA may or may not be integrated (covalently linked) into the genome of the cell. In prokaryotes, yeast, and mammalian cells for example, the transforming DNA may be maintained on an episomal element such as a plasmid. With respect to eukaryotic cells, a stably transformed cell is one in which the transforming DNA has become integrated into a chromosome so that it is inherited by daughter cells through chromosome replication. This stability is demonstrated by the ability of the eukaryotic cell to establish cell lines or clones comprised of a population of daughter cells.
containing the transforming DNA. A “clone” is a population of cells derived from a single cell or common ancestor by mitosis. A “cell line” is a clone of a primary cell that is capable of stable growth *in vitro* for many generations.

“Grain,” “seed,” or “bean,” refers to a flowering plant’s unit of reproduction, capable of developing into another such plant. As used herein, especially with respect to coffee plants, the terms are used synonymously and interchangeably. As used herein, the term “plant” includes reference to whole plants, plant organs (e.g., leaves, stems, shoots, roots), seeds, pollen, plant cells, plant cell organelles, and progeny thereof. Parts of transgenic plants are to be understood within the scope of the invention to comprise, for example, plant cells, protoplasts, tissues, callus, embryos as well as flowers, stems, seeds, pollen, fruits, leaves, or roots originating in transgenic plants or their progeny.

**Description:**

In one of its aspects the present invention features nucleic acid molecules from coffee that encode a variety of proteins involved in the phenylpropanoid and flavonoid biosynthetic pathways. Representative examples of nucleic acid molecules encoding proteins that comprise the phenylpropanoid and flavonoid biosynthetic pathways were identified from databases of over 47,000 expressed sequence tags (ESTs) from several *Coffea canephora* (robusta) cDNA libraries made with RNA isolated from young leaves and from the grain and pericarp tissues of cherries harvested at different stages of development. Overlapping ESTs were identified and “clustered” into unigenes (contigs) comprising either complete or partial coding sequences. The unigene sequences were annotated by performing a BLAST search of each individual sequence against the NCBI (National Center for Biotechnology Information) non-redundant protein database.

BLAST searches of the coffee EST databases using biochemically characterized protein sequences from public databases revealed gene sequences representing several important enzymes of the phenylpropanoid and flavonoid biosynthetic pathways in the coffee plant. The full open reading frame for some of these sequences were obtained, and a partial open reading frame was obtained for several other sequences. In some cases, the partial coding sequence data were used as the starting point for experiments to isolate the full coding sequence by either 5’ RACE or the genome walking technique. The cDNAs obtained and their encoded proteins are referred to herein as follows:
<table>
<thead>
<tr>
<th>Enzyme</th>
<th>cDNA</th>
<th>(SEQ ID NO.)</th>
<th>encoded protein</th>
<th>(SEQ ID NO.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenylalanine Ammonia Lyase</td>
<td>CaPAL1</td>
<td>1</td>
<td>CaPAL1</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>CaPAL3</td>
<td>2</td>
<td>CaPAL3</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>CcPAL2</td>
<td>3</td>
<td>CcPAL2</td>
<td>22</td>
</tr>
<tr>
<td>Trans Cinnamate-4-Hydroxylase</td>
<td>CcC4H</td>
<td>4</td>
<td>CcC4H</td>
<td>23</td>
</tr>
<tr>
<td>4-Coumarate-CoA Ligase</td>
<td>Ca4CL2</td>
<td>5</td>
<td>Ca4CL2</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Cc4CL1</td>
<td>6</td>
<td>Cc4CL1</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Cc4CL2</td>
<td>7</td>
<td>Cc4CL2</td>
<td>26</td>
</tr>
<tr>
<td>Chalcone Synthase</td>
<td>CcCHS</td>
<td>8</td>
<td>CcCHS</td>
<td>27</td>
</tr>
<tr>
<td>Chalcone Reductase</td>
<td>CcCHR1</td>
<td>9</td>
<td>CcCHR1</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>CcCHR2A</td>
<td>10</td>
<td>CcCHR2A</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>CcCHR2B</td>
<td>11</td>
<td>CcCHR2B</td>
<td>30</td>
</tr>
<tr>
<td>Chalcone Isomerase</td>
<td>CcCHI</td>
<td>12</td>
<td>CcCHI</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>CcCHI-like</td>
<td>13</td>
<td>CcCHI-like</td>
<td>32</td>
</tr>
<tr>
<td>Flavanone 3-Hydroxylase</td>
<td>CcF3H</td>
<td>14</td>
<td>CcF3H</td>
<td>33</td>
</tr>
<tr>
<td>Flavonoid 3',5'-Hydroxylase</td>
<td>CcF3'5'H</td>
<td>15</td>
<td>CcF3'5'H</td>
<td>34</td>
</tr>
<tr>
<td>Dihydroflavonol-4-Reductase</td>
<td>CcDFR</td>
<td>16</td>
<td>CcDFR</td>
<td>35</td>
</tr>
<tr>
<td>Leucoanthocyanidin Dioxygenase</td>
<td>CcLDOX</td>
<td>17</td>
<td>CcLDOX</td>
<td>36</td>
</tr>
<tr>
<td>Leucoanthocyanidin Reductase</td>
<td>CcLAR</td>
<td>18</td>
<td>CcLAR</td>
<td>37</td>
</tr>
<tr>
<td>Anthocyanidin Reductase</td>
<td>CcANR</td>
<td>19</td>
<td>CcANR</td>
<td>38</td>
</tr>
</tbody>
</table>

Although polynucleotides encoding proteins that catalyze key steps of the phenylpropanoid and flavonoid biosynthetic pathways from *Coffeea arabica* and *Coffeea canephora* are described and exemplified herein, this invention is intended to encompass nucleic acids and encoded proteins from other *Coffeea* species that are sufficiently similar to be used interchangeably with the *C. arabica* and *C. canephora* polynucleotides and proteins for the purposes described below. Accordingly, when the term polypeptides or proteins that “comprise the phenylpropanoid and flavonoid biosynthetic pathways” is used herein, it is intended to encompass all *Coffeea* proteins that have the general physical, biochemical, and functional features described herein, as well as the polynucleotides that encode them.

Considered in terms of their sequences, the polynucleotides of the invention that encode proteins that comprise the phenylpropanoid and flavonoid biosynthetic pathways include allelic variants and natural mutants of SEQ ID NOs: 1-19, which are likely to be found in different varieties of *C. arabica* and *C. canephora*, and homologs of SEQ ID NOs: 1-19 likely to be found
in different coffee species. Because such variants and homologs are expected to possess certain
differences in nucleotide and amino acid sequence, this invention provides isolated
polynucleotides encoding proteins that comprise the phenylpropanoid and flavonoid biosynthetic
pathways that have at least about 60%, preferably at least about 61%, 62%, 63%, 64%, 65%,
66%, 67%, 68%, 69% or 70%, more preferably at least about 71%, 72%, 73%, 74%, 75%, 76%,
77%, 78%, 79%, or 80%, even more preferably 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%,
89%, and even more preferably 90%, 91%, 92%, 93%, 94%, 95%, and most preferably 96%,
97%, 98% and 99% or more identity with any one of SEQ ID NOs:20-38, and comprise a
nucleotide sequence having equivalent ranges of identity to any one of SEQ ID NOs:1-19.
Because of the natural sequence variation likely to exist among proteins that comprise the
phenylpropanoid and flavonoid biosynthetic pathways, and the genes encoding them in different
coffee varieties and species, one skilled in the art would expect to find this level of variation,
while still maintaining the unique properties of the polypeptides and polynucleotides of the
present invention. Such an expectation is due in part to the degeneracy of the genetic code, as
well as to the known evolutionary success of conservative amino acid sequence variations,
which do not appreciably alter the nature of the encoded protein. Accordingly, such variants and
homologs are considered substantially the same as one another and are included within the scope
of the present invention.

The gene regulatory sequences associated with genes encoding proteins that comprise the
phenylpropanoid and flavonoid biosynthetic pathways are of practical utility and are considered
within the scope of the present invention. Promoters and other gene regulatory sequences of
genes encoding proteins that comprise the phenylpropanoid and flavonoid biosynthetic pathways
from any coffee species may be obtained by the methods described below, and may be utilized
in accordance with the present invention. Promoters and regulatory elements governing tissue
specificity and temporal specificity of the expression of genes encoding proteins that comprise
the phenylpropanoid and flavonoid biosynthetic pathways may be used to advantage, alter or
modify the expression of proteins that comprise the phenylpropanoid and flavonoid biosynthetic
pathways toward the goal of enhancing the flavor and aroma of coffee products produced from
coffee beans comprising such modifications, among other utilities.

The following sections set forth the general procedures involved in practicing the present
invention. To the extent that specific materials are mentioned, it is merely for the purpose of
illustration, and is not intended to limit the invention. Unless otherwise specified, general
biochemical and molecular biological procedures, such as those set forth in Sambrook et al.,

- 24 -
Molecular Cloning, Cold Spring Harbor Laboratory (1989) or Ausubel et al. (eds), Current Protocols in Molecular Biology, John Wiley & Sons (2005) are used.

**Nucleic Acid Molecules, Proteins and Antibodies:**

Nucleic acid molecules of the invention may be prepared by two general methods: (1) they may be synthesized from appropriate nucleotide triphosphates, or (2) they may be isolated from biological sources. Both methods utilize protocols well known in the art.

The availability of nucleotide sequence information, such as the cDNA having SEQ ID NOs:1-19, enables preparation of an isolated nucleic acid molecule of the invention by oligonucleotide synthesis. Synthetic oligonucleotides may be prepared by the phosphoramidite method employed in the Applied Biosystems 38A DNA Synthesizer or similar devices. The resultant construct may be purified according to methods known in the art, such as high performance liquid chromatography (HPLC). Long, double-stranded polynucleotides, such as a DNA molecule of the present invention, must be synthesized in stages, due to the size limitations inherent in current oligonucleotide synthetic methods. Thus, for example, a long double-stranded molecule may be synthesized as several smaller segments of appropriate complementarity. Complementary segments thus produced may be annealed such that each segment possesses appropriate cohesive termini for attachment of an adjacent segment. Adjacent segments may be ligated by annealing cohesive termini in the presence of DNA ligase to construct an entire long double-stranded molecule. A synthetic DNA molecule so constructed may then be cloned and amplified in an appropriate vector.

In accordance with the present invention, nucleic acids having the appropriate level sequence homology with part or all of the coding and/or regulatory regions genes encoding proteins that comprise the phenylpropanoid and flavonoid biosynthetic pathways may be identified by using hybridization and washing conditions of appropriate stringency. It will be appreciated by those skilled in the art that the aforementioned strategy, when applied to genomic sequences, will, in addition to enabling isolation of coding sequences for genes encoding proteins that comprise the phenylpropanoid and flavonoid biosynthetic pathways, also enable isolation of promoters and other gene regulatory sequences associated with genes encoding proteins that comprise the phenylpropanoid and flavonoid biosynthetic pathways, even though the regulatory sequences themselves may not share sufficient homology to enable suitable hybridization.
As a typical illustration, hybridizations may be performed according to the method of Sambrook et al., using a hybridization solution comprising: 5X SSC, 5X Denhardt's reagent, 1.0% SDS, 100 μg/ml denatured, fragmented salmon sperm DNA, 0.05% sodium pyrophosphate and up to 50% formamide. Hybridization is carried out at 37-42°C for at least six hours. Following hybridization, filters are washed as follows: (1) 5 minutes at room temperature in 2X SSC and 1% SDS; (2) 15 minutes at room temperature in 2X SSC and 0.1% SDS; (3) 30 minutes-1 hour at 37°C in 2X SSC and 0.1% SDS; (4) 2 hours at 45-55°C in 2X SSC and 0.1% SDS, changing the solution every 30 minutes.

One common formula for calculating the stringency conditions required to achieve hybridization between nucleic acid molecules of a specified sequence homology (Sambrook et al., 1989):

\[ T_m = 81.5°C + 16.6 \log [Na^+] + 0.41(\% G+C) - 0.63 (\% \text{formamide}) - 600/\#\text{bp in duplex} \]

As an illustration of the above formula, using [Na+] = [0.368] and 50% formamide, with GC content of 42% and an average probe size of 200 bases, the Tm is 57°C. The Tm of a DNA duplex decreases by 1 - 1.5°C with every 1% decrease in homology. Thus, targets with greater than about 75% sequence identity would be observed using a hybridization temperature of 42°C. In one embodiment, the hybridization is at 37°C and the final wash is at 42°C; in another embodiment the hybridization is at 42°C and the final wash is at 50°C; and in yet another embodiment the hybridization is at 42°C and final wash is at 65°C, with the above hybridization and wash solutions. Conditions of high stringency include hybridization at 42°C in the above hybridization solution and a final wash at 65°C in 0.1X SSC and 0.1% SDS for 10 minutes.

Nucleic acids of the present invention may be maintained as DNA in any convenient cloning vector. In a preferred embodiment, clones are maintained in plasmid cloning/expression vector, such as pGEM-T (Promega Biotech, Madison, WI), pBluescript (Stratagene, La Jolla, CA), pCR4-TOPO (Invitrogen, Carlsbad, CA) or pET28a+ (Novagen, Madison, WI), all of which can be propagated in a suitable E. coli host cell.

Nucleic acid molecules of the invention include cDNA, genomic DNA, RNA, and fragments thereof which may be single-, double-, or even triple-stranded. Thus, this invention provides oligonucleotides (sense or antisense strands of DNA or RNA) having sequences capable of hybridizing with at least one sequence of a nucleic acid molecule of the present invention. Such oligonucleotides are useful as probes for detecting genes encoding proteins that
comprise the phenylpropanoid and flavonoid biosynthetic pathways or mRNA in test samples of plant tissue, e.g., by PCR amplification, or for the positive or negative regulation of expression genes encoding proteins that comprise the phenylpropanoid and flavonoid biosynthetic pathways at or before translation of the mRNA into proteins. Methods in which oligonucleotides or polynucleotides may be utilized as probes for such assays include, but are not limited to: (1) in situ hybridization; (2) Southern hybridization; (3) Northern hybridization; and (4) assorted amplification reactions such as polymerase chain reactions (PCR) (including RT-PCR) and ligase chain reaction (LCR).

Polypeptides encoded by nucleic acids of the invention may be prepared in a variety of ways, according to known methods. If produced in situ the polypeptides may be purified from appropriate sources, e.g., seeds, pericarps, or other plant parts.

Alternatively, the availability of nucleic acid molecules encoding the polypeptides enables production of the proteins using in vitro expression methods known in the art. For example, a cDNA or gene may be cloned into an appropriate in vitro transcription vector, such as pSP64 or pSP65 for in vitro transcription, followed by cell-free translation in a suitable cell-free translation system, such as wheat germ or rabbit reticulocytes. In vitro transcription and translation systems are commercially available, e.g., from Promega Biotech, Madison, WI, BRL, Rockville, MD or Invitrogen, Carlsbad, CA.

According to a preferred embodiment, larger quantities of polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways may be produced by expression in a suitable prokaryotic or eukaryotic system. For example, part or all of a DNA molecule, such as the cDNAs having SEQ ID NOs: 1-19, may be inserted into a plasmid vector adapted for expression in a bacterial cell (such as E. coli) or a yeast cell (such as Saccharomyces cerevisiae), or into a baculovirus vector for expression in an insect cell. Such vectors comprise the regulatory elements necessary for expression of the DNA in the host cell, positioned in such a manner as to permit expression of the DNA in the host cell. Such regulatory elements required for expression include promoter sequences, transcription initiation sequences and, optionally, enhancer sequences.

The polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways produced by gene expression in a recombinant prokaryotic or eukaryotic system may be purified according to methods known in the art. In a preferred embodiment, a commercially available expression/secretion system can be used, whereby the recombinant protein is expressed and thereafter secreted from the host cell, to be easily purified from the surrounding medium. If
expression/secretion vectors are not used, an alternative approach involves purifying the recombinant protein by affinity separation, such as by immunological interaction with antibodies that bind specifically to the recombinant protein. Such methods are commonly used by skilled practitioners.

The polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways of the invention, prepared by the aforementioned methods, may be analyzed according to standard procedures.

Polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways purified from coffee, or produced recombinantly, may be used to generate polyclonal or monoclonal antibodies, antibody fragments or derivatives as defined herein, according to known methods. Antibodies that recognize and bind fragments of the polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways of the invention are also contemplated, provided that the antibodies are specific for polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways. For example, if analyses of the proteins or Southern and cloning analyses (see below) indicate that the cloned genes belongs to a multigene family, then member-specific antibodies made to synthetic peptides corresponding to nonconserved regions of the protein can be generated.

Kits comprising an antibody of the invention for any of the purposes described herein are also included within the scope of the invention. In general, such a kit includes a control antigen for which the antibody is immunospecific.

Flavonoids play a role in many aspects of human health and wellness. Flavonoids have been demonstrated to be powerful antioxidants (Rice-Evans, C 1991), have been shown to inhibit oxidation of LDL in vitro (DeWhalley, CV et al. 1990; Frankel, EN et al. 1993; and Yan, LJ et al. 1995), are anti-hypertensive (Duarte, J et al. 2002), and may be anti-inflammatory (Yoshimoto et al. 1983; and Huang, MT et al. 1991). In addition, evidence indicates that dietary flavonoids reduce the risk of coronary heart disease mortality (Hertog, MG et al. 1993), and may be protective against certain cancers. (Yang, CS et al. 1998; Yamane, T et al. 1996; Gupta, S et al. 2001; Yamagishi, M et al. 2002; and Hertog, MG et al. 1994). This list of health benefits attributable to flavonoids is meant to be illustrative and not exhaustive, and it is presumed that there are many other beneficial health effects attributable to flavonoids presently unknown. Accordingly, the coffee polypeptides that comprise the biosynthetic pathways of phenylpropanoids and flavonoids described and exemplified herein are expected to find utility in a variety of food, health, and wellness applications. For example, the coffee polypeptides that
comprise the biosynthetic pathways of phenylpropanoids and flavonoids, or their respective flavonoid products, may be utilized as dietary supplements. In addition, the antioxidant properties of flavonoids may prove advantageous in both food and cosmetic products.

One or more of the aforementioned applications for the polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways may be pursued by exploiting the availability of the polynucleotides encoding polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathway described herein to generate significant quantities of pure protein using recombinant organisms (e.g., in the yeast *Pichia pastoris* or in food compatible *Lactobacilli*, or in plant cells), and then testing the proteins in new or established assays for antioxidant potential, antihypertensive potential, immunoproliferative potential, chemoprotective or chemotherapeutic potential, and the like. Similar testing may be carried out using the flavonoids produced by these proteins according to suitable means established or developed in the art. If specific purified proteins, or flavonoid products produced by such proteins are found to be particularly useful, natural versions of those proteins and their flavonoid products also may be isolated from coffee grains or other plant parts, or from tissues and organs of other plants enriched in those phenylpropanoid and flavonoid biosynthetic pathways enzymes.

**Vectors, Cells, Tissues and Plants:**

Also featured in accordance with the present invention are vectors and kits for producing transgenic host cells that contain a polynucleotide encoding polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways, or an oligonucleotide, or homolog, analog or variant thereof in a sense or antisense orientation, or a reporter gene and other constructs under control of cell or tissue-specific promoters and other regulatory sequences. Suitable host cells include, but are not limited to, plant cells, bacterial cells, yeast and other fungal cells, insect cells and mammalian cells. Vectors for transforming a wide variety of these host cells are well known to those of skill in the art. They include, but are not limited to, plasmids, phagemids, cosmids, baculoviruses, bacmids, bacterial artificial chromosomes (BACs), yeast artificial chromosomes (YACs), as well as other bacterial, yeast and viral vectors. Typically, kits for producing transgenic host cells will contain one or more appropriate vectors and instructions for producing the transgenic cells using the vector. Kits may further include one or more additional components, such as culture media for culturing the cells, reagents for performing transformation of the cells and reagents for testing the transgenic cells for gene expression, to name a few.
The present invention includes transgenic plants comprising one or more copies of a gene encoding a polypeptide that comprises the phenylpropanoid and flavonoid biosynthetic pathways, or nucleic acid sequences that inhibit the production or function of a plant's endogenous polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways. This is accomplished by transforming plant cells with a transgene that comprises part of all of a coding sequence for a polypeptide that comprises the phenylpropanoid and flavonoid biosynthetic pathways, or mutant, antisense or variant thereof, including RNA, controlled by either native or recombinant regulatory sequences, as described below. Transgenic plants from coffee species are preferred, including, without limitation, C. abeokutae, C. arabica, C. arnoldiana, C. aruwenensis, C. bengalensis, C. canephora, C. congensis C. dewevei, C. excelsa, C. eugenioides, and C. heterocalyx, C. kapakata, C. khasiana, C. liberica, C. moloundou, C. rasemosa, C. salvatrix, C. sessiflora, C. stenophylla, C. travencorensis, C. wightiana and C. zanguebariae. Transgenic plants of any species are also included in the invention; these include, but are not limited to, tobacco, Arabidopsis and other "laboratory-friendly" species, cereal crops such as maize, wheat, rice, soybean barley, rye, oats, sorghum, alfalfa, clover and the like, oil-producing plants such as canola, safflower, sunflower, peanut, cacao and the like, vegetable crops such as tomato, tomatillo, potato, pepper, eggplant, sugar beet, carrot, cucumber, lettuce, pea and the like, horticultural plants such as aster, begonia, chrysanthemum, delphinium, petunia, zinnia, lawn and turfgrasses and the like.

Transgenic plants can be generated using standard plant transformation methods known to those skilled in the art. These include, but are not limited to, Agrobacterium vectors, polyethylene glycol treatment of protoplasts, biolistic DNA delivery, UV laser microbeam, gemini virus vectors or other plant viral vectors, calcium phosphate treatment of protoplasts, electroporation of isolated protoplasts, agitation of cell suspensions in solution with microbeads coated with the transforming DNA, agitation of cell suspension in solution with silicon fibers coated with transforming DNA, direct DNA uptake, liposome-mediated DNA uptake, and the like. Such methods have been published in the art. See, e.g., Methods for Plant Molecular Biology (Weissbach & Weissbach, eds., 1988); Methods in Plant Molecular Biology (Schuler & Zielinski, eds., 1989); Plant Molecular Biology Manual (Gelvin, Schilperoort, Verma, eds., 1993); and Methods in Plant Molecular Biology - A Laboratory Manual (Maliga, Klessig, Cashmore, Gruissem & Varner, eds., 1994).

The method of transformation depends upon the plant to be transformed. Agrobacterium vectors are often used to transform dicot species. Agrobacterium binary vectors include, but are
not limited to, BIN19 and derivatives thereof, the pBI vector series, and binary vectors pGA482, pGA492, pLH7000 (GenBank Accession AY234330) and any suitable one of the pCAMBIA vectors (derived from the pPZP vectors constructed by Hajdukiewicz, Svab & Maliga, (1994) Plant Mol Biol 25: 989-994, available from CAMBIA, GPO Box 3200, Canberra ACT 2601, Australia or via the worldwide web at CAMBIA.org). For transformation of monocot species, biolistic bombardment with particles coated with transforming DNA and silicon fibers coated with transforming DNA are often useful for nuclear transformation. Alternatively, Agrobacterium “superbinary” vectors have been used successfully for the transformation of rice, maize and various other monocot species.

DNA constructs for transforming a selected plant comprise a coding sequence of interest operably linked to appropriate 5' regulatory sequences (e.g., promoters and translational regulatory sequences) and 3' regulatory sequences (e.g., terminators). In a preferred embodiment, a coding sequence encoding a polypeptide that comprises the phenylpropanoid and flavonoid biosynthetic pathways under control of its natural 5' and 3' regulatory elements is utilized. In other embodiments, coding and regulatory sequences are swapped (e.g., CcDFR coding sequence operably linked to the CcF3H promoter) to alter the protein content of the seed of the transformed plant for a phenotypic improvement, e.g., in flavor, aroma or other feature.

In an alternative embodiment, the coding region of the gene is placed under a powerful constitutive promoter, such as the Cauliflower Mosaic Virus (CaMV) 35S promoter or the figwort mosaic virus 35S promoter. Other constitutive promoters contemplated for use in the present invention include, but are not limited to: T-DNA mannopine synthetase, nopaline synthase and octopine synthase promoters. In other embodiments, a strong monocot promoter is used, for example, the maize ubiquitin promoter, the rice actin promoter or the rice tubulin promoter (Jeon et al., Plant Physiology. 123: 1005-14, 2000).

Transgenic plants with coding sequences to express polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways under an inducible promoter are also contemplated to be within the scope of the present invention. Inducible plant promoters include the tetracycline repressor/operator controlled promoter, the heat shock gene promoters, stress (e.g., wounding)-induced promoters, defense responsive gene promoters (e.g. phenylalanine ammonia lyase genes), wound-induced gene promoters (e.g., hydroxyproline rich cell wall protein genes), chemically-inducible gene promoters (e.g., nitrate reductase genes, glucanase genes, chitinase genes, etc.) and dark-inducible gene promoters (e.g., asparagine synthetase gene) to name only a few.
Tissue-specific and development-specific promoters are also contemplated for use in the present invention. Non-limiting examples of seed-specific promoters include Cim1 (cytokinin-induced message), cZ19B1 (maize 19 kDa zein), milps (myo-inositol-1-phosphate synthase), and celA (cellulose synthase) (U.S. application Ser. No. 09/377,648), bean beta.-phaseolin, napin, beta.-conglycinin, soybean lectin, cruciferin, maize 15 kDa zein, 22 kDa zein, 27 kDa zein, g-zein, waxy, shrunken 1, shrunken 2, and globulin 1, soybean 11S legumin (Bäumlein et al., 1992), and C. canephora 11S seed storage protein (Marraccini et al., 1999, Plant Physiol. Biochem. 37: 273-282). See also WO 00/12733, where seed-preferred promoters from end1 and end2 genes are disclosed. Other Coffea seed specific promoters may also be utilized, including but not limited to the oleosin gene promoter described in commonly-owned, co-pending PCT Application No. US2006/026121, the dehydrin gene promoter described in commonly-owned, co-pending PCT Application No. US2006/026234, and the 9-cis-epoxyvetanoid dioxygenase gene promoter described in commonly-owned, co-pending PCT Application No. US2006/34402. Examples of other tissue-specific promoters include, but are not limited to: the ribulose bisphosphate carboxylase (RuBisCo) small subunit gene promoters (e.g., the coffee small subunit promoter as described by Marracini et al., 2003) or chlorophyll a/b binding protein (CAB) gene promoters for expression in photosynthetic tissue; and the root-specific glutamine synthetase gene promoters where expression in roots is desired.

The coding region is also operably linked to an appropriate 3' regulatory sequence. In embodiments where the native 3' regulatory sequence is not use, the nopaline synthetase polyadenylation region may be used. Other useful 3' regulatory regions include, but are not limited to the octopine synthase polyadenylation region.

The selected coding region, under control of appropriate regulatory elements, is operably linked to a nuclear drug resistance marker, such as kanamycin resistance. Other useful selectable marker systems include genes that confer antibiotic or herbicide resistances (e.g., resistance to hygromycin, sulfonylurea, phosphinotricin, or glyphosate) or genes conferring selective growth (e.g., phosphomannose isomerase, enabling growth of plant cells on mannose). Selectable marker genes include, without limitation, genes encoding antibiotic resistance, such as those encoding neomycin phosphotransferase II (NEO), dihydrofolate reductase (DHFR) and hygromycin phosphotransferase (HPT), as well as genes that confer resistance to herbicidal compounds, such as glyphosate-resistant EPSPS and/or glyphosate oxidoreductase (GOX), Bromoxynil nitrilase (BXN) for resistance to bromoxynil, AHAS genes for resistance to
imidazolinones, sulfonylurea resistance genes, and 2,4-dichlorophenoxyacetate (2,4-D) resistance genes.

In certain embodiments, promoters and other expression regulatory sequences encompassed by the present invention are operably linked to reporter genes. Reporter genes contemplated for use in the invention include, but are not limited to, genes encoding green fluorescent protein (GFP), red fluorescent protein (DsRed), Cyan Fluorescent Protein (CFP), Yellow Fluorescent Protein (YFP), Cerianthus Orange Fluorescent Protein (coFp), alkaline phosphatase (AP), β-lactamase, chloramphenicol acetyltransferase (CAT), adenosine deaminase (ADA), aminoglycoside phosphotransferase (neo', G418') dihydrofolate reductase (DHFR), hygromycin-B-phosphotransferase (HPH), thymidine kinase (TK), lacZ (encoding α-galactosidase), and xanthine guanine phosphoribosyltransferase (XGPRT), Beta-Glucuronidase (gus), Placental Alkaline Phosphatase (PLAP), Secreted Embryonic Alkaline Phosphatase (SEAP), or Firefly or Bacterial Luciferase (LUC). As with many of the standard procedures associated with the practice of the invention, skilled artisans will be aware of additional sequences that can serve the function of a marker or reporter.

Additional sequence modifications are known in the art to enhance gene expression in a cellular host. These modifications include elimination of sequences encoding superfluous polyadenylation signals, exon-intron splice site signals, transposon-like repeats, and other such well-characterized sequences that may be deleterious to gene expression. Alternatively, if necessary, the G/C content of the coding sequence may be adjusted to levels average for a given coffee plant cell host, as calculated by reference to known genes expressed in a coffee plant cell. Also, when possible, the coding sequence is modified to avoid predicted hairpin secondary mRNA structures. Another alternative to enhance gene expression is to use 5' leader sequences. Translation leader sequences are well known in the art, and include the cis-acting derivative (omega') of the 5' leader sequence (omega) of the tobacco mosaic virus, the 5' leader sequences from brome mosaic virus, alfalfa mosaic virus, and turnip yellow mosaic virus.

Plants are transformed and thereafter screened for one or more properties, including the presence of the transgene product, the transgene-encoding mRNA, or an altered phenotype associated with expression of the transgene. It should be recognized that the amount of expression, as well as the tissue- and temporal-specific pattern of expression of the transgenes in transformed plants can vary depending on the position of their insertion into the nuclear genome. Such positional effects are well known in the art. For this reason, several nuclear transformants should be regenerated and tested for expression of the transgene.
Methods:

The nucleic acids and polypeptides of the present invention can be used in any one of a number of methods whereby the protein products can be expressed in coffee plants in order that the proteins may play a role in protecting the coffee plant from pathogens, and from herbivore or insect attack, and in the enhancement of flavor and/or aroma of the coffee beverage or coffee products ultimately produced from the bean of the coffee plant expressing the protein. Similarly, the polypeptides of the invention can be used in any one of a number of methods whereby the phenylpropanoids, flavonoids, and other such phytochemical products synthesized from the polypeptides may play a role in protecting the plant from pathogens, and from herbivore or insect attack, and in the enhancement of flavor and/or aroma of the coffee beverage or coffee products ultimately produced from the bean of the coffee plant containing the phenylpropanoids and flavonoids.

With respect to protection of the plant from attack by pathogens, herbivores, and insects, increasing evidence indicates that at least some flavonoids and some flavonoid secondary metabolites are produced at increased levels in certain plants as part of the defense response, or play a role in passive resistance to attack. (Lahtinen, M et al. (2004); Onyilagha, JC et al. (2004); Kotkar, HM et al. (2002); Lattanzio, V et al. (2000), and Christensen, AB et al. (1998)). Accordingly, the ability to manipulate production of polypeptides that comprise the biosynthetic pathway for phenylpropanoids and flavonoids in a plant, or even to use the polynucleotides and proteins of the invention to monitor such gene expression, will enable study and manipulation of the response of the coffee plant to pathogen, herbivore, or insect attack. This knowledge will enable the generation of modified coffee plants that are better equipped against disease or devastation by herbivores or insects.

With respect to flavor and aroma of roasted coffee grain, it is expected that the polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways exert some influence on the generation of coffee flavors via the Maillard reaction that occurs during roasting, by means of the content of the proteins themselves, or the products such as phenylpropanoids or flavonoids they produce. Proteins, and particularly protein degradation products (peptides and amino acids), represent an important group of flavor precursors (Spanier et al., 2004). Therefore, relatively abundant proteins such as those that comprise the phenylpropanoid and flavonoid biosynthetic pathways can be expected to make some contribution to the flavor generating reactions that occur during coffee roasting. Such a contribution may stem from the concentration of the proteins themselves in the coffee bean, or
the concentration of the phenylpropanoids or flavonoids ultimately produced from the proteins. The ability to monitor (e.g., through marker-assisted breeding) or manipulate protein expression profiles for polypeptides that comprise the phenylpropanoid or flavonoid biosynthetic pathway is provided by the polynucleotides of the present invention, in accordance with the methods described herein.

Thus, one aspect of the present invention features methods to alter the profile of polypeptides that comprise the phenylpropanoid or flavonoid biosynthetic pathway in a plant, preferably coffee, comprising increasing or decreasing an amount or activity of one or more polypeptides that comprise the phenylpropanoid or flavonoid biosynthetic pathway in the plant. For instance, in one embodiment of the invention, a gene encoding a polypeptide that comprises the phenylpropanoid or flavonoid biosynthetic pathway under control of its own expression-controlling sequences is used to transform a plant for the purpose of increasing production of that polypeptide in the plant. Alternatively, a coding region for a polypeptide that comprises the phenylpropanoid or flavonoid biosynthetic pathway is operably linked to heterologous expression controlling regions, such as constitutive or inducible promoters.

Loss-of-function (null) mutant plants may be created or selected from populations of plant mutants currently available. It will also be appreciated by those of skill in the art that mutant plant populations may also be screened for mutants that over-express or under-express a particular polypeptide that comprises the phenylpropanoid or flavonoid biosynthetic pathway, utilizing one or more of the methods described herein. Mutant populations can be made by chemical mutagenesis, radiation mutagenesis, and transposon or T-DNA insertions, or targeting induced local lesions in genomes (TILLING, see, e.g., Henikoff et al., 2004, Plant Physiol. 135(2): 630-636; Gilchrist & Haughn, 2005, Curr. Opin. Plant Biol. 8(2): 211-215). The methods to make mutant populations are well known in the art.

The nucleic acids of the invention can be used to identify mutant polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways in various plant species. In species such as maize or Arabidopsis, where transposon insertion lines are available, oligonucleotide primers can be designed to screen lines for insertions in the genes encoding polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways. Through breeding, a plant line may then be developed that is heterozygous or homozygous for the interrupted gene.

A plant also may be engineered to display a phenotype similar to that seen in null mutants created by mutagenic techniques. A transgenic null mutant can be created by
expressing a mutant form of a selected polypeptide that comprises the phenylpropanoid and flavonoid biosynthetic pathways to create a "dominant negative effect." While not limiting the invention to any one mechanism, this mutant protein will compete with wild-type protein for interacting proteins or other cellular factors. Examples of this type of "dominant negative" effect are well known for both insect and vertebrate systems (Radke et al., 1997, Genetics 145: 163-171; Kolch et al., 1991, Nature 349: 426-428).

Another kind of transgenic null mutant can be created by inhibiting the translation of mRNA encoding the polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways by "post-transcriptional gene silencing." The gene from the species targeted for down-regulation, or a fragment thereof, may be utilized to control the production of the encoded protein. Full-length antisense molecules can be used for this purpose. Alternatively, antisense oligonucleotides targeted to specific regions of the mRNA that are critical for translation may be utilized. The use of antisense molecules to decrease expression levels of a pre-determined gene is known in the art. Antisense molecules may be provided in situ by transforming plant cells with a DNA construct which, upon transcription, produces the antisense RNA sequences. Such constructs can be designed to produce full-length or partial antisense sequences. This gene silencing effect can be enhanced by transgenically over-producing both sense and antisense RNA of the gene coding sequence so that a high amount of dsRNA is produced (for example see Waterhouse et al., 1998, PNAS 95: 13959-13964). In this regard, dsRNA containing sequences that correspond to part or all of at least one intron have been found particularly effective. In one embodiment, part or all of the coding sequence antisense strand is expressed by a transgene. In another embodiment, hybridizing sense and antisense strands of part or all of the coding sequence for polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways are transgenically expressed.

In another embodiment, phenylpropanoid and flavonoid genes may be silenced through the use of a variety of other post-transcriptional gene silencing (RNA silencing) techniques that are currently available for plant systems. RNA silencing involves the processing of double-stranded RNA (dsRNA) into small 21-28 nucleotide fragments by an RNase H-based enzyme ("Dicer" or "Dicer-like"). The cleavage products, which are siRNA (small interfering RNA) or miRNA (micro-RNA) are incorporated into protein effector complexes that regulate gene expression in a sequence-specific manner (for reviews of RNA silencing in plants, see Horiguchi, 2004, Differentiation 72: 65-73; Baulcombe, 2004, Nature 431: 356-363; Herr, 2004, Biochem. Soc. Trans. 32: 946-951).
Small interfering RNAs may be chemically synthesized or transcribed and amplified in vitro, and then delivered to the cells. Delivery may be through microinjection (Tuschl T et al., 2002), chemical transfection (Agrawal N et al., 2003), electroporation or cationic liposome-mediated transfection (Brummelkamp TR et al., 2002; Elbashir SM et al., 2002), or any other means available in the art, which will be appreciated by the skilled artisan. Alternatively, the siRNA may be expressed intracellularly by inserting DNA templates for siRNA into the cells of interest, for example, by means of a plasmid, (Tuschl T et al., 2002), and may be specifically targeted to select cells. Small interfering RNAs have been successfully introduced into plants (Klahre U et al., 2002).

A preferred method of RNA silencing in the present invention is the use of short hairpin RNAs (shRNA). A vector containing a DNA sequence encoding for a particular desired siRNA sequence is delivered into a target cell by any common means. Once in the cell, the DNA sequence is continuously transcribed into RNA molecules that loop back on themselves and form hairpin structures through intramolecular base pairing. These hairpin structures, once processed by the cell, are equivalent to siRNA molecules and are used by the cell to mediate RNA silencing of the desired protein. Various constructs of particular utility for RNA silencing in plants are described by Horiguchi, 2004, supra. Typically, such a construct comprises a promoter, a sequence of the target gene to be silenced in the “sense” orientation, a spacer, the antisense of the target gene sequence, and a terminator.

Yet another type of synthetic null mutant can also be created by the technique of “co-suppression” (Vaucheret et al., 1998, Plant J. 16(6): 651-659). Plant cells are transformed with a copy of the endogenous gene targeted for repression. In many cases, this results in the complete repression of the native gene as well as the transgene. In one embodiment, a gene encoding a polypeptide that comprises the phenylpropanoid and flavonoid biosynthetic pathways from the plant species of interest is isolated and used to transform cells of that same species.

Mutant or transgenic plants produced by any of the foregoing methods are also featured in accordance with the present invention. Preferably, the plants are fertile, thereby being useful for breeding purposes. Thus, mutant or transgenic plants that exhibit one or more of the aforementioned desirable phenotypes can be used for plant breeding, or directly in agricultural or horticultural applications. They will also be of utility as research tools for the further elucidation of the participation of polypeptides that comprise the phenylpropanoid and flavonoid biosynthetic pathways in flavor, aroma and other features of coffee seeds associated with pigments and photosynthesis. Plants containing one transgene or a specified mutation may also
be crossed with plants containing a complementary transgene or genotype in order to produce plants with enhanced or combined phenotypes.

The present invention also features compositions and methods for producing, in a seed-preferred or seed-specific manner, any selected heterologous gene product in a plant. A coding sequence of interest is placed under control of a seed-specific coffee promoter and other appropriate regulatory sequences, to produce a seed-specific chimeric gene. The chimeric gene is introduced into a plant cell by any of the transformation methods described herein or known in the art. These chimeric genes and methods may be used to produce a variety of gene products of interest in the plant, including but not limited to: (1) detectable gene products such as GFP or GUS, as enumerated above; (2) gene products conferring an agronomic or horticultural benefit, such as those whose enzyme activities result in production of micronutrients (e.g., pro-vitamin A, also known as beta-carotene) or antioxidants (e.g., ascorbic acid, omega fatty acids, lycopene, isoprenes, terpenes); or (3) gene products for controlling pathogens or pests, such as described by Mourguès et al., (1998), TibTech 16: 203-210 or others known to be protective to plant seeds or detrimental to pathogens.

The following examples are provided to illustrate the invention in greater detail. The examples are intended illustrate, not to limit, the invention.

**Example 1**

**Materials and Methods for Subsequent Examples**

**Plant material.** Freshly harvested roots, young leaves, stems, flowers and fruit at different stages of development were harvested from Coffea arabica L. cv. Caturra T-2308 and young leaf tissues were harvested from Coffea canephora var. BP409 grown under greenhouse conditions at Tours (25°C, 70 RH). All other tissues from Coffea canephora BP-409 were grown in the field in East Java, Indonesia. The development stages are defined as follows: small green fruit (SG), large green fruit (LG), yellow fruit (Y) and red fruit (R). Fresh tissues were frozen immediately in liquid nitrogen, then stored at −80°C until used for RNA extraction.

**RNA Preparation.** Total RNA was extracted from the various tissues of Coffea arabica (T2308) and Coffea canephora (BP409) as described previously (Rogers et al. 1999). In the case of the coffee cherries from the Small Green (SG), Large Green (LG), Yellow (Y), and Red (R) stages, these were first separated into pericarp and grain tissues and then the RNA was extracted as described above.

**cDNA Preparation.** Four different methods of cDNA preparation were carried out:
Method 1: 1 μg of total RNA and 50 ng oligo dT$_{(18)}$ (Sigma) was dissolved in DEPC-treated water (12 μl final volume). This mixture was subsequently incubated at 70°C for 10 min and then rapidly cooled down on ice. Next, 4 μl of 5X first strand buffer (Invitrogen, Carlsbad, CA), 2 μl of DTT 0.1M (Invitrogen) and 1 μl of dNTP mix (10 mM each, Invitrogen), were added. The reaction mixes were incubated at 42°C for 2 min before adding 1 μl of SuperScript III Rnase H-Reverse transcriptase (200U/μl, Invitrogen). After addition of the enzyme, the reactions were incubated at 25°C for 10 min., then at 42°C for 50 min, followed by enzyme inactivation by heating at 70°C for 10 min. The cDNA samples generated were then diluted ten-fold in sterilized water and stored at −20°C. This cDNA was then used at different dilutions as noted in the various experiments outlined below.

Method 2: 1 μg total RNA sample (Table 1), plus 870 ng oligo dT (Proligo) was made up to a final volume of 13 μl with DEPC-treated water. This mixture was incubated at 65°C for 5 min to denature the nucleic acids, and the samples were then put on ice. Next, 4μl of 5x buffer Transcriptor RT Reaction Buffer (Roche), 10 U of Ribonuclease Inhibitor (Sigma), 1mM final of each dNTP (Roche) and 10 U of Transcriptor Reverse Transcriptase (20U/μL, Roche) were added. The 20μl final reaction mixes were mixed by vortexing and briefly centrifuged. The reactions were then incubated at 55°C for 50 min. 1 U of RNase H (Invitrogen) was then added to the reaction mixes, followed by an incubation at 37°C for 30 min. The samples were then stored at −20°C.

Method 3 for cDNA preparation was identical to Method 2, except different primers were used and 30 pmoles of Gene-Specific Primer (see Table 1 and Table 3) were substituted for the 870 ng oligo dT for the priming step. The samples were then stored at −20°C.

Method 4 for cDNA preparation closely followed the manufacturer’s protocol for First Strand cDNA Synthesis method in 5’ RACE system for Rapid Amplification of cDNA Ends kit (Invitrogen). In brief, 1 μg total RNA sample (Table 1), plus 10 pmoles Gene-Specific Primer (Tables 1 and 3) was diluted to a final volume of 15.5 μl with DEPC-treated water. This mixture was incubated at 70°C for 10 min to denature the nucleic acids, and the samples were then put on ice for 1 min and centrifuged. Next, 2.5μl of 10x PCR buffer (Invitrogen), 2.5 μl of MgCl$_2$, 25mM (Invitrogen), 2.5 μl of DTT 0.1M (Invitrogen) and 1 μl of dNTP mix (10 mM each, Invitrogen), were added to a final volume of 24 μl with DEPC-treated water. These reaction mixes were incubated at 42 °C for 1 min before adding 1 μl of SuperScript II Rnase H-Reverse Transcriptase (200U/μl, Invitrogen). After addition of enzyme, the reactions were incubated at
42 °C for 50 min, followed by enzyme inactivation by heating at 70 °C for 15 min. The tubes were then briefly centrifuged and 1 µL of RNase mix was then added to the reaction and incubated for 30 min at 37°C. The samples were then stored at -20°C.

**Table 1. cDNA Production for the Different 5’ RACE Experiments.**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>RNA origin variety + stage</th>
<th>cDNA synthesis method used</th>
<th>Primers Used for First Strand cDNA Synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>CcPAL1–5’ RACE</td>
<td>BP409, Flowers</td>
<td>Method 3</td>
<td>PAL1-RT1</td>
</tr>
<tr>
<td>CcPAL2–5’ RACE</td>
<td>BP409, Pericarp (mixed stages)</td>
<td>Method 2</td>
<td>oligo dT</td>
</tr>
<tr>
<td>CcPAL3–5’ RACE</td>
<td>BP409, Pericarp (mixed stages)</td>
<td>Method 4</td>
<td>PAL3-RT1</td>
</tr>
<tr>
<td>Cc4CL1–5’ RACE</td>
<td>BP409, Pericarp (mixed stages)</td>
<td>Method 3</td>
<td>4CL1-RT1</td>
</tr>
</tbody>
</table>

The sources of the RNA used, the cDNA synthesis method employed, and the primers used for first strand cDNA synthesis are given for each 5’ RACE Experiment. The DNA sequences of the primers are given in Table 3. In order to have a full representation of the pericarp RNA made at different stages, approximately equivalent amounts of pericarp tissue from stages small green, large green, yellow and red of total RNA were mixed after being prepared as described in the methods indicated in the table (see methods). BP409, Coffea canephora (robusta).

**5’ RACE Reactions (Rapid Amplification of cDNA Ends).** 5’ RACE reactions were carried out closely following the manufacturer’s protocol described in the kit for the 5’ RACE system for Rapid Amplification of cDNA Ends (Invitrogen). The cDNA preparations used in this experiment were first purified to remove any unincorporated nucleotides (as they would interfere in the dC tailing reaction) using S.N.A.P. columns (Invitrogen) according to the instructions given by the manufacturer. Once purified, the cDNA were recovered in 50µL of sterilized water and stored at -20°C before being used for 5’RACE PCR. The specific cDNA used for each reaction are noted in Table 1. The 5’ RACE experiments all began with a TdT tailing of each specific S.N.A.P. purified cDNA. The poly dC tailing reaction proceeded as follows: 25 µl reactions were set up with 5 µl of the purified cDNA, 11.5 µl of DEPC treated water, 5 ul 5x TdT tailing buffer (Invitrogen), and 2.5 µl of dNTPs. The reactions were then incubated at 94°C for 3 minutes, and chilled on ice. 1 µl of TdT was then added and the reaction
was incubated for 10 minutes at 37°C. The reactions were terminated by heating 10 minutes at 65°C and again placed on ice.

The first round 5' RACE PCR1 reactions were performed in a final 50 μl volume, as follows: 5μL of each tailed cDNA, 5 μl of 10 x Taq Polymerase Buffer (Stratagene buffer), 400 nM of the gene-specific-primer 1 (Tables 2 and 4) and of the Abridged Anchor Primer (AAP) (Table 4), 1 μl of dNTP mix (10 mM each, Invitrogen) and 2.5 U of Taq DNA polymerase (Stratagene). The first round PCR cycling conditions were: denaturing at 94°C for 2 min, then 94 °C for 1 min, annealing at the temperatures specified in Table 2 for 1 min 30 seconds, and extension at 72°C for 3 min, for 45 or 55 cycles (Table 2). An additional final extension step was carried out at 72°C for 7 min. The PCR products were then analyzed by agarose gel electrophoresis and ethidium bromide staining.

The second round of PCR reactions were performed in a final volume of 50 μl volume, as follows: 5μL of 1% diluted PCR1 (First Round) product; 5 μl of 10 x Taq Polymerase Buffer (Stratagene buffer), 200 nM of gene-specific-primer 2 (Tables 2 and 4) and of the Abridged Universal Amplification Primer (AUAP), (Table 4), 1 μl of dNTP mix (10 mM each, Invitrogen) and 2.5 U of Taq DNA polymerase (Stratagene). The reaction protocol was: denaturing at 94°C for 2 min, then 94 °C for 1 min, annealing at the temperature specified in Table 2 for 1 min 30, and extension at 72°C for 3 min for 45, 50, or 55 cycles (Table 2). An additional final extension step was carried out at 72°C for 7 min. PCR products were then analyzed by agarose gel electrophoresis and ethidium bromide staining.

Table 2. Primers and PCR Conditions Used for the Different 5' RACE Experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Gene Specific Primers (SEQ ID No.)</th>
<th>Annealing Temperature</th>
<th>Number of cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CcPAL1–5' RACE</strong></td>
<td>PAL1-GSP1 (103)</td>
<td>50°C</td>
<td>45</td>
</tr>
<tr>
<td>First round RACE PCR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CcPAL2–5' RACE</strong></td>
<td>PAL2-GSP1 (104)</td>
<td>55°C</td>
<td>45</td>
</tr>
<tr>
<td>First round RACE PCR</td>
<td>PAL2-GSP2 (105)</td>
<td>53°C</td>
<td>50</td>
</tr>
<tr>
<td>Second round RACE PCR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CcPAL3–5' RACE</strong></td>
<td>PAL3-GSP1 (106)</td>
<td>55°C</td>
<td>55</td>
</tr>
<tr>
<td>First round RACE PCR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second round RACE PCR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CcACL1–5' RACE</strong></td>
<td>4CL1-GSP1 (108)</td>
<td>55°C</td>
<td>45</td>
</tr>
<tr>
<td>First round RACE PCR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second round RACE PCR</td>
<td></td>
<td>4CL1-GSP2 (109)</td>
<td>53°C 50</td>
</tr>
</tbody>
</table>
The primers, annealing temperatures, and the number of cycles are given for the various 5' RACE PCR reactions. The DNA sequences of the primers are given in Table 4.

**Table 3. List of Gene Specific Primers used for First Strand cDNA Synthesis experiments.**

<table>
<thead>
<tr>
<th>Primers</th>
<th>Sequences</th>
<th>SEQ ID NO:</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAL1-RT1</td>
<td>$5^\prime$ GACGTAAGAGCTCCATCC $3^\prime$</td>
<td>98</td>
</tr>
<tr>
<td>PAL3-RT1</td>
<td>$5^\prime$ GGCCCTCAAGTCTCCTC $3^\prime$</td>
<td>99</td>
</tr>
<tr>
<td>4CL1-RT1</td>
<td>$5^\prime$ CATACTTATCCACCACAGG $3^\prime$</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 4. List of primers used for 5'Race PCR experiment**

<table>
<thead>
<tr>
<th>Primers</th>
<th>Sequences</th>
<th>SEQ ID NO:</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAP</td>
<td>$5^\prime$ GGCCACACGCTCGACTAGTACGGGGG $3^\prime$</td>
<td>101</td>
</tr>
<tr>
<td>AUAP</td>
<td>$5^\prime$ GGCCACACGCTCGACTACG $3^\prime$</td>
<td>102</td>
</tr>
<tr>
<td>PAL1-GSP1</td>
<td>$5^\prime$ CTCGCGTGGCTATCCATC $3^\prime$</td>
<td>103</td>
</tr>
<tr>
<td>PAL2-GSP1</td>
<td>$5^\prime$ CCTCGAAATGCCCTACGATC $3^\prime$</td>
<td>104</td>
</tr>
<tr>
<td>PAL2-GSP2</td>
<td>$5^\prime$ CCTGATTGTGGTTCGGGC $3^\prime$</td>
<td>105</td>
</tr>
<tr>
<td>PAL3-GSP1</td>
<td>$5^\prime$ TTCAATATGTTCACACGCTTCTG $3^\prime$</td>
<td>106</td>
</tr>
<tr>
<td>PAL3-GSP2</td>
<td>$5^\prime$ TGACGTCTTGTTGGTTGTGCTC $3^\prime$</td>
<td>107</td>
</tr>
<tr>
<td>4CL1-GSP1</td>
<td>$5^\prime$ GCAATTGTACCTATATATTTTGCAG $3^\prime$</td>
<td>108</td>
</tr>
<tr>
<td>4CL1-GSP2</td>
<td>$5^\prime$ CCACCAAGCAACACAGAAATTGAC $3^\prime$</td>
<td>109</td>
</tr>
</tbody>
</table>

**DNA Sequencing.** Plasmid DNA was purified using Qiagen kits according to the instructions given by the manufacturer. Plasmid DNA and PCR products were sequenced by GATC Biotech AG (Konstanz, Germany) using the dideoxy termination method (Sanger et al., 1977). In some cases, the unique PCR fragments produced from the 5' RACE and genome walking experiments were directly sequenced, without purification or cloning, using the same primers as in the PCR amplification reactions. Computer analysis was performed using Laser Gene software package (DNASTAR). Homologies with sequences in the public GenBank database were verified using different BLAST programs (Altschul et al. 1990).

**Northern Blotting.** Total RNA was extracted from the various tissues of *Coffea arabica* (T2308) as described (Rogers et al. 1999). In the case of the coffee cherries from the Small
Green (SG), Large Green (LG), Yellow (Y), and Red (R) stages, these were first separated into pericarp and grain tissues and then the RNA was extracted from each.

A total of 5 μg of RNA was run on a 1.2% (w/v) denaturing RNA gel containing formaldehyde. The total RNA samples from each plant tissue were heated at 65°C for 15 min in presence of 7 μL “RNA Sample Loading Buffer” (without ethidium bromide, Sigma), and then put immediately on ice for 2 minutes before being loaded onto the 1.2% RNA gel. The gels were run at 60 Volts for 5 hours. The gel was then soaked twice in 10× SSC for 20 min. The RNA in the gel was transferred overnight by capillary transfer to a “Positive TM Membrane” (Qbiogene) in 10× SSC, and the RNA was fixed by heating the blot for 30min at 80°C. Probes were generated using “Rediprime™ II random prime labelling system” kit (Amersham) in the presence of (P³²) dCTP. Hybridization was carried out at 65°C for 24 h in hybridization solution (5X SSC, 40μg/ml Denatured Salmon Sperm DNA, 5% [w/v] SDS, and 5× Denhardt’s solution). Then, the membrane was washed twice at 65°C using 2X SSC, 0.1% SDS [w/v] and 1X SSC, 0.1% SDS [w/v] for 30 minutes each wash.

The probes used were hybridized with the appropriate ³²P-dCTP labeled PCR fragment obtained after PCR amplification with T7 (³⁷-TAATACGACTCAGTATAGGG) (SEQ ID NO: 110) and T3 (³⁷-ATTAACCCCTCACTGAGGGA) (SEQ ID NO: 111) primers and the corresponding clone.

**Semi-quantitative RT-PCR.** RNA was extracted from the various tissues of Coffea arabica and Coffea canephora, and cDNA was prepared according to Method 1, described above. The cDNA samples generated were then diluted ten-fold and 1 μl samples of the diluted material were used for semi-quantitative RT-PCR under the following reaction conditions (final volume 50 μl): 1 x PCR reaction buffer (Stratagene), 2 % DMSO (v/v), gene specific primers at 600 nM each (Table 5), 2.5 units of Taq DNA polymerase (Stratagene). The cycling protocol was as follows: 94°C for 2 min; then 94°C for 1 min, 60°C for 1 min 30 seconds and 72°C for 1 min and 27, 30 or 35 cycles (number of cycles noted in Figure 17). An additional final step of elongation was carried out for 7 minutes at 72°C. 15 μl of the PCR products were analyzed by agarose gel electrophoresis and ethidium bromide staining. The oligonucleotides (CcCHIS, CcCHI, CcCHI-like, or GOS2 specific primers) used for the RT-PCR reactions and length of PCR products obtained are listed in Table 5.
Table 5. Sequences of the CcCHS, CcCHI, CcCHI-like and GOS2 gene-specific primers used to perform the PCR reactions steps for RT-PCR.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Primers</th>
<th>Sequences</th>
<th>SEQ ID NO:</th>
<th>Length of PCR product</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOS2</td>
<td>Gos26</td>
<td>5’TACCCGACCCGAAACCCCAATT 3’</td>
<td>112</td>
<td>485 bp</td>
</tr>
<tr>
<td></td>
<td>Gos27</td>
<td>5’ACACCAGATGAATGCACACTG 3’</td>
<td>113</td>
<td></td>
</tr>
<tr>
<td>CcCHS</td>
<td>cccp8j10-FWD</td>
<td>5’GATCCCCGTTCCGAAAGTAGAGAG 3’</td>
<td>114</td>
<td>503 bp</td>
</tr>
<tr>
<td></td>
<td>cccp8j10-REV</td>
<td>5’CATGATTACTTTGAATCGTGCGCCG 3’</td>
<td>115</td>
<td></td>
</tr>
<tr>
<td>CcCHI</td>
<td>cccp22k18-FWD</td>
<td>5’CCCACTGGGAGCTCTATTCGTTG 3’</td>
<td>116</td>
<td>272 bp</td>
</tr>
<tr>
<td></td>
<td>cccp22k18-REV</td>
<td>5’CCCCGTCGCGTCATGTTTGC 3’</td>
<td>117</td>
<td></td>
</tr>
<tr>
<td>CcCHI-like</td>
<td>cccp12o15-FWD</td>
<td>5’GCTATAATTTCGGGCCCCGGTGGAC 3’</td>
<td>118</td>
<td>497 bp</td>
</tr>
<tr>
<td></td>
<td>cccp12o15-REV</td>
<td>5’GAAGACATGGAATCCCAACACCAG 3’</td>
<td>119</td>
<td></td>
</tr>
</tbody>
</table>

Real time RT-PCR (QRT-PCR) experiments. The cDNA used for these experiments was prepared according to Method 1, described above. TaqMan-PCR was performed as recommended by the manufacturer (Applied Biosystems, Perkin-Elmer). All reactions contained 1x TaqMan buffer (Perkin-Elmer) and 5mM MgCl₂, 200μM each of dATP, dCTP, dGTP and dUTP, and 0.625 units of AmpliTaq Gold polymerase. PCR was carried out using 800nM of each gene specific primers, forward and reverse, and 200nM TaqMan probe. Primers and probes were designed using PRIMER EXPRESS software (Applied Biosystems) (Table 6). Reaction mixtures were incubated for 2 min at 50°C and 10 min at 95°C, followed by 40 amplification cycles of 15 sec at 95°C/1 min at 60°C.

Table 6. List of primers and TaqMan probes for the quantitative RT-PCR experiments.

<table>
<thead>
<tr>
<th>Primers and Probes</th>
<th>Sequences</th>
<th>SEQ ID NO:</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpl39-F1</td>
<td>5’GAACAGGCCCATCCCTTTG 3’</td>
<td>120</td>
</tr>
<tr>
<td>rpl39-R1</td>
<td>5’CGCGCGTTTGAGTTGTA 3’</td>
<td>121</td>
</tr>
<tr>
<td>rpl39-MGB1</td>
<td>5’ATGCACCTGACACA 3’</td>
<td>122</td>
</tr>
<tr>
<td>CcPAL1-F1</td>
<td>5’GTCAACACCTCTCACTTCCAAAA 3’</td>
<td>123</td>
</tr>
<tr>
<td>CcPAL1-R1</td>
<td>5’TGTAGAGAGACGCTTCTAGT 3’</td>
<td>124</td>
</tr>
<tr>
<td>CcPAL1-MGB1</td>
<td>5’ATTGTGCATTGAGG 3’</td>
<td>125</td>
</tr>
<tr>
<td>CcPAL2-F2</td>
<td>5’GCTCGGCTACCCCTGTGGAA 3’</td>
<td>126</td>
</tr>
<tr>
<td>CcPAL2-R2</td>
<td>5’CACCCAGAATAACCAACACTGAATCTG 3’</td>
<td>127</td>
</tr>
</tbody>
</table>
Samples were quantified in the GeneAmp 7500 Sequence Detection System (Applied Biosystems). Quantification of transcript levels was carried out using the method of relative quantification, using the constitutively expressed ribosomal protein *rpl39* as the reference. In order to use the method of relative quantification, it was necessary to show that the amplification efficiency for the gene sequences was roughly equivalent to the amplification efficiency of the reference sequence (*rpl39* cDNA sequence) using the specifically defined primer and probe sets. To determine this relative equivalence, plasmid DNA containing the appropriate cDNA sequences were diluted 1/1000, 1/10,000, 1/100,000, and 1/1,000,000, and using the Q-PCR conditions described above, the slope of the curve *Ct* = *f*(Log quantity of DNA) was calculated for each plasmid/primer/TaqMan probe set. Plasmid/primer/TaqMan probe sets giving curves with slopes close to 3.32, which represents an efficiency of 100%, are considered acceptable. The plasmid/primer/TaqMan probe sets used (Table 6) all gave acceptable values for *Ct* = *f*(Log quantity of DNA).

**Isolation of the 5' Coding Sequences of CcPAL2, Ca4CL2 and Cc4CL2 Using Primer-Assisted Genome Walking.** Total genomic DNA was extracted from fresh leaves of *C.*
*arabica* T2308 and *C. canephora* BP409 harvested from the greenhouse in Tours using the method of Crouzillat *et al.*, (1996). Primer-assisted walking was performed using the Universal GenomeWalker kit (BD Biosciences) according to the manufacturer’s protocol. The eight GenomeWalker libraries used here were constructed from the *C. arabica* T2308 and the *C. canephora* BP409 genomic DNA that had been digested with the restriction enzymes DraI, EcoRV, PvuI, StuI, and then blunt-end ligated to the GenomeWalker Adaptor of the Universal GenomeWalker kit. The genomic DNA digestions and the GenomeWalker Adaptor ligation reactions were carried out in accordance with the kit user manual.

The eight libraries were then employed as templates in PCR reactions using the PAL2-GSP gene-specific primers, PAL2-GW-GSP1 and PAL2-GW-GSP2 (SEQ ID NOs: 143, 144) or those for 4CL2-GSP (4CL2-GW-GSP1 and 4CL2-GW-GSP2 (SEQ ID NOs: 145, 146)) (Table 7). The PCR reaction mixtures contained 1 µL of GenomeWalker library template, 5 µl 10x PCR buffer (LA buffer II Mg²⁺ plus), 200 µM of each dNTP, 400 nM of each primer (AP1 and either PAL2-GW-GSP1 or 4CL2-GW-GSP1) and 0.5 U of DNA polymerase Takara LA Taq (Cambrex Bio Science) in a final volume of 50 µl. The following conditions were used for the first round of PCR: after denaturing at 94°C for 2 min, the first seven cycles were performed at 94°C for 25 seconds and annealing/elongation at 72°C for 3 min. A further 31 cycles were carried out at 94°C for 25 seconds and annealing and elongation at 67°C for 3 min. An additional final step of elongation was done at 67°C for 7 min. The second PCR reaction was set up exactly as described above for the first round, except the DNA substrate was 1 µl of the first amplification reaction which had been diluted 1/50. The PCR cycling conditions were as follows: 5 cycles of denaturing at 94°C for 25 seconds and annealing/elongation at 72°C for 3 min. A further 25 cycles were carried out at 94°C for 25 seconds, followed by 3 minutes at an annealing/elongation temperature of 67°C. An additional final step of elongation was carried out at 67°C for 7 min. The resulting PCR fragments were analyzed by agarose gel electrophoresis.

Three PCR products were obtained and cloned in pCR4-TOPO vector. One PCR product was obtained with the *C. canephora* BP409/DraI digested library (the sequence obtained was called GW1_CePAL2 (SEQ ID NO: 44), which was cloned into pCR4-TOPO to generate the plasmid pML18. A second PCR product was obtained with the *C. canephora* BP409/EcoRV digested library (the sequence obtained was called GW1_Ce4CL2), which was cloned into pCR4-TOPO to generate the plasmid pML21. A third PCR product was obtained with the *C.
arabica T2308/StuI digested library (the sequence obtained was called GW1_Ca4CL2), which was cloned into pCR4-TOPO to generate the plasmid pML22.

Table 7. List of primers used for GenomeWalker experiments.

<table>
<thead>
<tr>
<th>Primers</th>
<th>Sequences</th>
<th>SEQ ID NO:</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP1</td>
<td>5'GTAATACGACTCACTATAGGGGC3'</td>
<td>141</td>
</tr>
<tr>
<td>AP2</td>
<td>5'ACTATAGGGCCACGCGTGGT3'</td>
<td>142</td>
</tr>
<tr>
<td>PAL2-GW-GSP1</td>
<td>5'TCCCTCCATGCAAAGCTTTTCTCG3'</td>
<td>143</td>
</tr>
<tr>
<td>PAL2-GW-GSP2</td>
<td>5'GGTCCTAGGCTGCGGTGAAGTACGA3'</td>
<td>144</td>
</tr>
<tr>
<td>4CL2-GW-GSP1</td>
<td>5'GCCCACATAGCAGAAAATTGAGTCAGCA3'</td>
<td>145</td>
</tr>
<tr>
<td>4CL2-GW-GSP2</td>
<td>5'GGAAGAGGCAAAACGCAATCATCACT3'</td>
<td>146</td>
</tr>
</tbody>
</table>

Isolation of cDNA Containing the Complete ORF’s for Coffea arabica CaPAL1 and CaPAL3 Using Gene-Specific Primers. The existing canephora cDNA sequence (pccccwc22w18n3) (SEQ ID NO: 39) found in the database, and the new 5’ canephora sequence obtained from 5’ RACE (Race1_CcPAL1) (SEQ ID NO: 40) were used to design 2 specific primers in the 5’ and 3’ UTR sequences to amplify the complete ORF sequence of CaPAL1 (pML8) (SEQ ID NO: 1). To amplify the complete ORF sequence of CaPAL3 (pML14) (SEQ ID NO: 2), the cDNA sequence found in the database (pccc1611) (SEQ ID NO: 45), and the 5’ region of the published sequence from Coffea canephora PAL1 (GenBank Accession Number: AF460203) (SEQ ID NO: 47), which has high similarity to the pccc1611 and Race1_CcPAL3 sequences (SEQ ID NOs: 45, 46), served to design 2 specific primers in the 5’ and 3’ UTR sequences of this gene. The cDNA used to isolate the complete ORF sequences are noted in Table 8, and the sequences of the specific primers for each PCR reaction are given in Table 9.

The PCR reactions were performed in 50 μl reactions as follows: 5μl of cDNA from Coffea arabica T2308 (Table 8), 5 ul 10 x PCR buffer (La PCR Buffer II Mg++ plus), 600 nM of the each gene specific primer, 200μM of each dNTP, and 0.5 U of DNA polymerase Takara LA Taq (Cambrex Bio Science). After denaturing at 94°C for 2 min, amplification was carried out according to the following protocol: 40 cycles of denaturing at 94°C for 1 minute, 1 min at the specified annealing temperature (Table 8), and 2 min of extension at 72°C. An additional final step of extension was carried out at 72°C for 7 min. The PCR products were analyzed by agarose gel electrophoresis and ethidium bromide staining. Fragments of the expected size were then cloned in pCR4-TOPO using TOPO TA Cloning.
Kit for Sequencing (Invitrogen) according to the instructions given by the manufacturer. The cloned inserts were then sequenced.

Table 8. Isolation of cDNA sequences encoding the full length protein sequences for *Coffea arabica* PAL1 (CaPAL1), PAL3 (CaPAL3), 4CL2 (Ca4CL2) and *Coffea canephora* 4CL2 (Cc4CL2).

<table>
<thead>
<tr>
<th>Gene</th>
<th>cDNA (Source of RNA)</th>
<th>Gene Specific Primers</th>
<th>Annealing Temperature</th>
<th>SEQ ID NOs:</th>
</tr>
</thead>
<tbody>
<tr>
<td>CaPAL1</td>
<td>T2308*, G-Y</td>
<td>PAL1-FullUp1 / PAL1-FullLow1</td>
<td>47°C</td>
<td>147/148</td>
</tr>
<tr>
<td>CaPAL3</td>
<td>T2308*, G-Y</td>
<td>PAL3-FullUp1 / PAL3-FullLow1</td>
<td>52°C</td>
<td>149/150</td>
</tr>
<tr>
<td>Ca4CL2</td>
<td>T2308*, P-Y</td>
<td>4CL2-FullUp1 / 4CL2-FullLow1</td>
<td>55°C</td>
<td>151/152</td>
</tr>
<tr>
<td>Cc4CL2</td>
<td>BP409*, P-Y</td>
<td>4CL2-FullUp1 / 4CL2-FullLow1</td>
<td>56°C</td>
<td>151/152</td>
</tr>
</tbody>
</table>

The specific cDNA, primers, and PCR annealing temperatures used to amplify the complete ORF sequences are presented. G-Y, grain at the yellow developmental stage; P-Y, pericarp at the yellow developmental stage. These cDNA were synthesized using Method 1.

**Isolation of cDNA Containing the Complete ORF of 4CL2 from *Coffea canephora* and *Coffea arabica* Using Gene-Specific Primers.** The existing 4CL cDNA sequences found in the database, and the new 5’ sequences obtained from primer assisted genome walking experiments were used to design 2 gene-specific primers in the 5’ and 3’ UTR sequences of the coffee 4CL2 gene (Tables 8 and 9) to amplify the complete ORF sequence from *Coffea arabica* and *Coffea canephora* (Table 8). The amplified products were Ca4CL2 (pGCl) (arabica) (SEQ ID NO: 5) and Cc4CL2 (pGC3) (canephora) (SEQ ID NO: 6). The cDNA used to isolate the complete ORF’s are noted in Table 8, and the sequences of the specific primers for each PCR reaction are given in Table 9.

The PCR reactions were performed using the Takara LA Taq DNA polymerase (Cambrex Bio Science). The PCR reactions were performed in 50 µl reactions as follows: 3µL of cDNA (Table 8), 5 µl 10 x PCR buffer (La PCR Buffer II Mg⁺⁺ plus), 300 nM of the each gene specific primer, 200µM of each dNTP, and 0.5 U of DNA polymerase Takara LA Taq (Cambrex Bio Science). After denaturing at 94°C for 2 min, the amplification reaction was carried out as follows: 25 cycles of 1 min at 94 °C, 1 min at annealing temperature (noted in Table 8) and extension for 2 min 30 seconds at 72°C. An additional final step of extension was carried out at 72°C for 7 min.
The PCR products were analyzed by agarose gel electrophoresis and ethidium bromide staining. Fragments of the expected size were then cloned in pCR4-TOPO using TOPO TA Cloning Kit for Sequencing (Invitrogen) according to the instructions given by the manufacturer. The cloned inserts were then sequenced.

Table 9. Sequences of the primers used for the amplification of cDNA sequences encoding the full length protein sequences of CaPAL1, CaPAL3, Ca4CL2 and Cc4CL2.

<table>
<thead>
<tr>
<th>Primers</th>
<th>Sequences</th>
<th>SEQ ID NO:</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAL1-FullUp1</td>
<td>5'-CACCACGTCTACTGCTTC-3'</td>
<td>147</td>
</tr>
<tr>
<td>PAL1-FullLow1</td>
<td>5'-ACATTGAAGGATTGATAA-3'</td>
<td>148</td>
</tr>
<tr>
<td>PAL3-FullUp1</td>
<td>5'-ATGGAATGGCTAATGGGAAATG-3'</td>
<td>149</td>
</tr>
<tr>
<td>PAL3-FullLow1</td>
<td>5'-TTCAAACATTATGCGAAACGAAC-3'</td>
<td>150</td>
</tr>
<tr>
<td>4CL2-FullUp1</td>
<td>5'-TAGCTCGTACATTAACCTCAACA-3'</td>
<td>151</td>
</tr>
<tr>
<td>4CL2-FullLow1</td>
<td>5'-TCGACAATCACCACAGTAACTCG-3'</td>
<td>152</td>
</tr>
</tbody>
</table>

EXAMPLE 2

Isolation and Characterization of Coffea cDNA Clones Encoding Phenylalanine Ammonia Lyases (PAL)

To find cDNA encoding coffee phenylalanine ammonia lyases, the protein sequences of biochemically characterized PAL proteins *Petroselinum crispum* PAL1 (GenBank Accession Number CAA68938, Appert et al., 1994) (SEQ ID NO: 54) and of *Zea Mays* PAL1 (GenBank Accession Number AAL40137, Röslter et al., 1997) (SEQ ID NO: 50) were used as the query sequences for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm. The first search with the *P. crispum* PAL1 protein sequence uncovered 3 unigenes exhibiting relatively high levels of homology: #121018 (e value=0), #120370 (e value =e-123) and #119778 (e value= e-116). The second search with PAL1 from *Zea Mays* uncovered the same 3 unigenes #121018 (e value=0), #120370 (e value =e-85) and #119778 (e value= e-81).

*Coffea arabica* CaPAL1 (Full ORF). The first unigene found with high similarity to plant PAL proteins was unigene #121018. The longest cDNA of this unigene (pccw22w18n3) (SEQ ID NO: 39) was isolated from the 22 weeks whole cherries library and sequenced. The insert of pccw22w18n3 (SEQ ID NO: 39) was found to be 1637 bp long, and to encode a partial protein sequence of 470 amino acids. Because the full length *P. crispum* protein was 716 amino acids long, it was assumed that the coffee PAL encoded by pccw22w18n3 (SEQ ID NO: 39) was missing over 246 amino acids at the N terminal. To find the missing 5’ coding
region of this gene, specific primers were designed from the 5’ end of the sequence in pcccwc22w18n3 (SEQ ID NO: 39) for use in the well-established technique of 5’ RACE PCR.

A unique fragment of approximately 1050 bp was obtained using cDNA prepared from C. canephora (BP-409) flower RNA and Method 3 described in Example 1, above. The primers and the RACE PCR1 conditions are noted in Tables 1-4 and in Example 1. The PCR fragment obtained was sequenced directly (without purification) using the specific PCR amplification primer PAL1-GSP1 and an additional primer (5’ GTGCCCTTGTCTACTCTCCATC 3’) (SEQ ID NO: 153). The resulting sequence (Race1_CcPAL1) (SEQ ID NO: 40) was 1037 bp long and, as expected, overlapped the 5’ end of the sequence in pcccwc22w18n3 (SEQ ID NO: 39) (Figure 2, 140 bp of overlapping sequence).

This C. canephora sequence (Race1_CcPAL1) (SEQ ID NO: 40) and the insert of pcccwc22w18n3 (SEQ ID NO: 39), allowed the design of two new primers capable of specifically amplifying the complete ORF sequence of the coffee CaPAL1. The amplification of the complete ORF was carried out using cDNA made by Method 1 from RNA of C. arabica (T2308) grain (yellow stage; see Table 8 and Table 9) and following PCR conditions noted in Example 1 and in Table 8. This experiment generated the plasmid pML8 (Figures 2A and 2B), which contained the cDNA sequence (SEQ ID NO: 1) for CaPAL1 (SEQ ID NO: 20). Sequence analysis of the pML8 insert (SEQ ID NO: 1) indicated that this cDNA was 2344 bp long, with a complete ORF of 2136 bp that encoded a polypeptide (SEQ ID NO: 20) of 711 amino acids having an estimated molecular weight of 77.10 kDa.

An alignment of the complete ORF encoded by pML8 (CaPAL1) (SEQ ID NO: 1) and two complete Coffea canephora PAL protein sequences available in public databases, CcPAL1 (GenBank Accession Number AAN32866) (SEQ ID NO: 48) and CcPAL2 (GenBank Accession Number AAN32867) (SEQ ID NO: 49) is presented in Figure 5A. These data demonstrate that CaPAL1 (SEQ ID NO: 20) has 83.1% and 74.3% identity at the amino acid level with the full length Coffea canephora CcPAL1 and CcPAL2 sequences (SEQ ID NOs: 48, 49), demonstrating that CaPAL1 is a new unique coffee PAL sequence. In addition, two partial PAL sequences from Coffea arabica were present in the NCBI database (GenBank Accession Numbers AAF27654 and AAF27655).

An alignment of CaPAL1 (pML8) (SEQ ID NO: 1) with biochemically-characterized PAL proteins (Figure 5B) shows that the newly isolated CaPAL1 coffee sequence exhibits 67.9%, 81.6%, 80.3%, 80 %, 85%, 84.8% and 85.2% identity with Zea Mays PAL1 (SEQ ID NO: 50), Arabidopsis thaliana PAL1, PAL2, PAL4 (SEQ ID NOs: 51, 52, 53) and the P.
*crispum* PAL1, PAL2 and PAL3 sequences (SEQ ID NOs: 54, 55, 56), respectively. These data confirm the annotation of the newly discovered ORF of pML8 as a coffee PAL protein (Figure 5B). This alignment also shows that at protein level, the CaPAL1 sequence (SEQ ID NO: 20) presented here shares 69.8% and 83.1% identity with the isolated partial CaPAL2 (described below) and the complete CaPAL3 contained in pML14 (SEQ ID NO: 2) (described below).

**Coffee canephora CcPAL2 (partial ORF).** The second coffee unigene sequence found with high similarity to the plant PAL proteins was unigene #119778. It was not possible to recover the longest cDNA of this unigene (pcccl25c13) from the libraries, so the original database EST sequence for this EST was used. The next longest EST clone pcccp19k7 was isolated and sequenced. The 840bp EST for pcccl25c13 (SEQ ID NO: 42) and the 673 bp cDNA contained in pcccp19k7 (SEQ ID NO: 41) were aligned using the Seqman software (DNASTAR). The resulting contiguous sequence showed that the two sequences were identical over the 431bp overlapping region, thereby confirming they both belong to the same cDNA (Figure 3).

The length of the unique sequence resulting from the contig of these two sequences was found to be 1082 bp long, and to encode a partial ORF sequence of 261 amino acids. Because the full length *P. crispum* PAL1 protein was 716 amino acids long, this coffee cDNA was assumed to be missing over 454 amino acids at the N terminal. To isolate the 5' end of this coffee PAL, specific primers were first designed for a 5' RACE PCR reaction. Using cDNA prepared from *C. canephora* (BP-409) pericarp RNA (all stages mixed, Table 1) according to the cDNA synthesis Method 2 described in Example 1, above, the primers PAL2-GSP1 and PAL2-GSP2 (SEQ ID NOs: 104, 105) (Tables 1-4), and the RACE PCR conditions described in Example 1 and in Table 2, a unique fragment of approximately 400 bp was obtained. The PCR fragment was sequenced without purification using the specific primer PAL2-GSP2 (SEQ ID NO: 105) that was employed for the amplification of this fragment. The resulting sequence (Race1_CcPAL2) (SEQ ID NO: 43) was 406 bp long and, as expected, overlapped the 5' end of the sequence ccccl25c13 (SEQ ID NO: 42) (Figure 3B, 63 bp of overlapping sequence).

Because the newly isolated *C. canephora* 5' end sequence (Race1_CcPAL2) (SEQ ID NO: 43) was still missing the 5' end, a genome walker experiment was performed on eight libraries to try to recover the 5' sequence. These experiments used the primers PAL2-GW-GSP1 and PAL2-GW-GSP2 (SEQ ID NOs: 143, 144) (Table 7) designed from the 5' end of the *C. canephora* Race1_CcPAL2 sequence (SEQ ID NO: 43) and were carried out following Primer-
Assisted Genome Walking PCR conditions described in Example 1. A unique PCR fragment of approximately 2100 bases pair (estimation of length in gel) was obtained with the *C. canephora* BP409/DraI digested library, cloned in the pCR4-TOPO vector and sequenced using the T3 and T7 universal primers and two gene specific internal primers. This resulting plasmid was called pML18, and the insert sequence (GW1_CcPAL2) (SEQ ID NO: 44) was found to be 2118 bp long and to overlap the sequence Race1_CcPAL2 (SEQ ID NO: 43) by 67 bp (Figure 3A).

The various *canephora* PAL2 sequences noted in Figure 3A were aligned using Seqman software (DNASTAR). The alignment showed that there was 100% identity in the overlapping regions (Figure 3B), meaning that these sequences are from the same gene. The gene associated with the full 3476 bp contig generated was named *CcPAL2* (SEQ ID NO: 3). The longest ORF in the *CcPAL2* contiguous sequence was found to be 1752 bp (position 1430 to 3181) because the first 1429 first bp of the genomic sequence GW1_CcPAL2 (SEQ ID NO: 44) were associated with a large intron. The partial ORF of the CcPAL2 contig sequence encoded a partial protein of 583 aa. Based on the alignment of this 583 amino acid ORF with the full length *P. crispum* PAL1 protein (SEQ ID NO: 54) (716 aa) using Clustal-W (Figure 5B), it is assumed that the coffee cDNA for *CcPAL2* (SEQ ID NO: 3) is a partial, and its ORF is missing over 134 amino acids at the N-terminal end. Additional 5'RACE or an additional primer assisted genome walk procedures are expected to isolate the missing 5' coding sequence of the *CcPAL2* (SEQ ID NO: 3) gene.

The alignment of the partial *CcPAL2* protein sequence (SEQ ID NO: 22) with the two *Coffea canephora* PAL sequences available in the NCBI databank and with the two other coffee PAL sequences described here, shows that CcPAL2 (SEQ ID NO: 22) is a unique coffee PAL (Figure 5A). The identity scores from this alignment show that the partial CcPAL2 protein sequence (SEQ ID NO: 22) has 85.2% and 85.2% identity with the publicly available full length *Coffea canephora* PAL1 (GenBank Accession Number AAN32866) (SEQ ID NO: 48) and PAL2 (GenBank Accession Number AAN32867) (SEQ ID NO: 49) sequences, and 69.8% with the CaPAL1 (SEQ ID NO: 20) and 85.1% with the CaPAL3 (SEQ ID NO: 21) described herein.

Another alignment (not shown) revealed that the CcPAL2 protein sequence (SEQ ID NO: 22) also shares 84.5% and 83.8% identity with the two partial PAL publicly available from *Coffea arabica* (GenBank Accession Numbers AAF27654 and AAF27655, respectively) over the region in common for all three sequences. These data confirm that the newly discovered partial ORF of CcPAL2 represents a new coffee PAL protein sequence.
Another alignment between the partial CcPAL2 protein sequence, several biochemically characterized PAL proteins, and the two other coffee PAL sequences presented here is shown in Figure 5B. The newly discovered partial CcPAL2 sequence exhibits 69.6%, 81.6%, 81.5%, 82%, 83.5%, 83.4% and 83.9% identity with Zea Mays PAL1 (SEQ ID NO: 50), Arabidopsis thaliana PAL1, PAL2, PAL4 (SEQ ID NOs: 51, 52, 53) and the P. crispum PAL1, PAL2 and PAL3 (SEQ ID NOs: 54, 55, 56) sequences, respectively. The alignment data strongly supports the argument that the partial ORF of CcPAL2 represents a coffee PAL protein sequence. It also clearly shows that CcPAL2 (SEQ ID NO: 22) is a different gene product from the new CaPAL1 (SEQ ID NO: 20) and CaPAL3 (SEQ ID NO: 21) presented here, exhibiting 69.8% and 85.1% identity with these sequences.

*Coffea arabica CaPAL3 (Full ORF).* The third coffee unigene sequence found with high similarity to the plant PAL proteins was unigene #120370. The longest cDNA of this unigene (pccep161l) (SEQ ID NO: 45) was isolated and sequenced. The insert of the clone pccep161l (SEQ ID NO: 45) was found to be 1060 bp long, and to encode a partial ORF sequence of 261 amino acids. Because the full length P. crispum protein was 716 amino acids long, it was assumed that the coffee PAL encoded by pccep161l (SEQ ID NO: 45) was missing over 455 amino acids at the N terminal end. To recover the missing 5’ end, specific primers were designed for use in 5’ RACE PCR.

cDNA was prepared from *C. canephora* (BP-409) pericarp RNA (all stages mixed, Table 1), according to the cDNA synthesis Method 4. Using the primers and PCR conditions described in Tables 1-4 and in Example 1, a fragment of approximately 400 bp was obtained. This fragment was cloned in the pCR4-TOPO vector and sequenced. The resulting plasmid (pML19) had an insert sequence (Race1_CcPAL3) of 377 bp (SEQ ID NO: 46). As expected, the insert of pML19 overlapped the 5’ end of the sequence in pccep161l (SEQ ID NO: 45) (Figure 4, 107 bp of overlapping sequence), but the sequence Race1_CcPAL3 (SEQ ID NO: 46) was still missing the 5’ end of the coding sequence of the coffee PAL3 sequence (SEQ ID NO: 2).

When the 1330 bp contig from pccep161l (SEQ ID NO: 45) together and the RACE fragment Race1_CcPAL3 (SEQ ID NO: 46) was aligned with a published *Coffea canephora* PAL sequence (GenBank Accession Number AAN32866) (SEQ ID NO: 48), it was determined that these three sequences were nearly identical, showing 99.2% identity over the regions in common. This fact allowed amplification of the complete ORF of the coffee PAL3 by using the
published *Coffea canephora* PAL1 sequence (GenBank Accession Number AAN32866) (SEQ ID NO: 48) to design a 5’ specific primer for this gene and a 3’ specific primer from the sequence in pcccpl611 plasmid. Using cDNA made by Method 1 from RNA of *C. arabica* (T2308) grain (yellow stage, see Tables 8 and 9) and following PCR conditions described in Example 1 and Table 8, this experiment resulted in the generation of the cDNA sequence CaPAL3 (SEQ ID NO: 2) contained in the plasmid pML14 (Figures 4A and 4B). Sequence analysis of the pML14 insert (SEQ ID NO: 2) indicated that this cDNA was 2318bp long, with a complete ORF of 2154 bp, which encodes a polypeptide (SEQ ID NO: 20) of 717 amino acids having an estimated molecular weight of 77.93 kDa.

The alignment in Figure 5A demonstrates that sequence of CaPAL3 (pML14) (SEQ ID NO: 2) is different from the new coffee PAL sequences CaPAL1 (SEQ ID NO: 1) and CcPAL2 (SEQ ID NO: 3) presented here (83.1% and 85.1%, respectively). However, this alignment also confirms that CaPAL3 (SEQ ID NO: 2) is almost identical to PAL1 and PAL2 from *Coffea canephora* (both clones have very similar sequences and thus appear to be allelic). This observation indicates that CaPAL3 (SEQ ID NO: 2) is actually a new, potentially arabica-specific allele of the coffee PAL 3 gene. It is believed that all three sequences are alleles of the *Coffea* PAL3 gene. The two other previously described partial arabica PAL protein sequences (GenBank Accession Numbers AAF27654 and AAF27655) are 99.3% and 95.9% identical to the CaPAL3 sequence (SEQ ID NO: 21) (over the region in common with the publicly available sequences), respectively, indicating that these are alleles of the CaPAL3 product. Alignment of the complete CaPAL3 protein sequence (SEQ ID NO: 21) was also carried out with several biochemically-characterized PAL protein sequences (Figure 5B). This alignment demonstrates that the CaPAL3 coffee sequence exhibits high similarity with plant PAL proteins, showing 67.6%, 80.2%, 80.3%, 79.6%, 82%, 81.7% and 82.6% identity with *Zea mays* PAL1 (SEQ ID NO: 50), *Arabidopsis thaliana* PAL1, PAL2, PAL4 (SEQ ID NOs: 51, 52, 53) and the *P. crispum* PAL1, PAL2 and PAL3 (SEQ ID NOs: 54, 55, 56) sequences, respectively. This observation confirms previous claims that this sequence encodes a PAL protein.

**EXAMPLE 3**

**Isolation and Characterization of a *Coffea canephora* cDNA Clone Encoding Cinnamate-4-Hydroxylase (CcC4H)**

*Coffea canephora* CcC4H (Full ORF). To find cDNA encoding coffee cinnamate-4-hydroxylase (C4H), two protein sequences encoding biochemically characterized C4H, AtC4H (*Arabidopsis thaliana*, GenBank Accession Number # BAA24355, Mizutani *et al.*, 1997) (SEQ
ID NO: 57) and PbC4H (*Populus balsamifera*, GenBank Accession Number #AAG50231, Ro et al., 2001) (SEQ ID NO: 58) served as query sequences for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm.

Using *Arabidopsis thaliana* AtC4H protein sequence, only one unigene (#124550) was found to exhibit a very high level of identity (% identity = 77%, e value = 0, score = 785). The other hits presented too low percent of identity with AtC4H (≤29% identity, e-value ≥ 3e-52, scores ≤ 200) to potentially encode C4H.

The second search with PbC4H (SEQ ID NO: 58) from *Populus balsamifera* uncovered the same best hit #124550 (e value=0, score = 869). The longest cDNA representing the 5’ end of the unigene #124550 (pccl27h22) that seemed to encode a complete coding sequence of a C4H, was then isolated from the leaves library, sequenced, and named CcC4H (SEQ ID NO: 4). The insert of pccl27h22 was determined to be 1668 bp long, and to encode an ORF (coding) sequence of 1518 bp. The deduced protein sequence (SEQ ID NO: 23) shows a protein of 505 amino acids having a predicted molecular weight of 58.05 kDa.

An alignment of the protein sequence (SEQ ID NO: 23) of pccl27h22 with the C4H protein sequences from *A. thaliana* (SEQ ID NO: 57), *P. balsamifera* (SEQ ID NO: 58) and MsC4H (*Medicago sativa*, GenBank Accession Number P37114) (SEQ ID NO: 158) demonstrated that pccl27h22 encoded protein (CcC4H) (SEQ ID NO: 23) shares 85%, 85.8% and 89.9% identity, respectively, with these protein sequences (Figure 6). The alignment data indicates that the pccl27h22 plasmid contains a full length cDNA (SEQ ID NO: 4) encoding a *C. canephora* cinnamate-4-hydroxylase (CcC4H) (SEQ ID NO: 23). An alignment of the complete coding DNA sequence (5’UTR-ORF-3’UTR) contained in pccl27h22, with the complete coding DNA sequences (5’UTR-ORF-3’UTR) of C4H sequences from *A. thaliana* (D78596) and *P. balsamifera* (AF302495) was performed using ClustalW method in MegAlign software. The alignment demonstrates that the CcC4H DNA sequence (SEQ ID NO: 4) contained in pccl27h22 shares 71% and 75.6 % identities with the respective public DNA sequences at DNA level. In this analysis, the complete DNA coding sequences were aligned (i.e., 5’ UTR, complete ORF, and 3’ UTR sequences were included).

**EXAMPLE 4**

Isolation and Characterization of *Coffea* cDNA Clones Encoding 4-Coumarate:Coenzyme A Ligases (C4CL1, C4CL2, and C4CL2)

*Coffea canephora* Cc4CL1 consensus (partial ORF). To find cDNA encoding coffee 4-coumarate:coenzyme A ligases, two protein sequences encoding biochemically-characterized
4CL activities served as query sequences for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm: Nt4CL1, *Nicotiana tabacum*, (GenBank Accession Number O24145, Lee et al., 1996) (SEQ ID NO: 67), and At4CL1, *Arabidopsis thaliana*, (GenBank Accession Number Q42524, Ehting et al., 1999) (SEQ ID NO: 63). The first search with the 4CL1 protein sequence from *N. tabacum* revealed 3 unigenes; #119670 (e value= e-151), #128581 (e value = e-102) and #123098 (e value = e-101) that exhibit high levels of homology to Nt4CL1 (SEQ ID NO: 67). The second search with 4CL1 sequence from *A. thaliana* (SEQ ID NO: 63) uncovered the same three unigenes #119670 (e value= e-132), #128581 (e value = 2e-71), and #123098 (e value= 2e-86).

The longest cDNA representing the 5' end of the unigene #119670 (pcccp27d21) (SEQ ID NO: 59) encoding the partial coding sequence of a 4CL, was isolated from the pericarp library and sequenced. The insert of pcccp27d21 (SEQ ID NO: 59) was found to be 1124 bp long, and to encode a partial ORF of 316 amino acids. Because the full length *N. tabacum* protein (SEQ ID NO: 67) was 547 amino acids long, it was assumed that the coffee 4CL encoded by the cDNA contained in pcccp27d21 (SEQ ID NO: 59) was missing over 231 amino acids at the N terminal.

To recover the 5' missing end, specific primers were designed for use in the well established technique of 5' RACE PCR. cDNA was prepared from *C. canephora* (BP-409) pericarp tissue, all stages mixed, (Table 1) (small green, large green, yellow and red) by Method 3 described in Example 1, above. Using the primers and the PCR conditions described in Tables 1-4 and in Example 1, a unique fragment of approximately 350 bp was obtained (estimation by length in gel), cloned in pCR4-TOPO vector (TOPO TA Cloning Kit for Sequencing (Invitrogen), and sequenced using the universal primer T3. The resulting sequence (Race1_Cc4CL1) (SEQ ID NO: 60) was 335 bp long and, as expected, overlapped the 5' end of the sequence in pcccp27d21 (SEQ ID NO: 59) (Figure 7, 55 bp of overlapping sequence), although the newly isolated Race1_Cc4CL1 (SEQ ID NO: 60) still did not contain the full 5' end of this gene.

The consensus sequence resulting from the alignment of Race1_Cc4CL1 (SEQ ID NO: 60) with sequence contained in pcccp27d21 (SEQ ID NO: 59) was 1404 bp long, and called Cc4CL1 (SEQ ID NO: 6). The longest ORF found in Cc4CL1 (SEQ ID NO: 6) was found to be 1230 bp long and encoded a partial protein sequence of 409 amino acids (SEQ ID NO: 25).

Based on length of *Nicotiana tabacum* 4CL1, it is likely another 139 aa at the N terminal must be recovered to obtain the full length Cc4CL1. Additional 5'RACE or an additional primer
assisted genome walker will be employed to isolate the full 5' missing end of the *Cc4CL1*
sequence.

A multiple alignment of the partial *Cc4CL1* protein sequence (SEQ ID NO: 25) was
done with biochemically-characterized 4CL sequences from *Arabidopsis thaliana* and *Nicotiana
tabacum* (SEQ ID NOs: 63, 67), available in GenBank. (Figure 9). This alignment confirmed
the initial annotation of this coffee sequence using BLAST, i.e., the partial ORF of *Cc4CL1*
(SEQ ID NO: 6) encodes a coffee 4CL protein (SEQ ID NO: 25). The alignment shows 81.9 %
identity between partial Cc4CL1 sequence (SEQ ID NO: 25) and the Nt4CL2 protein sequence.

*Coffeea arabica Ca4CL2 (complete ORF) and Coffeea canephora Cc4CL2 (complete
ORF).* The longest cDNA representing the 5' end of the unigene #123098 (pcccs30w16n14)
encoding a partial coding sequence for a 4CL, was isolated from the 30 weeks grain library (30
weeks after flowering) and sequenced. The insert of pcccs30w16n14 was found to be 946 bp
long, and to encode a partial ORF sequence of 217 amino acids. Further sequence analysis
suggested that the sequences in unigene #128581 (singleton pcccl24i21) and in unigene #123098
could actually belong to the same gene. A subsequent alignment of the completely sequenced
plasmids pcccl24i21 (insert length is 1778bp) and pcccs30w16n14 sequences confirmed that the
sequences of both plasmids could be assembled into a single contig sequence.

The alignment between the DNA sequences encoding the ORF regions contained in these
plasmids show the protein sequences are 98.9% identical in the overlapping region, with the
differences being due to single nucleotide polymorphisms. This result indicates that the
sequences contained in the plasmids pcccl24i21 and pcccs30w16n14 represent two alleles of the
same gene termed Cc4CL2.

Based on sequences comparisons with other plant 4CL proteins, it was determined that
the ORF of pcccl24i21 was missing approximately 44 amino acids at the N terminal end. Using
the 5' end sequence of the insert in pcccl24i21, specific primers were designed for a primer
assisted genome walking to isolate the missing 5' end of this gene (primers 4CL2-GW-GSP1
and 4CL2-GW-GSP2 (SEQ ID NOs: 145, 146), Table 7), following the PCR conditions
described in Example 1.

A unique PCR fragment of approximately 1100 bases pair (estimation of length in gel)
was obtained with the *C. canephora* BP409/EcoRV digested library (sequence obtained called
GW1_Cc4CL2) (SEQ ID NO: 62). A second unique PCR fragment of approximately 1550
bases pair (estimation of length in gel) was obtained with the *C. arabica* T2308/StuI digested

- 57 -
library (sequence obtained called GW1_Ca4CL2) (SEQ ID NO: 61). These two fragments (GW1_CC4CL2 and GW1_Ca4CL2) (SEQ ID NOs: 62, 61) were then cloned in pCR4-TOPO vector and sequenced. The fragment GW1_CC4CL2 (SEQ ID NO: 62) was 1102 bp long and overlapped the sequence contained in pcccl24i21 over 579 bp (Figure 8). The second fragment, GW1_Ca4CL2 (SEQ ID NO: 61), was 1547 bp long and overlapped the sequence contained in pcccl24i21 over 571 bp (Figure 8). Over 800 bp of DNA sequence upstream of the 4CL2 start codon (ATG) was recovered in GW1_Ca4CL2 (SEQ ID NO: 61). This upstream sequence includes the 5' UTR sequence of the 4CL2 gene as well as a significant portion of the promoter sequence for this gene. These newly isolated 5' end 4CL sequences and the sequence in the cDNA pcccl24i21 allowed the design of two new primers (see Table 8 and Table 9) capable of specifically amplifying the complete ORF of the coffee 4CL2.

Using cDNA generated by Method 1 from C. arabica (T2308) pericarp RNA (yellow stage: see Table 8) and primers and PCR conditions described in Example 1 and Tables 8 and 9, a plasmid named pGC1 was produced containing a fragment of 1771 bp whose sequence was named Ca4CL2 (SEQ ID NO: 5) (Figures 8A and 8B). Sequence analysis indicated that pGC1 contained a complete ORF of 1626 bp that encoded a polypeptide (SEQ ID NO: 24) of 541 amino acids having an approximate molecular weight of 59.4 kDa.

Using cDNA generated by Method 1 from C. canephora (BP409) pericarp RNA (yellow stage: see Table 8) and primers and PCR conditions described in Example 1 and Tables 8 and 9, a plasmid named pGC3 was produced containing an insert of 1771 bp long whose sequence was called Cc4CL2 (SEQ ID NO: 7) (Figures 8A and 8B). Sequence analysis indicated that pGC3 contained a complete ORF of 1626 bp that encoded a polypeptide (SEQ ID NO: 26) of 541 amino acids having an approximate molecular weight of 59.5 kDa.

A manually-optimized alignment was then carried out between the protein sequences (SEQ ID NOs: 24, 26) encoded by pGC1 (Ca4CL2) and pGC3 (Cc4CL2) (SEQ ID NOs: 5, 7) and biochemically-characterized 4CL sequences from Arabidopsis thaliana and Nicotiana tabacum, available from GenBank (Figure 9). This alignment shows identity levels between 54.3% to 81.1% for the two coffee sequences and the other database 4CL sequences. The identity between the arabica and canephora sequences was determined to be 98.3%, supporting the argument that these sequences are alleles of the same coffee 4CL gene (i.e., 4CL2). The multiple alignment also shows 81.1% identity between the Ca4CL2 and Nt4CL1 protein sequences, and 81% identity between the Cc4CL2 and Nt4CL1 protein sequences. Moreover, the alignment shows that Ca4CL2 and Cc4CL2 (SEQ ID NOs: 5, 7) encode protein different
from the Cc4CL1 protein described above, as the multiple alignment shows 77.8% of identity between Ca4CL2 (SEQ ID NO: 24) and Cc4CL1 (SEQ ID NO: 25) and between Cc4CL2 (SEQ ID NO: 26) and Cc4CL1 (SEQ ID NO: 25) proteins.

**EXAMPLE 5**

**Isolation and Characterization of a *Coffeea canephora* cDNA Clone Encoding Chalcone Synthase (CcCHS)**

To find coffee cDNA encoding chalcone synthase, two protein sequences encoding chalcone synthase (CHS) served as query sequences for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm: HaCHS, *Hypericum androsaemum*, GenBank Accession Number AAG30295 (SEQ ID NO: 69); and LeCHS *Lycopersicon esculentum*, GenBank Accession Number CAA38981 (SEQ ID NO: 70). Using the *Hypericum androsaemum* HaCHS protein sequence, only one unigene (#123889) was found to exhibit a very high level of identity (% identity = 90%, e value = 0, score = 723). The second search with PhC4H from *Lycopersicon esculentum* LeCHS uncovered the same best hit #123889 (% identity = 91%, e value=0, score = 721).

A cDNA representing the 5' end of the unigene #123889 (pcccp8j10) (SEQ ID NO: 97) was isolated from the pericarp library and sequenced. The insert of pcccp8j10 (SEQ ID NO: 97) was found to be 1397 bp long and to encode a complete ORF sequence of 1176 bp, and was named CcCHS (SEQ ID NO: 8). The deduced protein sequence encodes a protein of 391 amino acids (SEQ ID NO: 27) having a predicted molecular weight of 42,92 kDa. A multiple alignment (ClustalW) of the deduced protein sequence (SEQ ID NO: 27) encoded by pcccp8j10 (SEQ ID NO: 97) with the highly related CHS protein sequences from *Hypericum androsaemum* and *Lycopersicon esculentum* (SEQ ID NOs: 69, 70) and the related characterized stilbene synthases VSS from *Vitis* (GenBank Accession Number AAB19887) (SEQ ID NO: 71) and PsSTS from *Pinus strobes* (GenBank Accession Number CAA87013) (SEQ ID NO: 72) demonstrates that CcCHS shares 90%, 90.5%, 75.7% and 70.8% identity with these protein sequences, respectively (Figure 10), and supports the argument that pcccp8j10 (SEQ ID NO: 97) encodes a CHS protein rather than a VSS protein.

An alignment of the complete coding DNA sequence (5'UTR-ORF-3'UTR) contained in pcccp8j10 (SEQ ID NO: 97) with the complete coding DNA sequences (5'UTR-ORF-3'UTR) of CHS sequences from *Hypericum androsaemum* (AF315345) and *Lycopersicon esculentum* (X55195) was performed using ClustalW method in MegAlign software. The alignment demonstrates that the CcCHS complete coding sequence contained in pcccp8j10 (SEQ ID NO:
97) shares 72.8% and 70.2 % identity with the respective public DNA sequences at DNA level. In this analysis, the complete DNA coding sequences were aligned (i.e., 5' UTR, complete ORF and 3' UTR sequences were included).

**EXAMPLE 6**

**Isolation and Characterization of *Coffea canephora* cDNA Clones Encoding Chalcone Reductases (CcCHR1, CcCHR2A, and CcCHR2B)**

To find cDNA encoding chalcone reductases, two protein sequences encoding chalcone reductases (CHR) served as query sequences for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm: SrCHR, *Sesbania rostrata*, GenBank Accession Number CAA11226 (SEQ ID NO: 73); and PICHR *Pueraria Montana* var. *lobata*, GenBank Accession Number AAM12529 (SEQ ID NO: 74). (Joung *et al.* (2003)). Using the SrCHR protein sequence from *Sesbania rostrata*, 3 unigenes with high levels of homology were uncovered: unigenes #125260 (e value=e-118), #125256 (e value =e-116) and #125257 (e value= 3e-94). The second search with PICHR from *Pueraria Montana* var. *lobata* uncovered the same 3 unigenes: #125256 (e value=2e-84), #125260 (e value =4e-83), #125257 (e value= 1e-66).

*Coffea canephora* CcCHR1 (full ORF). A cDNA representing the 5' end of the unigene #125256 (pcccp24e9) was isolated from the pericarp library and sequenced. The insert of pcccp24e9 was found to be 1271 bp long and to encode a complete ORF sequence of 972 bp, which was named CcCHR1 (SEQ ID NO: 9). The deduced protein sequence reveals a protein of 323 amino acids (SEQ ID NO: 28) having a predicted molecular weight of 36.02 kDa. A manually optimized alignment of the deduced protein sequence encoded by pcccp24e9 was performed with the CHR protein sequences SrCHR from *Sesbania rostrata* (GenBank Accession Number CAA11226) (SEQ ID NO: 73), PICHR from *Pueraria Montana* var. *lobata* (GenBank Accession Number AAM12529) (SEQ ID NO: 74) and MsCHR from *Medicago sativa* (GenBank Accession Number AAB41555 (SEQ ID NO: 75), biochemical and crystallographic information available Bomati *et al.* (2005b)). This alignment demonstrates that CcCHR1 protein (SEQ ID NO: 28) shares 60.7% and 48.9% and 48.3% identity with the protein sequences SrCHR, PICHR and MsCHR (SEQ ID NOs: 73, 74, 75) noted above, respectively, (Figure 11), and supports the initial annotation of pcccp24d9 as a *C. canephora* chalcone reductase.

*Coffea canephora* CcCHR2A AND CcCHR2B (full ORF). The longest cDNA representing the 5' end of the unigene #125257 (pcccl28k6), which appeared to encode a coffee
1205 bp long and to encode a nearly complete ORF sequence of 972 bp (only the initial MET appears to be missing). As this sequence was different from CcCHR1 (SEQ ID NO: 8), it was called CcCHR2A (SEQ ID NO: 10). The deduced protein sequence is a polypeptide (SEQ ID NO: 29) of 323 amino acids having a predicted molecular weight of 36.11 kDa. Figure 11 shows an alignment of the deduced protein sequence (SEQ ID NO: 29) encoded by pcccl28k6 (CcCHR2A) (SEQ ID NO: 10) with the CcCHR1 protein from Coffea canephora (SEQ ID NO: 28 described above) and CHR protein sequences SrCHR from Sesbania rostrata (GenBank Accession Number CAA11226) (SEQ ID NO: 73), PCHR from Pueraria Montana var. lobata (GenBank Accession Number AAM12529; activity characterized by Joung et al. (2003) (SEQ ID NO: 74) and MsCHR from Medicago sativa (GenBank Accession Number AAB41555) (SEQ ID NO: 75). This alignment shows that the protein encoded by pcccl28k6 shares 61.6%, 49.8% and 49.2% identity with the protein sequences SrCHR, PCHR and MsCHR (SEQ ID NOs: 73, 74, 75), respectively. This alignment also indicates that the ORF in pcccl28k6 is nearly complete, missing only the N-terminal amino acid (met) and thus, with the addition of this met, can be considered full length. It is noted that the CcCHR2A protein sequence (SEQ ID NO: 29) exhibits a higher identity with CcCHR1 (79.5%) than with the publicly available CHR sequences presented in Figure 11.

The longest cDNA representing the 5' end of the unigene #125260 (pcccl26f18) was isolated from the leaf library and sequenced. The insert of pcccl26f18 was found to be 1377 bp long, and was also found to contain one unusually short insertion of 24 bp in an otherwise complete ORF. Because this 24bp insertion is not in frame, and when spliced from the sequence results in a complete ORF, it is believed to be an intron sequence. The ORF generated after in-silico splicing is 987 bp long, and encodes a protein sequence of 328 amino acids having a predicted molecular weight of 36.73 kDa.

Alignment of the spliced insert DNA sequence of pcccl26f18 with the insert DNA sequence of pcccl28k6 (CcCHR2A) (SEQ ID NO: 10) using ClustalW revealed that these two sequences exhibit 99.1% identity, and therefore are presumably allelic. Accordingly, pcccl26f18 was called CcCHR2B (SEQ ID NO: 11). This alignment also shows that within the 1207 bp overlapping sequence, there are only nine single nucleotide differences in the ORF sequence and only 2 single nucleotide differences and a 2 base pair insertion in the 3' UTR of pcccl26f18 (SEQ ID NO: 11).
Figure 11 shows the alignment of the protein sequence (SEQ ID NO: 30) encoded by pcccl26f18 (CcCHR2B) (SEQ ID NO: 11) with the CcCHR1 (SEQ ID NO: 28) and the CcCHR2A (SEQ ID NO: 29) protein sequences from *Coffea canephora* (described above) and CHR protein sequences SrCHR, PICH and MsCHR (SEQ ID NOs: 73, 74, 75). The alignment demonstrates that CcCHR2B (SEQ ID NO: 30) protein shares 61.3%, 48.9% and 48.3% identity with SrCHR, PICH and MsCHR (SEQ ID NOs: 73, 74, 75) protein sequences, respectively, and supports the initial annotation of pcccl26f18 as a *C. canephora* chalcone reductase. This multiple alignment also shows that the CcCHR2B (SEQ ID NO: 30) protein has 79.5% and 97.3% identity with CcCHR1 (SEQ ID NO: 28) and CcCHR2A (SEQ ID NO: 29), respectively. The complete coding DNA sequence CcCHR1 (5'UTR-ORF-3'UTR) (SEQ ID NO: 9) contained in pcccp24e9, the partial coding DNA sequence CcCHR2A (partial ORF–3'UTR) (SEQ ID NO: 10) contained in pcccl28k6, and the complete CcCHR2B (5'UTR-ORF-intron-ORF-3'UTR) (SEQ ID NO: 11) contained in pcccl26f18 were aligned with public Chalcone reductases complete DNA coding sequences (5'UTR-ORF-3'UTR) from *Sesbania rostrata* (AJ223291), *Pueraria montana* var. *lobata* (AF462632) and *Medicago sativa* (U13994) using ClustalW method in MegaAlign software. This alignment shows 99.1% identity between CcCHR2A (SEQ ID NO: 10) and CcCHR2B (SEQ ID NO: 11), supporting the hypothesis they are alleles of the same gene. It is interesting to note that there is an unusual 24bp insertion in the ORF of CcCHR2B (SEQ ID NO: 11) that is absent in all the other sequences in this alignment. The significance of this insertion sequence is not currently known. The CcCHR1 (SEQ ID NO: 9) and CcCHR2B (SEQ ID NO: 11) complete coding sequences share 80.5% identity and the alignment permits the confirmation that they are two different genes. When compared with public sequences, CcCHR1 shows 58.4%, 45.2% and 47.4% identity with SrCHR, PICH and MsCHR, respectively, at the DNA level. CcCHR2B shares 54.9%, 43% and 45.5% identity with the complete DNA sequences from *Sesbania rostrata*, *Pueraria montana* var. *lobata* and *Medicago sativa*, respectively.

EXAMPLE 7

Isolation and Characterization of *Coffea canephora* cDNA Clones Encoding Chalcone Isomerases (CcCHI)

To find cDNA encoding chalcone isomerases, two protein sequences encoding biochemically characterized chalcone isomerases (CHI) served as query sequences for BLAST searches against the Nestlé/Cornell unigene set 5 using the tblastn algorithm (GmCHI-2, *Glycine max* Type I CHI, GenBank Accession Number AAT94360, (Ralston et al. (2005b) (SEQ ID NO:
and LjCHI-1, *Lotus corniculatus* var. *japonicus*, Type II CHI, GenBank Accession Number BAC53983 (Shimada et al. (2003)) (SEQ ID NO: 77). Using the GmCHI-2 protein sequence from *Glycine max*, two unigenes were found. The first unigene #124216 exhibited higher level of identity (e value = 2e-65) than the second unigene #124635 (e value = 4e-18). The second search with LjCHI-1 protein sequence from *Lotus corniculatus* var. *japonicus* (SEQ ID NO: 76), uncovered the same two unigenes; unigene #124216 exhibited higher level of identity (e value = 2e-53) than the second unigene #124635 (e value = 2e-16).

**Coffeea canephora CcCHI (full ORF)**. The longest cDNA representing the 5' end of the unigene #124216 (pcccp22k18) was isolated from the pericarp library and sequenced. The insert (SEQ ID NO: 12) of this clone was found to be 926 bp long and to encode a complete ORF sequence of 780 bp. The deduced protein sequence (SEQ ID NO: 31) encoded by this ORF was 259 amino acids (SEQ ID NO: 31) having a predicted molecular weight of 27.88 kDa, and this sequence was named CcCHI (SEQ ID NO: 31). A manually optimized alignment of the deduced protein sequence (SEQ ID NO: 31) encoded by pcccp22k18 (CcCHI) (SEQ ID NO: 12) was performed with the CHI protein sequences PhCHI from *Petunia x hybrida* (PIR ID: ISPJA1) (SEQ ID NO: 76), LjCHI-2 (Type I CHI, GenBank Accession Number BAC53984) (SEQ ID NO: 77), from *Lotus corniculatus* var. *japonicus*, and GmCHI-2 (Type I CHI, GenBank Accession Number AAT94360) (SEQ ID NO: 78) from *Glycine max*, available in public databanks (Figure 12A).

This alignment demonstrates that the CcCHI protein (SEQ ID NO: 31) shares 62.5%, 63.4% and 57.7% identity with the Type I CHI protein sequences PhCHI, LjCHI-2 and GmCHI-2, (SEQ ID NOs: 76, 77, 78) respectively, and supports the annotation of the polypeptide (SEQ ID NO: 31) encoded by pcccp22k18 (SEQ ID NO: 12) as a *C. canephora* chalcone isomerase. Comparison of CcCHI (SEQ ID NO: 31) with the well characterized Type I and Type II CHI protein sequences in the database using an optimal alignment revealed that CcCHI was closer to the Type I CHI than Type II CHI sequences.

An alignment of the complete coding DNA sequence (5'UTR-ORF-3'UTR) contained in pcccp22k18 with the complete coding DNA sequence GmCHI-2 (5'UTR-ORF-3'UTR) from *Glycine max* (AY595415) was performed using ClustalW method in MegAlign software. The alignment demonstrates that the CcCHI complete coding sequence contained in pcccp22k18 shares 57.2 % identity with the DNA sequence GmCHI-2.
from unigene #124635 (pcccp12o15) was isolated from the pericarp library and sequenced. The insert of pcccp12o15 (SEQ ID NO: 13) was found to be 1073 bp long and to encode a complete ORF of 642 bp. The deduced protein sequence (SEQ ID NO: 32) is a protein of 213 amino acids having a predicted molecular weight of 23,82 kDa, and this sequence was named CcCHI-like (SEQ ID NO: 32). An optimized alignment of the deduced protein sequence (SEQ ID NO: 32) encoded by pcccp12o15 (SEQ ID NO: 13) was performed with full length CHI protein sequences GmCHI-4A from *Glycine max* (GenBank Accession Number AAT94362) (SEQ ID NO: 79) and GmCHI-1A from *Glycine max* (GenBank Accession Number AAT94358) (SEQ ID NO: 80) (Figure 12B). This alignment shows CcCHI-like (SEQ ID NO: 32) shares 64.3% identity with GmCHI-4A (SEQ ID NO: 79) and 29.1% identity with GmCHI-1A (SEQ ID NO: 80).

Using Clustalw Method in MegAlign software, the complete coding DNA sequence (5'UTR-ORF-3'UTR) contained in pcccp12o15 (SEQ ID NO: 13) was aligned with the complete coding DNA sequences (5'UTR-ORF-3'UTR) GmCHI-4A (AY595417) and GmCHI-1A (AY595413) from *Glycine max*. The alignment shows that the CcCHI-Like sequence shares 55.8% and 45.2% identity with GmCHI-4A and GmCHI-1A, respectively, at the DNA level. Thus, the CcCHI-Like complete coding sequence is more closed to GmCHI-4A than to GmCHI-1A. It is also interesting to note that CcCHI-Like full protein shares 64.3% identity at protein level with GmCHI-4A, whereas their respective DNA sequences share lower level identity (55.8%). In contrast, CcCHI-Like full protein shares 29.1% identity at protein level with GmCHI-1A, whereas their respective DNA sequences have a higher level of identity (45.2%).

**EXAMPLE 8**

**Isolation and Characterization of *Coffee canephora* cDNA Clone Encoding Flavanone 3-Hydroxylase (CcF3H)**

To find cDNA encoding flavanone 3-hydroxylase, two protein sequences encoding partially characterized flavanone 3-hydroxylases (F3H) served as query sequences for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm (AtF3H; *Arabidopsis thaliana*, GenBank Accession Number #AAC6858, knock-out mutants done by Wissman *et al.*, 1998; and GmF3H, *Glycine max*, GenBank Accession Number #AAT94365. Using the GmF3H protein sequence from *Glycine max*, only one unigene (#123808) was found to exhibit a high level of identity (e value = e-142). The second search with the AtF3H protein sequence from *Arabidopsis thaliana* uncovered the same best Unigene hit #123808 (e value= e-145).
A cDNA representing the 5' end of the unigene #123808 (pcccp5120) was isolated from pericarp library and sequenced. The insert of pcccp5120 was found to be 1286 bp long and to encode a complete ORF of 1092 bp (SEQ ID NO: 14). This sequence was named CcF3H (SEQ ID NO: 14). The deduced ORF encodes a protein (SEQ ID NO: 33) of 363 amino acids having a predicted molecular weight of 40.88 kDa. An alignment of the deduced protein sequence (SEQ ID NO: 33) encoded by pcccp5120 (SEQ ID NO: 14) was made with the F3H protein sequences from *Glycine max* and *Arabidopsis thaliana* cited above. This alignment demonstrates that the CcF3H protein sequence of pcccp5120 has 82.4% and 81.8% identity with these protein sequences, respectively (Figure 13A), supporting the annotation of pcccp5120 (SEQ ID NO: 14) as a coffee flavanone 3-hydroxylase.

An alignment of the complete CcF3H coding DNA sequence (5'UTR-ORF-3'UTR) contained in pcccp5120 was performed using ClustalW (MegAlign software) with the complete coding DNA sequence of GmF3H (5'UTR-ORF-3'UTR) from *Glycine max* (AY595420). The alignment demonstrates that the CcF3H complete coding sequence shares 70.1% identity with the public complete coding DNA sequence for GmF3H.

**EXAMPLE 9**

*Isolation and Characterization of a Coffea canephora cDNA Clone Encoding Flavonoid 3',5'-Hydroxylase (CcF3'5'H)*

To find cDNA encoding coffee flavonoid 3',5'-hydroxylase (F3'5'H), two protein sequences encoding biochemically-characterized flavonoid 3',5'-hydroxylase served as query sequences for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm: CrF3'5'H, *Catharanthus roseus*, GenBank Accession Number CAA09850 (SEQ ID NO: 83), enzyme activity characterized by Kaltenbach *et al.* (1999); and GtF3'5'H, *Gentiana triflora*, GenBank Accession Number Q96581 (SEQ ID NO: 84), enzyme activity characterized by Tanaka *et al.* (1996)). Using CrF3'5'H protein sequence from *Catharanthus roseus* only one unigene (#130482) was found to exhibit significant homology (over 60% identity). The second search with GtF3'5'H protein sequence from *Gentiana triflora* uncovered the same unigene hit #130482 (over 58% identity).

The single cDNA representing unigene #130482 (pccwc22w23n18) (SEQ ID NO: 15) was isolated from the 22 weeks whole cherry library and fully sequenced. The sequence obtained for pccwc22w23n18 (SEQ ID NO: 15) is 1350 bp long, and contains a partial ORF. Two unusual features can be seen in this sequence. First, the premier 78 bp of this cDNA (5' end) appears to contains an intron sequence because this sequence does not match the
homologous proteins, and has no homologs in the GenBank database. Second, the main (partial) ORF contains a stop codon that leads to a break in this ORF. It is believe that this stop codon is due to a mutation (TGA) induced during the production of this cDNA clone. This proposed explanation will be verified in the future by re-amplifying this region of the sequence, and can be done simultaneously with the recovery of the missing 5’ sequence by 5’ RACE PCR. In the current sequence this TGA is replaced by TGN in the sequence.

The partial ORF of pccw22w23n18 (SEQ ID NO: 15) is 552bp long and encodes a polypeptide sequence of 183 amino acids (SEQ ID NO: 34). This sequence was named CcF3’5’H (SEQ ID NO: 15). Based on an alignment with the complete ORF of CrF3’5’H (512 aa), it was assumed that the CcF3’5’H protein was missing over 331 amino acids at the N terminal end. An alignment of the 183 amino acids ORF in pccw22w23n18 (CcF3’5’H) (SEQ ID NO: 15) with the same region of the F3’5’H protein sequences from Catharanthus roseus (CrF3’5’H) (SEQ ID NO: 83) and Gentiana triflora (GtF3’5’H) (SEQ ID NO: 84) (Figure 13B) strongly indicates that this coffee ORF represent a coffee F3’5’H because it shares 67.8% and 60.7% identity with CrF3’5’H and GtF3’5’H protein sequences, respectively. In comparison, the complete ORF’s of CrF3’5’H and GtF3’5’H are 74% identical. This partial sequence will enable the cloning of the remaining portion of the coding sequence of CcF3’5’H upon generation of specific DNA primers from the DNA of pccw22w23n18 (SEQ ID NO: 15), and using these primers in the well-known techniques of 5’ RACE and primer-assisted genome walking.

**EXAMPLE 10**

**Isolation and Characterization of a *Coffee canephora* cDNA Clone Encoding Dihydroflavonol-4-Reductase (CcDFR)**

To find cDNA encoding a coffee dihydroflavonol-4-reductase, two protein sequences encoding biochemically characterized dihydroflavonol-4-reductases (DFR) served as query sequences for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm: MtDFR1 and MtDFR2 from *Medicago truncatula*, GenBank Accession Numbers AAR27014 and AAR27015 (SEQ ID NOs: ), respectively, activities characterized by Xie et al. (2004). Using the MtDFR1 protein sequence from *Medicago truncatula*, only one unigene (#122897) was found to exhibit a high level of identity (e value = e-133). The second search with MtDFR2 protein sequence from *Medicago truncatula* uncovered the same best unigene hit #122897 (e value = e-137).

The longest cDNA representing unigene #122897 (pcccp5115) was isolated from the pericarp library and sequenced. The insert of pcccp5115 (SEQ ID NO: 16) was found to be 1398
bp long and to encode an ORF sequence of 1110 bp. This sequence was named CcDFR (SEQ ID NO: 16). The deduced polypeptide (SEQ ID NO: 35) is 369 amino acids long and has a predicted molecular weight of 40.91 kDa. An alignment of the protein sequence (SEQ ID NO: 35) encoded by pcccp5115 (SEQ ID NO: 16) with the protein sequences MtDFR1 and MtDFR2 from *Medicago truncatula* (#AAR27014 and #AAR27015, (SEQ ID NOs: 85, 86) respectively) demonstrates that polypeptide (SEQ ID NO: 35) encoded by pcccp5115 (CcDFR) (SEQ ID NO: 16) shares 65.2% and 67.6% identity with those protein sequences, respectively (Figure 14). This alignment data indicates that pcccp5115 plasmid contains a full length cDNA encoding a *C. canephora* dihydroflavonol-4-reductase (CcDFR).

Using the ClustalWMethod in the MegAlign software, the complete coding DNA sequence (5'UTR-ORF-3'UTR) of CcDFR contained in pcccp5115 was aligned with the complete coding DNA sequence (5'UTR-ORF-3'UTR) of MtDFR1 from *Medicago truncatula* (AY389346). CcDFR and MtDFR1 are 59.4% identical

**EXAMPLE 11**

Isolation and Characterization of *Coffeea canephora* cDNA Encoding Leucoanthocyanidin Dioxygenase (CcLDOX)

To find cDNA encoding a coffee leucoanthocyanidin dioxygenase, the protein sequence encoding biochemically characterized LDOX (also known as Anthocyanidin Synthase (ANS)) from *Perilla frutescens* PfANS served as query sequence for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm (PfANS, from *Perilla frutescens*, GenBank Accession Number #O04274 (SEQ ID NO: 88), activity characterized by Saito et al. (1999)). Using the PfANS protein sequence, two unigenes were found to exhibit significant homologies; unigene #131577 (e value = 2e-73) and unigene #122116 (e value = 7e-46).

*Coffeea canephora* CcLDOX (full ORF). The single cDNA representing unigene #131577 (pcccl21b21) was isolated from the leaf library and sequenced. The insert of pcccl21b21 (SEQ ID NO: 17) was found to be 1430 bp long and to encode a complete ORF sequence of 1128 bp. This sequence was called CcLDOX (SEQ ID NO: 17). The deduced protein sequence is a protein of 375 amino acids (SEQ ID NO: 36) having a predicted molecular weight of 42.52 kDa.

Figure 15A shows an optimized alignment of the protein sequence (SEQ ID NO: 36) encoded by pcccl21b21 (CcLDOX) (SEQ ID NO: 17) with the protein sequences of AtLDOX from *Arabidopsis thaliana* (#CAD91994) (SEQ ID NO: 87), PfANS from *Perilla frutescens*
(#004274) (SEQ ID NO: 88), FiANS from *Forsythia x intermedia* (#CAA73094) (SEQ ID NO: 89) and InANS from *Ipomoea nil* (BAB71810) (SEQ ID NO: 90). This alignment demonstrates that the CcLDOX protein (SEQ ID NO: 36) shares 69.3%, 70.7%, 72.3% and 73.1% identity with these protein sequences (SEQ ID NOs: 87, 88, 89, 90), respectively. This alignment data indicates that the pccc121b21 plasmid contains a full length cDNA (SEQ ID NO: 17) encoding a *C. canephora* leucoanthocyanidin dioxygenase (CcLDOX) (SEQ ID NO: 36).

When the sequence for the singleton cDNA representing unigene #12216 (pcccp14i24, pericarp library) was obtained, it was observed that this sequence could be assembled with the insert sequence of pccc121b21 into a unique contig that had an overlapping region of 588bp. Only two single nucleotide differences were observed in this overlapping region, and they were localized in 3'UTR. Based on these observations, it was concluded that these two cDNA represent allelic sequences of CcLDOX. The insert of pcccp14i24 is 659 bp long and contains a partial ORF of 321 bp coding for 106 amino acids.

The complete coding DNA sequence CcLDOX (5'UTR-ORF-3'UTR) contained in pccc121b21 was aligned with the complete coding DNA sequence (5'UTR-ORF-3'UTR) of PfANS from *Perilla frutescens* (AB0037779) using ClustalW method in MegAlign software. This alignment shows there is 65.5% identity between CcLDOX and the well characterized public sequence from *Perilla frutescens*.

**EXAMPLE 12**

**Isolation and Characterization of a *Coffea canephora* cDNA Clone Encoding Leucoanthocyanidin Reductase (CcLAR)**

To find cDNA encoding a coffee leucoanthocyanidin reductase, the protein sequence encoding a biochemically characterized leucoanthocyanidin reductase (LAR) served as query sequence for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm: DuLAR, *Desmodium uncinatum*, GenBank Accession Number CAD79341; activity characterized by Tanner et al. (2003). Using the protein sequence DuLAR, one unigene #132429 was found to exhibit a high level of homology (e-value = 3e-45). Currently, the only sequence data for the singleton cDNA representing unigene #132429 (pcccl21e8) is the EST DNA sequence in the Nestlé/Cornell database. The available sequence for the insert of pcccl21e8 is 648 bp long and encodes the N-terminal region of an ORF. This partial ORF is 561 bp long, and codes for 187 amino acids. The sequence has been named CcLAR.
Figure 1D shows an alignment of the available protein sequence (SEQ ID NO: 37) encoded by pcccl21e8 (CcLAR) (SEQ ID NO: 18) with the full protein sequences VvLAR from *Vitis vinifera* (GenBank Accession Number CA126309) (SEQ ID NO: 91), LuLAR from *Lotus uliginosus* (GenBank Accession Number AAU45392) (SEQ ID NO: 92), and DuLAR from *Desmodium uncinatum* (GenBank Accession Number Q84V83) (SEQ ID NO: 93). This alignment demonstrates that the available protein sequence (SEQ ID NO: 37) of pcccl21e8 (CcLAR) (SEQ ID NO: 18) represents the 5' region of the protein, and that it shares 59.9%, 55.6% and 54% identity with the foregoing protein sequences, respectively. This alignment data indicates that the plasmid pcccl21e8 (SEQ ID NO: 18) represents a *C. canephora* leucoanthocyanidin reductase (CcLAR) (SEQ ID NO: 37), and potentially encodes a complete cDNA for this gene.

**EXAMPLE 13**

**Isolation and Characterization of a *Coffea canephora* cDNA Clone Encoding Anthocyanin Reductase (CcANR)**

To find cDNA encoding a coffee anthocyanidin reductase, a protein sequence encoding the biochemically characterized anthocyanidin reductase from *Medicago truncatula*, served as the query sequence for a BLAST search against the Nestlé/Cornell unigene set 5 using the tblastn algorithm: MtANR; GenBank Accession Number, AAN77735, activity characterized by Xie et al. (2003b). The ANR protein of *Medicago* is encoded by the BAN gene.

Using the MtANR protein sequence from *Medicago truncatula* as the search query, only one unigene (122851) was found to exhibit a high level of homology (e value = e-108). Currently, the only sequence data for the longest cDNA representing unigene #122851 (pccwc22w14g7) is the EST DNA sequence from the whole cherries library in the Nestle/Cornell database. The available sequence data for the insert of pccwc22w14g7 from the Cornell database was found to be 1048 bp long and to encode a complete ORF sequence of 1014 bp. The deduced protein (SEQ ID NO: 38) is 337 amino acids long, having a predicted molecular weight of approximately 36.33 kDa. The sequence was named CcANR (SEQ ID NO: 38).

The CcANR protein sequence (SEQ ID NO: 38) was aligned with the protein sequences MtANR (*Medicago truncatula*, GenBank Accession Number AAN77735) (SEQ ID NO: 94), AtANR (*Arabidopsis thaliana*, GenBank Accession Number AAF23859 (SEQ ID NO: 95), encoded by AtBAN gene GenBank Accession Number AF092912, activity characterized by Xie et al. (2003c)); and CsLAR (*Camellia sinensis*, GenBank Accession Number AAT68773) (SEQ
ID NO: 96). This alignment showed that the coffee ANR protein (pcccl21e8, CcANR) (SEQ ID NO: 38) shares 71.5%, 63.2% and 77.4% identity with the other ANR protein sequences, respectively (Figure 15c). This alignment data indicates that the pccwc22w14g7 (SEQ ID NO: 19) plasmid encodes a C. canephora anthocyanidin reductase (CcANR) (SEQ ID NO: 38).

Using ClustalW (MegAlign software), the complete ORF DNA sequence contained in pccwc22w14g7 was aligned with the complete ORF DNA sequences of MtBAN (AY184243) from Medicago truncatula and AtBAN (AF092912) from Arabidopsis thaliana, that encode well-characterized Anthocyanidin reductases MtANR and AtANR. The alignment shows that the coffee pccwc22w14g7 ORF has 71.8% and 63.4% identity with the respective public DNA sequences MtBAN and AtBAN, at the DNA level.

EXAMPLE 14

Northern Blot and RT-PCR Analysis of the Expression of CHS, CHI, DFR and F3H in Grain, Pericarp, and Other Tissues of Coffea arabica

Northern blot analysis of the expression of the CHS, CHI, DFR, and F3H genes is shown in Figure 16. The data presented indicates that the CHS gene is most strongly expressed in the late pericarp (yellow and red stages), weakly expressed in flowers and leaves, and very weakly expressed in roots and stems.

CHS expression was not detected by northern blotting in the grain at any stage, or in the root and stem. This expression pattern was confirmed by using RT-PCR except expression in stems and roots (Figure 17). By Northern blotting, CHI was found to be expressed only in the late stages of pericarp development. No CHI expression was detected by Northern blotting in roots, stems, leaf, early pericarp tissue, or in any of the grain samples. Using the more sensitive RT-PCR technique however, low levels of CHI transcripts could now be detected in the leaf, stem, and root. The results from the RT-PCR suggest that CHI could also be very weakly expressed in the last three stages of grain development.

Expression of the CHI-like gene was also analysed by RT-PCR (Figure 17). This experiment showed that the CHI-like gene is weakly expressed in the root, stem, leaf, late pericarp (yellow and red), and possibly very weakly in the first three stages of grain development (small green, large green, and yellow). No expression was observed in the early pericarp (small and large green) or red grain samples.

A similar expression pattern to CHI was observed for the DFR gene by Northern blotting, i.e., a high level of DFR transcripts were detected during the late stages of pericarp development/maturation, and no expression was detected in the other tissues. Northern blot
analysis shows that the F3H gene is also very strongly expressed in the late pericarp developmental/maturation stages. Very high levels of expression of F3H is observed in the flowers, and unlike CHI and DFR, a lower, but significant level of expression was also detected for F3H at all stages in the grain, in the two early stage of pericarp (small green and large green), as well as in the root, stem and leaf.

EXAMPLE 15

Expression of Early Phenylpropanoid and Flavonoid Genes

The number of ESTs associated with a particular unigene gives an estimation of the expression level of the associated gene in each library (in each tissue). Therefore, an examination of the number of ESTs within the different unigenes of the phenylpropanoid and flavonoid genes discussed above can give a broad overview of the expression of these genes. All the unigenes discussed herein, and the number of ESTs in each library for these unigenes, are provided in Table 10.

Table 10. In silico distribution of ESTs in the unigenes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>#Unigene Number</th>
<th>EST/cDNA Name</th>
<th>In silico expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ccc1</td>
</tr>
<tr>
<td>CcPAL1</td>
<td>#121018</td>
<td>cccwc22w18n3 (partial)</td>
<td>5</td>
</tr>
<tr>
<td>CcPAL2</td>
<td>#119778</td>
<td>ccc25c13, cccp19k7 (partial)</td>
<td>1</td>
</tr>
<tr>
<td>CcPAL3</td>
<td>#120370</td>
<td>cccp161 (partial)</td>
<td>1</td>
</tr>
<tr>
<td>CcCAL1</td>
<td>#124559</td>
<td>ccc127h22 (full)</td>
<td>2</td>
</tr>
<tr>
<td>CcCAL2</td>
<td>#119670</td>
<td>cccp27d21 (partial)</td>
<td>1</td>
</tr>
<tr>
<td>CcCAL3</td>
<td>#128581</td>
<td>ccc2421 (partial)</td>
<td>1</td>
</tr>
<tr>
<td>CcCAL4</td>
<td>#123589</td>
<td>cccp8j10 (full)</td>
<td>5</td>
</tr>
<tr>
<td>CcCHR1</td>
<td>#125256</td>
<td>cccp24e9 (full)</td>
<td>18</td>
</tr>
<tr>
<td>CcCHR2A</td>
<td>#125257</td>
<td>ccc28k66 (nearly full)</td>
<td>5</td>
</tr>
<tr>
<td>CcCHR2B</td>
<td>#125260</td>
<td>ccc26f18 (full)</td>
<td>7</td>
</tr>
<tr>
<td>CcCHI</td>
<td>#124216</td>
<td>cccp22k18 (full)</td>
<td>2</td>
</tr>
<tr>
<td>CcCHI-like</td>
<td>#124635</td>
<td>cccp12a15 (full)</td>
<td>2</td>
</tr>
<tr>
<td>CcF3H</td>
<td>#123808</td>
<td>cccp5120 (full)</td>
<td>1</td>
</tr>
<tr>
<td>Gene</td>
<td>Accession #</td>
<td>Description</td>
<td>Count</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>CcFRS</td>
<td>#130482</td>
<td>cccwc22w23s18 (partial)</td>
<td>1</td>
</tr>
<tr>
<td>CcDFR</td>
<td>#122897</td>
<td>cccp51115 (full)</td>
<td>19</td>
</tr>
<tr>
<td>CcLDOX</td>
<td>#131577</td>
<td>ccc21b21 (full)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>#122116</td>
<td>cccp14124 (full, 100% identity</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with Unigene #131577 in CDS</td>
<td></td>
</tr>
<tr>
<td>CcLAR</td>
<td>#132429</td>
<td>ccc21e8 (full)</td>
<td>1</td>
</tr>
<tr>
<td>CcANR</td>
<td>#122851</td>
<td>cccwc22w14g7 (full)</td>
<td>2</td>
</tr>
</tbody>
</table>

In silico distribution of ESTs in the unigenes containing the partial CcPAL1, partial CcPAL2, partial CcPAL3, full length CcC4H, partial Cc4CLI, partial Cc4CL2, full length CcCHS, full length CcCHR1, quasi-full length CcCHR2A, full length CcCHR2B, full length CcCHI, full length CcCHI-like, full length CcF3H, partial CcF3'5'H, full length CcDFR, full length CcLDOX, full length CcLAR, and the full length CcANR Coffea canephora cDNA. For each unigene the name(s) of the representative EST(s) that are partially or fully sequenced is indicated. Parentheticals denote if the 5' end coding sequence of the EST is full or partial. PAL1, PAL3 and 4CL2 genes were found in unigenes that contained only their partial coding sequences, but further experiments permitted isolation of the full length sequence in arabica or robusta.

**Expression of PAL1, PAL2, and PAL3.** Overall, the data in Table 10 suggests that CcPAL1 may be the most expressed of the three PAL genes, and that it is most highly expressed in whole small green cherries (grain and pericarp). In order to obtain more accurate data on the expression of these three genes, specific primers and TAQMAN probe sets (Table 6) were prepared for each gene, and used to measure the transcript levels for each PAL gene in several different coffee tissues of arabica and robusta. (Figure 18). The different cDNA used for these experiments were prepared by Method 1 with RNA isolated from roots, stems, leaves, flowers, and from the grain and pericarp tissues isolated from 4 different stages of development of arabica (T2308) and robusta (BP409) coffee cherries as described in Example 1 above. The RQ value for each tissue sample was determined by normalizing the transcript level of this sample versus the rpl39 transcript level in that tissue.

The expression of PAL1 is very low in the grain at all four stages tested, with the levels being similar for arabica and robusta (RQ's 0.02-0.05), except at the mature red stage where the levels of PAL1 seem to be higher in arabica (0.28) relative to robusta (0.03). In the pericarp, PAL1 appears to be more highly expressed in robusta relative to arabica, and the transcript levels in the robusta pericarp seem to peak at the large green stage (RQ 0.51 for robusta), and then fall off slightly. A relatively low level of PAL1 transcripts are also observed in the roots (arabica RQ 0.20 and robusta RQ 0.11), and stems (arabica RQ 0.09 and robusta RQ 0.23) of both arabica and robusta. Interestingly, the leaf tissue showed very significant differences in PAL1 expression, with an RQ of 0.73 observed for arabica and an RQ of only 0.04 for robusta. A large
difference in PAL1 transcript levels is also observed in the flowers, with arabica exhibiting a much higher RQ (16.33) than robusta RQ (0.89).

The TAQMAN expression data for PAL2 shows that transcripts for this gene are barely detectable in both arabica and robusta grain for all stages tested, with the exception of the small green robusta grain, which is known to be at an earlier stage than the small green arabica grain used in the experiments presented here. (Hinniger et al. manuscript submitted for publication; co-pending U.S. Provisional Application No. 60/696,890). In contrast, PAL2 is clearly expressed in the cherry pericarp from the large green stage, with robusta showing higher levels of expression than arabica at the large green and red stages. PAL2 is also highly expressed in the pericarp at the small green stage of robusta (RQ of 4.09), whereas no expression is detected in the pericarp of the arabica tissue at this physically defined stage. This expression difference is consistent with the earlier developmental stage of the robusta small green grain sample.

The high level of PAL2 expression in the pericarp of robusta at the small green stage (RQ of 4.09) suggests that expression of this gene could be important for the significant expansion of the pericarp tissue that occurs during, and just after, this period. Expression of PAL2 was not detected in the robusta root (low expression in arabica root), and was not detected in the robusta and arabica stem tissue. PAL2 transcripts were detected at relatively high levels in the arabica leaf and flower tissues, but not in these tissues in robusta.

The TAQMAN data for PAL3 shows that this gene is expressed at relatively low levels in the grain (RQ 0.09 for large green robusta grain) (Figure 18). The more immature small green grain of robusta shows a higher level of expression relative to the arabica small green grain (RQ 0.81 versus RQ 0.04). The mature arabica grain has a higher level of PAL3 expression than the mature robusta grain (RQ 0.72 versus RQ 0.16). In the pericarp, PAL3 expression is very low for arabica in the small green and large green stages, but increases somewhat in the last two stages. In contrast, although PAL3 expression levels are low in robusta, these levels appear to be more similar at all the developmental stages examined. In stems, PAL3 expression was relatively high in robusta (RQ 0.24), but relatively very low in arabica (RQ 0.03). In contrast, in the leaf, the levels of PAL3 were higher in arabica than robusta (RQ 2.57 versus RQ 0.03). In the flower samples, a slightly higher level of PAL3 transcripts were observed in arabica versus robusta. The fact that each of the three PAL genes are expressed at different levels in the arabica and robusta leaf and flower samples indicates that the these two tissues are expressed at different developmental stages and their expression may be controlled differently depending on the environmental stress imposed.
Expression of 4CL1, 4CL2, and 4CL4 in Robusta and Arabica. C4H is suggested that this gene is expressed in all the tissues examined (leaf, seed, pericarp; Table 10). This observation was confirmed using a TAQMAN assay for this gene (Figure 18). In the robusta grain, transcripts for C4H are low in the small green stage (RQ=0.40), and then appear to fall progressively to a very low level in the mature grain stage (RQ=0.03). In the arabica grain, transcripts for C4H are generally higher than in the robusta grain, especially at the mature red stage. The one exception is in the yellow stage where the arabica/robusta levels seem relatively similar. In the arabica pericarp, relatively similar levels of C4H transcripts are observed at all stages. In contrast, the C4H transcript levels vary more in the different stages of robusta pericarp, with small green and large green stages having significantly higher levels of C4H transcripts than the large green arabica pericarp. In the yellow and mature (red) arabica pericarp, the levels of C4H transcripts are much more similar to those seen in the robusta samples. Similar levels of C4H transcripts are also observed in the roots of arabica and robusta, although in the stems, there is a significantly higher level of C4H transcripts in robusta relative to arabica. In the leaf and flower samples, higher levels of C4H transcripts are observed in the arabica sample relative to the robusta sample.

Expression of 4CL1 and 4CL2. The number of ESTs for the 4CL unigenes indicate that these two genes are expressed at relatively similar levels in at least some robusta tissues. For more information on the expression of these genes, specific TAQMAN assays were developed for each gene. The QRT-PCR results obtained for these two genes are presented in Figure 18.

In robusta, relatively low levels of 4CL1 transcripts are detected in nearly all the tissues except in the pericarp at the large green (0.08 of RQ) and yellow stages (0.04 of RQ), where slightly higher levels of transcripts were detected, and the roots and leaves where no significant levels of 4CL1 expression were detected. Using the 4CL1 set of TAQMAN probes, 4CL1 expression was only detected at low levels in the roots (0.01 of RQ) of arabica. The fact that expression can be detected in at least one arabica sample suggests that this 4CL1 allele exists in the arabica genome, but that the expression of this gene is very different between the robusta and arabica varieties presented herein.

The QRT-PCR results obtained for 4CL2 indicate that this gene shows a variable expression pattern in robusta, with nearly undetectable levels in the small green grain (0.01), roots (0.01), stem (0.07), and leaf samples (0.06) (Figure 18). Slightly higher levels are seen in robusta for the large green grain (0.59), red (mature) grain (0.65), for the large green (0.65),
yellow (U.95), and red pericarp (0.88). In contrast, the transcript level is lower in yellow grain (0.12) and in the flowers (0.19). A spike of 4CL2 expression in robusta was detected in the small green pericarp tissue (RQ=3.49), and this may be associated with the fact that these cherries are at an earlier developmental stage than the visually comparable arabica cherries used in this experiment (as indicated by the absence of endosperm specific transcripts in the small green robusta cherries. See copending U.S. Provisional Application No. 60/696,445). For some arabica tissues, the expression of 4CL2 is relatively similar to that observed in robusta. For example, 4CL2 expression is detected at similar low levels in arabica large green (0.19), yellow (0.23), and red grain (0.38) samples and in araboica yellow (0.27) and red pericarp samples (RQ 0.27), and at very low levels in the stem samples (RQ=0.114). In contrast to robusta, however, 4CL2 expression is slightly detected in arabica small green pericarp (0.02) (approximately equivalent to large green arabica pericarp), and the expression of 4CL2 is significantly higher in leaves (RQ 1.86) and flowers (RQ 4.84) of the arabica samples relative to the robusta samples.

**EXAMPLE 16**

**Production of the Ca4CL2 and Cc4CL2 Proteins in E. Coli**

To verify that the *Ca4CL2* and *Cc4CL2* cDNAs (SEQ ID NOs: ) can be used to produce functional Ca4CL2 and Cc4CL2 proteins (SEQ ID NOs: ), the ORFs for these cDNA were cloned into an *E. coli* expression vector and overexpressed.

In order to over-express the proteins Ca4CL2 and Cc4CL2 (SEQ ID NOs: ) encoded by pGC1 and pGC3 the ORF sequences were sub-cloned using PCR into the expression vector pET28a+ (Invitrogen). To facilitate this cloning reaction, restrictions sites were added at the ends by PCR (primers and PCR conditions shown in Table 11). A BglII site was generated immediately 5' to the ATG codons, and a HindIII site was added just after the stop codon. The PCR reactions were performed in 50 μl reactions as follows: 2μl of pGC1 or pGC3 plasmid (1/50 diluted), 5μL 10 x PCR buffer (La PCR Buffer II Mg++ plus), 300 nM of each gene specific primer (Table 11), 200μM each dNTP, and 0.5 U of DNA polymerase Takara LA Taq (Cambrex Bio Science). After denaturing at 94°C for 5 min, the amplification consisted of 25 cycles of 1 min at 94°C, 1 min at 55°C (Table 11) and 2 min 30 at 72°C for elongation. An additional final step of elongation was carried out at 72°C for 7 min. The specific PCR products generated were purified on an agarose gel and extracted using the Qiagen MiniElute kit according to the manufacturer's protocol, with the DNA ending in a final volume of 12μl. Five microliters of each purified PCR product was then digested at 37°C overnight with 5 units of
BglII and HindIII restriction enzymes in the appropriate buffer. The plasmid pET28a+ was also digested with BamHI and HindIII under the same conditions (BamHI and BglII generate compatible sites). Afterward, digestion products were purified on agarose gels and extracted using Qiagen MiniElute kit. Next, 1 μl of each digestion product was ligated with 1 μl of the digested plasmid pET28a+ using T4 DNA Ligase (Promega) overnight at 16°C. Top 10 competent cells (Invitrogen) were then transformed with 1 μl of each ligation mixture according to the manufacturer's protocol. Screening of colonies for those with the appropriate inserts was performed by PCR using the same primers employed to clone the ORF. Selected plasmids were then purified and the 4CL2 inserts were sequenced. The pET28a+ plasmid containing the Ca4CL2 sequence was named pGC5 and the pET28a+ plasmid containing the Cc4CL2 sequence was named pGC8. Finally, expression competent cells Bl21AI were transformed according to manufacturer's protocol (Invitrogen) with pGC5 or pGC8 in order to produce the protein.

For protein expression, a pre-culture of each recombinant Bl21AI cells containing either Ca4CL2 or Cc4CL2 was grown over night at 37°C in 5 ml of LB medium containing 50 μg/ml of kanamycin. One milliliter of each pre-culture was then used to inoculate two cultures (50 ml of LB medium with 50 μg/ml of kanamycin), and cells grown until the OD 600nm reached 0.6. The induction of expression was then performed with 1.5 mM of IPTG and 0.2% of L-arabinose and the cultures were grown for a further 4h at 37°C. A control culture was also established for each transformed strain that was not subject to induction by IPTG. After the induction treatment, the cells were pelleted, and then resuspended in four ml of lysis buffer (50 mM Tris-HCl pH 7.9, 300 mM NaCl, 10% Glycerol, 1% Triton X100, imidazole 20 mM). Lysis was carried out by three cycles of freeze/thaw (-180°C/42°C) and sonication (Bioblock scientific 88155) during one minute. The lysed cells were centrifuged (30min 10,000g), and the supernatant was applied to a Ni-Nta media (Ni-Nta superfloss, Qiagen) for 1 hour. The medium was then transferred on a chromatography column (Invitrogen), and then washed twice with 4 ml of washing buffer (50 mM Tris HCl pH 7.9, 300 mM NaCl, 20 mM imidazole, 10% Glycerol). The his-tagged protein was eluted in five distinct fractions of 0.5 ml elution buffer (50 mM Tris HCl pH 7.9, 300 mM NaCl, 10% Glycerol, 250 mM imidazole). After pooling fractions containing the protein band of interest, the pooled fractions were dialyzed overnight against 4L of 100 mM Tris HCl pH 7.5 2.5 mM MgCl2, 10% Glycerol using Silde A lyser 3.5K cassette (Pierce). The production of the recombinant proteins and their purification was followed by analysis on 12% SDS-PAGE gel (Novex NuPage precast Gel Invitrogen).
Results are shown in Figure 19. As can be seen, both recombinant overexpression vectors pGC5 and pGC8 produced proteins of the expected size when the corresponding bacterial cultures were subjected to induction with IPTG. The size of recombinant proteins were estimated from the gel (Figure 19A) to be between 60 and 65 kDa, values close to the predicted sizes of the fusion proteins HisTag-Ca4CL2 and HisTag-Cc4CL2 (approximately 62.7Kda).

Table 11. List of specific primers and PCR conditions used to sub-clone Ca4CL2 and Cc4CL2.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene Specific primer</th>
<th>Primer sequence</th>
<th>Annealing temperature</th>
<th>Number of cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ca4CL2</td>
<td>Ca4CL2-Bgi2</td>
<td>5' GAAGATCTGCTGTCAAAAACAAAGCAGAAAG 3'</td>
<td>55°C</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Cx4CL2-Hind3</td>
<td>5' CCCAACGCTTTTATTTTGGCAGCCAGGACG 3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cc4CL2</td>
<td>Cc4CL2-Bgi2</td>
<td>5' GAAGATCTGCTGCCCAAAAACAAAGCAGAA 3'</td>
<td>55°C</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Cx4CL2-Hind3</td>
<td>5' CCCAACGCTTTTATTTTGGCAGCCAGGACG 3'</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1, 2, 3, 4 = SEQ ID NOs: 154, 155, 156, 157, respectively.

References:


Sequence: (SEQ ID NO:1)

--------

<213> OrganismName : Coffea arabica
<400> PreSequenceString :
caccactgtc actgttctcgc ttcttttttt cactttgtttt acatgttaa actgctactc 60
cctcttgga taactacgca ttagtttctgt ggctgacctat ctgctacgtc tagccatattt 120
tctgtttccgg tggcactcaat cctgagcagta gaagaacgtca agatggaataa ttgctactac 180
gaagcgcgtaaat acgtctcggga gtttctgtttt aatcagatc ctctgaaactg gggagttgcca 240
gctgactcactg tcatggaagaa tcattaagttacg ggaatggaac gatggtgactc tgaagttagg 300
aagcgggtgttg taaatagccgtgg cggtgaagtc ttgaccggtttt tctcagggtggtt cggagttgacc 360
gccaaggttgcattcaggttgtc gaagatagctctggcgcggag tgggctgaggag tggaggaggtg 420
gcaagcagcactggtgtgct gggagatagatgc aacccaagcagct gggataacgagtaa cgggagttacc 480
actggttgggtgt ggtgcttaccct acacacacctgg acgggagcgtggg ccacatagctg 540
cctattgtat ttctgagcctt cgggaattcctc ggaacgcagcta gggaaacctgg ccaatcagctg 600
ccacactcag ccacacacgca agccgctgatact gttgaaactaca acacccctcct ccaaggttat 660
tcgggagctc gattcgaata cttggacaagct acacccactct tctttaaccaca ccaacactcacc 720
ccatgtctgc cttctctggag tacaacactac gctctctgcgtg atctttgcctc tttgtctac 780
attgctcgttg ctactacgggag aacacactccttg cggacgcacacgtg ccggaatggtg cgggaggtct 840
ctcaatgccc aagagcatg ggcctctgcc ggcctctgcgtc tggcgttctctct tcggatgctg 900
ccttcacagag gctgtgctctg ttgtaagggag cagcgagagtt gttctctggtt ggcctctatt 960
gttcatttg aagctcactgt gttgctgctctttatctctgtgagccgttgcagc aatctttgtc 1020
gaaggtgatc aagagcagcgc ccaagtcgctt aatgcttttaatc cggatatgtt cagcgtttttct 1080
cggttaataa cttgagagcg acgctctatcagtcgctttcttgagtttatgctttctac 1140
aagctgctc aaaaaaaaaa gtagttggtg cccotgcaaa acggaaggac gacgagttcg 1200
gcttccagag cgttcattggc cgggttgtttgg ccaacactcctg cgggactcag 1260
aaatgttcct gaaagagagat caatttctgcctttact ccagatcttg cgggtgcagttc 1320
aacaagctgctc tatactggttgg caatcctctct ggagaactttgctgagct cagagctgctg 1380
ggtggacgc gtcatacatct aattggcagact ctctgaaaatcg gagagttcg 1440
aatgatcact tcaacactctcc gttagtgctc attctctgctc aaggtgttga aacacgccctg 1500
gactaggtat cttcagggagc ttagattgctt atgggtgcaac ttaaattttctctcactgtt 1560
ttgggcaatc cagtgacaa ccagtgcagc aggtcagagc aacacaaacc aagctcaac 1620
ctctttggat taatcttttc aagaaaaacg gcagagccca ttgatatatt gaagctatag 1680
tcataacttt atttgggtgc acctttgcaaa caaatgtcatt tgaagttttt ggaagaaaaac 1740
ttgaaatatcg tggtaaaagaa tattttgcag caagtgccca aagcgaactct gacaatgggt 1800
gtctatgag aacctgcatcc ttcagcgttt tgtgaagaaggt atttgtcag agttgtggac 1860
gcgaataac gcctccgtcata ttgggtagac ccttgcaagcg ctacctatac attaatgcaaa 1920
aagtttaggca aagtgtctgtgt ggtactattg tgtgaagaaggt gtagtgagag gagagactgc 1980
aacaacctca ctctcccaaaa gattgtgctca ttggagatgc aactgaagggctg tctctatca 2040
aagaaagtgg agagtggcag aagcgtccgtg ggaagtggaa atccagccac ccctctacgg 2100
ataagggagtt gcagacgatca cccatggtaac aagtcgtcctc gagaagtgtct gggacagga 2160
ctggctggacg gcagagaaagc tcagcactct gttgagcagt tggacacaggt gttcacaagca 2220
atggagccag gcgacgtatgt gatgcttttg tgtgaagatgc tccaagaaattg aatgggtgct 2280
cctctcccaaaa tctgttgatt taccctctac ccttcaaaac aatttttaac taaatctctca 2340
atgt 2344

<212> Type: DNA
<211> Length: 2344
SequenceName: CaPAL1 (cDNA in pML8)
SequenceDescription:

Sequence: (SEQ ID NO:2)

------
<213> OrganismName: Coffea arabica
<400> PreSequenceString:
atggagtgcg ctaatggaaaa tgtgaatgac cttgagggaga cctttttgtac tcaacgggcc 60
gggccagcag cccagcctgtta aatattgaac gcagcogcgg agctccctaa aagagaatcatt 120
tggagtgaag tcaacgcogat gcctcgagag ttcagagggc ccctggtgcog gctaggttgg 180
gagagcgtgta cgatacgtcga gttgggagggc atagcgtgctt cctcggagaog agcgggtttaag 240
gtggagcgtgta cggagccggtg cagggcgggcgtcaagacca gcagtgacgt ggttacetggag 300
agctcgaagga agggtactga tagttcgttggt atccaccacg gctttggtgc aacgctcagc 360
aggagaaca aacaagccgg agctcgtcag gagggagctca tccgtcttctt gaagctcggaa 420
atgttggcca acggacgacga aacatgtcctg acattgtccgc atctggccacg agggtctttct 480
atgttgggtc gcatacacaac ccttcctcctag gggtactcttg gcatacagatt tggatctcgt 540
gagctatta cacaaggtctct ccacaataat atccaccccat gtttgcctct ccgcccggacg 600
atcagcgcct cgggtgtattt ggttcggtcttc ctctcataatg tttgattattt aacaggtcga 660
ccaaatttcca agctgttggg acctgaogga aaattttgca atgctactga agcatccagc 720
cggcggga ttgacactgg attttttagg ctgcaggcaaa aagaagggct tgcacctgttg 780
aatgcaacgt ctgctggctc tcgcctgtgc tccatggttc ttctttgagc taattttctt 840
gctgtttctg tctgaagttct tccagagaatt tttgctgaga ttcagccagg gcacggtcag 900
attacaggcc atttgactca aataactgaa ccacactccgc gatcaattga gccgctgagct 960
atttgggaac acatatggga tggaggttca ttgtgtaagg agctcgagag gttctcatgaa 1020
ttcagccttt tgcagaagcc caaacaagat cgttagcact tcgacacactc ccaaataggg 1080
ttaggtcact tggattgaagt gataagggtt ttcgcanaac acaatcggag agagtagcct 1140
ttgctgaatc acaatccttt gatgacttgt tccggaataa aggccttaca tgggaggaca 1200
ttcacaagcta caccataattg agtctctcatg gataaccacc gtctggccatg tgcatacaatt 1260
ggttaacacttca tttgcggact gctgtaattg acttttacca caatgtggag 1320
ccctcacaac tatccggtggg acgtatcactg agtttggact atgggcttcag agggtcgtgaa 1380
atgtccatgg cagctttactg ctctgacactt cagttctctag ccaatcctgt caacaaacat 1440
gttgcaaaagt cagagcaaca caaccaagat gttacactcat tagattgtat ctcatacaga 1500
aaacacagcgg aagctctgatg cctagctgcag ccatctactt gttgctactt 1560
tggcagcggaa ttgacctggg gcacctctgg agagactgtg gaaacacacg 1620
gttacccattg tgcacaaggaa agtgctaaca atgggctaca aatggaagtt gcacccatct 1680
agattctgag aaaaaagactt gttcaaaagtg gtggacagag agcatgttttt tgcctacatt 1740
gatgaccctt gtagtggca acctccctctg atgcaggagtc taacggcagtt cctatgctggg 1800
caaccttttag caaatggtggga caaggagag gacgcaacca ctctaaatgg gcaacagatc 1860
ggtgccttttg aggtgaacct aagagcccttt tgcacaagag aagtggagag tgcattgagt 1920
gagttggaga atgggaaggc accgtcctgc accgtacacta aggtttgca gttctattc 1980
ttgtcacaagt tttgtgaggag agagtgtgggg acaaatcccc tgcagtttgtg gaagttggga 2040
tcatctgtgg agagatattcga aaggtttttc actgcttatc atgggggtgg gtgattagtt 2100
ccatttgctgg atttttgtgaa agagtttgga gttgtctccc gtcacatgctct ctagatcctt 2160
ttttcatcct ttgtggtcatc ttataaaatt tttgtaacat gtttttttct tctagtgttg 2220
ccccccctttt gaaattgtgta tccaagttggt cactagcttg ttttttttttt ttttttttttt 2280
gttctattct gatatttgcttg gttggccattca atgttgggaa 2318

<212> Type : DNA
<211> Length : 2318
SequenceName : CaPAL3 (cDNA in pML14)
SequenceDescription:

Sequence: (SEQ_ID NO:3)

<213> OrganismName: Coffea canephora
<400> PreSequenceString:
aaagaaacgt taatattctca aatgcagcag aatatgttga aacggaggc agtaaatat
  acacgaaccg acagggtaggg aaaaagggg ggtgaactggtt ggggtaaaag agttgataag 60
tatttggttg tgttttatgg ccgaggggga aattgttttt gcaggccaga cagttaaagg 120
gaaaaagcga gagggtgtct ttcagagggg tggtaataaa ttaagaggtt ggggaaagaag 180
gatgaaagat cccctcctgt tttctctctc tttatgtcct taccatcctt tcagcccttt 240
tctttaagtgg tttggttttc gcgttaaatgc tttttttgt tttttttgtt aaaaagtttagg 300
tttttactct tttatggaac aatttttagg taccgtttgc ttcaccataat gagaaccttg 360
tgccccacaac aacaggagca gctttataaa gttctgtgtg attatttctt ttgacttgca 420
ttttgtccga ggatcctcttg gaatattcct cagcatcttg tttatctttg tagccggtcc 480
gtacccctca gatccctat atataagacta tagaatttat cggtgtagaga aaaaaagaag 540
aagagagagac tttttgtcaaa aaaaagttgaa taattgagc gccattgctta ttattttataa 600
ttttattttt ttgctagg gcgtcttcat aacgatcttt gcggaaaaaa aacaaaaagat 660
ggaataataa aaaaaaa atatattctg agtaaagtta gatactcaca aaggtggggc 720
ggggttgttg ggaacacttg gaaacacttt ttggtttttt cgggtgtagaca caaggtggga 780
ggggacacgag gggggcggcc ccggagggga aaaaaaagt agtactatct cttacccaac 840
ccccttttga ctgtaggca gatcggccca aagggcattt atcggggatt gttcctttgga 900
gacgggctgc ttgatggcaag tagagagtct tgcttattag gggacactac cctctgtttga 960
aaaazzzaaa aaazzzaaac taaaactaa aagtagtagg ccggagtagc cttctattaa 1020
tagcggatit ttttgcctata aaaaaattttg gaaaaaacac tttttgattt aagttgcttt 1080
aaaaaaaaaa ataaaaaacaact taaaaactaa aagtagtaggg cccggagttt aatcataata 1140

tttttttttttttttttttt gggttaactga ttgacctttta tccctattct cgggggcatt 1200
gttctgtttt gatttggacca taggttatga gacagccatc ggtttatcctcc tggaaataa 1260
atgtcattgc aagatggttt ccctaccatt ttttgcattgt aaaaaaaa ccttttcttt 1320
gggacattttt agcggttagtat attatgtgaa atatatatat gtaatattgt gtaaatcata 1380
tttgttaacc cacacgaaaa gaaactattt gacccaatgc tgggtgaaaa tccgggatttt 1440
gagctcctg gatcgggcct ttcgaggcgac agttggccac atacacgagc 1500
gaggagctgca atgtattgga ggttaaacac cttctatcag gggtattcgg ggtcagatt 1560
tatgatatct ctgcctagag caagatatttg gtcaczggct ggtggccttg caaacaaca 1260
gagacactga gggagccaga ggagttcaga ccacagaaag ttgttgagga gggacttcaa 1320
gtgatgcca atggcacagc cctgctgat gttcggtgtag gagaagctgg 1380
cctggactca tctctgact gccaaactct cgtatcacttg tgggacgctt gttgaggact 1440
ttgagctgttg tcctctttcc agggccacttc aagattctgc atgcacagaa gggggcagaa 1500
ctcctggctc acatctttgg accaacactg actgttctga aagccagac gcccactagcg 1560
tgattgcatac tgggtgcatac aatggctttta tggcaaccttg cctttgaaaaa aaaaaaaa 1620

<212> Type : DNA
<211> Length : 1668
SequenceName : CcC4H (cDNA in pccl27h22)
SequenceDescription :

Sequence: (SEQ ID NO:5)

<213> OrganismName : Coffea arabica
<400> PreSequenceString :
tagctcgtag taacccctccaa ccaacggccttt gtcacatgcc tgtcaaaaca aagcaagaag

aatacatatt ccgatacaagg ctcctgtgata tttactatgc aaaaaacttg ccctgcaca 60
cattacgttt cgaagacccct ctaagttcgca gatcacaagg tgtttttgata aatggg TTC
ccagatgaaat ttaacttttct gaaaagttg agctcaacag cagaagatt gcactcggggc 120
ctaacaatgg tgtggtcagc aagggacata cggctatgat cctgtctgca aacctcggccg 180
aatggtgttt ccctctctct ggtgacacttt cccgggggac cattaccaac aggcccaatc 240
catatctccac ctcgctggga tgtcataagc aaggccagga atcaacagcg ccagctcatca 300
tcaacagcgg ctttgattgct gaaacagcggc ggacattactg atgtgaaat gggttgaaag 360
tcggtgtcct gcaccctctgg ccggagagtt gtttatcactt ctcggagctca acogagggctc 420
atgaaaggca aatgccccag gtcagacagc caccatatct gttggtgccc ctcgggtact 480
cctgccggac caactggtagcc tctgaggggg tgaatctgac ccacagggga cttggctact 540
ggtgggcac aaagttggac ggagagaacc caaattttcta tataaccaat caagtgtgatga 600
ttgtgcgttt gtctctgttcc cactataatt cgcctgaactt aatggttgcct tgtgggttgta 660
gggcggccac acaatatttt atcatgcaga aattgtcatt aatgcgttct tgtgagttga 720
tctaaaaata taaatgccaca aatgggcccct tgtggcccacc aattgctctg gcctagccga 780
aaagcctgag ctttgctaaa tattgaccttt cgtgctgtgaa gactgctatc ctgaggaggg 840

- 88 -
cgcctacgg gaagagcctt gaagatgcct ttgaaaccaa atttcttaag gcacaaactt
1020
gtccaggtaa tggagtaca aagcgccccg ctgtgctacg aatgtgtcma gcattttgcta
1080
agatctcccct tagggttaaa tccggccggat gtggttccgt tggttgaatatgctgaat
1140
agatttgtaga tcccggaact gttctcctt taccocggaa ccaacctgga aaatctgca
1200
tcagaggtgga ccaatctcg aagggctatc tggatgaccc tgaagccaca aaagccaaaca
1260
tgacgaaaga tgggtggtta cagacagttg atgtaggtota cattgagcag gatgatggaac
1320
tttctagtgt tgagcgccct aagagcctaa tcaagctcaca aaggtcaccg aatgcacocctg
1380
cagaacctgga agcccttgcc tcggctcact tgcacatctc agatgtgtct gttgacctca
1440
tgagagtga cggagcgagcc gaagttcccg tttgctttgtt tgtgaaatca aaagctccca
1500
aticcaccga ggatgaatct aagcaaatata tcaagacaaca ggttatattc tcaagacagc
1560
taaccgtgt ttttttctttt gatgcccattt cgaagtcacc actcaggaaacttcttggaa
1620
aggaatgttgag agaagacgata gctgtgctgcc gcggcacaata agcaggcaca acagcatact
1680
ggcagcataa ccagagggcata tttatagcaca aacacatcctg gatatttttgc aacaaggtctct
1740
gcacgctgg ccattctggtct tgtgtttgctg a
1771

<212> Type : DNA
<211> Length : 1771
SequenceName : C4A4L2 (cDNA in pGCl)
SequenceDescription :

Sequence: (SEQ ID NO:6)
----------
<213> OrganismName : Coffea canephora
<400> PreSequenceString :
tttgtagaag gtgatggtgct ttgctaaagga aataatgttc aaagctacgtg gcacgtatgc
60
cctctccggag gtttggtgtgc atttttcggaa gttgctgctg gctgacgaa aaagctcattcc
120
agcggtgaaa atcattccaa acagatgcctg tgcactgctcc tattctcag gcacccaactgg
180	cttaccgaaa gggctctgctgc tgcaccgcaca aagggttgttc acaagtggttg gtcacagctgg
240
tgatggagaa aatcccaactt ttttttttca caaggaagat gtgatattgtg gcgttttggct
300	ttgctccag ctatatcctac tgaatctctgt gttgcttttg gtggctcagaa tttttggtct
360
aatttttgac atcgaagattg ttagattaa tgcactaatg gacgcttgctg aaaaaattaa
420
gtgacaaatt gttcctagtg tcggcccaact tggttttggaa atggcacaata gtcctggctgtttg
480
ggataagta gatctttcct ccatagacat ggtgatgcct gcggccggaac ccattaaggaa
540
ggagctcgag gacacgcttc gagcataagt ccacaaaagcc gttctcagagc aggatacagg
600
cattagggag gcaggaccttc tggctgtcat gtggcttagcg ttgcaaaagg agcctatttga
660

- 89 -
tgtcaaatca  ggtgcttgcg  ggacagtgtg  gaggaatgct  gaatgaaaaa  ttgtagatcc  720
ccaaaacta  atctcctcctc  cccgcatact  agctgagagaa  atttgccatca  gagccgacca  780
gatcatgaa  ggtctacctta  atgccggaga  ggcaactcag  atataaatcg  acacaagagg  840
atggtgacct  aacagagaga  taaaggctatg  tggatgatg  gatgaaatct  ttcataagag  900
coagattgac  gaaattacat  agtataagaa  gtttcaagtg  gcacctgcag  agctggaagc  960
catgctcctt  ttctaacctctg  gtaattctga  tgcacgtgtt  gttcctcatgta  aagatgagcc  1020
agctggagaa  gttccctgttg  ctttttggtg  gagaagcaagt  gtttccaaaaa  tttcgcagga  1080
tgagataca  caacttatct  caaaccaggt  gattttttat  aagogaatcc  atcggtgttt  1140
ttctcatggt  aaaaattctta  aagcctccatc  cggccaaataa  ttgagaagag  acctaaagac  1200
taaagctgca  gctgaaagtg  cttgcaatta  gagatctgtga  ttatataact  aacagttctct  1260
acaacaagcc  tgttaatttg  tattcggttct  gggagaaaaag  gagagaaagt  agttgtatgttt  1320
tctttctgatc  ttggtcagaca  ttctctctca  ttctcaactc  aagttgatcc  ttttccttctt  1380
tttcctcaaaa  aaaaaaaa  aaaa  1404

<212> Type : DNA
<211> Length : 1404
SequenceName : Cc4CL1 (in silico consensus cDNA poccp27d21 and Race1_Cc4CL1)
SequenceDescription : Sequence: (SEQ ID NO:7)
---------------
<213> OrganismName : Coffea canephora
<400> PreSequenceString :
tagctcgtag  taaccccttca  acaagccctctt  tgtcaatggc  tgccaaacaa  aagcaagaag  60
aaatcatatt  ccgatcaagg  cttccctgata  ttatacatccc  aaaaacctcg  cccctgcaca  120
cttactgtttt  cgaagacctt  ctaagttgcta  gatcaacaggc  tttttttgata  aatggcgcac  180
ccgatgaasat  ttacccctctc  caaacaagttg  agctcaagcgc  cagaagagttg  gcatacgggccg  240
taaacaaagt  tgggtatacag  ccagggatac  cgatcagatg  ctgtctgcaca  aacctgccgg  300
aatcagtcttt  ogccttcctcttg  tggcgtctttctc  tccggtggagc  catttttacag  atggcccaatc  360
caatatccac  cttgctgcag  ctgatcagac  aagccagggc  atccaaagcag  aagctcatac  420
tcaacggaggg  ctgtgtagctc  aaaagaggtca  ttgactatgctg  atgtgaatatg  ggggttagag  480
tcggtgtcaat  cgcctcagctg  gttaagcttttc  cttggaggctca  acgcaggcccg  540
atgaaagga  aatgtgaggtc  atgcctgagtag  gcctcgaaaga  tgggttgccgctgtcgtact  600
cctccgggac  tacagtgcagttctagggg  tggatggtgac  ccacaaggggg  cttgctcactc  660
gcggtgccaac  caggttggag  ggaggagacc  caattttctct  tatatacaaat  caagttgatga  720

-90-
<212> Type: DNA
<211> Length: 1771
SequenceName: Cc4CL2 (cDNA in pGC3)
SequenceDescription:

Sequence: (SEQ ID NO: 8)

---

<213> OrganismName: Coffea canephora
<400> PreSequenceString:

gccagctca aagctagcag catctgcaac aacttaatttt ttctcttttt
cccgacaaa atgttagacg tcgaggaagt tagagggct caaaggcccg aagcccgagc
gagcagcatg gccatcgga cagctagcgg acaaatatttc gtcagcagaa gcaacctgcc
ggattattat tttcggatca ctgatagtga gcaataaacg ggctagcaag aaaaattttaa
ggcgtagctg gacaatccaa tgattagacgt cagctacagt tactgacagag aaaaaattttaa

-91-
ggttcaagcc aagtcgaaga tcacccatct agtctttcgtt accaccagtg gttgagcacat 480
gcctggaag caactatcag tcaccaaatc cttgggacct cggccgctcg tcaagcgctt 540
catgatgtca caacagggtt gttttgcggg tggtgacggtc ctcccggttag ccaaggactt 600
ggctgagac aacaaagggg cccgtgtgctt gctcgtctgc tcagaaatca tgcaggttac 660
attcctggtgc ccaagtgatt cggctttgga tcagccctgtg ggcacagccc tgtttgaga 720
tggggccagct gcctatcatta ctggccggcc tccgcgtccc gaagttgaga ggcctttgtt 780
tgcgcctgtgta cagccagccc aacaaccttc ttccagacgt caaggggcta tcagcggca 840
tctctgttga gttggccctta cgtcctcact tctcaaggtt gttccccgggt taatctccaa 900
gaacctgaa aagatcccgtg aagaagcatt tgaagcctctc ggtatattgtg attggaactc 960
actctcttgg atctgcacact cttggtggcc tccgaatttta gaccagggtg agcacaacct 1020
gctctttaaa cccgaaaaat taoqgggactc taggcattgtg cttgagtgatg atggaataat 1080
gtcaagtggc tgtgtctgtt tgtatcttga cgaagattgga aaggcctcgag caagagattg 1140
attcacaacc caaggggaag gtagaacatgg ggggtgtctc tttgtttttg gcctggaact 1200
caacgttgag acagtgtgttct ttcagcagcc caagttccaa aagaatcatc tgttgaactc 1260
ttcagatga attcacaattc tgtattacta tggagacta agtaaatatttt tttttatgtg 1320
tgccttaagt ttatcgaaca aatcagttga aataaggttg ttacaccttg ttctggcaca 1380
aaaaaaaaaa aaaaaaan 1397

<212> Type : DNA  
<211> Length : 1397  
SequenceName : CcCHS (cDNA in pccc8j10)  
SequenceDescription :  
Sequence: (SEQ ID NO:9)  

---<213> OrganismName : Coffea canephora  
<400> PreSequenceString :  
agaccatggga aaaagctggga aaaaacatcc cagaaatagt gtttaataca ggccacaaaa 60

tgccattggt ggggttttggga tgtgcaagctc agcctttggc accagcaaga caacagtat 120

cacatctttat tgatgcaatgt gagatttgtg aacaggtcatt tgatacagca gcagtgattg 180
gcactgagga agctcttggtt aagaagttgag cttaagcact agagatttgga ttaattaaga 240
gcagggatga attgttctatc acttctaaac tttgtgtgatct tgatgtctgtg catgacccctg 300
ttcgcgcgct cccaaacaaa actcttgagg agtggtgggtc agagtatttg gatcctttatt 360
tgattcactg gcctcttaggg ctaaagcagg gtactgagat gctcaatttc accaaagatg 420
ciaattctccc ttttgacactg gtaggacgct ggagagccat ggaagaatgc agcaaaatttgg 480

-92-
gcttgacaaa gtctatatggt ttgagcaact tcacctgtga aaaaatctcc aatcctccaaag 540
aaagtgcac cactctccaa gcagtttaaet agtggaggat gatgttgggt tggcagccagc 600
gaaaatttgtt accatttgcc aaagagaaaga gaattcacaat aagtgtcctgg ttctctcttgt 660
gaggttatgg taacctttgg ggtagcaatg cagctcatgga gatctcaatc atccaaaaata 720
ttgctgaetc aagaacacaa accggttgcag aggtgtggcatt gataggggta tatcctgcaag 780
gagcaacgct cattgtgaag agcttcagca aggaaagggat gaacagaac ctcgaagtat 840
ttgaattggga actcacaacag gaagaatgg atcacaattct gcagattcct cagcgttagag 900
ccccctcggac ggaagcggttt ttgtgactca cagggcataa caaattttgg gaggaaatttt 960
ggggtgagca cgtctgaatt ttctggacaat ttggaacctca actctcactat gatgtacacac 1020
attaccaaa ttctgtttgg ggattctcaat atcaaatagg gagaagcagc aaaccaccaaaat 1080
cgtatgtcact taccttttac caaaccacccac tagtatgggtt gattactcgt ttgcaatgtt 1140
ggattgttctatatcgat catataagat ttaaaaccat ataatatata caccgcggaaaka 1200
agaaataaaa tgattatgta gagaagtaaa aaaaaaaaaa aaaaaaaaaaa aaaaaaaaaaa 1260

<212> Type : DNA
<211> Length : 1271
SequenceName : CcCHR1 (cDNA in pccc24e9)
SequenceDescription :
Sequence: (SEQ ID NO:10)

<213> OrganismName : Coffea canephora
<400> PreSequenceString :
ggaagaagtt gaaacagaaaaa tccctggaat agcagttatta aatccgggcc aacaatgcccc 60
agttgggcggc ttaggtcttg cgcgacatcc ttggccacca ttagagcaat tagtaacaac 120
ttttattgtg gcacatggaga ttggatacag gcatttgcagc acagcagcat gctatggcac 180
agaggagggcc cttggttagag ctgttgctaa agcactagag attggtattaa ttagagcacg 240
ggatgaattg ttcatacatt ctaaacttgg tagcaactgtg gcgtatcagc acctttgctca 300
gccggccctt aaaccacaactc ttgggaaagtt gggttctggag ttagtggatactttatagggt 360
tcaaatgcgtc aacagattcag ccacatgtcc cattccggcca aagatgcaaat 420
tcccttctctt gatattccagga acgttgccga gcccattggga gatgcacacaa aatattggttt 480
gcagacatgc atcggctctga gcaacattcgct ttgtgagaaaa atctgcaaaac ttcataagacat 540
tgcctaccctt cccaggctgta ttagatgcat gcgtgtggcc agcagagaaa 600
attcgggcctcttg cagagatggatt ccgttttcgcc gcgtgcttccttgcctcctcctgttgcaae - 93 -
ttatggtggt ctggggggca acagtgcaagt catggagaat ccaagtgcctaa aggatatggtc
720
tgcctccaaa acgcagtcgc ccgggcaggt tggcatttgga tggatatatac ccgcaagggcc
780
aacgttctttgg gctgagaggt tcagcaggya aagagtaaaa cccaacccccc aataatattga
840
tgggaaacttc aacagaggaag aatagtcatc caattgcacg accttctcag gcggagggctt
900
tgcctggggaa tgtggtagtttc atccaaacgag gcaataaaaa tcggtgcagg aacctttggga
960
tggcgcacacc tgtaatatttc ctttggtggta aaataaatca agattgaaag tagttatntta
1020
ttttactctactt aagtttattgg gatgcgaagtt ttttggatttt gtggtagatta atttgaagt
1080
acgatattttg ttggatgtgg ggtgatgcag gttcggttggg tttaaacaaag gatgtcatgg
1140
ttttgtagag aatccacaca gatgatatct aagggtgtgg caccctccaa aaaaaaaaaa
1200
aaaaaaaa
1205

<212> Type : DNA
<211> Length : 1205
SequenceName : CcCHR2A (cDNA in pc0c128k6)
SequenceDescription :

Sequence: (SEQ ID NO:11)

----------
<213> OrganismName : Coffea canephora
<400> PreSequenceString :
actttttcat caagttggaga tggataaga taattgcata gacagaaacct tcttgtataa
60
agttgctgca ctcttttcct gttgttaaaaat catgccgtag acaattgaaa
120
aagtgaaca aaaaaacctt gaaatagcag tatattacgc gggcacaaca aatgcagttg
160
cgggattagg atgtgcgcgca catccttttc gacccagaga acaattagta acaactttta
240
tgtgatcaat ggagatggga tagcggcact tgtgcattca agcatctgtat gggcagggg
300
agcgcctttgg tagagctttg gtcataagac tagagattgg attaattaag aggagggagt
360
aattgttcat cactttcata cttttggtca ctgtgctgta tcagccttct cttgtgctcg
420
cctccaaaaa aaccttgggt tgtgataaccc tctactcagg taggaagttg gggtgtgagg
480
atttggatot ttattttggtc actgtgccag teaggttaaa gcaggtgtct gagaagttca
540
atttgcacca agatgaaatt ctctctttttg atacccatgg aacgttcggag ggcacaggaag
600
aatgcacca aattggttgtt ccagagtcga tcggtctcag cacacctact tgtgagaaaa
660
tcctgtatac cctagaatt gcctacccct cccagcagt taacgcgggt gagaatgagtg
720
tttggtggcag cagagaaaa ttggtgcctt tgtccaaagaa cagagggatt cttatatattg
780
ccttgtctcc tctttgcatct tctttggtgt ctttggggca cagtgcaagtc atggagaatc
840
cagtgctcaa ggatattgct gcctcaaaaaa gcgaatgcggt ggacacaggtt gcttttcggat
900
ggatatatca gcagaggca gactgttggc gcgaaggtctt caacaagagaa aaggatgaaac
960
aaaacctcaa aattattgatt tgggaacttca ccagagagaga aataggatcaaa attttgca
1020
ttcctcagcg gagagctttc cggcggagaa tgggtggggtca tccacacaggg ccctacaat
1080
cggctcgagga aatcttgggagt gcgcagacctt gaatattcag tgggtttgttaa aataaatac
1140
gattagaagt atttatatttt ttttactcatc tagttttttatt ggagtcaggt ttgggagttag
1200
tggttattatt aatattagaag taatgtatttt tttttgtggtat tgggtgatga agctgttcatg
1260
gttataaaac ggaagtaattt gtttcctgaga gatctcaaca aagtgatatac tagtttggatt
1320
gtaacctcaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa
1377

<212> Type: DNA
<211> Length: 1377
SequenceName: CcCHR2B (cDNA in pcocl26f18)
SequenceDescription:

Sequence: (SEQ ID NO:12)

caatttctca aagagccgcg gaaatctctg agtgctagtct acaatatgtc cctctctgctc
60
ttcggttggct gagttctagtt tgaagccgag ccaggggccc taatttgggagc aatgggcac
120
ggtactgatct aacactttttg tctgggagggt gcaggggccc caagttgggtta aatgtaagggc
180
aacgtttttaa aatccacgggct taattttgcgga gacagcccttt cccgtcagtctg
240
gccgtaatagt ggacaggccg caactgcgctg gaattaactgga aatccgcttct gttctgcca
300
gcacctgtaa cagggcctctt tgaagatattt ataatgggtc gaagtattttct gcccttaacgg
360
ggatccgctt capgtacagag ggttgcaggg aatgtctctg tattggagagc aagatgggtga
420
attacactctc atccagaggg taaagccatt ggaatgcttc ggcatattct cccgtataggg
480
ttcctccctt ccaggtgtccct gattttctttt aaccagctccc cttcgggtattc attaagcttt
540
aggttcctgga aaggatgttc aatcaaggtg gctagcaactg cagttgagtc gaacaagcta
600
cgtcagagag ccggttgctt gcacctattt gcggaaaaatg ggtggcttcag tcgaacaaggg
660
aagagtggta caatcagttt atcagaatatt ctaagacggttt gttgcacaatg gctaataaat
720
gtcactgctg ataataagaa cctgtgggctc gacgagttata ttgctcgggga aagccggggtgcc
780
gaagagccag ttaattgctgc gcagaggtcct caatcagttt atcagaactctg atgcattgctg
840
tttgggacact tccattggtcat tgtcttttgttga aagagtataaa cttcagactttc cttcagttc
900
ttttgataa aaaaaaaaaaaaaaaaaaa
926
<212> Type : DNA
<211> Length : 926
SequenceName : CcCHI (cDNA in pccp22k18)
SequenceDescription :

Sequence: (SEQ ID NO:13)
--------
<213> OrganismName : Coffea canephora
<400> PreSequenceString :
gttatatata cctctctttcct atttccgttg aatattagtagt aaccagcaac actagattgt  60
tcttcggccaa taagaggtct gccgcaggtt tctttcgaac taatatttcag atggaactg 120
aagtgtgaa ggtggtgtaaa attcccccttc ctctccaaagt caccacctctt accaccaaacg 180
cactatcctt ggtgagccat ggtataactg acattgagat tcaacctcttt caaataaat 240
tcactgcaat cggggttttat atggatcccg aaattgtcac ctatattaaca caatggaagg 300
gcaagaagtg cactgatcct gcagaagagg atgacatctct tggccctata attctgcccc 360
cgtgacaaa gtattctgaga atcgttagta ttaaagagat ccagggctcg caatattggcg 420
tcgagctaga aagcgcgggtg agggtagcac ttgcaagctga tgatatgatat gnagatgagg 480
aggaagcagc tcttttgagga cttattgagt ttttcacacc taaatatttc aaggaagatt 540
catctttgac atattatctt cctgcgctgt ttctctgcctc cgtctgagatt gcatccacaac 600
cagggaaa aaggaataca aagataagg tggagaatgc aacggtgtgc gagacgatca 660
agaaaatttta ccttggttgg aaccagaggg ttggcagcag caccatcttc tctttgacca 720
atactttgac ttggttggta tcctaagagat gagaagttaa acgtcttagtg ccttggcttg 780
gaatctttgt cttggatttt tttgacaggt ttggttaccttg gtgggtggtat ctatggtcttt 840
cctattttta tagtttaatttttccatct ctctttacttg tggactattt aaccgggtgaga 900
aacccctctgt actattgctgt ttatggatct cttttagcca atttggacct ctgctgctgaa 960
tatattttaaga agttggagac agctttcttc ttttttttcag aaaaaaaaaa aaaaaaaaa 1020
aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaa aaaa 1073

<212> Type : DNA
<211> Length : 1073
SequenceName : CcCHI-like (cDNA in pccp12o15)
SequenceDescription :

Sequence: (SEQ ID NO:14)
--------
<213> OrganismName : Coffea canephora
<400> PreSequenceString :
cgccgataat taatccttg tttcttagtgt ggccaaaaacg aagaagagag cttgacaaaact  60
gaaacagtga gacagatggc tccctcaact ctcacagetc tagcagagga gaagacgctt 120
cagtcagtt tcatcaggga tgaagatgaa cgtcccaagag tggctctacaa ccaatctcgc 180
aatgagattg ctatcattttct taaaagggt ctttagtagca ctagaaggtgc aagacctgaa 240
atttgtaaaa agattgtgga agcctgcca gactgaggagag ttttcaaggt tattgactat 300
gggggttgag aacaacactatt tttgcaagttt ccccgcttgc aagaaagcctt acataacttg 360
ccacctgaag agaagcttcgg atccgataatg ttcagagcga aagaaagggcg atttatctgc 420
tcaagcactc tacaggttgaa aacttgccga gactggtcgtg agatgtagac atatattcgc 480
tatccagttc gggcaagggcc taactcaagaa ggggctgcccc agaaaaggggg atatatttcc 540
gtggacagaga aatatgtgta gaagttgtagtag gacgtcagct ataatctttg ggaagtttgtg 600
tcagagggca tggggattgga aaagagacgct ttaactaatgg cttggtggts catgagacca 660
aaagttgtgg tagaatctttt cccaaaggtgtct cccaaagctacttttg ggggccaaa 720
cgccacaccc acctgtgagc cactactctc ttggtacagg accaagtggc tggactctcag 780
gccacccggag aaggttgccaa gcctgtgattt accgtcaggct ctgtgtaagg tcggtttttgt 840
gttaaccttg gtggaccatgg ccactacttg aggcaagggga gtttcaagaga ccagcatcat 900
cagcgttggt taatactccaa ttggtcaggg cccagattaac ccacatttca gacccgggca 960
ccagagccgg caagttgatcc accagaagact cggtaaggggt agaaggcgggt gctttgatgtgct 1020
cccaatcgct ttcagagatgt tacagaggaag aaatgagca aagacacattg tgtgacgggg 1080
ctggagaagc tgtgccagagga gcaagatggtg caagcatgtc agaaagccaa tgtgagggcc 1140
aagcccaattg agcagatcttt tcgggtcagc ac tatcagatttt gttgttgccac tacattac 1200
acttttttgt gtattgcaatt cctcaaaaatta gtctttgagga ttctcatgaat aagaatgtat 1260
gtggccagaa gcgaaaaaaaa aaaaaa 1286

<212> Type : DNA
<213> Length : 1286
<215> SequenceName : CcF3H (cDNA in pcccp5120)
<220> SequenceDescription :
<223> (SEQ ID NO:15)
<225> OrganismName : Coffea canephora
<400> PreSequenceString :
gttctgctgta aggtagataag tttgtagataca attgcaataaa aaatatataaa gattcagta 60
tttagatgga tagatggagc acaagaggag atggatagag tcactgtgtag aacccggaaga 120
attgttgatt gntttgagtc ggtatatccct agctccgat cctcccaagc tttatatcgg 180
gagcatattt ggagacacct ttgtagcacagct ttaaatctttc ctagaattgct atctccagga 240
tgtgaagtaga atccctagct actatcacaag aacactggac ttagcgtgaa tatttgggca 300
attggaagag atcttgagtt gttgggaata ccactagatt tccacccctga tagtttcttt 360
agccgaaac atgcaaaacct gaccccccaca cccccagttta atgattttgca gttgtttcat 420
tccatttgggg ccctgagaga gaattttgcg tggagccaga aagttggtag tgcctggttgtg 480
gatatactag taggacagtt gttgcactca ttcgactgga aacgctcgcg tgaagttgcct 540
gggtgaaca ttgaggaaga cttttgcttg gcttgccctc ttaagccaa agttggtcc 600
agactggctc taaatttctca ttacagttca gtccttggct ccggtgtcaca gattttccaa 660
caatcagtttt ttgtaatttt gaatgagtac gaatttgata gtaggaaa taacaggttaa 720
agatattca cacccaaagat ctcctcttttt cggagagcga atggcccaat caagaatagt 780
ttggtgtgcg gctctgacctc gatgcattaatt tgtgtaaatc acaagagaa 840
ggagacacaa aagagatcag agcgcttccc tccggtgtcg agtgctgactc cgatagatag 900
cttaatttcc ttcgaggttg ttagaaacct cggggccgta gattttcttg gattttcttg 960	taacctgttat cggggccgag aggatatctt gtatggtggtt aagagaggtg ccccatcattct 1020
ttgtataacttc gctatgaaatg tgtctgctcta attgcaaga tccaatttca taagcaggac 1080
acaacatgaac ccacactctc ttaactgttta ctagaagctcc gcttgtctcaat tgttgccgca 1140
ccaaactcaga ttggtgtgaca tggcaaggg tgaacgcctga tcaattttc tccaaattt 1200
tttccaaataac tagtaaatgct ttatactgttt tggacagtct tataaatgttc taaaaatattat 1260	tttccttttag atactacttttt ttgtaaatatt ttgatcataca acattataat 1320
acagtttttt tagttttaaa aaaaaaaaaa 1350
</212> Type : DNA
</211> Length : 1350
SequenceName : CcF3'5'H (cDNA in pcccwc22w23n18)
SequenceDescription :  
Sequence: (SEQ ID NO:16)
-
<213> OrganismName : Coffea canephora
<400> PreSequenceString :
cttttttctt gcatatataa tagccattag aagtttccccgc ttagattttttt atctctctacc 60
 ttctggaataa ttggaagaga taattgcaaccc tctgcgcgacg ctaagggcag cggtttggttc 120
agcggagcgc ctgttttctt cggcatcagtt cttggtatatga gctttccttg aagttggctat 180
gttgtccgg ccacgttgcgg ggaacgcggg aatagggga aagttggatgct tttgtagttgat 240
ttgcaccaag cccagcagcgca tttgacgttg tggagccggc acatagaagc agaaggaagtt 300	ttcagagcg ccacccacgca ctgacgcttg tggagccggc acatgaagc agaaggaagtt 360

- 98 -
gactctaaag acctgtgaaga cgaataatct aagcccaacga tcaatggggc ttttaaccatc 420
ataagatcat gtgtcaaggg aaaaactctc aagaggctgg tttacacttc atctgcttga 480
accgtcaacg ttcaagaaca ccaacagccc gttcacgacg agaccaactg gagttgattg 540
gacttttat attcgacaaa gatgcagaga tggagatatt tcgcttctaa gccttttggct 600
gagaagaaga cctgggaggt ctcgaaacag agcaacattg atttccataag cattaacca 660
acgctagtct tagtgctcatt catcagctgc aacctcccaac ccagccctaa atctgacattt 720
tctttgatac ctgggaatga agctcactac tcaattatta ggcaaggcga gttgcttcat 780
gtggatgtac tttgtgaggc ccataatatt ttgtagaggg atctctacgc cgaaggaaga 840	 tacatctgtc cttctcatga tgccacattc catgatttgg caaaattgat cgccggagaaa 900
tggcagcagt actcttatccc cagtttaag ggcgtgagaca aggacatacc cg tgtttttccc 960	 tttttccccca aagaattgtt aggcaagggc ttccaatata agtacacctt ggagacatg 1020
ttcgccccgg ccattgcacac atgccgtaag aaggagttgc ttcccttatc tattccaaacc 1080
cagaaaaatg gaaagaagaaga aagacacattct tttggtctaa caaagacagc ccgagcgggc 1140
cagaagcagc cccacctttt ccattaagc cccaaatatgc accattgact ttttttaagt 1200
aacaagaataa agatggagag ggcgcaaggg caaaagcttt ttccagcggtt tatattcatt 1260
tctgagctat ctcacaacga ggatgtagaa gcaggtgtgc gctttttgtaa cattttgcaa 1320	tttaattca atccaagactc tactagagtt gcggggtttt tgggtgtaaa aaaaaaaaaa 1380

<212> Type: DNA
<211> Length: 1398
SequenceName: CcDFR (cDNA in pcaccpSl5)
SequenceDescription:
Sequence: (SEQ ID NO:17)
-----------
<213> OrganismName: Coffea canephora
<400> PreSequenceString:
caggctttag atataccccgt ttaaatagaa aaagcgatat caaagtaggg taactcgggtat 60
gagtaacata caggaatalg caccactgtc gcggtgacct caagagttga gaggttgccc 120
aaccagtggga ttcacattaacc aacaaagagc tattcagggc aagacgcctgc gaacatggtt 180
aacgcttttg aggaaagaga aacaacagca gg gcccttcaagg tgcgaaacat tgaattggg 240
gacattgaag cagagcagca agctgcttctg gagagagttc aacaagagtt gaagaaggtct 300
gccatggagt gggaggattg gcaacacagttt acacctggcga tatcaatga gcctattataa 360
tcgagcagg cggctggaga gccctttctc aatccaccc ttgagaaaaa gggagaatttt 420
gccaatgatc aggcacgagg caaatcgcaaa ggatacggcga gcaaaactagc aaacaactgcc 480
agtggctcgc ttaaatgggga ggactacttt ttctcttttg tttcccccga ggcaagaaggg 540
gattgtgca ttggtcccaaa gaccccaaga gactatatact ctcgacagcag tgtgatgca 600
aagcaattga ggggctttgcc aaccaagctta ttaagctgtgc ttgtcttttgg ttgggatta 660
geagaaagcaga gctcagaaaa ggaagggtttg gtgatcgaag agttaattct gcaaatagag 720
attaactact acccccaactg ccctcaaacgc gacctagact tgtggtgtgga agctcaactc 780
gatgtcaggt ctctggacctt cactcttccc cattatgttgcc ctgggctgca atttgtctat 840
agggacaagt ggattactgc caaatgtgdc cccaatctccaa ttatatcgca cattggagac 900
accattgga gccatagcaaa ttgaaaaaaaaa aagagcaaacgct catatagagc ccttggtgac 960
aaagacaaag taagattactt tttggcacgtt ttcttgtaaag cacccaaagga gaagatcactc 1020
ctcaacaccac tggctctgagac atggcttgaag actgagccccct ctcgtaacctc gctctgaacg 1080
tttggcgcag ccattttgacc taacctgttgg ccagaaaaacc gcggagctgtg tgaagaaaaat 1140
caatccactg aggacaataa actctctgtgaa gcaaatcaccaac ctcgatgaag tcggtaaacgt 1200
tcgtgaatgg agatcctctcc taactttttcct gtaaattccaa gttgcacccaa gtaattaat 1260
aacccttttct actgttgcttc cttttgggagc ctcgggctcagc actgctttcag ggcccctttttt 1320
gcgtgtgcttc attactttttg gaagttgtggcg ttttgtgtctt ctaagtttac tacggtcttt 1380
gctttgcatc gaataaaatct atgtatattt taaaaaaaaaa aaaaaaaat 1430
<211> Length : 1430
SequenceName : ColDOX (cDNA in pcccl21b21)
SequenceDescription : 
Sequence: (SEQ ID NO:18) 
<213> OrganismName : Coffea canephora
<400> PreSequenceString : 
gacgtatagct ctattttcag acaatttttac ttggtctctact tttttttcag 60
gcggagccttt aattagccag aaaaaaatgctc ggtgctctcttt ctgtccagatgt gacaagccag 120
aagagaacg agaatattttg tcacagccct tatttttttt tatttgcttc tattgcaaa 180
agcaaatgcctt gctctccgact ctacagctttttatgtgtcttact ttttttttttt 240
tttttttttttttt ttttctgtccattttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
tgtggcgtc attaaggact tcttgccctc cagttttggg ctagatgttg acagacgga 540
gccagtggag ccagggctga ccatgtaaca tcggaaaagc agagtcgcaa ggtgtatgg 600
ggatcagga attacatata cttactaatt gctgaacact atgtgttc 648

<212> Type : DNA
<211> Length : 648
SequenceName : CcLAR (cccl21e8 5' Cornell sequence)
SequenceDescription :

Sequence: (SEQ ID NO:19)
---------------
<213> OrganismName : Coffee canephora
<400> PreSequenceString:
attcatagc aataagacttg atggagcctc atgagtgcag agaggtagcg 60
tctagttgag tcgtgctgaa atcagaaaaa gataaactac cattgcccac 180
tacagagttt ggagactta aagagtttta aagcagatct tacagatgaa gcaagcttttg 240
atgagccctgt agcaggtcgt gaccttgctt ttggtgtggc gcggcagtc aacctttgct 300
tgaggatatc agaagatgt atgtaaaaac cgcaatctca aggaggtgac cattctctaa 360
agctgtggtg aaaaagcggga tcagtgaaac gcttaaattt taccctctca gctgcagctg 420
tacccatcag cgaatattac cgtcaggcgc tgaattatgt tgaagggac tgcagagatg 480
tttggtttcag tatttctgca aaaccactca cttgggggtta tctgtctc ccagaacttag 540
tctgaagggagc tttgctgaa agaagaagat tgaattattc aactagcttc 600
cagttctcat tggctgctct cccctgacac cagacgcttc cttcccaggtc aacttcctga 660
ttgctctgat cagaggaacc gatttctcaa ttatggtctt gaaagggag cagatgtcgg 720
caggtccag tctcatttca catttggaag atgtctttga aagccacatt tttttggtct 780
agaaaaatacg ggtctcgctg cttcactatt gcgtgcctgc caatactagt ctttgctggc 840
tcgcaacattc tctggtgtaaa agataccgag actacaaat ccaaaacagaa tccgaggttt 900
tcccacatca agcaagttg atcattttca cagagaagct tatcaaaag aggtttcaatt 960
cttacctcag aactggagct aagccctttgc ttatctcaag gcacaggttt 1020
tatgtcagcatttttatttattttttttt 1048

<212> Type : DNA
<211> Length : 1048
SequenceName : CoaNR (cccwc22w14g7 5' Cornell sequence)
SequenceDescription :

Sequence: (SEQ ID NO:20)
---------------
<213> OrganismName : Coffea arabica
<400> PreSequenceString:
MENHIDEGVK VSEFCLKSQP LNQVGAESSL MGSHLDEVKV MAVFRKPRVY KLGGSLTVVA
QVAIAAKGQ QGVYVIWIEED ARAGVASSD WMVESMNKGT DSQGYTFTGF AGTSHHRTRNQG
GALQKELIRP DNQAGIFQNTG BTCHMLPHSA TRAAMLVRI INULLQQYSGR FIILARHITT
LHNENFTCLP LRGTITASGD LVPLSYIALGL LTGRPNCKAV GPGSADFNAE RAFRLAGLSSG
GFFIIQPKGO LALVNTAGV SGASIVLFLV ANVLAVLGVV LSAILPAEVNM GKPETFIDHLT
HKLKHPGQI EBAAMIEHIL DQSSYVKAQ KLHELDPLQK PKQDROYALRT SPQWLGPQIE
VIRAATKMIR RINSVNDNP LIDVSRNKL HGGNPOQTPN GVSMMDARLA IAASIGKLMFA
QFSELVNDYY NNGLPSNLSSG GRNPSDLGYF KGAARIAAY CSEQLQYLGNP VTNHVSABQ
HNQVDNSSLG ISSRKTASAI DLKLMMSSTY LVACMQAIIDL RFELENNLKNA VKNTIVSQQAK
RTLMTNGANG LHPRSRCSKD LRVVDREYFA FAYVYDDPQSA YPIILMQKLQR LVLDHALKNG
DQEKNVNTSI PKXIAAEFEDE LKAVLPKEHE SARSASGEQ1 PAIPNRIERE RSPLYKYFVR
EVLGTLGTLTG EKAOQSGPGEV DQVPAMSGQ QIVDPLLECL QEWSNAGPLPI C

<212> Type : PRT
<211> Length : 711
SequenceName : CaPAL1 (Protein encoded by pML8)
SequenceDescription : 
Sequence: (SEQ ID NO: 21)

----------
<213> OrganismName : Coffea arabica
<400> PreSequenceString :
MECANGNND LATEYCTQRA GPAPDPLNWN AAAAAELKSSH LDEVKRMVDS FRRPLVRLGG
ETLITAQVAA IAASSDAAVK VELSEGARAG VKASSDWVME SMRKGTDKSY ITTGFQATSH
RRTQKGGALQ EELIRPLNAG IFPQNTBTCH TLPHSATRAS MLVRINTLQQ GYSQKREIFF
RAITKLLNNN ITPCPLPLGT ITASGGDLVPL SYIYVOLLGQR PNSKAVGPDG KFVPNATBAPS
LAGIDTGFPE LQAKGRGLALV NGTAVGSALR SMVLFPAHNL AVLAVILGSG FAEVWNGKPE
PTDHLTHLKKL HHPQGIEAAA IMEHILGSS FVKEAQRVHC FDPLQKPKDQ RYALRTSPOW
LGPLIEVIRA STKSIERHN SVNDNPLIDV SRNKALHQQN PQQTPIGVSM DNTRLAIAIS
GKLMPAQFSE LVNDFNYNQL PSKLSGRGNP SLQYGPKGAE IAMAAYCESL QPLENPVTNH
VQSAEOHNQD VNSELGLISSR KTBAVDILK LMSSTYLVAC QAIDLRHLLE ENLKASVKNT
VSLVXAVKLTV MGYNGLHSPS RFCEKDLKLV VDEHVFYAI DDPSCSYTPL MQQKLQVLVLE
HSIANDDKEK DATTSPQKIQ GAPHDLKAL LPKKEVESARC ELENGKPGIA NRKIDCRSYS
LYKFVREGLG TNFLTGKVR SPGEFDPKVFR TAICEBKLID PLNCDLKEVWN GAPRPPIC

<212> Type : PRT
<211> Length : 717
SequenceName: CaPAL3 (Protein encoded by pHL14)
SequenceDescription:

Sequence: (SEQ ID NO: 22)

<213> OrganismName: Coffea canephora
PreSequenceString:
IRFLNAGIFIC NGTWSGHITLP HSATRAAMLAV RINTLQLGYS GIRFEILEAL TKLLAHNITP 60
CLPLRSTITA SDDLVPLYSY AGLTSGRPSNS RCVQDPNQSL DATEAPQIQAG MNSGPFBLQP 120
KBGLAYQIGT AVVSGLASV MVRHALLVIL SEYVFSAIFFAE ORMKPEPVID HMHHILKHHIP 180
QGIEAAAIMBE HILDGSSYVK AAKALHETDP IQQPKQDREYA LRTSFQWLGP LIVEIRSSSTK 240
SIEEINSVNI DNFIVDVRQK KALGGGNNQQG TPQVSMMDNT RLAIAISQKLM MFAQPFSLVN 300
DHYNLQPSN LSGGRRPSDL YGFKGAESIAM AAYCSEKLQPL ANFPVTNHVQS AROHNQDVNS 360
LGLISSRKAA RAVDILKAIMS STYIWLACQA IDLRHLHELN KNAKVSTVHQ VAKKVL5TTG 420
NGELHPSRFCC EKDLLKQVVER EYVFTYIDDWP CSATYPLAQK LRQVLVQHDAL INNEDLPNP 480
SIFLKVKAFE EELKTTLPIE VERSARNACH GNPAYFPNNRIK KCRSYPYKF VRELDLATQFL 540
TGIAKSKPGE RFDKVSAIC DGMVDPPLLE CKLWNGAPL PLC 583

<212> Type: PRT
<211> Length: 583
SequenceName: CcPAL2 (Partial Protein encoded by in silico Partial CcPAL2)
SequenceDescription:

Sequence: (SEQ ID NO: 23)

<213> OrganismName: Coffea canephora
PreSequenceString:
MIDLLELEKTL LGLFAAIVTA IVVSKLHGGK PKLPPGPPIPVP PIFGANWLQVG DDLNRHNRLTD 60
YAKKFGKHFL LRMQ QRNLVUV VSSPELAXDV LHQVGVQGFS RTRNQVDIPF TGKQDMVFT 120
VYGHMRKRMR RINTVPFPPTN KVVQYQRHGQ RWAEARVVED VKNPENSSTN GIVLRRRLQQL 180
MMYNNMYRIM PDYRFBSEDDB PLFNKLKALN GERSLQAQSF EYNYGDFIPRI LRPFLRQYLLK 240
ICKEVKERRL QLFPKHDQFVE RKKLASSSTDM DSHSLKCAID HILBAQQQGE INEDNLYIV 300
EMNVAAEIET TLWSIEWIGIA BLVNHQFQVR KLQRHBDTVL GPGQVQETED TLXLPYLIQAV 360
VKEOTLRMA IPLLVPQHMNL NBEALGGYDI PASIKLQMA WLLLANNPENN RKPEFRPER 420
PLEEBSKVDQA NGNDFRYLPFF VQGRRSCPGI ILALFILGHT LGRLVQNFEL LPPFGQSKID 480
TAELQGQQSFL HILKHHSTIVL KPRSL 505

<212> Type: PRT
<211> Length: 505
SequenceName: CcC4H (Protein encoded by pcccl27h22)
SequenceDescription:
Sequence: (SEQ ID NO:24)

---------

<213> OrganismName : Coffea arabica
<400> PreSequenceString :
MAVTVKQSEI IFRSRLDPDIY IPKHLPLHTY CFEDLPKFRS QAELINGATD EIYTPEQVEL  60
TARRVASGLN KVGGQQCDTV MILLPNSSEP VFAFLGASFR GAISTMANFY FTSABVQKQA  120
KASHAKLIIT QCYVEKVRD YACENGVKV CIDSAPGEGC HFSMTBADE REMPDEVEISP  180
DDLVFAVFSS GTTQLPGGVM LTHKGLVITSV AQVQGDHNPN FHYHNNVMMV VLSLPHIYSL  240
NSILLCGILRA GTTILMQKDF DIIPPLHLIQ KKYVTGPFVF PPVVLAIKSA PEVSDKLYLSS  300
VKTVMDGAAP LGKELEDAVR TKPPKAQKLQ QGQMTRAGPV LAMCSAFKAD PFEVKSGGCC  360
SVVRKNAEMKI VDPRTGSSLP RNQPGSRICIR GDQIMKGYLD DPEATKATID EDGWLHTGDV  420
GYIDDEDELNF IVDLKELKIK YKGQPQAPAE LEALLLLASHD ISDAAVVPKD DDAAGEVPPA  480
FVVKSKDQNI TDEIKEYIK KQVIFYKRN RVFFVDAIPK SPSGKLRKRD LRARLAAAGVP  540
K

541

<212> Type : PRT
<211> Length : 541
SequenceName : Ca4CL2 (Protein encoded by pGC1)
SequenceDescription : 

---------

<213> OrganismName : Coffea canephora
<400> PreSequenceString :
VEKVMDFPAKE NNVKVMCTDA PPEGCLHFSE LSSADSKVIP AVKINPNDAV ALFYSSGTTG  60
LPKGVMLTHK GLVTSVAQQV DGGPNLYLFH KEDVILCVLP EFHIYLSNSV LLLGRLVQAA  120
ILIMQKFEIN ALMBELVQKYK VTIAPPVPPI VLEIAKSPVV DKYDLSRIRM VMQGAPMGGK  180
ELEDTRVAKL PKAVLQGQYGY MTAEGFLLSM CLAPAKEFDF VQGACGTVV RENMKIVWD  240
ETINLSPRNQ AGESICRGIDQ IMKQYLDPE ATENTIDKKE WLTHTGDIY DDDDEIFIVD  300
RLKELIKYYK PQVAPAELEA MLSSHPIGSD AAVSVMKDEA AGSVPVAFVV RASGSKISED  360
RKQFIQSNQ IFYKRIRHRVF FMQKIPKAPS GKLRLKDRA KLAAEVACN  409

<212> Type : PRT
<211> Length : 409
SequenceName : Cc4CL1 (Partial Protein encoded by in silico Partial Cc4CL1)
SequenceDescription : 

---------

<213> OrganismName : Coffea canephora
<400> PreSequenceString :
MAAKTYQESEI IFRSRLDPDIY IPKHLPLHTY CFEDLPKFRS QAELINGATD EIYTPEQVEL  60

- 104 -
TARRVASGLN KVGIQQGDTI MILLPSFEP VFALGASPR GAISTMANPY PTSÆVIKQA
KASNAKLIT QGWYVEKVM YACENGVKVY CIDSÆPECGL RFSSSLTEADE REMLDEVEISP
EDVVALPYSS GTTGLPKGVM LTHKGLVTSN AQVQDSEPN FYIHNOQVMC VLPNFLHYSL
NSILLCGRLA GTTILM9KFQ DIIPFLELIQ KYKVVTHGPFV PPIVLAIAKS PPEVVKDLSS
VKTVMGAAAP LGKELEDAVR TKFPPAKLQQ GYGMTEAGPV LAMCSAFAKD PPEVVKSGCG
SVVRNABMKI VDPETGSSLP RNQPSGECIR GDQIMKGYLD DPEATKATID EDGWLATGDV
GYIDEDEBLPI IVDRKLKLIK YKGFPQVAPAE LEALLAHSD ISDAAVVPMK DDAAG3YVPVA
FVVKSKDNSI TEDRIKYIKI KQVIFYKRIN RyPVFDAPFK SPSKILKSLD LARLSAQVP K

<212> Type: PRT
<211> Length: 541
SequenceName: Cc4CL2 (Protein encoded by pGC3)
SequenceDescription:

Sequence: (SEQ ID NO:27)

<213> OrganismName: Coffea canephora
<400> ProSequenceString:

MVTVEERVRQAQPATAMAIATATPPNC VBQSTYPDDY PRITDSHHKT ELKKEFPRMC

DESHMIRRKYM YLTERILKIN PNICAYMPAS LDARQDMVVV EYQKLGKBEA QAQAIKENWQP

KSKITHLVFCT TTSVMPGGA DYSQTLKSLGL RFSVKKRMYY QQQFCAGGTG LRLAKDLAEN

NKGARVLLVCC SEITAVTFRQ PSDSHLDLSLV QGAQLPDDQAA AIITGADVPV EVRPLFLSLSV

TAATLIPDDS HGADHILRLR VGLTFHLLKK VPEGLSKNIE KSLKEAFEPIL GISDWNLFW

IAEPGGPAIL DQVEQKLARK PLLKLRATRH VSEYGNMSSA CVVFILDEMR KASAKDGCNNT

TGEQLDQGVL FGFQGQLTVVE TVVLHNASITH K

<212> Type: PRT
<211> Length: 391
SequenceName: CcCHS (Protein encoded by pccc83j10)
SequenceDescription:

Sequence: (SEQ ID NO:28)

<213> OrganismName: Coffea canephora
<400> ProSequenceString:

MEKAGETIPS IVLNTGHKMP LVGFQCAAQLP LQPSEQLQVST PIDAMEIGYR HFDTAACYG

REALGKAVAK ALEIGLRKS DELFITSKLW CTDAHDLVL PALKQTLGKL GLEYLDVYLI

HWPLRKLQGT EMLNFTKDAI LPPFDHMGTWK AMERCSKLGFL TKSIGLSNFT CEKISKLQES

ATILPNAVQV EMNNVGWQQRK LVVFALKEGI HISAWSPLGG YGTWSGSNAV MESSPIKSNIA

- 105 -
DSRNKTVAEV ALRWWYLQGA SVIVKSFSKE RMKQNLQVFDP WELTKHEMDQ ILQIPQRRAP 300
GEALVDFTG PYKSLLEEPWD GDV 323

<212> Type: PRT
<211> Length: 323
Sequence Name: CcCHR1 (Protein encoded by cccp24e9)
Sequence Description:

Sequence: (SEQ ID NO:29)
--------
<213> Organism Name: Coffea canephora
<400> PreSequenceString:
EKVQSIPEET AVLNSGHKMP VAGLGCAAHP LPPFLQVLVT FIDAMEIGYR HDFTAACYGT 60
EEALGRAVAK ALEIGLILKR SELFTSKLW CTDADHDLVQ PALKQTLGKL GLYGLDLVLV 120
HNPVRVKHGA EKFNPAKDEI LPPFHIWTQ AMEGCTKLGL TKSILGSNFT CEKICKLLEI 180
ATIPPAVQNOV RMNVGQWQQRK LVPPAKDRGI RICAWSPLAS YGGLWGSNAV MENFVLKDA 240
ASKSKSVAQV ALRWHYQQQA SFAVAKSNKE RMKQNLQIFD WELTKHEMDQ ILQIPQRGIF 300
AGEVVFVHPTG PYKSVEELWD GDV 323

<212> Type: PRT
<211> Length: 323
Sequence Name: CcCHR2A (Partial Protein encoded by ccc128k6)
Sequence Description:

Sequence: (SEQ ID NO:30)
--------
<213> Organism Name: Coffea canephora
<400> PreSequenceString:
MPETMKVEG KIPRIRALNS GHKMPVAGLG CAHPLPPSSE QLVTTFIDAM EIGYRHFDTA 60
ACYGTEERALG RAVAKALEIG LIKSRDELFI TSKLMCDTAD HDLVLPALKQ TLKGLGLEYL 120
DLYLVHHPVIR VHKSGEPFNF AKREIBLPFDI HGTVQAMEDC TLKGLTKSIG LSNTFCKEKIC 180
KLKHEIATIPP AVNVIGMVWQ WQQRKLPQPA KORDIGICAW SPLASYGGLW GNSAVMENPV 240
LKDIASSKSK SVAQALRMI YQQGASFVAK SPOKBEMKQN LQIPFDWELTK EMDQLILQIP 300
QRRGFAGEVF VHPTGYPKSV BELWGDGD 328

<212> Type: PRT
<211> Length: 328
Sequence Name: CcCHR2B (Protein encoded by pcccl26f18)
Sequence Description:

Sequence: (SEQ ID NO:31)
--------
<213> Organism Name: Coffea canephora
<400> PreSequenceString:
MSLSDVYGBV HDGSHVPPAA AKPPGSDQN FFLGGAGARG LEISGKFI KF TAIGVYMEET 60
AIFSLAVKWK GTAEELTES VEPFRDIVTG PPEKFIHVTM ILQPFGRQYS EKVAEYNCSAY 120
WKAVGITYTDA EGKAIEMFLD IFQNESFPPG ASILFTQSPPL GSLTISFSKD SSIPQSVNAV
VENKLLGEAV LEIIIGGNV SPDKKSLAI RLSELHVFDP NNNNNVTADN KKnHADGAIa
ASAPGEKQVN GVQVPVQVP

<212> Type: PRT
<211> Length: 259
SequenceName: CcCHI (Protein encoded by pcccp22k18)
SequenceDescription: 

Sequence: (SEQ ID NO:32)
---------------
<211> OrganismName: Coffea canephora
<400> PreSequenceString: MGTEVVKVDE IPFPIQVTPS TTKPLELGLH GITLIEIEHFL QIKPTAIGVY MDSIEIVTYLQ

<211> Length: 213
SequenceName: CcCHI-like (Protein encoded by pcccp12o15)
SequenceDescription: 

Sequence: (SEQ ID NO:33)
---------------
<211> OrganismName: Coffea canephora
<400> PreSequenceString: MAP8TLTALA EETLQSSPFI RDEDERPKVA YNQFSNEIFPI ISLKGGLDDTD GARBPICKKI

VSACBDWQIF QVIDHGVDSDK LISDMTRLAR EFYPALPPEEK LRFDMSGGGK GGFIIVSHLQ
GETVQDWRBI VYTFSYPIRA RDIYSHFDRP EAWRAVTEKY SEKLMRACK LLEVLSRAGM
LEKSAALTNC VDMQKVQVVNI FYPKCPQPDLD TLGLKRHTDP GTITLLLQDQ VGLQATRDRG

<211> Length: 363
SequenceName: CoF3H (Protein encoded by pcccp5120)
SequenceDescription: 

Sequence: (SEQ ID NO:34)
---------------
<211> OrganismName: Coffea canephora
<400> PreSequenceString: AQSEBDRVGTG NRRRVDXGLS SDIPKLYLQ AICKBAPWHQ LSAPLNLPRI ASQACREVNGY

CIPKNKDSLVIWAIAGRPDP VWENPLDFPHD DRFLSGKHAK LDPPPPVDVF ELFHSIWGLE

- 107 -
The present invention is not limited to the embodiments described and exemplified above, but is capable of variation and modification within the scope of the appended claims.
What is Claimed:

1. A nucleic acid molecule isolated from coffee (Coffea spp.), having a coding sequence that encodes a phenylpropanoid or flavonoid biosynthetic pathway enzyme selected from phenylalanine ammonia lyase, trans cinnamate-4-hydroxylase, 4-coumarate:CoA ligase, chalcone synthase, chalcone reductase, chalcone isomerase, flavanone 3-hydroxylase, flavonoid 3',5'-hydroxylase, dihydroflavonol-4-reductase, leucoanthocyanidin dioxygenase, leucoanthocyanidin reductase, and anthocyanidin reductase.

2. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a phenylalanine ammonia lyase.

3. The nucleic acid molecule of claim 2, wherein the phenylalanine ammonia lyase has an amino acid sequence greater than 85.2% identical to SEQ ID NO:20.

4. The nucleic acid molecule of claim 3, wherein the phenylalanine ammonia lyase has an amino acid sequence comprising SEQ ID NO:20.

5. The nucleic acid molecule of claim 2, wherein the phenylalanine ammonia lyase has an amino acid sequence greater than 83.9% identical to SEQ ID NO:21.

6. The nucleic acid molecule of claim 5, wherein the phenylalanine ammonia lyase has an amino acid sequence comprising SEQ ID NO:21.

7. The nucleic acid molecule of claim 2, wherein the phenylalanine ammonia lyase has an amino acid sequence greater than 82.6% identical to SEQ ID NO:22.

8. The nucleic acid molecule of claim 7, wherein the phenylalanine ammonia lyase has an amino acid sequence comprising SEQ ID NO:22.

9. The nucleic acid molecule of claim 2, wherein the coding sequence comprises SEQ ID NO:1, SEQ ID NO:2, or SEQ ID NO:3.
10. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a trans cinnamate-4-hydroxylase.

11. The nucleic acid molecule of claim 15, wherein the trans cinnamate-4-hydroxylase has an amino acid sequence greater than 89.9% identical to SEQ ID NO:23.

12. The nucleic acid molecule of claim 11, wherein the trans cinnamate-4-hydroxylase has an amino acid sequence comprising SEQ ID NO:23.

13. The nucleic acid molecule of claim 10, wherein the coding sequence comprises SEQ ID NO:4.

14. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a 4-coumarate:CoA ligase.

15. The nucleic acid molecule of claim 14, wherein the 4-coumarate:CoA ligase has an amino acid sequence greater than 81.1 % identical to SEQ ID NO:24.

16. The nucleic acid molecule of claim 15, wherein the 4-coumarate:CoA ligase has an amino acid sequence comprising SEQ ID NO:24.

17. The nucleic acid molecule of claim 14, wherein the 4-coumarate:CoA ligase has an amino acid sequence greater than 81.9% identical to SEQ ID NO:25.

18. The nucleic acid molecule of claim 17, wherein the 4-coumarate:CoA ligase has an amino acid sequence comprising SEQ ID NO:25.

19. The nucleic acid molecule of claim 14, wherein the 4-coumarate:CoA ligase has an amino acid sequence greater than 81% identical to SEQ ID NO:26.

20. The nucleic acid molecule of claim 19, wherein the 4-coumarate:CoA ligase has an amino acid sequence comprising SEQ ID NO:26.
21. The nucleic acid molecule of claim 14, wherein the coding sequence comprises
SEQ ID NO:5, SEQ ID NO:6, or SEQ ID NO:7.

22. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a
chalcone synthase.

23. The nucleic acid molecule of claim 22, wherein the chalcone synthase has an amino
acid sequence greater than 90.5% identical to SEQ ID NO:27.

24. The nucleic acid molecule of claim 23, wherein the chalcone synthase has an amino
acid sequence comprising SEQ ID NO:27.

25. The nucleic acid molecule of claim 33, wherein the coding sequence comprises
SEQ ID NO:8.

26. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a
chalcone reductase.

27. The nucleic acid molecule of claim 26, wherein the chalcone reductase has an
amino acid sequence greater than 60.7% identical to SEQ ID NO:28.

28. The nucleic acid molecule of claim 27, wherein the chalcone reductase has an
amino acid sequence comprising SEQ ID NO:28.

29. The nucleic acid molecule of claim 26, wherein the chalcone reductase has an
amino acid sequence greater than 61.6% identical to SEQ ID NO:29.

30. The nucleic acid molecule of claim 29, wherein the chalcone reductase has an
amino acid sequence comprising SEQ ID NO:29.

31. The nucleic acid molecule of claim 26, wherein the chalcone reductase has an
amino acid sequence greater than 61.3% identical to SEQ ID NO:30.
32. The nucleic acid molecule of claim 31, wherein the chalcone reductase has an amino acid sequence comprising SEQ ID NO:30.

33. The nucleic acid molecule of claim 26, wherein the coding sequence comprises SEQ ID NO:9, SEQ ID NO:10, or SEQ ID NO:11

34. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a chalcone isomerase.

35. The nucleic acid molecule of claim 34, wherein the chalcone isomerase has an amino acid sequence greater than 63.4% identical to SEQ ID NO:31.

36. The nucleic acid molecule of claim 35, wherein the chalcone isomerase has an amino acid sequence comprising SEQ ID NO:31.

37. The nucleic acid molecule of claim 34, wherein the chalcone isomerase has an amino acid sequence greater than 64.3% identical to SEQ ID NO:32.

38. The nucleic acid molecule of claim 37, wherein the chalcone isomerase has an amino acid sequence comprising SEQ ID NO:32.

39. The nucleic acid molecule of claim 34, wherein the coding sequence comprises SEQ ID NO:12, or SEQ ID NO:13

40. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a flavanone 3-hydroxylase.

41. The nucleic acid molecule of claim 40, wherein the flavanone 3-hydroxylase has an amino acid sequence greater than 82.4% identical to SEQ ID NO:33.

42. The nucleic acid molecule of claim 41, wherein the flavanone 3-hydroxylase has an amino acid sequence comprising SEQ ID NO:33.
43. The nucleic acid molecule of claim 40, wherein the coding sequence comprises SEQ ID NO:14.

44. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a flavonoid 3',5'-hydroxylase.

45. The nucleic acid molecule of claim 44, wherein the flavonoid 3',5'-hydroxylase has an amino acid sequence greater than 67.8% identical to SEQ ID NO:34.

46. The nucleic acid molecule of claim 45, wherein the flavonoid 3',5'-hydroxylase has an amino acid sequence comprising SEQ ID NO:34.

47. The nucleic acid molecule of claim 44, wherein the coding sequence comprises SEQ ID NO:15.

48. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a dihydroflavonol-4-reductase.

49. The nucleic acid molecule of claim 48, wherein the dihydroflavonol-4-reductase has an amino acid sequence greater than 67.6% identical to SEQ ID NO:35.

50. The nucleic acid molecule of claim 49, wherein the dihydroflavonol-4-reductase has an amino acid sequence comprising SEQ ID NO:35.

51. The nucleic acid molecule of claim 48, wherein the coding sequence comprises SEQ ID NO:16.

52. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a leucoanthocyanidin dioxygenase.

53. The nucleic acid molecule of claim 52, wherein the leucoanthocyanidin dioxygenase has an amino acid sequence greater than 73.1% identical to SEQ ID NO:36.
54. The nucleic acid molecule of claim 53, wherein the leucoanthocyanidin dioxygenase has an amino acid sequence comprising SEQ ID NO:36.

55. The nucleic acid molecule of claim 52, wherein the coding sequence comprises SEQ ID NO:17.

56. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a leucoanthocyanidin reductase.

57. The nucleic acid molecule of claim 56, wherein the leucoanthocyanidin reductase has an amino acid sequence greater than 59.9% identical to SEQ ID NO:37.

58. The nucleic acid molecule of claim 57, wherein the leucoanthocyanidin reductase has an amino acid sequence comprising SEQ ID NO:37.

59. The nucleic acid molecule of claim 56, wherein the coding sequence comprises SEQ ID NO:18.

60. The nucleic acid molecule of claim 1, wherein the coding sequence encodes a anthocyanidin reductase.

61. The nucleic acid molecule of claim 60, wherein the anthocyanidin reductase has an amino acid sequence greater than 77.4% identical to SEQ ID NO:38.

62. The nucleic acid molecule of claim 61, wherein the anthocyanidin reductase has an amino acid sequence comprising SEQ ID NO:38.

63. The nucleic acid molecule of claim 60, wherein the coding sequence comprises SEQ ID NO:19.

64. The nucleic acid molecule of claim 1, wherein the coding sequence is an open reading frame of a gene.

65. A mRNA molecule produced by transcription of the gene of claim 64.
66. A cDNA molecule produced by reverse transcription of the mRNA molecule of claim 65.

67. An oligonucleotide between 8 and 100 bases in length, which is complementary to a segment of the nucleic acid molecule of claim 1.

68. A vector comprising the coding sequence of the nucleic acid molecule of claim 1.

69. The vector of claim 68, which is an expression vector selected from the group of vectors consisting of plasmid, phagemid, cosmid, baculovirus, bacmid, bacterial, yeast and viral vectors.

70. The vector of claim 69, wherein the coding sequence of the nucleic acid molecule is operably linked to a constitutive promoter.

71. The vector of claim 69, wherein the coding sequence of the nucleic acid molecule is operably linked to an inducible promoter.

72. The vector of claim 69, wherein the coding sequence of the nucleic acid molecule is operably linked to a tissue specific promoter.

73. The vector of claim 72, wherein the tissue specific promoter is a seed specific promoter.

74. The vector of claim 73, wherein the seed specific promoter is a coffee seed specific promoter.

75. A host cell transformed with the vector of claim 69.

76. The host cell of claim 75, selected from the group consisting of plant cells, bacterial cells, fungal cells, insect cells and mammalian cells.
77. The host cell of claim 76, which is a plant cell selected from the group of plants consisting of coffee, tobacco, Arabidopsis, maize, wheat, rice, soybean barley, rye, oats, sorghum, alfalfa, clover, canola, safflower, sunflower, peanut, cacao, tomatillo, potato, pepper, eggplant, sugar beet, carrot, cucumber, lettuce, pea, aster, begonia, chrysanthemum, delphinium, petunia, zinnia, and turfgrasses.

78. A fertile plant produced from the plant cell of claim 77.

79. A method of modulating flavor or aroma of coffee beans, comprising modulating production or activity of one or more phenylpropanoid or flavonoid biosynthetic pathway enzymes within coffee seeds, wherein the phenylpropanoid or flavonoid biosynthetic pathway enzymes are selected from phenylalanine ammonia lyase, trans cinnamate-4-hydroxylase, 4-coumarate:CoA ligase, chalcone synthase, chalcone reductase, chalcone isomerase, flavanone 3-hydroxylase, flavonoid 3',5'-hydroxylase, dihydroflavonol-4-reductase, leucoanthocyanidin dioxygenase, leucoanthocyanidin reductase, and anthocyanidin reductase.

80. The method of claim 79, comprising increasing production or activity of the one or more phenylpropanoid or flavonoid biosynthetic pathway enzymes.

81. The method of claim 80, comprising increasing expression of one or more endogenous phenylpropanoid or flavonoid biosynthetic pathway enzyme genes within the coffee seeds.

82. The method of claim 80, comprising introducing a phenylpropanoid or flavonoid biosynthetic pathway enzyme-encoding transgene into the plant.

83. The method of claim 79, comprising decreasing production or activity of the one or more phenylpropanoid or flavonoid biosynthetic pathway enzymes.

84. The method of claim 83, comprising introducing a nucleic acid molecule into the coffee that inhibits the expression of one or more of the phenylpropanoid or flavonoid biosynthetic pathway enzyme-encoding genes.
Phenylalanine

PAL

\[ \text{Cinnamic acid} \]

C4H

\[ \text{p-Coumaric acid} \]

4CL

\[ \text{p-Coumaroyl CoA} \]

Salicylic acid
Coumarins

Lignin

Flavonoids
Isoflavonoids
Anthocyanins Stilbenes
Condensed tannins
Chlorogenic acid

Fig. 1a
Fig. 1b
Fig. 2a
Fig. 3a
Fig. 3b
1 of 4
Fig. 3b
4 of 4
pccp16II

Race1_CcPAL3
(pML19)

Coffea canephora PAL1
AF460203

↓ Full Length cDNA recovery in Coffea arabica T2308

pML14

Fig. 4a
Fig. 7a
Fig. 8a
Fig. 13a
Fig. 13b
Fig. 15c
Fig. 16
Fig. 17
Fig. 18
Fig. 19