6-28-2006

The development of molecular markers for use across all plant species using expressed sequence tags

Stephanie Douglas
Florida International University

DOI: 10.25148/etd.FI15101221

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

THE DEVELOPMENT OF MOLECULAR MARKERS FOR USE ACROSS ALL PLANT SPECIES USING EXPRESSED SEQUENCE TAGS

A thesis submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

in

BIOLOGY

by

Stephanie Douglas

2006

To: Interim Dean Mark Szuchman
College of Arts and Sciences

This thesis, written by Stephanie Douglas, and entitled The Development of Molecular Markers for use Across all Plant Species using Expressed Sequence Tags, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

Ray Schnell

Javier Francisco-Ortega

David N. Kuhn, Major Professor

Date of Defense: June 28, 2006

The thesis of Stephanie Douglas is approved.

Interim Dean Mark Szuchman
College of Arts and Sciences

Interim Dean Stephan L. Mintz
University Graduate School

Florida International University, 2006

ACKNOWLEDGMENTS

I would first like to thank Dr. David Kuhn, who sponsored me in spite of my lack of knowledge about molecular biology. I am very appreciative for the amount of patience and time he took to improve my knowledge. I would also like to thank my other committee members: Dr. Raymond Schnell and Dr. Javier Francisco-Ortega for their support and assistance throughout my research and writing my thesis.

I would also like to sincerely thank all of the staff at the USDA for use of the facilities during my research and for providing a pleasant and helpful working environment. My sincere thanks especially to Kyoko Nakamura, Cecile Olano, James Borrone, Wilbur Quintanilla, Katherine Cariaga, and all of the lab technicians who taught me everything about lab work; without their assistance I could not have conducted this research. I would also like to thank my friends and many of the professors at FIU who have made my graduate school experience extremely pleasant and memorable.

I would also like to thank Judith Evans-Parker for funding my fieldwork. Without her help I would never have been able to undertake my fieldwork halfway across the world. I would like to thank all those who helped me obtain my collecting permits and assisted in my fieldwork in New Caledonia, especially my sister Julie Douglas-Riso, Jorn Therkauf, and Jerome Muntzinger at the Institute Research and Development in Noumea.

Finally, I would like to thank FIU and Silvia Smith for the funding to design and obtain my primers.

ABSTRACT OF THE THESIS

THE DEVELOPMENT OF MOLECULAR MARKERS FOR USE ACROSS ALL PLANT SPECIES USING EXPRESSED SEQUENCE TAGS

by

Stephanie Douglas

Florida International University, 2006

Miami, Florida

Professor David N. Kuhn, Major Professor

There are over a half a million plant species on earth, and we use them in virtually every aspect of our lives. Little or no genomic information exists about the vast majority of these plants. This study investigated the use of Expressed Sequence Tags (ESTs) to locate highly conserved sequences from which to design a set of universal molecular markers for all plant species. Plant species for this study were chosen to representative of the plant kingdom. This was done by sampling several individuals of at least one species from all of the major terrestrial plant groups.

Conserved sequences are generally found in a wide range of plants species and often in all plant species. A set of eight degenerate primers was designed specifically to detect Single Nucleotide Polymorphisms (SNPs) using capillary array electrophoresis-single stranded conformational polymorphism (CAE-SSCP). The results of this research confirmed that homologous regions of the genome could be used to design universal molecular markers for all plant species.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

I. INTRODUCTION

There are approximately over a half a million plant species on earth. Little or no genomic information exists for the vast majority of these plants. This study investigated the use of Expressed Sequence Tags (ESTs) to locate highly conserved sequences from which to design a set of universal molecular markers for all plant species. A set of universal molecular markers for all plants would considerably reduce the cost and time it currently takes to develop genetic markers separately for each plant species. Plant species for this study were chosen to be representative of the entire plant kingdom. This was accomplished by sampling several individuals of at least one species from all of the major terrestrial plant groups in the plant kingdom.

Currently, the most commonly used methods to detect genetic variations in plants are Polymerase Chain Reaction (PCR) based methods. The most frequent type of genetic variations that occur in plants are single nucleotide polymorphisms (SNPs). These occur consistently throughout the plant genome. A PCR-based method known as, Capillary Array Electrophoresis-Single Stranded Conformational Polymorphism (CAE-SSCP) is a simple and reproducible method that does not require prior sequence information to detect SNPs. This method detects differences based on the mobility of single-stranded DNA caused by sequence differences. There are several other methods that can be used to identify SNPs but many of these require prior sequence information, are difficult to conduct and analyze, or are not reproducible.

A set of universal molecular markers developed for CAE-SSCP analysis from conserved ESTs could have many applications. Some of these include, species

identification, identifying syntenic regions of the genome, population/conservation studies, extending and creating linkage maps, or marker-assisted selection (MAS).

Expressed Sequence Tags are a generous source of publicly available genomic information through the National Center for Biotechnology Information (NCBI) website. Expressed Sequence Tags represent fragments of coding regions of the genome. Many of these "tags" are highly conserved and thus can be mined for homologous regions from which to design a set of universal plant primers. Homology searches such as Basic Local Alignment Search Tool (BLAST), coupled with any of several primer design programs have made it fairly simple and economical to design a set of universal plant primers.

The following research describes the possibility of designing molecular genetic markers from conserved Expressed Sequence Tags (ESTs), for genetic analysis to work in any plant species. Eight primer pairs were designed to be tested in a representative sample of the plant kingdom for amplification. Following successful amplification the plant species will them be tested for polymorphism using Capillary Array Electrophoresis-Single Stranded Conformational Polymorphism (CAE-SSCP) analysis.

Plant Classification

Green plants have been around for about a billion years (Mishler 2000; Wolf et al., 2005), and fossil records indicate that land plants have been in existence for half of that time (Karol et al., 2001; Judd et al., 2002; Stern 2000). Presently, there are an estimated half a million extant plant species (Judd et al., 2002; Wolf et al., 2005). As a result of newly discovered plant species, and the increasing availability of molecular data,

the classification of plants changes frequently. However, there are still certain groups that are generally accepted as distinct within the plant kingdom.

The first major divergence occurs between the land plants including a few fresh water green algae groups (Charophyta), and the remaining green algae groups (Chlorophyta) (Karol et al., 2001). The land plants are collectively known as the Embryophytes. The Embryophytes can be divided into non-vascular (Bryophytes) and vascular (Tracheophytes) plants. The main distinction between these two groups is the presence of vascular tissue, (xylem and phloem). Some Bryophytes (mosses) do have water and food conducting vessels, which are similar to xylem and phloem, known as hydroids and leptoids respectively (Stern, 2000). However, these conducting vessels are much less efficient than xylem and phloem and the Bryophytes generally obtain most of their water through absorption (Stern, 2000).

The Bryophytes have a fossil record that extends back about 370 million years. Currently there are three groups in existence: liverworts, hornworts and mosses (Goremykin and Hellwig 2005). Bryophytes are thought to be relatives of the earliest land plants, but their relationship to embryophytes and to each other are still unclear and it is likely that they form a paraphyletic group (Goremykin and Hellwig 2005; Pryer et al., 2001; Kenrick and Crane 1997; Kato and Akiyama 2005; Karol et al., 2001; Judd et al., 2002). The mosses are the most diverse of the three groups and contain ~15,000 extant species (Judd et al., 2002; Stern 2000). Liverworts resemble mosses, but they lack stomata, and only have about 8000-9000 extant species (Judd et al., 2002; Stern 2000). Hornworts are the rarest of the non-vascular plants with only ~100 species, and differ

from mosses and liverworts by the ability of their sporophytes to carry out photosynthesis (Judd et al., 2002; Stern 2000; http://www.sirinet.net/~jgjohnso/seedless.html).

Vascular plants (Tracheophytes) can be divided into the seed plants and non-seed plants. Non-seed plants consist of four major divisions including: the Lycophytes (clubmoss and quillworts), Spenophytes (horsetails and scouring rushes), Psilophytes (whisk ferns), and Pterophytes (true ferns). The Lycophytes are a monophyletic clade that is sister to all other vascular plants (Kenrick and Crane, 1997; Wolf et al., 2005). Lycophytes thrived during the Carboniferous era and fossil records from this group indicate that modern species are morphologically similar to their ancient relatives (Stern, 2000). Today there are about 1,200 known species of lycophytes. Division Psilophyta lacks extensive fossil records, but was previously thought to be the most primitive of the ferns and their allies, because it is the only division of vascular non-seed plants that lack leaves or roots (Judd et al., 2002; Stern, 2000). However, other molecular evidence has shown them to be more closely related to the seed plants (Kenrick and Crane, 1997). Presently, the Psilophytes have approximately 15 species remaining. Division Spenophyta, which also thrived during the Carboniferous era, has only a single genus (*Equisetum*), with about 15 species remaining today (Judd et al., 2002; Stern, 2000). Division Pterophyta also thrived during the Carboniferous era and today are still a very diverse group of plants with more than 11,000 extant species known (Judd et al., 2002; Stern, 2000).

The main difference between the non-seed and seed plants is the protection and food supply that seeds provide for the embryos that spores do not (Judd et al., 2002). This protection can defend the embryo against drought, fire, pest invasion and freezing

(Stern, 2000). The seed plants have two major divisions commonly known as the "gymnosperms" (naked seed) or cone plants and "angiosperms" (vessel seed) or flowering plants (Judd et al., 2002; Stern, 2000). Gymnosperm seeds form on the surface of the sporophyll while angiosperms bear their seeds in a fruit. The first seed plants were morphologically similar to the ferns. Fossil records indicate that gymnosperms appeared more than 320 million years ago (mya) and angiosperms more than 130 mya (Chaw et al., 2000; Stern, 2000). The four divisions within the gymnosperms are: Cycadophyta, Ginkgophyta, Gnetophyta, and Pinophyta. Division Cycadophyta contain ~130 species today and is considered to be the most basal of the extant gymnosperms, appearing more than 300 mya and flourishing during the Mesozoic era (Judd et al., 2002; Stern, 2000). Division Ginkgophyta has survived since the Jurassic (~ 170 mya) and has only a single species (*Ginkgo biloba*) today (Brenner et al., 2005; Judd et al., 2002; Stern, 2000). Much debate exists about the Division Gnetophyta and its relationship within the seed plants. The "anthophyte" hypothesis regards the Gnetales as sister to the angiosperms on the basis of morphological features (Bowe et al., 2000; Chaw et al., 2000; Doyle 1998). However, some molecular data refutes this claim and places the Gnetophyta firmly within the gymnosperms, sister to the Pinophyta (Bowe et al., 2000; Chaw et al., 2000; Doyle 1998; Soltis et al., 2002). Currently, there are approximately 75 extant species of Gnetophytes (Judd et al., 2002). The Division Pinophyta dates to over 300 mya and thrived during the Carboniferous era (Judd et al., 2002; Stern, 2000). An estimated 600 species remain today (Judd et al., 2002).

Sometime around the mid-Cretaceous (~100 mya) angiosperms began to dominate the earth (Kendrick and Crane 1997; Pryer et al., 2001; Qiu et al., 1999).

Currently, flowering plants make up more than 75% of all the plant species. Originally this group of plants was split into two groups: "monocots" or "dicots". Currently, molecular and morphological data reveal the monocots as a monophyletic group, but the dicots are a paraphyletic group (Zimmer et al., 2000; Soltis et al., 2000). Classification of angiosperms changes frequently based on the latest molecular and morphological data. Because of this diversity it is very difficult to place all of the angiosperms families (table 1) into irrefutable clades, so there still remains much ambiguity in classification. A group of plant systematists known as the Angiosperm Phylogeny Group (APG II, 2003) was formed in 1998 to resolve and maintain a consensus on the classification of angiosperms.

Genetic Variation in Plants

Recombination during meiosis is the most common cause of genetic variation in plants (Hartl and Jones, 2002; Judd et al., 2002). Mutations can also occur from environmental factors such as radiation or sunlight. Substitutions, insertions/deletions (indels), duplications, inversions, and losing or gaining entire chromosomes are all types of mutation (Hartl and Jones, 2002; Judd et al. 2002). Variation occurs less often in genes that are necessary for the survival or reproduction of an organism such as, housekeeping genes, genes involved in meiosis, or those that code for metabolic functions (Hartl and Jones, 2002; Judd et al., 2002). These genes are said to be conserved and are sometimes termed "orthologs", which are genes that have the same or a comparable function and can be traced back to a common ancestor (Hartl and Jones, 2002; Judd et al., 2002). There is a possibly that these orthologs exist relatively unchanged in a wide variety of organisms simply because any change in these necessary genes would likely kill the organism before reproduction, and any chance of the mutation

being passed on would be lost. While it is likely that these conserved regions exist in all plant species, there is also the possibility that these homologous regions could have more than one copy within the genome. Consequently, this could cause added complications when detecting genetic variations. For example, if more than one fragment is amplified by the same set of PCR-primers, it would be impossible to determine which alleles belong to which locus without sequencing. In this case determining homozygous and heterozygous individuals would become difficult if not impossible.

Previously, detecting genetic variation was done using difficult methods such as, restriction fragment length polymorphism (RFLPs) or isozymes. Polymorphism using RFLPs is detected by differences in the size of the fragments cleaved by restriction enzymes. However, this method is lengthy, uses radioactivity and requires large amounts of high quality DNA (McCallum et al., 2001). Isozymes can be an accurate and useful method to detect polymorphism, however, the process is lengthy and complicated (Hillis and Moritz, 1990). With the development of polymerase chain reaction (PCR) in 1983, detecting genetic variations became a great deal easier (Mullis, 1986). PCR allows particular fragments of DNA to be amplified and then copied several thousand times (Mullis, 1990). These fragments can then be ananlyzed for sequence variations by a variety of different methods. If the amplicons demonstrate polymorphism across individuals, they can be used as molecular genetic markers. Polymorphism is the presence of two or more different forms of an allele that occurs in more than 1% of the population. However, PCR is temperature dependent on the composition of the oligonucleotides. This can possibly cause complications when using degenerate primers. Degeneracy is the insertion of all the nucleotides at several positions in the sequences so

that if there are differences across species then the primers will still anneal to the template DNA and generate product. Degeneracy reduces the specificity of the primers, and is generally used when amplifying homologous sequences from different species. To ensure optimal annealing a temperature gradient should be run for each primer when testing a new species with degenerate primers, because there will likely be sequence variations across species, especially at the 3rd base position of codons. Sequence variations can cause the specific annealing temperature to vary for the same set of primers. This is because each nucleotide contributes different properties that can affect the annealing temperature. In primers that have several degenerate positions this could make a significant difference in the annealing temperature from one species to another. There is also the possibility of differences within species and this could also cause slight differences in annealing temperature. Once a gradient is done it is best to choose the highest temperature that has good amplification product. This can reduce the nonspecific binding that can occur with lower temperatures, and ensure better resolution when running the product under CAE-SSCP conditions.

PCR methods

There are several PCR-based methods used to detect genetic differences. The most common methods are Random Amplified Polymorphic DNAs (RAPDs), Amplified Fragment Length Polymorphism (AFLPs), Microsatellites (SSRs), and Single Stranded Conformational Polymorphism (SSCP).

Random amplified polymorphic DNAs (RAPDs) is a PCR-based method that uses non-specific primers to amplify anonymous fragments of DNA. These are "fingerprints" of DNA where the bands can be scored as present or absent. This method is quick and

economical, but it is not as informative as other PCR-based molecular markers and is not easily reproducible (Gallego et al., 2005; Garcia et al., 2004; Mukhtar et al., 2002). Previous genomic information is not needed for this method, but it is very susceptible to small changes in the processing method (Garcia et al., 2004). Also, because RAPDs are dominant markers, heterozygous and homozygous states cannot be determined (Nair et al., 1995).

Amplified Fragment Length Polymorphism (AFLP) uses genomic DNA fragments previously digested by restriction enzymes for amplification by PCR, and determines polymorphism based on length differences. This method also uses arbitrary primers and does not require prior genomic information. AFLPs can potentially identify a large number of loci, and is commonly used to map a segregating population (Gallego et al., 2005, Harry et al., 1998, Mohan et al., 1997 and Vos et al., 1995). Although AFLPs are more reliable than RAPDs, it is very costly and produces complex patterns because of the high number of amplified fragments. Gels yield several bands and sometimes the patterns can be difficult to interpret (Garcia et al., 2004; Vidal et al., 2000). Also, AFLP markers are dominant markers, and so as with RAPDs different allelic forms cannot be distinguished.

Microsatellites (SSRs) are simple sequence repeats usually consisting of two to six nucleotides. Microsatellites occur in both coding and noncoding regions of the genome, but are more abundant in introns, and can be found in all eukaryotes (Gupta and Varshney 2000). Polymorphism is determined by length differences that are caused by the number of repeats present in an individual. Microsatellites can be used for genotyping individuals, genomic mapping, and population studies (Cordeiro, 2001).

However, developing these makers is very expensive and time consuming because they usually have to be designed separately for each species (Chen et al., 2002; Liewlaksanneeyanaein et al., 2004). With the development of ESTs libraries the cost of designing SSR markers has decreased significantly. Recently, SSRs have been developed that have amplified DNA polymorphisms among directly related plant species (Bandopadhyay et al., 2004; Chen et al., 2002; Cordeiro et al., 2001; Fraser et al., 2005; Glazmann et al., 1997; Liewlaksanneeyanaein et al., 2004; Saha et al., 2004). If SSRs designed from EST libraries to amplify across genera are successful, this could be very beneficial because they could be associated with candidate genes (Bandopadhyay et al., 2004; Chen et al., 2002; Gupta and Varshney 2000; Saha et al., 2004). However, SSRs are less likely to be found in coding areas of the genome, thus limiting the use of ESTs for this method (Decroocq et al., 2003)

Point mutations are the most common type of genetic variation found in the plant genome (Bundock et al., 2006; Kuhn et al., 2005 Schneider et al., 2001; Torjek et al., 2003). These mutations usually result from substitutions or insertion/deletions (indels) and can be termed single nucleotide polymorphisms (SNPs) (Bundock et al., 2006; Kuhn et al., 2005; Sato and Mishio, 2003; Schneider et al., 2001; Torjek et al., 2003). The more common SNPs are known as transitions, which are the substitution of a purine (adenine or guanine) for a purine or the substitution of a pyrimidine (thymine or cytosine) for a pyrimidine. Transversions are less common and are a substitution of a purine for a pyrimidine and vice versa. Transitions are more common than transversions because of the size of the bases, and the difference in the number of hydrogen bonds between bases. SNPs occur most frequently at the 3rd nucleotide position of a codon because changing

the base pair at this position usually does not change the code for the amino acid sequence. It is for this reason that to design successful primers for amplification across species there must be some degeneracy in the primers. Most of the degeneracy will be at the 3rd base position of the codons because this is where most of the variation will likely occur. SNPs appear in both coding and noncoding regions of the plant genome at a rate of about every 140 nucleotides (Kuhn et al., 2005; Salmaso et al., 2004; Schneider et al., 2001). There are several methods used to detect SNPs, but most of them are technically demanding, need extremely controlled conditions, or require previous information about the SNP or corresponding sequence, (Kirk et al., 2002; Konieczny and Ausubel 1993; Kuhn et al., 2005; Plomion et al., 1999). SNP detection methods are reviewed in Kwok (2001), Gupta et al (2001), or Gut (2001). A reproducible and accurate method that easily detects SNPs without sequencing is CAE-SSCP analysis. This method does not use size differences; instead it detects conformational changes of single-stranded DNA by differences in mobility. After DNA is denatured, single-stranded DNA undergoes intrastrand base pairing and as a result the molecule has a complex 3-D configuration that results in looping and compression. These configurations may display different mobility depending on their shape when electrophoresed under specific conditions. Thus, DNA that has a change of even one base pair, but has the same length, may fold into a different conformation and may appear very different when analyzed under non-denaturing conditions by electrophoresis (Arakawa et al., 1996; Kong et al., 2003; Kuhn et al., 2005; Slabaugh et al., 1997). These conformations are sequence and temperature dependent. Thus polymorphism might be detected at one temperature but not another, so the product should always be run at two different injection temperatures. In the instance of gel-based

systems the temperature of the gel, product, buffer, and voltage can be adjusted to ensure two different running temperatures. Optimally, the size of the fragment to be analyzed by SSCP should be between 150-250 base pairs (bp), because accuracy and sensitivity decreases for larger fragments (Kuhn et al., 2005), although fragments as large as 700bp have been analyzed successfully using gels (Bryja et al., 2005; Jordanova et al., 1997; Orti et al., 1997; Sunnucks et al., 2000). Originally, SSCP was developed as a method to discover SNPs in humans (Orita et al., 1989) and since its development, over a million SNPs have been identified in the human genome (Rafalski, 2002). SSCP markers are codominant, meaning that if amplifying a single biallelic locus then homozygous and heterozygous states can be determined, and thus segregation patterns can be determined. Even if more than one locus is being amplified pattern differences across individuals or across species can still be determined. Thus, SSCP markers have several different applications including; population studies, mapping ESTs, constructing saturated linkage maps, finding relationships between genotype and phenotype, and determining disease susceptibility (Baba et al., 2003; Kong et al., 2003; Kuhn et al., 2005; Sunnucks et al., 2000). Degenerate primers have been successfully developed to identify SNPs within WRKY and for resistance gene homologues (RGHs) genes by CAE-SSCP analysis (Borrone et al., 2003; Kuhn et al., 2003). Genetic markers for SSCP analysis have also been recently developed from ESTs for use in SNP detection in pearl millet with great success (Bertin et al, 2005).

Applications for Molecular Markers

CAE-SSCP molecular markers have several applications such as, population/conservation studies, marker assisted selection (MAS), establishing

synteny/comparative genomic mapping, creating or extending linkage maps, and identifying candidate genes (Baba et al., 2003; Kuhn et al., 2003; Liewlaksanneeyanawin et al., 2004; Mohan et al., 1997).

Universal plant genetic markers could be employed to genotype all individuals within a population. This would determine the amount of diversity and the frequency of heterozygosity within that population. In a study of rare or endemic species this information could be very useful. For instance, if all of the individuals in the population showed the same genotype, this could indicate that the entire population was a result of clonal reproduction or that the individuals were self-pollinated. If this happened to be the only known population of that particular plant, then that species would be in high danger of becoming extinct. This type of information could aid in determining which natural areas were in the most need of protection.

PCR-based markers can also be used to identify syntenic regions of the genome. Synteny is the assumption that the organization of genes is comparable in all plant species (Devos and Gale, 1997; Galzmann et al., 1997). If the location of specific genes is known for one plant, it can be presumed that the arrangement will be the same in another species (Kuhn et al., 2003). Syntenic regions can be conserved across species, and in some cases can be identified in distantly related species (Gualtieri et al., 2002; Yan et al., 2004). Previous research has demonstrated high amounts of conservation in genome comparison studies of several species in the Solanaceae family (Schmidt, 2000). Detecting syntenic regions in a species with a complete genomic map such as in *Arabidopsis thaliana* could possibly correspond with similar regions of the genomes of related species, and even in all plant species. This would make it less difficult to create

linkage maps for species that have little or no information about their genome, and aid in the identification of candidate genes (Acarkan et al., 2000; Devos and Gale 2000; Grant et al., 2000; Kuhn et al., 2003; Paterson et al., 2000; Salse et al., 2002; Yan et al., 2004).

Traditionally, farmers have selected the individuals from their crops that demonstrated the most desirable traits for breeding purposes. Now genetics and traditional selection can be utilized together by a technique known as marker assisted selection (MAS). MAS is the use of genetic markers associated with certain traits of interest (phenotypes) to identify individuals with the desired genotype for breeding purposes (Dirlewanger et al., 2004; Mohan et al., 1997; Steele et al., 2006). Using molecular markers for MAS can propose a quicker and more effective way for selecting the best cultivars within a population (Dirlewanger et al., 2004; Fazio et al., 2003; Liewlaksanneeyanaein et al., 2004; Mohan et al., 1997). One major advantage to this technique is that the plant can be tested with molecular markers for the preferred traits at any time during its life, thus saving years of waiting for the plant to reveal certain characteristics in the field (Dirlewanger et al., 2004 and Mohan et al., 1997). Another advantage to this is that markers are not likely to be affected by environmental factors and so growing conditions usually do not influence the results (Mohan et al., 1997). Many important agronomical traits are quantitative traits, meaning that they are not just coded for by a single gene, but rather involve interactions among several loci across the genome (Lande and Thompson, 1990). Thus, several molecular markers are needed for an organism before useful information can be obtained. A large source of genomic information for development of molecular markers is publicly available at the NCBI website (http://www.ncbi.nlm.nih.gov).

Bioinformatics

Several million EST sequences are available from the National Centre for Biotechnology Information (NCBI). This is a rich source of information from which molecular genetic markers can be developed to identify SNPs. ESTs are small coding regions of the genome, usually about 200-700bp in length (Murray et al., 2005). They commonly only represent segments of genes and not the complete coding sequence. However, they are still specific enough to identify the gene that is "tagged". Production of ESTs is quick and relatively easy and needs only one sequencing reaction for each complementary DNA (cDNA) generated (Pappas et al., 2005). They can be cloned randomly or in one direction and sequencing errors do not need to be checked as mistakes do not usually prevent identification of the gene from which the gene was derived (Murray et al. 2005). ESTs are made by first isolating messenger RNA (mRNA), and then using reverse transcription to generate cDNA. Complementary DNA is synthesized from mRNA, so it does not include introns because they are spliced out. Then the cDNA is cloned into a vector to produce a cDNA library. These sequences can then be accessed through the NCBI website. However, many of the sequences available are repetitive copies of the same EST and can contain errors from genomic or vector contamination (Murray et al., 2005). Although there is a high error rate in sequences in EST libraries, alignment searches can be used to screen out those errors caused from vector contamination (Murray et al., 2005). Also, because of the high amount of redundancy and error associated with the EST database it is more efficient to do alignment searches with EST sequences against the non-redundant database rather than the EST database. Using the non-redundant database reduces the likelihood of inaccurate matches. It also

offers the likelihood of receiving matches with genomic DNA that might contain introns. If the goal is to include an intron within the sequence from which to design primers, then using that type of database can be helpful. Also, it is very informative to do alignment searches against more than one database and then to compare the results.

ESTs can be downloaded into a local database and then subjected to homology searches. Sequence comparison programs such as BLAST (basic local alignment search tool), can aid in the discovery of highly conserved sequences occurring in genetically distant plants. BLAST alignment searches can be done at the nucleotide or protein level of a specified database. This is done by using a maximal segment pair (MSP) score (Altschul et al., 1990). MSP chooses the pair of sequences of the same length with the highest homology score (Altschul et al., 1990). Parameters can also be changed to utilize gaps, or reading frames (Altschul et al., 1990; Altschul et al., 1997). A reading frame is the specific point in mRNA where the sequence is "read" in groups of three nucleotides known as a codon. In an incomplete sequence there are three possibilities from which start reading the sequence and so there are three possible reading frames. Generally the reading frame is initiated from the start codon AUG. From these BLAST reports the actual sequence can be checked over and conserved regions can be determined for primer design.

Several programs are freely available to the public for primer design. One of these is from the Joint Center for Structural Genomics (JCSG) (www.jcsg.org/prod/primer). This particular program designs primers based on melting temperature, GC content, clamp score and length of sequence. The melting temperature is derived based on sequence length, GC content and salt concentration (Canaves et al.,

2004). The higher the GC content the higher the annealing temperature will be. Clamp score is the existence of a G or C at the 3' end of the primer, and is calculated using the last three base pairs of each primer. The presence of a G or C at this position helps ensure accurate binding due to the stronger hydrogen bonds of cytosine to guanine. The higher the clamp-score, the better the primer sequence will anneal to the DNA template. Sequence length should be at least 18 bp to ensure specific binding and GC content should be around 45-55%. A lower percentage of GC content generally does not cause PCR failures if the primers have an adequate clamp score (Canaves et al., 2004). This program does not measure primer-dimer or self-annealing, so to be sure that the oligonucleotides do not have these properties it is wise to double check them in any of a number of other primer design programs. One such program available through GCG accelrys software is PRIMEPAIR*. This program evaluates sequences based on melting temperature, GC content, primer length, clamp score, self-annealing, and primer dimer potential. It essentially measures the same parameters as JCSG, with the addition of annealing properties of the oligonucleotides. Both programs are simple to use and it is informative to compare results from more than one program.

Since ESTs only represent the coding regions of the genome, the presence of introns cannot be identified. A program known as SPIDEY at the NCBI website uses an mRNA sequence and aligns it with a genomic sequence to detect introns. A report is returned with the expected position and the size of the introns within the mRNA sequence as determined from the genomic sequence. Since only a few plant genomes have been sequenced it is likely that the mRNA and genomic sequences will be from different species and because of this there is potential for error.

PCR-based primers between 150-250bp in length, developed from ESTs for CAE-SSCP analysis should exhibit polymorphism in most cases since SNPs occur frequently and consistently throughout the plant genome. Since these primers will be designed from highly conserved homologous regions of the genome there is a high potential that these primers will amplify any plant species. Also, because of the high occurrence of SNPs within the plant genome, most of these plants should also exhibit polymorphism. The development of a set of universal molecular markers would greatly reduce the cost and time it takes to develop them separately for each plant species.

Objectives

The objective of this study was to develop eight PCR-based primers for use as universal genetic markers in plants. Once the primers were developed, I tested them for amplification using plants from each of the major groups across the plant kingdom. A very wide range of plant species were included in this study to ensure diversity. I sampled commercially important plant species, plant species from different climates, habitat types and endemic species. Endemic species were chosen from New Caledonia because of the high rate of endemism that occurs there. The inclusion of some endemic species from this country would further strengthen the results because of the isolation of the species while undergoing evolution. If the first set of eight species amplified the expected PCR-product then subsequent species were tested for amplification. Following successful amplification all species were then tested for SNPs using CAE-SSCP.

II. MATERIALS AND METHODS

Primer Design

This research was conducted at the USDA in Miami, Florida. Much of the research at this facility focuses on *Theobroma cacao*. Since a large amount of *T. cacao* DNA and genomic information was available, the sequences used to design these universal primers came from *T.cacao*. *Gossypium hirsutum* was initially used as a basis for sequence comparison because it was the closest relative to *T. cacao* that had large amounts of sequence data available to the public. *Arabidopsis thaliana* was used as a third point of reference for homologous sequences because it has had its entire genome mapped, and thus the function of the genes from the sequences was known. This was beneficial when looking for sequences with different functions from which to design primers.

Over 17,000 *Gossypium hirsutum* EST sequences were downloaded into SeqLab (GCG accelrys software), and made into a local database. Following that, approximately 7000 *Theobroma cacao* sequences were also downloaded into SeqLab for use as query sequences for BLASTN (Altschul et al., 1997) searches against the created local *G. hirsutum* database. Several hundred high scoring results were returned. These results were then subjected to NETBLAST (Altschul et al., 1997) searches against the non-redundant database at the NCBI website. Following that, NETBLAST results were filtered for high scoring alignments and the inclusion of *A. thaliana* in the report. Over 300 sequences were returned that fit the criteria. Sequences were chosen based on BLAST scores, and which species had high homology with *T. cacao* and *Gossypium*. If

the sequences had high homology with *Arabidopsis* then they were a candidate for selection. Primers were designed by identifying the most conserved regions of the sequences from the alignments on the BLAST reports. Regions of 150-250 nucleotides (nt) in length that had the least number of mismatches, especially near the 3' end were chosen. Final choices were put into the JCSG (joint center for structural genomics) primer design program. (http://www.jcsg.org/scripts/prod/primer/primer_input_form.cgi). This program designs forward and reverse primers based on temperature compatibility, GC clamp score, and primer length. This program does not measure primer-dimer, self-annealing, or hairpins, so primers were checked again with the PrimePair* program in SeqLab. Primers were then ordered with each strand fluorescently labeled with a different dye for SSCP analysis. Primers were initially diluted to 100uM with 1xTE buffer (10mM Tris and 1mM EDTA), according to the nmol concentration of each primer. The primers were further diluted to 10uM with nuclease free water. Following the testing of all the primers, a program known as SPIDEY (http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/) from the NCBI website was used to determine the presence of introns. This program uses an mRNA-genomic DNA alignment system. Genomic sequences were taken from *A. thaliana* and the mRNA sequence from the corresponding *T. cacao* sequence was used.

Sampling Strategy

Plant species were chosen to represent all major terrestrial plant groups (Table 1). Representatives were chosen from tropical, subtropical and temperate climates, as well as a number of endemic species. Species collected from New Caledonia (permit n° 6034-

3782), were chosen for their level of endemism, three of the four endemic species were endemic to New Caledonia at the family level. Leaf tissue from eight to twelve individuals was collected from each species (Table 1). Reasons for choosing first eight and then twelve individuals was because 96-well plates were used to carry out the PCR and electrophoresis reactions. These plates have twelve wells across and eight wells down, so initially the setup used the eight wells down for the different individuals, then as the research progressed and temperature gradient tests were added a step in the process, the setup was changed and the twelve wells across were used for the different individuals of the same species. This was set up this way to be able to test the species at twelve different temperatures. This also increased the sample size so that determination of polymorphism was more likely with more individuals.

DNA Extraction

Both fresh and dried leaf tissue was used for DNA extraction. DNA extraction was done with 75-200mg of leaf tissue depending on whether the leaf material was fresh or dried. The fast DNA kit (BIO 101, INC; Carlsbad Calif.) procedure was used to extract the DNA following the manufacturers' instructions. Tissue was homogenized with the Garnet Matrix and two 1/4 –inch spheres as the Lysing Matrix combination using the FastPrep FP120 instrument on speed 5 for 30 seconds, repeated three times. DNA quantification was performed using an automated spectrophotometer (Spectramax 190) in conjunction with Softmax Pro 4.3.1 LS software package (Molecular Devices; Sunnyvale, Calif). DNA was then diluted to approximately 5ng/ul.

PCR amplification

Initially, a temperature gradient from was run for each species against all eight primers to determine the best annealing temperature. Amplifications were executed in 20ul volumes including, 2ul BSA (10mg/ml), 2ul ThermoPol Buffer 10x (BioLabs inc.), 0.4 ul (10mM) dNTPs, 0.4 ul of both forward and reverse primers (10 um), 0.14ul Amplitaq (5U/ul), and 2ul (5ng/ul) of genomic DNA. PCR amplification reactions were conducted using the following thermocycling profile: 2 minutes of denaturation at 94°, 33 cycles of denaturation at 94° for 30 seconds, 1 minute of annealing with the temperature gradient of 45-65°, and 1 minute of extension time at 72°; followed by a final extension period of 5 minutes at 72°, and then held at a temperature of 4°. All thermocycling reactions were done on a PTC-225 DNA engine tetrad thermal cycler. Once the optimal temperature was established for each species for each primer set PCR reactions were conducted as accounted above with the exception that the annealing temperature step in the thermocycling profile was modified to be optimal for each species and primer (tables 3-10). Following amplification, each product was prepared for electrophoresis by combining 1ul of PCR product, 0.1ul of Genescan 500 ROX size standard (Applied Biosystems, Inc.), and 20ul of dH_2O. Samples were then denatured for 30 seconds at 95°, and then chilled on ice. Capillary electrophoresis was executed with an ABI Prism 3730 Genetic Analyzer using Performance Optimized Polymer (POP) 7 polymer (Applied Biosystems, Inc.). Samples were injected electrokinetically at 2kV for 10 seconds and run at 15kV for 20 minutes at 66°C. Analysis was done using GeneMapper 3.7 software (Applied Biosystems inc.), to determine success of amplification and

fragment size. Once it was concluded that fragments were of the expected size with sufficient product, samples were then prepared for CAE-SSCP. The amount of PCR product to be added to the mix was determined based on electrophoresis data from the ABI 3730, and varied from 1ul to 8ul. The remaining volume was made up with dH2O to 10ul total. The samples were first denatured at 94° for 3 minutes and then snap-cooled on an ice-slurry for 5 minutes. Following denaturation, 10ul of a 1:100 dilution of GeneScan ROX 2500 (Applied Biosystems, Inc.), was added to each well containing the denatured PCR product. Capillary electrophoresis was conducted on the ABI 3100 Genetic Analyzer. Products were separated on a 36cm capillary containing 5% GeneScan polymer (Applied Biosystems Inc.), in 1x TBE, and 10% glycerol. Injection was set for 22 seconds at 1.5 kV and all plates were run at both 22° and 28°C. Analysis for polymorphism was done using GeneScan v.3.7 and Genotyper v.3.7 software (Applied Biosystems Inc.).

Cloning and Sequencing

Cloning reactions were done using the TOPO4© Cloning Kit (Invitrogen Corp., Baltimore, MD) following the instructions of the manufacturer. Plates for transformations were prepared with LB media using 10g of Bacto tryptone, 5g Bacto yeast extract, and 10g NaCl dissolved in 1 liter of dH2O and then autoclaved. Before pouring plates, 2 mls Ampicillin (50mg/ml), is added for a final concentration of 100ug/ml. Transformations were plated out and incubated overnight at 37°C. Individual colonies were then picked and transferred to 96-well plates with SOC broth containing 100ug/ml ampicillin, incubated overnight at 37°C, with mild agitation at 225rpm. SOC broth was prepared to a final volume of 1 liter with the following ingredients: 10ml KCl

(250mM), 20g tryptone, 5g yeast extract, 0.5g NaCl, 5ml $MgCl_2$ (2M), 20ml glucose (1M) and 950ml dH2O. Following overnight incubation plates are centrifuged to pellet cells and the broth removed. Cells are then resuspended in 10mM Tris-HCL pH 8.0. Subsequently, the M13 insert amplification was prepared for a 12ul PCR reaction with the following recipe; 1.5ul buffer (10x), 0.6ul dNTPs, 0.3ul (10um) of both forward and reverse M13 primers, 0.06ul Taq (5U/ul) and 9.24ul dH2O. Thermocycling profile as follows; denaturing at 94° for 5 minutes, 30 cycles of denaturing at 94° for 30 seconds, 60°C for 30 seconds of annealing, and 2 minutes of extension at 72°C, a final extension at 72° for 10 minutes was performed, followed by a holding temperature of 4°. Following amplification an exonuclease reaction with 10ul of the following reaction mixture added to each well: 6ul of buffer (1X), 0.075 exonuclease enzyme (0.025U/ul), and 3.925 dH_2O. The plates were covered with foil tape and inverted several times to mix the solution and then were left to stand at room temperature for 15 minutes. The plates were then centrifuged @ 1700 rpm for 30 minutes, then inverted and centrifuged again @ 700 rpm for 1 minute and then left to air dry for 15 minutes. This was followed by cycle sequencing using the T7 primer. The mix consisted of 10ul reactions each with 1ul of T7 primer (1.6pmol/ul), 1.75ul 5X buffer (Applied Biosystems), 0.5ul Big Dye Terminator v3.1 (Applied Biosystems), and 6.75ul dH_2O. Thermocycling profile was 26 cycles each of, 96°C for 10 seconds, 50°C at 5 seconds, and 60°C for 4 minutes. Ethanol precipitation followed, by adding 32uls of 95% ethanol and 8ul dH2O per well and mixing the plates by inverting several times. Plates were placed in a dark place so products could precipitate for more than 1 hour. Plates were then centrifuged @ 1810 rcf for 45 minutes to pellet the product, and then inverted and centrifuged at 50 rcf for 30

seconds to remove supernatant. The plates were then air dried for at least 15 minutes and the product was resuspended in 20ul of dH2O. Plates then were run on the sequencing program on the 3730 ABI Genetic Analyzer. Samples were then analyzed in Seq Lab software (GCG accelrys software).

Plasmid Isolation

A bacterial loop was sterilized with ethanol and heat and inserted into wells containing plasmid DNA and then plated out. Colonies were picked into 1ml aliquots of SOC broth containing ampicillin. Colonies were then grown up overnight in 15ml centrifuge tubes at 37°C with agitation at 350 rpms. Cells were then pelleted by centrifugation at 9000x g for 10 minutes at 4°C. The colonies were then transferred into 100ml of SOC broth containing ampicillin in 500ml Erlenmeyer flasks. These were incubated overnight at 37°C with agitation at 350 rpms. Purification was done using the Wizard® Plus midipreps DNA Purification System. Cells were resuspended in 3ml of cell resuspension solution. Following that 3ml of cell lysis solution was added and the solution was gently inverted several times to mix. Then 3ml of neutralization solution was added and also inverted several times to mix. This mixture was centrifuges at 9000x g for 20 minutes at 4°C. The supernatant was carefully poured off and DNA was resuspended in 10ml resin and swirled to mix. A midicolumn was then attached to a vacuum manifold and the liquid was vacuumed out. Following this 15ml of washing solution was added to each column and the liquid was vacuumed out. This step was repeated once. Midicolumn was then placed into a 1.5ml microcentrifuge tube and centrifuged at 10000 x g for 2 minutes. Midicolumn was then placed into a new 1.5ml microcentrifuge tube and 300ul of water preheated to 65°C and allowed to sit for 2

minutes. This was then centrifuged at 10000x g for 20 seconds to elute the DNA. DNA was then quantified using an automated spectrophotometer (Spectramax 190) in conjunction with Softmax Pro 4.3.1 LS software package (Molecular Devices; Sunnyvale, Calif). DNA was then diluted to 20ng/ul and used for PCR reactions. PCR master mix was prepared as previously stated in the PCR amplification paragraph. PCR product was then sequenced on the using the sequencing module on the ABI Prism 3730 Genetic Analyzer using Performance Optimized Polymer (POP) 7 polymer (Applied Biosystems, Inc.).

Sequencing

Sequencing of all PCR products for primers four and seven was conducted as follows. All PCR products were subjected to an exonuclease reaction with 10ul of the following reaction mixture added to each well: 6ul of buffer (1X), 0.075 exonuclease enzyme (0.025U/ul), and 3.925 dH_2O. The plates were covered with foil tape and inverted several times to mix the solution and then were left to stand at room temperature for 15 minutes. The plates were then centrifuged @ 1700 rpm for 30 minutes, then inverted and centrifuged again @ 700 rpm for 1 minute and then left to air dry for 15 minutes. This was followed by cycle sequencing using the T7 primer. The mix consisted of 10ul reactions each with 1ul of T7 primer (1.6pmol/ul), 1.75ul 5X buffer (Applied Biosystems), 0.5ul Big Dye Terminator v3.1 (Applied Biosystems), and 6.75 ul dH_2O. Thermocycling profile was 26 cycles each of, 96°C for 10 seconds, 50°C at 5 seconds, and 60°C for 4 minutes. Ethanol precipitation followed, by adding 32uls of 95% ethanol and 8ul dH2O per well and mixing the plates by inverting several times. Plates were placed in a dark place so products could precipitate for more than 1 hour. Plates were

then centrifuged @ 1810 rcf for 45 minutes to pellet the product, and then inverted and centrifuged at 50 rcf for 30 seconds to remove supernatant. The plates were then air dried for at least 15 minutes and the product was resuspended in 20ul of dH2O. Plates then were run on the sequencing program on the 3730 ABI Genetic Analyzer. Samples were then analyzed in Seq Lab software (GCG accelrys software).

III. RESULTS

Primer 1

Primer 1 was designed specifically for one locus on a gene that codes for a heat shock protein in *Arabidopsis*. Although this is a large gene family, a BLAST search of the EST sequence against the *Arabidopsis* genome on the NCBI website revealed only two hits. Both of these hits were for single-copy genes at the same physical location, only a few thousand base pairs apart, but on opposite strands of DNA on chromosome five in *Arabidopsis*. Both the forward and reverse primer sequences were also blasted against the *Arabidopsis* genome to determine whether the 3' end of both primers had hit both loci, and therefore had the potential to amplify both sequences. This did not appear to be the case. Only one of the sequences consistently matched up with the entire oligonucleotide sequences. The e score for the original BLASTN search with the entire *T. cacao* EST sequence against the non-redundant database gave an e score of 8e-173 for the alignment with *Gossypium* and 2e-40 for the alignment with *Arabidopsis* (figure 1). The location of the primers is indicated in figure 1. The results indicate that more than one locus was amplified for some of the species tested (figure 2), but this was usually at the lower temperatures, which can also cause non-specific binding. Non-specific binding is when the primers bind to several different places along the sequence. In addition it might not be possible to tell if more than one locus was being amplified if all fragments were the same size. The only way to resolve this question is to clone and sequence the PCR products from some of the species.

Only one position in each of the forward and reverse primers was degenerate in primer pair number one. The clamp score was 1 for both forward and reverse primers.

The suggested annealing temperature for this primer pair from Sigma Genosys was about 63°, and the fragment was designed to be 203 bp in length. This was consistent with the optimum temperature and fragment size within a few base pairs for *T. cacao*, *Equisetum giganteum*, and a *moss species*. This was unexpected as these species are not closely related. However, the temperature gradient data for most of the species showed a trend of generally much lower optimal annealing temperatures (figure 3). This primer did show amplification for all of the species that were tested (table 3).

SPIDEY data obtained from *Arabidopsis* and *T. cacao* alignments (figure 4), determines that the forward primer should actually fall within the first intron and a second intron should fall between the primers making the size of the fragment approximately 349bp. Genomic sequence information was unavailable in some cases for *T. cacao* so genomic sequences from *Arabidopsis* were used for the alignment in the SPIDEY program. Also because *Arabdopsis* has its entire genome sequenced, the presence of all of the introns is known. Since *Arabidopsis* was used for the genomic sequence and *T. cacao* for the mRNA sequence in SPIDEY, mistakes might occur, because two different species are being aligned. There was some variability in fragment size, although only *Mangifera indica* and *Iris hexagona* exhibited a fragment close to 349 bp, and both of these species amplified more than one product. The three closest relatives of *Arabidopsis* included in this analysis are *T.cacao, Litchi chenensis* (litchi), and *M. indica* (mango), and these all amplified fragments of approximately 203bp, although mango did have a second fragment at 350bp for some temperatures. Some species had amplified products of different sizes depending on which annealing temperature was used.

Primer 1 showed polymorphism in *T. cacao* using CAE-SSCP analysis (figure 5 & 6). Figures five and six only show the electropherogram for the blue strand in this case because it is not necessary to include both the blue and green strands. If there are differences in mobility of either of the strands this indicates polymorphism. In the case that fluorescence is low for one of the strands, the other one can be used in analysis. There was better resolution of the alleles at 22°. At 28°C, only three different alleles were apparent for the same four individuals. At 22°C individual AMA-4 is homozygous, but at 28°C it is heterozygous. This is why it is important to run each analysis at two different temperatures. Since *T. cacao* is a diploid the presence of four alleles in one individual as is the case in IMC67 at 22°C is likely due to amplification of more than one locus, which the BLAST search against the *Arabidopsis* genome indicated was possible. Or each peak is not an allele, but one of the peaks is a "plus a" product from the PCR reaction, and there are really only two alleles for TC533. Another less likely option is that it is possible for DNA strands to have two stable conformations when run under CAE-SSCP conditions. The only way to determine with certainty that the expected fragment is being amplified is by cloning and sequencing the PCR product. The remaining species were not tested using CAE-SSCP because of the variability of fragment sizes across species and because some species amplified more than one product, indicating that the amplification product might not be the expected fragment. Since all species that were tested did amplify a product and showed polymorphism in *T. cacao*, this set of primers could potentially be optimized for use as a molecular genetic marker for any plant species.

Primer 2

Primer 2 was designed specifically to amplify one locus for a gene that codes for putative kinase on chromosome four in *Arabidopsis*. The original BLASTN report against the non-redundant (nr) database (figure 7) indicated a high homology (3e-82) with the mRNA sequence of *Arabidopsis* for protein kinase. When the *T. cacao* EST sequence was blasted against the Arabidopsis genome, no hits were returned. This could be the result of the presence of several introns in that sequence (figure 10). Thus it was difficult to ascertain if this is a single-copy gene. According to SPIDEY there are three introns within this fragment. The original fragment was designed to be 233 bp. However, most of the species amplified fragments much larger than this (table 4, figure 8). The SPIDEY data indicates that the expected size would be about 492 bp. This is close to the size of the product for *T. cacao* and for litchi, which was 510bp, and was the exact size of the fragment for mango and *Calocarpum sapota* (mamay).

The suggested annealing temperature from Sigma Genosys was 60°. This coincided correctly with *T. cacao*, but the remaining species all amplified at lower temperatures (figure 9). The clamp score was 2 for both the forward and reverse oligonucleotides and there was one degenerate position for the forward sequence and 4 positions for the reverse sequence. Amplification occurred in 71% of all species that were tested with this primer (table 4). A large amount of variability for this primer occurred for the size of the fragment and the optimal amplification temperatures. The sizes of the fragments were large for SSCP analysis. However, we were able to determine polymorphism in spite of this (figure 11). Resolution was better at 28°C than 22°C. If the alleles are to be determined by the green strands then it appears that more

than one locus amplified. However, it is more likely that the small peaks beyond the true alleles are artifact or pull-up peaks because these correspond directly with the blue peaks. These artifacts or pull-up peaks can result from an overlap in the fluorescence at a particular wavelength. Since there were clear distinction between homozygotes and heterozygotes we were able to map this locus in an F2 population of *T. cacao*. It is likely that this primer set could be used as a marker for most plant species with optimization.

Primer 3

Primer 3 was designed to specifically amplify one locus for a gene that codes for a proteasome subunit on chromosome three in *Arabidopsis.* The e score for the BLASTN against the nr database was 7e-83 for the alignment with *Arabidopsis* (figure 12). This appears to be a single-copy gene, but when the sequence was blasted against the *Arabidopsis* genome it had two significant matches located on chromosomes three and four. This suggests the possibility of amplification of more than one locus. However, this was not indicated in the results. To be completely sure the fragment being amplified is the expected sequence it would need to be cloned and sequenced.

There was one degenerate position in the forward sequence and none in the reverse sequence. The clamp score for the forward primer was two and one for the reverse primer. Of all the species tested against this primer pair, approximately 61% demonstrated amplification (table 5). The optimal annealing temperature suggested for *T. cacao* was 55°C and this corresponded to the observed annealing temperatures for most species tested (figure 14). At lower temperatures there were usually several products that amplified probably due to non-specific binding (figure 13). The size of the fragment was designed to be 221 bp, and according to SPIDEY this fragment did not

have any introns (figure 15). The fragments all exhibited the expected size fragment at higher temperatures except for one moss species. This primer did not exhibit polymorphism in *T. cacao* (figure 16). As seen in figure 17 the resolution at both temperatures was about the same and all individuals showed the same pattern of alleles.

Primer 4

Primer 4 was designed to amplify a single locus of a gene that codes for polyubiquitin in *Arabidopsis* on chromosome one. The original BLASTN report (figure 17) e score for alignment with Arabidopsis was 2e-79, indicating that it was a good candidate for primer design. However, when the EST sequence was blasted against the *Arabidopsis* genome, three significant matches were returned. This suggests that this gene exists at more than one locus. The forward and reverse primer sequences were also blasted against the Arabidopsis genome. Both sequences returned three potential places where the primers could bind. Thus the possibility of hitting more than one locus for this primer pair was likely. Because of the great success of this primer an endemic species from New Caledonia (*Myodocarpus lanceolata*) was cloned and sequenced to determine the relationship between the SSCP pattern and the number of alleles. I selected *M. lanceolata* because very little genomic information is available for this plant and it displayed polymorphism with this primer. Results from the sequencing data did conclude that it was highly likely that at least two loci of the same fragment size were being amplified. All individuals from all species were also sequenced and evidence from this data further strengthened the opinion that this primer pair was amplifying more than one locus. On the other hand, this might not have been the case for all of the species.

The clamp scores for both the forward and reverse sequences was very good at three and two respectively. The fragment for this primer pair was designed to be 215 bp, and there was only one degenerate position for both oligonucleotides. According to SPIDEY there were no introns in this fragment and results agreed with this consensus (figure 20). This was one of two primers that amplified fragments of the expected size in virtually every plant species that was available to test (table 6 & figure 18). The suggested annealing temperature was 58°C, and this was consistent with the optimal annealing temperature observed for *T. cacao*. The remaining species seemed to favor slightly lower temperatures (figure 19). Because of the notable success of this primer all species were tested for polymorphism by CAE-SSCP (table 6). The trend for this primer was that polymorphism was more likely observed at 22°C than at 28°C. This was the case for *Myodocarpus lanceolata* and the presence of four green peaks in one of the individuals suggests that these markers are hitting more than one locus (figure 21 & 22). Though this primer pair seems to be amplifying more than one locus it can still be utilized to genotype individuals, or to distinguish between species. Since it seems to amplify fragments of the same size for any plant from liverworts to angiosperms it must be an extremely conserved sequence. Thus this primer pair can likely be used for all plant species.

Primer 5

Primer 5 was designed to a single-copy gene that codes for methionine synthase protein on chromosome five in *Arabidopsis*. A BLAST search was done with the EST against the *Arabidopsis* genome and only one significant hit was returned. The forward and reverse primer sequences were also blasted and the results indicated implied that

there was only one locus for the primers to bind for that particular sequence. The original BLASTN score was 5e-103 for the alignment with *Arabidopsis*, illustrating high homology (figure 23).

The clamp scores were one for the forward sequence and three for reverse, and this primer pair did not have any degeneracy. The fragment size was intended to be 269 bp, however SPIDEY data showed one intron of 99 bp long (figure 26). This would make the expected fragment size to be about 368 bp. Several species including *T. cacao* had fragments that were very close to this. Some species such as *Eleocharis cellulosa, Zamia intergrifolia* and a moss species amplified a fragment of the expected length without an intron (figure 24). This indicates that the trend for plants further down the evolutionary scale did not have an intron inserted into that particular sequence. There were also some plants that amplified more than one product. This did not seem to be associated with a lower temperature for this primer pair. Since many plants amplified more than one product with primer 5 this implies that the primers are hitting more than one locus in those species. Of all the species that were tested, 80% demonstrated amplification (table 7). However, there was an extensive amount of fragment size variability. The recommended annealing temperature from Sigma Genosys was 65°C. This did not agree with the optimal annealing temperature for *T. cacao*, which was 62°C (figure 25), but there was still very high fluorescence at that temperature. Most of the other species had optimal annealing temperatures at or around 55°C.

Polymorphism was evident in *T. cacao* (figure 27) at both temperatures, but the alleles were more easily distinguished at 28°C. Only one or two alleles were apparent for

any individual. This implies that in *T. cacao* only one locus is being targeted. Thus this primer pair does have potential for use as a molecular marker for some species.

Primer 6

Primer six was designed to a gene that codes for ubiquitin-conjugating enzyme E2 on chromosome five in *Arabidopsis*. The BLAST search with the EST sequence and each of the primer sequences against the *Arabidopsis* genome revealed only one locus that the primers were likely to target. The original BLASTN search of the *T. cacao* EST sequence against the nr database displayed considerable homology with *Arabidopsis* (figure 28).

Only one position in the reverse primer was degenerate at the 5' end and the clamp score was three for the forward sequence and one for the reverse sequence. The optimal temperature as recommended by Sigma Genosys was 65°C. None of the species amplified at this temperature. *Theobroma cacao* amplified best at 47°C, and the best temperature was comparable for the other species as well (figure 30). It is also possible that this primer pair had an optimal annealing temperature that was higher than the temperature gradient tested for. Fragment sizes varied considerably (table 8), and many species had numerous products. Amplification was over 75% for all the species tested for this primer, but most of that amplification was non-specific (figure 29). The primers were made to amplify a fragment of 220bp, but according to SPIDEY one intron was present within the sequence (figure 31). This would make the expected fragment size to be 383bp. Only one species *Saccharum barberi* (sugarcane) was close to that size. It is not likely that the primers amplified the expected region for any of the species tested. All individuals of *T. cacao* that were tested for polymorphism showed that same genotype

(figure 32), so this primer set could not be used as a marker. However, this homologous region does have potential for primer development, and primers could potentially be redesigned to this region with success.

Primer 7

Primer seven was designed to target a single locus that of a gene on chromosome one that codes for alpha-tubulin in *Arabidopsis*. In order to determine if this primer set could target other loci, the EST and primer sequences were blasted against the *Arabidopsis* genome. Results from this indicated that there was only one locus on chromosome one that the primers were likely to bind to. The original BLASTN reports revealed this sequence to have a very high scoring match with *Arabidopsis* (figure 33). As was the case in primer four, *Myodocarpus lanceolata* was cloned and sequenced for this primer set. The remaining species were all sequenced as well. The sequencing results validated the results from the BLAST against the *Arabidopsis* genome.

Clamp score for the forward oligonucleotides was three, and two for the reverse sequence. Only one position was degenerate in the reverse primer and the suggested annealing temperature from Sigma Genosys was 62°C. Temperature gradient data showed that *T. cacao* amplified best at 51°C or 65°C, and most of the remaining species preferred temperature around 55°C (figure 35). The differences in optimum temperatures for *T. cacao* could be due to preparation errors in the PCR or differences in DNA concentrations. The fragment was designed to be 257 bp, and SPIDEY predicted the absence of introns (figure 36). All species amplified fragments of this expected size (table 9).

All species were then tested for polymorphism using CAE-SSCP. Polymorphism was clear for more than half of the species tested (table 9), and for this primer 28°C seemed to give better resolution of the alleles in most cases. In figure 37 the same individuals of *Selaginella species* are shown at both temperatures. In this particular example the resolution is better at 22°C for these specific individuals, but polymorphism was visible at both temperatures.

Primer 8

Primer eight was also designed to amplify a single-copy gene that codes for alpha-tubulin on chromosome four in *Arabidopsis*. A BLAST search of the EST and each primer sequence against the *Arabidopsis* genome returned only one possible match. The initial BLASTN search using a *T. cacao* EST sequence against the nr database returned very high matches with *Arabidopsis* (figure 38). Degeneracy was only present at two positions in the forward primers and one in the reverse. Clamp scores were very good with two for the forward and three for the reverse. Optimal annealing temperature as recommended by Sigma Genosys was about 61°C. This was in agreement with the best annealing temperature for *T. cacao* (figure 40). The trend for the other species was better amplification at lower temperatures, as seems to be the case for most of the primers. Over 64% of all the species tested amplified for this primer set (table 10). The expected fragment size was 266 bp, and SPIDEY predicted the absence of introns (figure 41). Only one species (*Eleocharis cellulosa*) actually amplified the expected fragment size (figure 39). The remaining species amplified several different fragment sizes, and many had more than one product. Different fragment sizes signify that either the sequence contains introns of different sizes in different species, or that the primes are

amplifying another locus than expected. The latter explanation is more likely. However, SSCP data does show polymorphism (figure 42) in *T. cacao*. Polymorphism is more evident at 22°C. Thus this primer can be used as a molecular marker in *T.cacao* and likely in other species as well. However, if it is necessary to know exactly what is being amplified cloning and sequencing would be necessary to determine this.

IV. DISCUSSION

The concept underlying universal primers is relatively simple and logical. Since all organisms contain certain genes necessary to sustain life, it is possible that these homologous genes may exist as similar regions across all plant species. These conserved regions of the genome could be a considerable resource from which to design universal primers for all plant species. My research demonstrated that molecular genetic markers can be made from these conserved ESTs that are likely to amplify in any plant species. Currently, developing molecular markers is an expensive process because they are usually developed separately for each species. A set of universal primers for all plant species could save considerable amounts of time and money. Finding conserved sequence information for genetically distant plant species has been made quick and easy by accessibility to EST databases couples with homology alignment programs such as BLAST to determine how many species these regions match up with. Programs like SPIDEY also add to the likelihood of success by identifying potential introns within the sequences of interest. In five out of eight primers in this study, the SPIDEY prediction was fairly accurate. Success would probably be even higher if the genomic and mRNA sequences being compared were from plant species that were more closely related to each other than *Arabidopsis* is to *T. cacao*. Once the desired sequence information is obtained degenerate primers can be designed from any of several programs freely available online. The program used in this study (JCSG) worked very well and all of the primers were successful for the species *(T.cacao)* from which they were designed.

For this project eight primer pairs were developed and two of them (primer 4 & 7) amplified the expected fragment size in 100% of all the species tested. The plant species

that were tested covered a range of plants from the entire plant kingdom with the exception of only a few from which samples were unable to be obtained (table 1). Approximately half of the species tested with both primers four and seven displayed polymorphism (table 6 & 9). Increasing the sample size would likely also increase the rate of polymorphism. Also, some of the species tested are known to commonly reproduce clonally, in which case polymorphism would not be expected, so it is likely that the success rate is even higher that thought from these results. Since SNPs are common throughout the plant genome, you would still expect to see them, even in these highly conserved sequences. They often occur at the 3rd base pair of the codon, since changing a nucleotide at this location does not change the amino acid being coded for. This might also explain why the optimum annealing temperature was different for the same primers in different plant species. SNPs are commonly found in both introns and exons at about every 140bp (Kuhn et al., 2005, Salmaso et al., 2004 and Schneider et al., 2001). Thus, with a larger sample size you would expect to see some at least some polymorphism in fragments between 150-300 bp (Kuhn et al., 2005, Salmaso et al., 2004 and Schneider et al., 2001).

An endemic species from New Caledonia, *Myodocarpus lanceolata* was cloned and sequenced for both primer four and seven to verify the association between the SSCP data and allele number. From this data it was determined that primer four was actually amplifying more than one locus, even though this was not evident from the SSCP pattern. This would make it difficult to distinguish between homozygous and heterozygous states because the presence of two alleles could signify a heterozygous individual or that the individual was homozygous at two different loci. Nevertheless, this primer pair can still

be used as a molecular marker by determining pattern differences across individuals. This is known as the diversity index and has often been done in the past for plants with different ploidy levels (Milbourne et al., 1997). Using a diversity index is essentially just distinguishing between individuals based on the amount of polymorphism that is evident, this is often used for plants with high ploidy levels (Milbourne et al., 1997). Primer four could also be used to distinguish between closely related species. Evidence for this was seen with two different *Zamia* species that were tested in this study. Both *Zamia* species are closely related and yet they displayed entirely different patterns with the same primer when analyzed with CAE-SSCP.

Sequence data from primer 7 confirmed that only one locus was targeted. This together with the high occurrence of polymorphism makes primer seven an ideal universal molecular marker for plants. Primer seven exhibited unusual trends in the temperature data (figure 36). Some individuals from the same species amplified better at higher temperatures and vice versa. This was also seen in some cases for the other primers. This might be because one of the individuals might have had a very high DNA concentration compared to the others and would have very high amplification at any temperature. The question of the concentration of the DNA is very important. The Spectrophotometer was not always accurate due to the presence of phenols and proteins in some of the DNA extracts. This could be determined by observing a graphical output of the reading. If a large peak was observed at a wavelength of 230 nanometers (nm), and a very low or no peak at 260nm then it was assumed that a large amount of phenols was present because 230nm is the wavelength a which phenols are absorbed, and 260nm is the wavelength that DNA is absorbed. In that case the Spectrophotometer readings

were deemed inaccurate. It was observed in some cases when a particular species or individual was difficult to amplify if the concentration was increased from the stock DNA the PCR reaction would usually work much better.

The other six primers amplified over 50% of the plant species that were tested for each primer. The fragment sizes were not as expected in many cases for the other six primers, but this could be the result of introns not foreseen with the SPIDEY program or the primers could be amplifying a different locus than expected. Even if that is the case, it does not prevent determining polymorphism among individuals, and the markers could still be used in population studies or to distinguish between closely related species or hybrid species. If it was necessary to identify the gene that was be amplified then a cloning and sequencing reaction could be done to determine which locus was actually targeted. In some cases the CAE-SSCP results displayed difficult allele patterns, and homozygous and heterozygous states were not clear. In that case pattern differences were used to determine genotype. It is likely that these difficult patterns were the result of the primers targeting more than one locus. In plants that are known polyploids, such as *Saccharum barberi*, and *Amborella*, difficult patterns would be expected. However, in *S. barberi* polyploidy was not evident from the SSCP data, but in *Amborella* it was.

Several of the species displayed several amplification products many of the primers at lower temperatures. This is probably the result of non-specific binding that commonly occurs at lower annealing temperatures. In these cases it is necessary to raise the annealing temperature to ensure that only one product is being generated. If several products are being produced then it would be very difficult to analyze under CAE-SSCP conditions. It would also be inaccurate because the products being amplified from one

individual might not be the same as the next and thus there are patterns from two different loci.

There did not appear to be any pattern between plant species that amplified better for one primer compared to another. Sometimes for a particular primer the only species that would not amplify was *Halesia tetraptera* (a dicot) and moss, which are not close relatives on the evolutionary scale. In other cases all species would amplify for a particular primer but the optimal annealing temperature would be the same for plant species on the opposite end of the evolutionary scale such as *Psilotum nudum* and *Jacquemonita reclinata*. Fragment size also did not seem to follow a pattern in most cases.

Future Prospects

Currently, developing molecular markers is expensive and takes a lot of time. There are hundreds of thousands of EST sequences that potentially contain large numbers of conserved sequences. For minimal cost and little time, these sequences can possibly be converted into universal molecular markers for plants. This could save several thousand dollars that it typically takes to develop molecular markers for each plant species.

These universal markers have many potential applications. For instance, in taxonomy and systematics studies, sometimes determining whether two plants are the same species or very closely related distinct species can be very difficult based solely on morphological data. This can be especially difficult when environmental factors affect the morphology. These markers could be used to distinguish among different species, among hybrid species, or even to describe a new species. Pattern differences were very

distinct across the different species that I examined. A relatively new idea has been recently examined using certain sequences as species identifiers among several animal species. This idea has recently been proposed to be put into use for plants as well. The use of a universal tag for use in animal biodiversity studies involves the use of a small region from the cytochrome *c* oxidase (CO1) gene as a "DNA barcode", or a species identification tool (Kress et al., 2005). This type of tag is further being discussed for use in the plant kingdom by using the nuclear internal transcribed spacer region, and the plastid *trnH-psbA* intergenic spacer for application as a "barcode" (Kress et al., 2005). These regions contain a high amount of variability, but yet are easily amplified across most plants (Kress et al., 2005). The same outcome could likely be accomplished by using conserved genes, as was done in this study.

Another potential important use for these molecular markers is identifying candidate genes or associations between genotype and phenotype to be used for MAS. A good deal of our food and medicine comes from plants. An example of this is taxol, which is very effective in the treatment of some types of cancer. This chemical comes from the bark of the Pacific Yew tree (Erdemoglu and Sener, 2000). Unfortunately, this natural compound is extremely difficult to synthesize in a lab and must be extracted directly from the plant. Since most plants grow in specific habitat types this could limit the availability of some important medicines. Universal markers could determine which genotypes were associated with high secondary chemical production, and MAS could aid in selecting for the desired plants.

All primers were designed from *T. cacao* sequences and all of them were able to be utilized for CAE-SSCP analysis in this species. Since all eight also amplified

fragments in over 50% of species tested, with some redesigning all of them might be able to work in any plant species. Increasing the amount of degeneracy, or ensuring a very strong clamp score would likely add to the success of this method. For this study a very small percentage of the possible conserved sequences were used to develop these primers. Since the success rate of this method of finding universal molecular markers for plants was high there is reason to believe that there are several more sequences with great potential to be converted into PCR-based primers for use as molecular markers.

TABLE 1
Species list

Division	Clade	Order	Family	Genus	Species	Source of Plant Material	Ploidy level
Hepaticophyta						T.M. Sperry herbarium Pittsburg	
		Jungermanniales	Porrellaceae	*Porella*	*pinnata*	State University	diploid
Anthocerotophyta		Unable to obtain					
Bryophyta		Unknown	Unknown	Moss	*spp.*	Christine Casado-Acorn DERM	unknown
		Bryales	Thuidiaceae	*Thuidium*	*spp.*	Christine Casado-Acorn DERM	diploid
Psilophyta		Psilotes	Psilotaceae	*Psilotum*	*nudum*	Fairchild Tropical Garden	
Pterophyta		Pteridales	Pteridaceae	*Pteris*	*bahamense*	Singletary propery of FIU	diploid
Spenophyta						Chad Husby, Florida International	
		Equisetales	Equisetaceae	*Equisetum*	*giganteum*	University	diploid
Lycophyta		Selaginellales	Selaginallaceae	*Selaginella*	*armata eatonii*		diploid
Pinophyta		Cycadales	Zamiaceae	*Zamia*	*portoricensis*	USDA Miami, FL	diploid
		Cycadales	Zamiaceae	*Zamia*	*integrifolia*	USDA Miami, FL	diploid
						Fairchild Tropical Garden & The	
		Gnetales	Gnetaceae	*Gnetum*	*gnemon*	Kampong	diploid
		Ginkoales	Ginkoaceae	*Ginko*	*biloba*	Nebraska Statewide Arboretum	diploid
						Edinburgh Centre for Tropical	
		Coniferales	Araucariaceae	*Araucaris*	*columnaris*	Forests	diploid
Magnoliphyta	Basal families	Amborellales	Amborellaceae	*Amborella*	*trichopoda*	New Caledonia permit # 6034-3782	polyploid
		Nymphales	Nymphaceae	*Nymphea*	*odorata*	Jenny Richards permit #	diploid
	Magnoliid complex	Laurales	Lauraceae	*Persea*	*americana*	USDA Miami, FL	tetr/diploid
	Monocots	Aspargales	Iridaceae	*Iris*	*hexagona*	USDA Miami, FL	diploid
		Aspargales	Amaryllidaceae	*Phaedranassa*	*tungaraguae*	USDA Miami, FL	
	Commelinds	Cyperales	Cyperaceae	*Eleocharis*	*cellulosa*	Singletary property of FIU	polyploid
						(25°23'47'N,80°28'06W)	
		Poales	Poaceae	*Saccharum*	*barberi*	USDA Miami, FL	polyploid
		Arecales	Arecaceae	*Cocos*	*nucifera*	USDA Miami, FL	tetraploid
	Eudicots/Basal						
	Tricolpates	Unable to obtain					
	Eudicots/Caryophyll					USDA Miami, FL	
	id	Caryophyllales	Chenopodiaceae	*Beta*	*Vulgaris*		diploid
	Rosids	unassigned	Strasburgeriaceae	*Strasburgeria*	*robusta*	New Caledonia permit, # 6034-3782	
	Eurosids I	Unable to obtain					diploid
	Eurosids II	Malvales	Malvaceae	*Theobroma*	*cacao*	USDA Miami, FL	diploid
		Sapindales	Anacardiaceae	*Mangifera*	*indica*	USDA Miami, FL	diploid
		Sapindales	Sapindaceae	*Litchi*	*chinensis*	USDA Miami, FL	diploid
	Asterids	Ericales	Styracaceae	*Halesia*	*tetraptera*	USDA Miami, FL	diploid
		Ericales	Sapotaceae	*Calocarpum*	*sapota*	USDA Miami, FL	diploid
	Euasterids I	unassigned	Oncothecaceae	*Oncotheca*	*balansae*	New Caledonia, permit # 6034-3782	
		Solanales	Convolvulaceae	*Jacquemontia*	*reclinata*	Fairchild Tropical Garden	diploid
		Gentianales	Apocyanaceae	*Plumeria*	*spp.*	USDA Miami, FL	diploid
	Euasterids II	Apiales	Myodocarpaceae	*Myodocarpus*	*lanceolata*	New Caledonia, permit # 6034-3782	diploid

Table 2
Primer information

		Closely-related protein sequence	Expected Size of Fragment (bp)	Expected Size with Introns in *Arabidopsis* (bp)	Annealing Tm as given by Sigma Genosys (˚C)	e-score	Forward primer (5'-3')	Reverse primer (5'-3')
1	gi - 38499237 CF973477	heat shock protein	203	349	63	1e-133	AGTGGNGGCATGTCAGCAT	NGAGCTGAAAAGAGAGCTGAT
2	gi – 26052545 CA795469	putative kinase	233	492	60	2e-82	GTGTNTGGCGAGCTATAAATAAG	TTTNNCNATCNTTCATAAGTTGGTAAAG
3	gi – 26052517 CA795441	proteasome subunit	221	221	55	5e-83	CNAAGCTTCTTCAAACCTC	GACCACCAAACTGTGTGTA
4	gi – 26955451 CA798365	polyubiquitin	215	215	58	7e-79	CANCAGAGGCTTATCTTTGC	ACCNTCCTTGTCCTGAATC
5	gi – 38498673 CA972913	methionine synthase protein	269	368	65	2e-98	TACCAGATTGCTTTGGCCATC	AATGGTGATGACATCAGCATCC
6	gi – 26052348 CA795272	ubiquitin conjugating enzyme E2	220	383	65	1e-102	GTGGCTGAAGACATGTTTCATTGG	NATGGTGAGGGCAGGGCT
7	gi – 26055742 CA798656	alpha-tubulin	257	257	62	1e-113	ACCCAAGAATCCACTTCATGC	GNCACCAGTCCACAAACTG
8	gi – 26052566 CA795490	alpha-tubulin	266	266	61	1e-123	ANAANTCCAAGTTGGGTTTCACTG	CATCNACTTCAAGGCACC

Table 3

Primer 1 amplification table

Genus	Specific epithet	Ampification	Optimal Tm	allele size
Porella	*pinnata*	not tested		
Moss	*spp.*	yes	62	205
Thuidium	*spp.*	not tested		
Psilotum	*nudum*	not tested		
Pteris	*bahamense*	yes	51	203
Equisetum	*giganteum*	yes	63	203
Selaginella	*armata*	not tested		
Zamia	*portoricensis*	yes	48	203
Zamia	*integrifolia*	yes	47	203,378
Gnetum	*gnemon*	not tested		
Ginko	*biloba*	yes	57	203
Araucaria	*columnaris*	not tested		
Amborella	*trichopoda*	not tested		
Nymphea	*odorata*	not tested		
Persea	*americana*	yes	47	203
Iris	*hexagona*	yes	47	206,340
Phaedranassa	*tungaraguae*	yes	47	203, 209
Eleocharis	*cellulosa*	yes	47	206
Saccharum	*barberi*	yes	47	242
Cocos	*nucifera*	yes	47	197
Beta	*vulgaris*	yes	57	203
Strasburgeria	*robusta*	not tested		
Theobroma	*cacao*	yes	62	203
Mangifera	*indica*	yes	55	200,350
Litchi	*chinensis*	yes	47	206
Halesia	*tetraptera*	yes	47	203
Calocarpum	*sapota*	yes	47	206
Oncotheca	*balansae*	not tested		
Jacquemontia	*reclinata*	yes	57	203
Plumeria	*spp.*	yes	49	203
Myodocarpus	*lanceolata*	not tested		

Table 4

Primer 2 amplification table

Genus	Specific epithet	Amplification	Optimal Tm	Allele size
Porella	*pinnata*	not tested		
Moss	*spp.*	yes	45	213
Thuidium	*spp.*	not tested		
Psilotum	*nudum*	no		
Pteris	*bahamense*	yes	45	262
Equisetum	*giganteum*	yes	47	271
Selaginella	*armata*	not tested		
Zamia	*portoricensis*	not tested		
Zamia	*integrifolia*	yes	47	292
Gnetum	*gnemon*	no		
Ginko	*biloba*	yes	45	211
Araucaris	*columnaris*	not tested		
Amborella	*trichopoda*	not tested		
Nymphea	*odorata*	not tested		
Persea	*americana*	yes	46	625
Iris	*spp.*	yes	45	480,360
Phaedranassa	*tungaraguae*	yes	47	344,363
Eleocharis	*cellulosa*	no		
Saccharum	*barberi*	yes	45	487
Cocos	*nucifera*	yes	48	486
Beta	*vulgaris*	no		
Strasburgeria	*robusta*	not tested		
Theobroma	*cacao*	yes	61	510
Mangifera	*indica*	yes	45	492
Litchi	*chinensis*	yes	55	511
Halesia	*tetraptera*	no		
Calocarpum	*sapota*	yes	45	491
Oncotheca	*balansae*	not tested		
Jacquemontia	*reclinata*	no		
Plumeria	*spp.*	yes	45	232
Myodocarpus	*lanceolata*	not tested		

Table 5

Primer 3 amplification table

Genus	Specific epithet	Ampification	Optimal Tm	allele size
Porella	*pinnata*	not tested		
Moss	*spp.*	yes	51	410
Thuidium	*spp.*	not tested		
Psilotum	*nudum*	no		
Pteris	*bahamense*	yes	46	221
Equisetum	*giganteum*	yes	57	221
Selaginella	*armata*	not tested		
Zamia	*portoricensis*	not tested		
Zamia	*integrifolia*	yes	54	221
Gnetum	*gnemon*	no		
Ginko	*biloba*	yes	61	221
Araucaria	*columnaris*	not tested		
Amborella	*trichopoda*	not tested		
Nymphea	*odorata*	not tested		
Persea	*americana*	yes	54	221
Iris	*hexagona*	no		
Phaedranassa	*tungaraguae*	yes	51	221
Eleocharis	*cellulosa*	no		
Saccharum	*barberi*	yes	45	221
Cocos	*nucifera*	no		
Beta	*vulgaris*	no		
Strasburgeria	*robusta*	not tested		
Theobroma	*cacao*	yes	54	221
Mangifera	*indica*	yes	55	221
Litchi	*chinensis*	yes	55	221
Halesia	*tetraptera*	yes	55	221
Calocarpum	*sapota*	no		
Oncotheca	*balansae*	not tested		
Jacquemontia	*reclinata*	yes	47	221
Plumeria	*spp.*	no		
Myodocarpus	*lanceolata*	not tested		

Table 6

Primer 4 amplification and polymorphism table

Genus	Specific epithet	Amp-lification	best Tm	allele size	Polymor-phism 22°C	Polymor-phism 28°C
Porella	*pinnata*	yes	57	215	unsure	unsure
Moss	*spp.*	yes	47	215	yes	yes
Thuidium	*spp.*	yes	55	215	yes	no
Psilotum	*nudum*	yes	57	215	no	no
Pteris	*bahamense*	yes	51	215	yes	yes
Equisetum	*giganteum*	yes	47	215	yes	yes
Selaginella	*armata*	yes	46	215	unsure	unsure
Zamia	*portoricensis*	yes	55	215	no	no
Zamia	*integrifolia*	yes	55	215	no	no
Gnetum	*gnemon*	yes	57	215	no	no
Ginko	*biloba*	yes	48	215	yes	yes
Araucaria	*columnaris*	yes	57	215	yes	no
Amborella	*trichopoda*	yes	47	215	yes	yes
Nymphea	*odorata*	yes	49	215	yes	yes
Persea	*americana*	yes	48	215	yes	yes
Iris	*hexagona*	yes	57	215	yes	yes
Phaedranassa	*tungaraguae*	yes	48	215	no	no
Eleocharis	*cellulosa*	yes	49	215	no	no
Saccharum	*barberi*	yes	55	215	yes	yes
Cocos	*nucifera*	yes	62	215	no	no
Beta	*vulgaris*	yes	57	215	no	no
Strasburgeria	*robusta*	Yes (low)	55	215	unsure	unsure
Theobroma	*cacao*	yes	57	215	no	no
Mangifera	*indica*	yes	55	215	yes	yes
Litchi	*chinensis*	yes	55	215	no	no
Halesia	*tetraptera*	yes	55	215	no	no
Calocarpum	*sapota*	yes	55	215	not tested	not tested
Oncotheca	*balansae*	Yes (low)	46	215	unsure	unsure
Jacquemontia	*reclinata*	yes	57	215	yes	yes
Plumeria	*spp.*	yes	55	215	no	no
Myodocarpus	*lanceolata*	yes	49	215	yes	yes

Table 7

Primer 5 amplification table

Genus	Specific epithet	Ampification	best Tm	allele size
Porella	*pinnata*	not tested		
Moss	*spp.*	yes	54	269
Thuidium	*spp.*	not tested		
Psilotum	*nudum*	no		
Pteris	*bahamense*	no		
Equisetum	*giganteum*	yes	46	402
Selaginella	*armata*	not tested		
Zamia	*portoricensis*	not tested		
Zamia	*integrifolia*	yes	48	269
Gnetum	*gnemon*	no		
Ginko	*biloba*	yes	45	326
Araucaria	*columnaris*	not tested		
Amborella	*trichopoda*	not tested		
Nymphea	*odorata*	not tested		
Persea	*americana*	yes	61	375
Iris	*hexagona*	yes	55	446
Phaedranassa	*tungaraguae*	no		
Eleocharis	*cellulosa*	yes	55	269
Saccharum	*barberi*	yes	55	380
Cocos	*nucifera*	yes	55	460
Beta	*vulgaris*	yes	54	419
Strasburgeria	*robusta*	not tested		
Theobroma	*cacao*	yes	62	375
Mangifera	*indica*	yes	55	355
Litchi	*chinensis*	yes	55	357
Halesia	*tetraptera*	yes	55	362,370,376
Calocarpum	*sapota*	yes	55	379,382
Oncotheca	*balansae*	not tested		
Jacquemontia	*reclinata*	yes	54	355
Plumeria	*spp.*	yes	55	384,369
Myodocarpus	*lanceolata*	not tested		

Table 8

Primer 6 amplification table

Genus	Specific epithet	Amplification	best Tm	allele size
Porella	*pinnata*	not tested		
Moss	*spp.*	yes	47	511
Thuidium	*spp.*	not tested		
Psilotum	*nudum*	no		
Pteris	*bahamense*	no		
Equisetum	*giganteum*	yes	47	253
Selaginella	*armata*	not tested		
Zamia	*portoricensis*	not tested		
Zamia	*integrifolia*	yes	55	435
Gnetum	*gnemon*	yes	47	574
Ginko	*biloba*	yes	45	475
Araucaria	*columnaris*	not tested		
Amborella	*trichopoda*	not tested		
Nymphea	*odorata*	not tested		
Persea	*americana*	yes	46	246
Iris	*hexagona*	yes	47	281
Phaedranassa	*tungaraguae*	yes	47	505
Eleocharis	*cellulosa*	no		
Saccharum	*barberi*	yes	55	343,373
Cocos	*nucifera*	no		
Beta	*vulgaris*	yes	48	370
Strasburgeria	*robusta*	not tested		
Theobroma	*cacao*	yes	47	455
Mangifera	*indica*	yes	47	322
Litchi	*chinensis*	yes	47	322
Halesia	*tetraptera*	no		
Calocarpum	*sapota*	yes	55	337
Oncotheca	*balansae*	not tested		
Jacquemontia	*reclinata*	yes	45	239
Plumeria	*spp.*	yes	48	428
Myodocarpus	*lanceolata*	not tested		

**This primer showed a lot of non-specific binding at lower temperatures.

Table 9

Primer 7 amplification and polymorphism table

Genus	Specific epithet	Amplification	best Tm	allele size	Polymor-phism 22C	Polymor-phism 28C
Porella	*pinnata*	yes	50	257	unsure	unsure
Moss	*spp.*	yes	54	257	no	no
Thuidium	*spp.*	yes	57	257	yes	yes
Psilotum	*nudum*	yes	54	257	yes	yes
Pteris	*bahamense*	yes	47	257	yes	yes
Equisetum	*giganteum*	yes	55	257	no	yes
Selaginella	*spp.*	yes	60	257	yes	yes
Zamia	*portoricensis*	yes	57	257	yes	yes
Zamia	*integrifolia*	yes	48	257	not tested	not tested
Gnetum	*gnemon*	yes	54	257	no	no
Ginko	*biloba*	yes	57	257	yes	yes
Araucaris	*columnaris*	yes	50	257	unsure	unsure
Amborella	*trichopoda*	yes	54	257	yes	yes
Nymphea	*odorata*	yes	50	257	no	no
Persea	*americana*	yes	45	257	yes	yes
Iris	*hexagona*	yes	54	257	no	no
Phaedranassa	*tungaraguae*	yes	57	257	yes	yes
Eleocharis	*cellulosa*	yes	60	257	no	no
Saccharum	*barberi*	yes	55	257	no	no
Cocos	*nucifera*	yes	54	257	no	no
Beta	*vulgaris*	yes	54	257	yes	yes
Strasburgeria	*robusta*	yes	57	257	yes	yes
Theobroma	*cacao*	yes	51	257	yes	yes
Mangifera	*indica*	yes	55	257	yes	yes
Litchi	*chinensis*	yes	55	257	no	no
Halesia	*tetraptera*	yes	54	257	yes	yes
Calocarpum	*sapota*	yes	55	257	not tested	not tested
Oncotheca	*balansae*	yes	46	257	no	no
Jacquemontia	*reclinata*	yes	54	257	no	yes
Plumeria	*spp.*	yes	54	257	yes	yes
Myodocarpus	*lanceolata*	yes	54	257	yes	yes

Table 10

Primer 8 amplification table

Genus	Specific epithet	Ampification	Best annealing Tm	allele size
Porella	*pinnata*	not tested		
Moss	*spp.*	yes	45	366,387
Thuidium	*spp.*	not tested		
Psilotum	*nudum*	not tested		
Pteris	*bahamense*	yes	51	340
Equisetum	*giganteum*	yes	48	275,451
Selaginella	*armata*	not tested		
Zamia	*portoricensis*	not tested		
Zamia	*integrifolia*	yes	45	460
Gnetum	*gnemon*	not tested		
Ginko	*biloba*	no		
Araucaria	*columnaris*	not tested		
Amborella	*trichopoda*	not tested		
Nymphea	*odorata*	not tested		
Persea	*americana*	yes	47	286
Iris	*hexagona*	no		
Phaedranassa	*tungaraguae*	yes	55	301
Eleocharis	*cellulosa*	yes	55	266
Saccharum	*barberi*	yes	55	366,370
Cocos	*nucifera*	yes	55	307,357
Beta	*vulgaris*	not tested		
Strasburgeria	*robusta*	not tested		
Theobroma	*cacao*	yes	61	350
Mangifera	*indica*	yes	55	360,369
Litchi	*chinensis*	no		
Halesia	*tetraptera*	no		
Calocarpum	*sapota*	no		
Oncotheca	*balansae*	not tested		
Jacquemontia	*reclinata*	not tested		
Plumeria	*spp.*	no		
Myodocarpus	*lanceolata*	not tested		

```
BLASTN 2.2.13 [Nov-27-2005]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. SchC$ffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           3,854,774 sequences; 17,018,580,683 total letters
Query=  /bighd/home/sdougl/.seqlab-genome1/input_268.rsf{gi-38499237}
gi|38499237|gb|CF973477.1|CF973477
PSU_blonup4-7_A05_blon4-7_034 Defense-related
Length=693
                                                                    Score     E
Sequences producing significant alignments:                         (Bits)  Value

gb|AF531369.1|  Gossypium barbadense fiber protein Fb9 mRNA, part   615    8e-173
gb|AY368907.1|  Lycopersicon esculentum molecular chaperone Hs...   519    4e-144
gb|M96549.1|TOMHSC80P  Tomato heat shock cognate protein 80 gene,   519    4e-144
gb|AY519499.1|  Nicotiana tabacum heat shock protein 90 (OINtH...   502    9e-139
dbj|AK101323.1|  Oryza sativa (japonica cultivar-group) cDNA c...   254    3e-64
gb|U55859.1|TAU55859  Triticum aestivum heat shock protein 80 mRN   254    3e-64
ref|NM_124642.2|  Arabidopsis thaliana HSP81-1 (HEAT SHOCK PRO...   174    2e-40
dbj|D00710.1|ATHHSP81  Arabidopsis thaliana HSP81-1 gene for 8...   174    2e-40
dbj|AB025606.1|  Arabidopsis thaliana genomic DNA, chromosome 5,    174    2e-40
gb|AF521007.1|  Hevea brasiliensis heat shock protein mRNA, compl   167    6e-38

>ref|NM_124983.3| Arabidopsis thaliana HSP81-3; ATP binding / unfolded protein
binding AT5G56010 (HSP81-3) mRNA, complete cds
Length=2423
 Score =  482 bits (243),  Expect = 8e-133
 Identities = 534/630 (84%), Gaps = 6/630 (0%)
 Strand=Plus/Minus

Query  60    TTAGTCGACTTCCTCCATCTTGCTGCCCTCGGCCTCTGCATCAGCATCCTCCAGTGGGGG  119
             |||||| ||||||||||||||||| || || || || ||||||   || || ||||| ||
Sbjct  2205  TTAGTCAACTTCCTCCATCTTGCTACCTTCAGCATCGGCATCA---TCTTCAAGTGGAGG  2149

Query  120   CATGTCAGCATCGGCTTCACCTGCATCCTC---ATCAATGCTCAATCCAAGTTTCAACAT  176
             |||||||||||||||||| |  |||| ||   |||||||||||||||| ||||||||
Sbjct  2148  CATGTCAGCATCGGCTTCAACAACATCATCGTCATCAATGCTCAATCCAAGCTTCAACAT  2089

Query  177   CCTGTGGATTCTGTTGCCAAAGGTGTTGGGGTCATCAAGGCTGAAGCCAGAAGTTAGGAG  236
             |||||| || ||| | || || |||||||| ||||| |||||||| |||||||| || ||
Sbjct  2088  CCTGTGAATCCTGCTCCCGAAAGTGTTGGGCTCATCGAGGCTGAAACCAGAAGTGAGAAG  2029

Query  237   GGCAGTCTCGAAAAGAAGGAGAACAAGGTCCTTGACAGACTTGTCATTCTTGTCAGCATC  296
              || ||||| || ||||| || |||||||||| |||||||||| |||||||| |||||
Sbjct  2028  AGCGGTCTCAAAGAGAAGAAGTACAAGGTCCTTCACAGACTTGTCGTTCTTGTCTGCATC  1969

Query  297   AGCTCTCTTTCTCAGCTCTTCCATGATGGGGTTCTCGGGGTTGATCTCCATGGTCTTCTT  356
             |||||||||||||||||| |||||||||| |||||| ||||| |||||||| ||||||||
Sbjct  1968  AGCTCTCTTTCTCAGCTCATCCATGATGGAGTTCTCTGGGTTAATCTCCATTGTCTTCTT  1909

Query  357   GCTTGACATGTATCCAGCCATGCTGTTATCCCTCAAAGCCTGGGCCTTCATGATTCTCTC  416
             ||| |||||||| ||| |||||||| | || ||||| || || || |||||||| |||||
Sbjct  1908  GCTCGACATGTAGCCACCCATGCTGCTGTCTCTCAAGGCTTGAGCTTTCATGATCCTCTC  1849
```

Figure 1. NETBLAST report showing the highest scoring matches of the *T. cacao* sequence for primer 1 against the non-redundant database at NCBI. The sequence alignment shows the nucleotide match between *T.cacao* and *A. thaliana*. Forward and reverse primers are labeled in bold type.

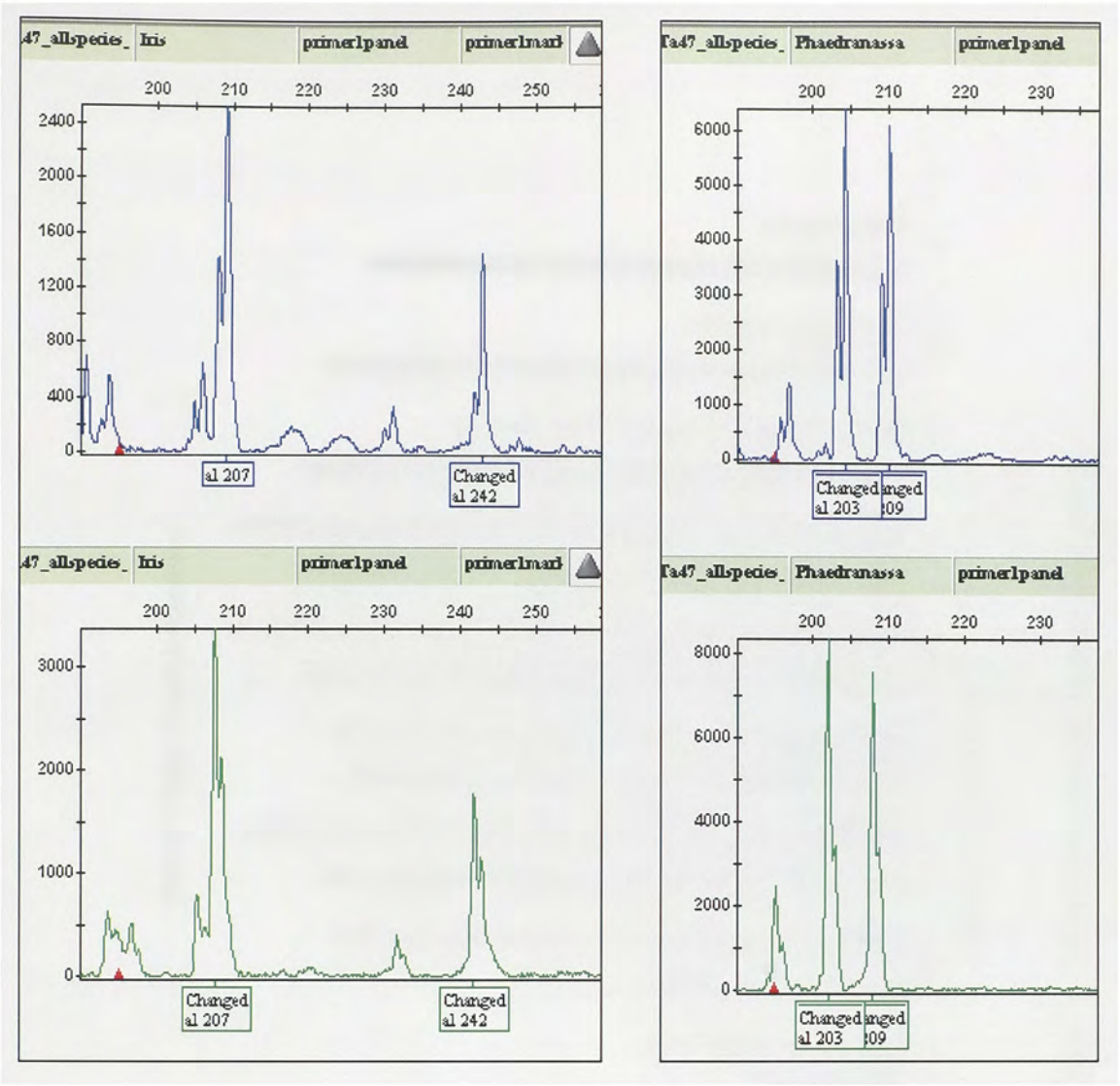

Figure 2. Electropherogram of multiple amplification products for *Iris hexagona* and *Phaedranassa tunguraguare* for primer 1. *Iris hexagona* (cultivar B2-20) at 47˚ with alleles at 207bp and 242bp, and *Phaedranassa tungaraguae* (cultivar 3pt6) at 47˚ with alleles at 203bp and 209bp.

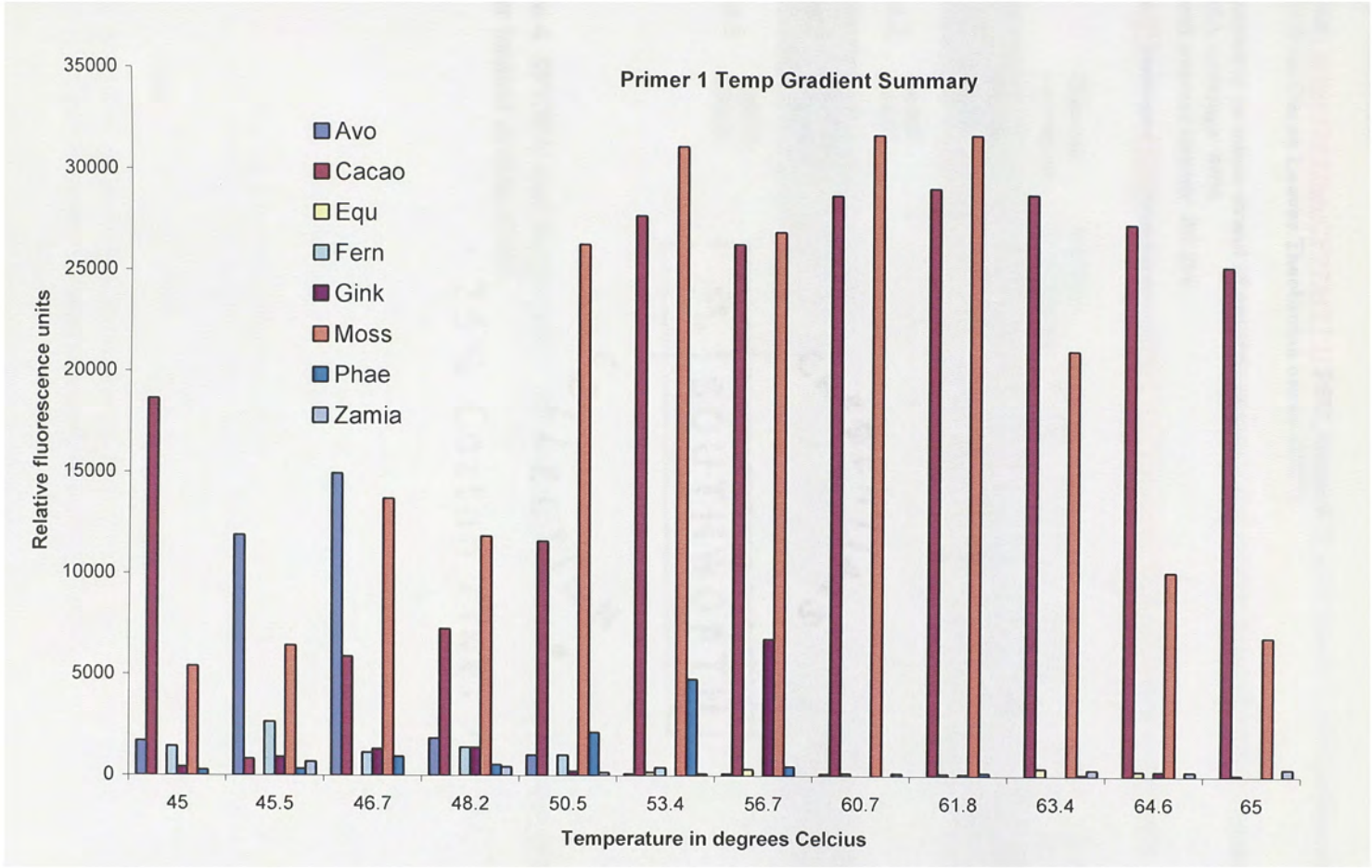

Figure 3. Temperature gradient summary for primer 1.

Genomic: gi|4589412|dbj|AB025606.1| Arabidopsis thaliana genomic DNA, chromosome 5, BAC clone:F6N7

mRNA: gi|38499237|gb|CF973477.1| PSU_blonup4-7_A05_blon4-7_034 Defense-related ESTs from Cacao Leaves Theobroma cacao cDNA

Alignment is on minus strand of genomic sequence and on plus strand of mRNA sequence
mRNA coverage: 46%
Overall percent identity: 86.2%

	Genomic coordinates	mRNA coordinates	length	identity	mismatches	gaps	Donor site	Acc. site
Exon 1	34462-34486	59-83	25	100.0%	0	0		
Exon 2	34366-34405	143-182	40	85.0%	6	0		
Exon 3	34042-34277	271-506	236	84.3%	37	0		
Exon 4	33970-33988	560-578	19	94.7%	1	0		

Figure 4. SPIDEY data for primer 1. Forward primer located in exon one and reverse primer located in exon three.

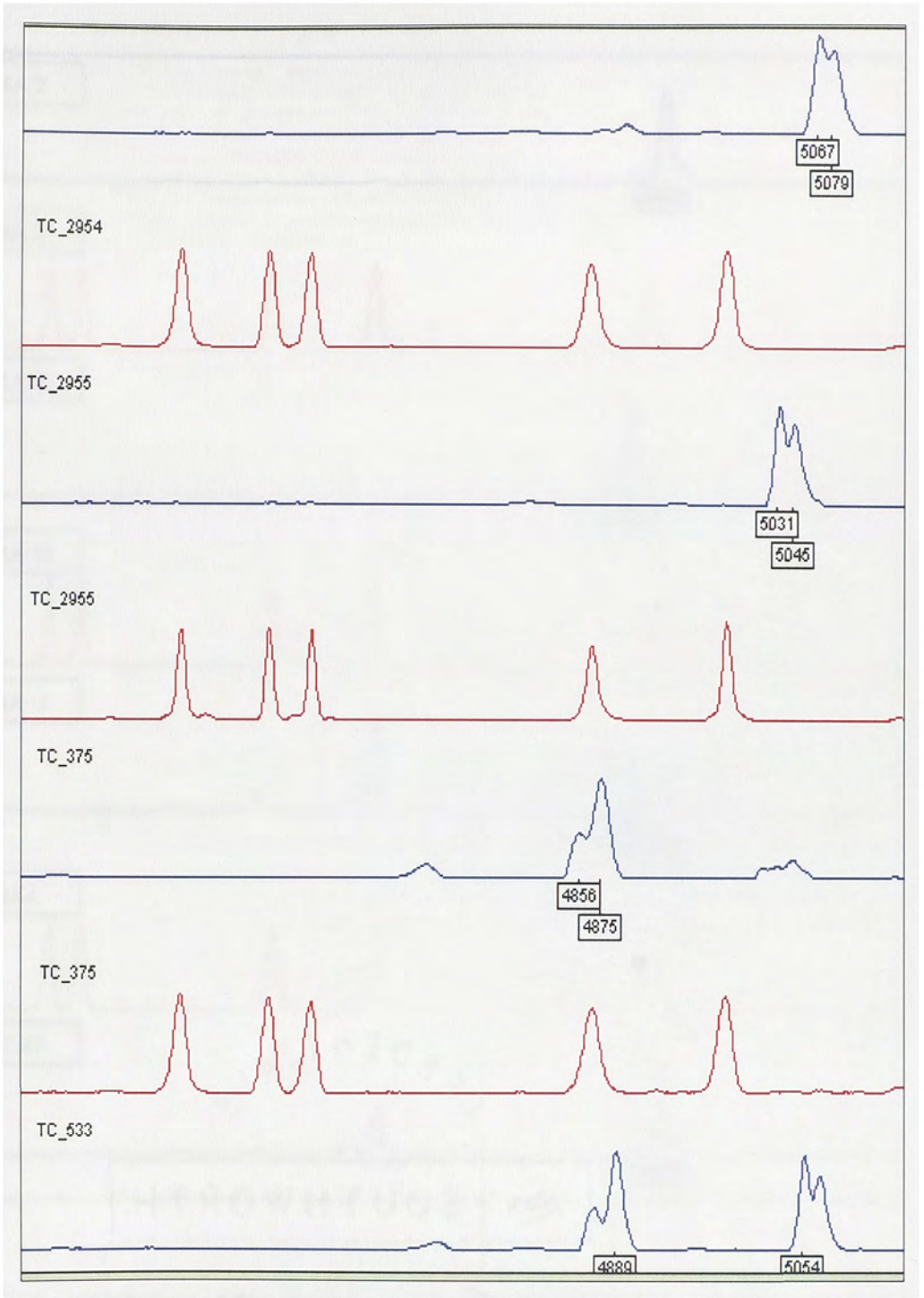

Figure 5. SSCP data of *Theobroma cacao* at 22˚C demonstrating polymorphism for primer 1. Cultivar names located in upper left corner of each electropherogram slide.

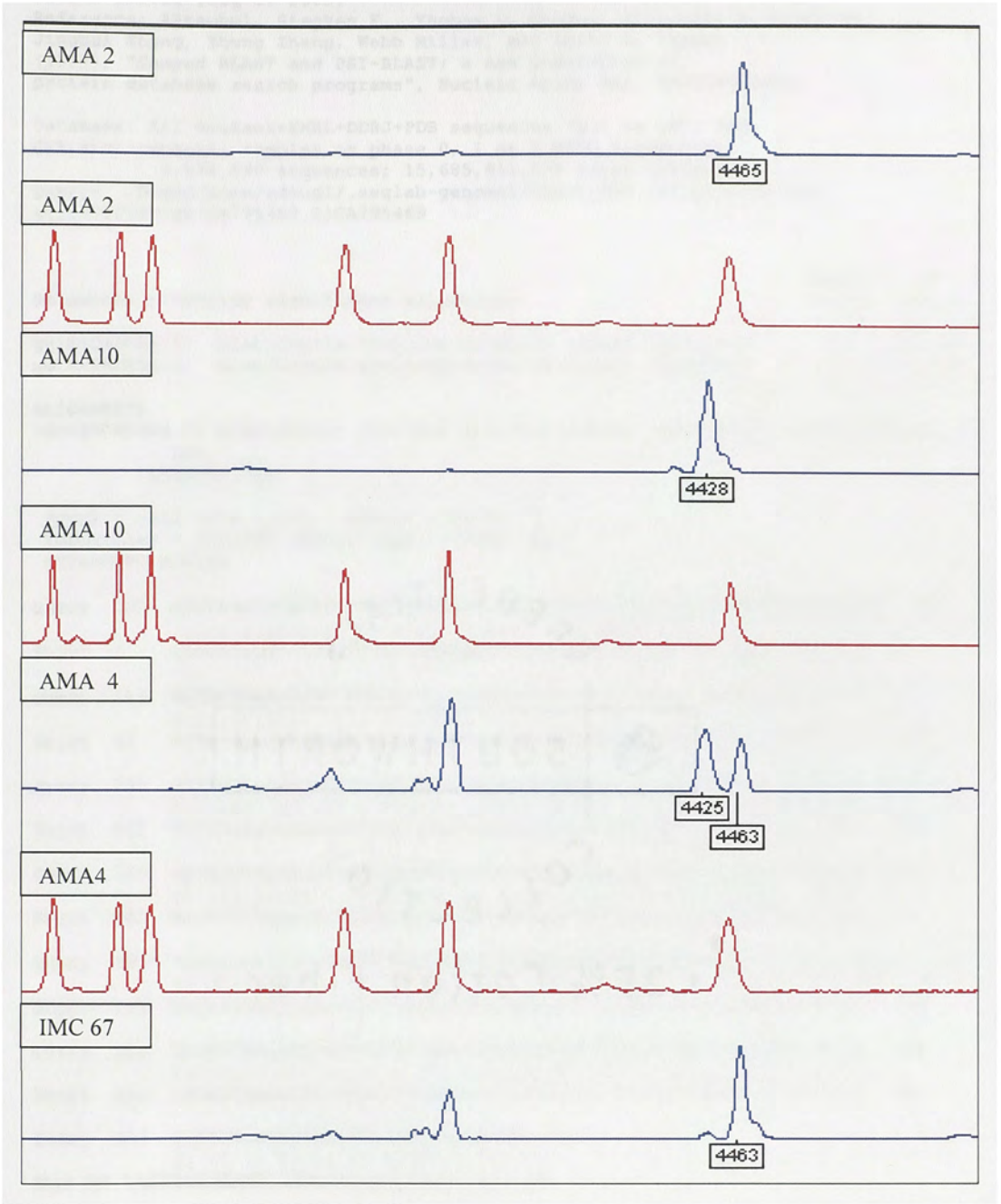

Figure 6. SSCP data of *Theobroma cacao* at 28˚C for primer 1. Culitivar names located in upper left corner of electropherogram panels.

```
BLASTN 2.2.12 [Aug-07-2005]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schdffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           3,534,590 sequences; 15,685,911,575 total letters
Query=  /bighd/home/sdougl/.seqlab-genome1/input_233.rsf{gi-26052545}
gi|26052545|gb|CA795469.1|CA795469


                                                                    Score     E
Sequences producing significant alignments:                         (Bits)  Value

gb|AY040004.1|  Arabidopsis thaliana putative kinase (At4g1911...   313    3e-82
gb|AF360190.1|  Arabidopsis thaliana putative kinase (At4g1911...   313    3e-82

ALIGNMENTS
>gb|AY040004.1| Arabidopsis thaliana putative kinase (At4g19110) mRNA, complete
           cds
          Length=1386

 Score =  313 bits (158),  Expect = 3e-82
 Identities = 332/390 (85%), Gaps = 0/390 (0%)
 Strand=Plus/Plus

Query  123  ATGGAGAGGTACAAGCTAATTAAGGAAGTTGGTGATGGAACATTTGGTAGTGTGTGGCGA  182
            ||||| ||||||||| ||||||| || |||||||||||||| |||||||||| ||||||
Sbjct  1    ATGGACAGGTACAAGTTAATTAAAGAGGTTGGTGATGGAACTTTTGGTAGTGTTTGGCGA  60

Query  183  GCTATAAATAAGTTGTCTGGTGAAGTTGTTGCAATAAAGAAAATGAAGAAGAAATATTAC  242
            |||||||||||  | | |||||||| |||||||| |||||||||| ||||| || |||
Sbjct  61   GCTATAAATAAGCAGACGGGTGAAGTCGTTGCAATTAAGAAAATGAAAAAGAAGTACTAC  120

Query  243  TCATGGGAAGAGTGTGTGAACCTGAGAGAAGTTAAGTCACTGCGGAGAATGAATCATCCC  302
            || ||||| || ||| | || |||||||||||||||||  ||  ||||||||||||||||
Sbjct  121  TCTTGGGATGAATGTATTAATCTGAGAGAAGTTAAGTCGCTCAGGAGAATGAATCATCCA  180

Query  303  AATATTGTGAAGCTTAAGGAAGTCATCAGGGAAAATGATATTCTTTACTTTGTGTTTGAA  362
            || ||||||||| | |||||||| ||| |||||||||||||| || |||||||| |||||
Sbjct  181  AACATTGTGAAGTTGAAGGAAGTAATCCGGGAAAATGATATCCTATACTTTGTCTTTGAG  240

Query  363  TATTTGGAGTGCAGCCTTTACCAACTTATGAAAGATAGGGAAAAGCTTTTTTCAGAAGTT  422
            ||  |||||||||  |||||||||||||||||| ||| |  |||||||||| |||||| |
Sbjct  241  TACATGGAGTGCAATCTTTACCAACTTATGAAGGATCGACAAAAGCTTTTTGCAGAAGCT  300

Query  423  GAAATCAGAAATTGGTGCTTCCAAGTCTTCCAAGGTCTTGCATACATGCACCAGCGTGGA  482
            || |||| |||||||||||| |||||||||||||| ||| | |||||||| ||||||||
Sbjct  301  GATATCAAAAATTGGTGCTTTCAAGTCTTCCAAGGCCTTTCTTACATGCATCAGCGTGGT  360

Query  483  TATTTTCATCGTGACCTAAAGCCAGAAAAT  512
            || || || || || || ||||||||||||
Sbjct 361 TACTTCCACCGCGATCTTAAGCCAGAAAAT 390
```

Figure 7. NETBLAST report showing the highest scoring matches of the *T. cacao* sequence for primer 2 against the non-redundant database at NCBI. The sequence alignment shows the nucleotide match between *T.cacao* and *A. thaliana*. Forward and reverse primers are labeled in bold type.

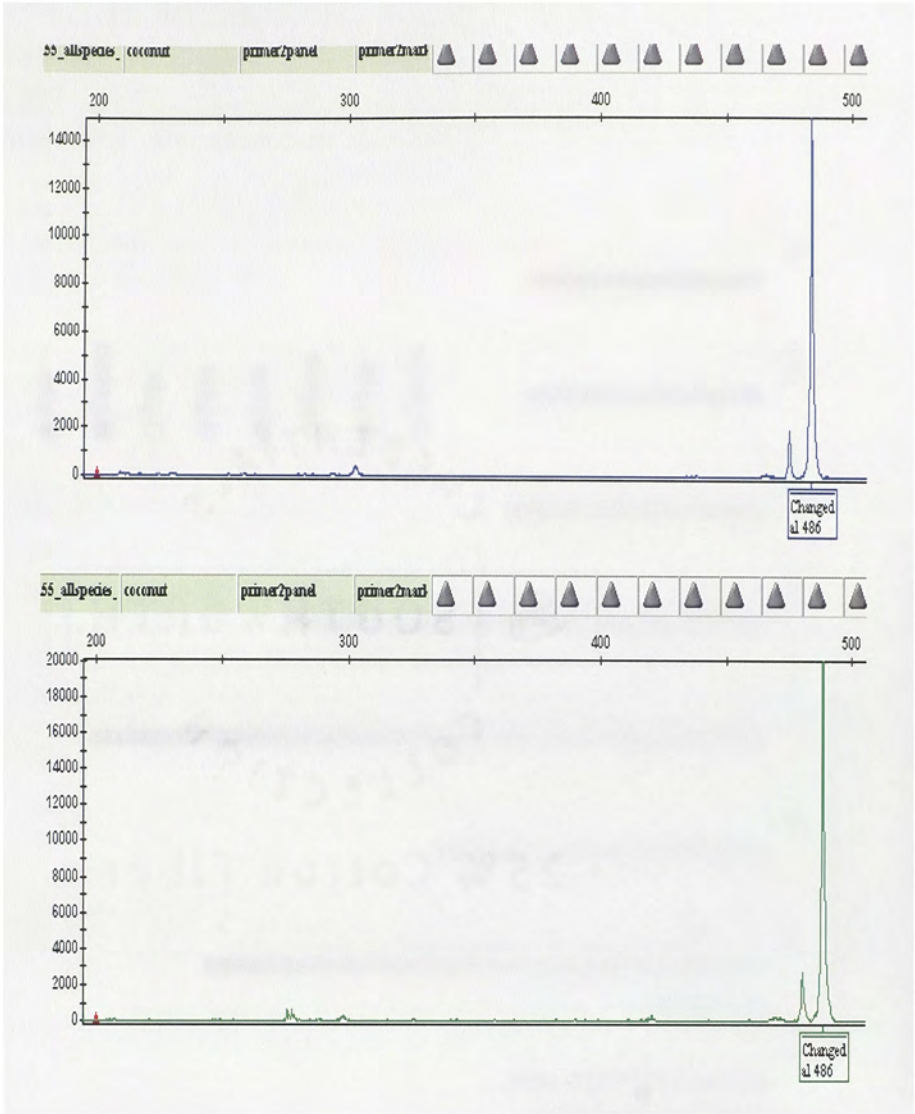

Figure 8. Electropherogram data for amplification product of *Cocos nucifera* for primer 2. Product size is 486 base pairs.

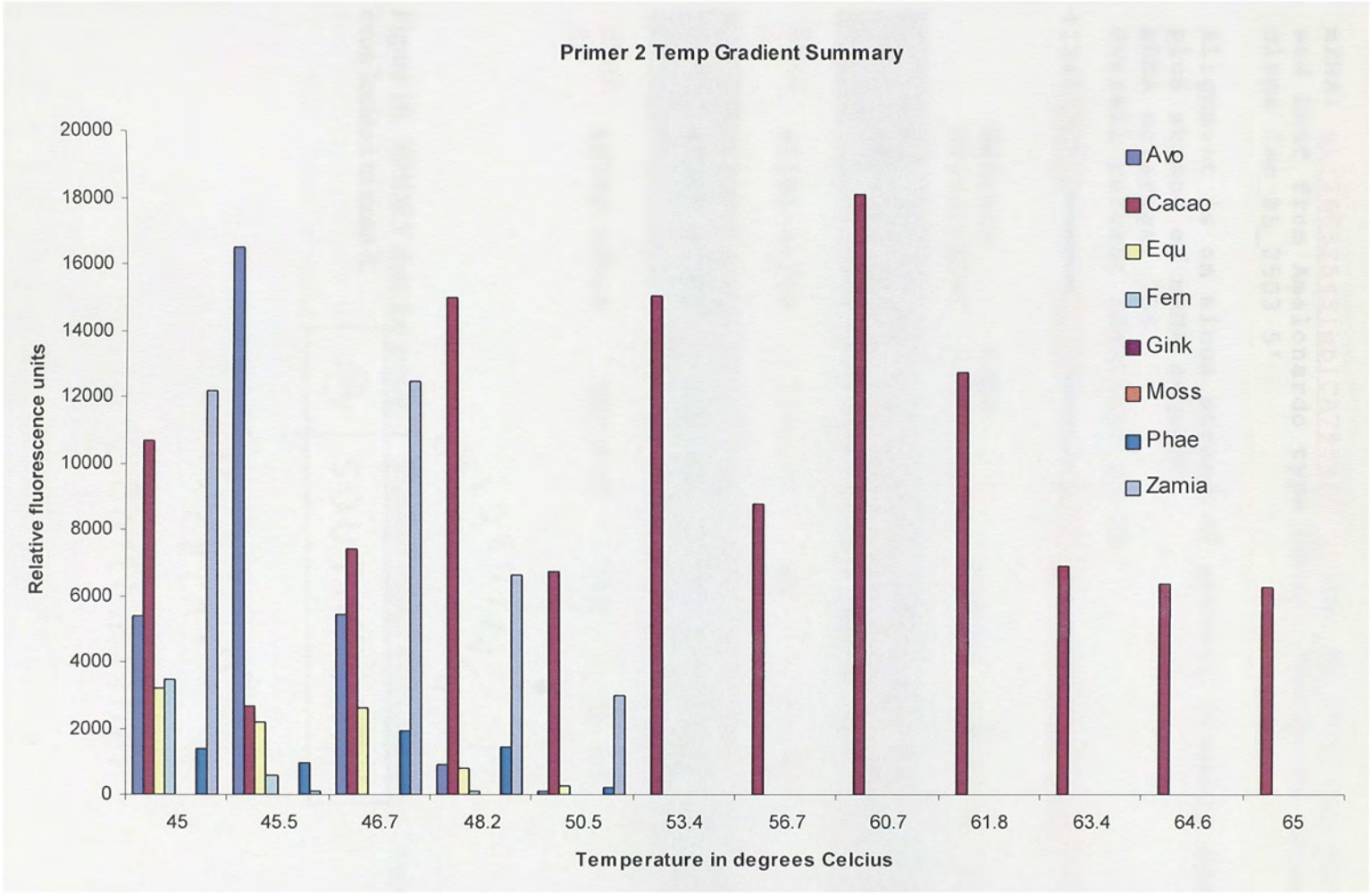

Figure 9. Temperature gradient summary for primer 2.

Genomic: gi|7268697|emb|AL161550.2|ATCHRIV50 Arabidopsis thaliana DNA chromosome 4, contig fragment No. 50

mRNA: gi|26052545|gb|CA795469.1| Cac_BL_2503 Cac_BL (Bean and Leaf from Amelonardo type Cacao) Theobroma cacao cDNA clone Cac_BL_2503 5'

Alignment is on minus strand of genomic sequence and on plus strand of mRNA sequence
mRNA coverage: 75%
Overall percent identity: 84.3%

41345 40929

	Genomic coordinates	mRNA coordinates	length	identity	mismatches
Exon 1	41345-41425	129-209	81	87.7%	10
Exon 2	41191-41259	210-278	69	85.5%	10
Exon 3	41019-41108	279-368	90	83.3%	15
Exon 4	40782-40929	369-516	148	82.4%	26

Figure 10. SPIDEY data for primer 2. Forward primer located in exon 1 and reverse exon located in exon 4.

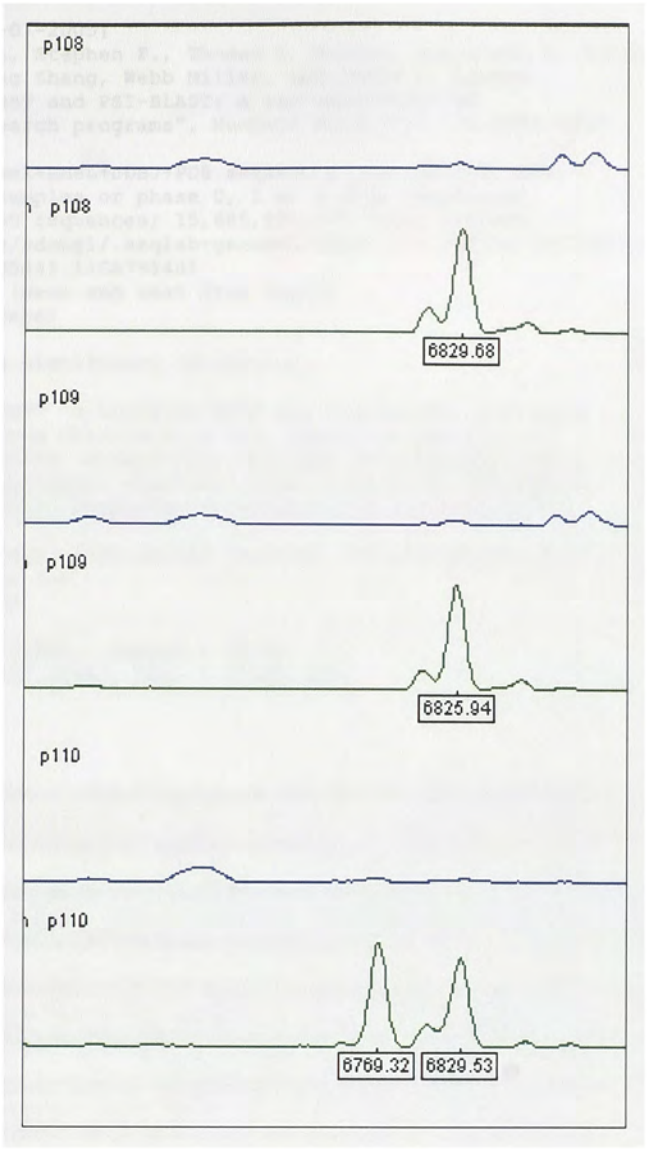

Figure 11. Electropherogram of SSCP data for primer 2 at 28˚C showing two homozygous individuals and one heterozygous individual. Cultivar names are located in the top left corner of each electropherogram panel.

```
BLASTN 2.2.12 [Aug-07-2005]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schdffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           3,534,590 sequences; 15,685,911,575 total letters
Query=  /bighd/home/sdougl/.seqlab-genome1/input_239.rsf{gi-26052517}
gi|26052517|gb|CA795441.1|CA795441
Cac_BL_2478 Cac_BL (Bean and Leaf from Amelon
          (524 letters)
                                                                  Score     E
Sequences producing significant alignments:                       (Bits)  Value

emb|X96974.1|SOPROTE37  S.oleracea mRNA for proteasome 37kD subun   331    1e-87
dbj|AP006077.1|  Lotus corniculatus var. japonicus genomic DNA...   323    3e-85
emb|BX822183.1|CNS0A7T8  Arabidopsis thaliana Full-length cDNA...   315    7e-83
gb|AY085914.1|  Arabidopsis thaliana clone 19620 mRNA, complete s   315    7e-83
gb|AF386953.1|AF386953  Arabidopsis thaliana 20S proteasome su...   315    7e-83

>gb|AF386953.1|AF386953 Arabidopsis thaliana 20S proteasome subunit PAC1 (MKA23.2) mRNA,
            complete cds
          Length=980

 Score =  315 bits (159),  Expect = 7e-83
 Identities = 357/423 (84%), Gaps = 0/423 (0%)
 Strand=Plus/Plus


Query  203  GGATGGTGTTGTCTTGGTTGGTGAAAAGAAAGTTACTTCCAAGCTTCTTCAAACCTCTAC  262
            || |||||||| ||| | ||||| |||||||| || || ||||||||||||||||| ||
Sbjct  214  AGACGGTGTTGTGTTGATCGGTGAGAAGAAAGTCACCTCAAAGCTTCTTCAAACCTCAAC  273

Query  263  ATCTACCGAGAAGATGTACAAGATTGATGATCATGTTGCTTGTGCTGTGGCGGGAATAAT  322
            ||| | ||||| ||||||||||||||||| |||||||| |||||||| || || |||||
Sbjct  274  CTCTGCTGAGAAAATGTACAAGATTGATGACCATGTTGCCTGTGCTGTTGCTGGTATAAT  333

Query  323  GTCTGATGCAAACATCCTCATCAACACTGCTAGGGTGCAAGCACAACGATACACATATGC  382
            ||||||||| ||||| || |||||||| ||| | || ||||| ||||| ||||| |
Sbjct  334  GTCTGATGCTAACATACTTATCAACACGGCTCGAGTTCAAGCTCAACGTTACACCTTCAT  393

Query  383  TTACCAAGAGCCAATGCCTGTCGAACAGTTAGTTCAATCTCTTTGTGACACAAAGCAGGG  442
            |||||||||| ||||| || || ||| | ||||||||||||||||||||||| || ||
Sbjct  394  GTACCAAGAGCCCATGCCGGTTGAGCAGCTGGTTCAATCTCTTTGTGACACAAAACAAGG  453

Query  443  TTACACACAGTTTGGTGGTCTCCCGCCCTTCGGTGTTTCATTTCTTTTTGCAGGATGGGA  502
            |||||||||||||||||| |  || || ||||| || ||||| ||||||||||||||
Sbjct  454  ATACACACAGTTTGGTGGTCTTCGCCCGTTTGGTGTCTCCTTTCTCTTTGCAGGATGGGA  513
```

Figure 12. NETBLAST report showing the highest scoring matches of the *T. cacao* sequence for primer 3 against the non-redundant database at NCBI. The sequence alignment shows the nucleotide match between *T.cacao* and *A. thaliana*. Forward and reverse primers are labeled in bold type.

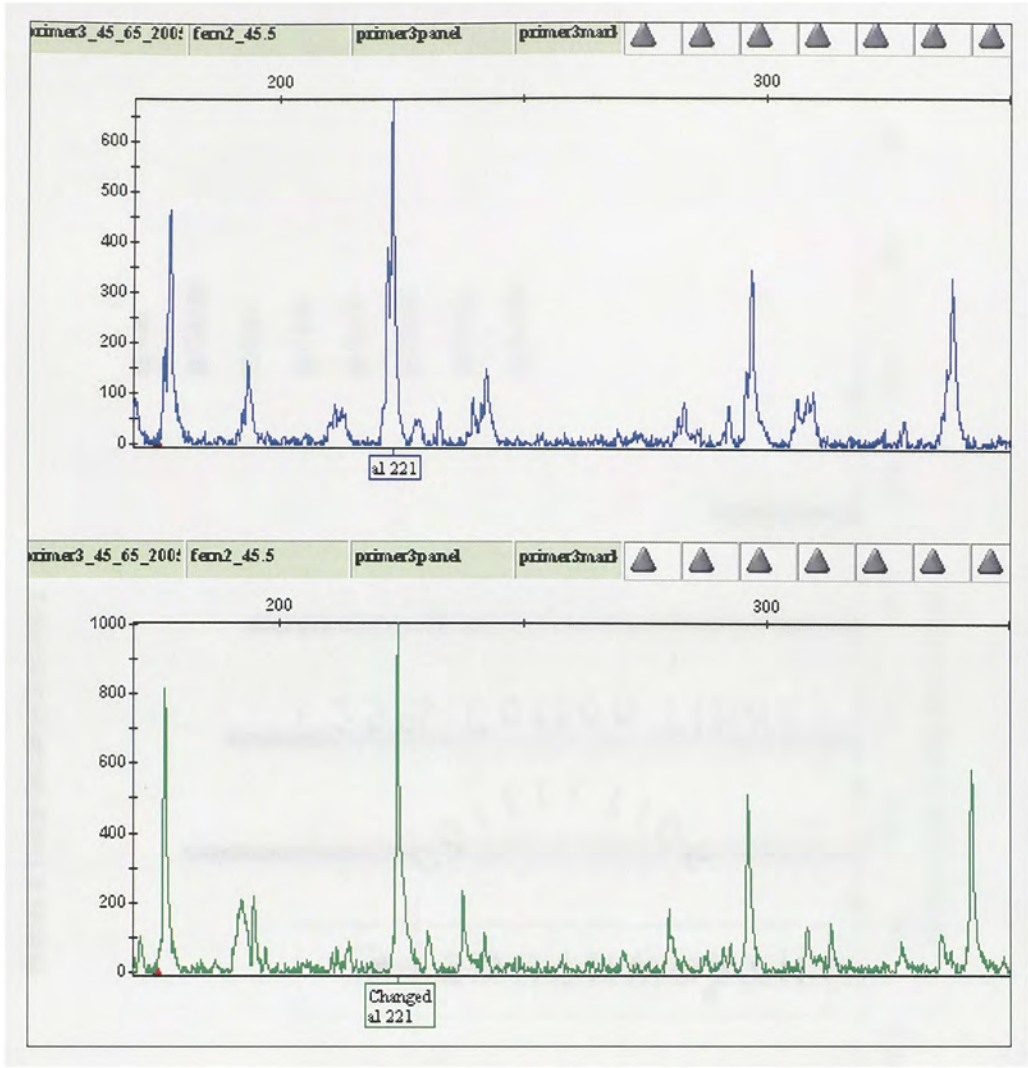

Figure 13. Electropherogram of amplification products for *Pteris bahamensis* with a main product at the size of 221 base pairs. The presence of several smaller products is due to the low annealing temperature of 45.5°C.

Figure 14. Temperature gradient summary for primer 3

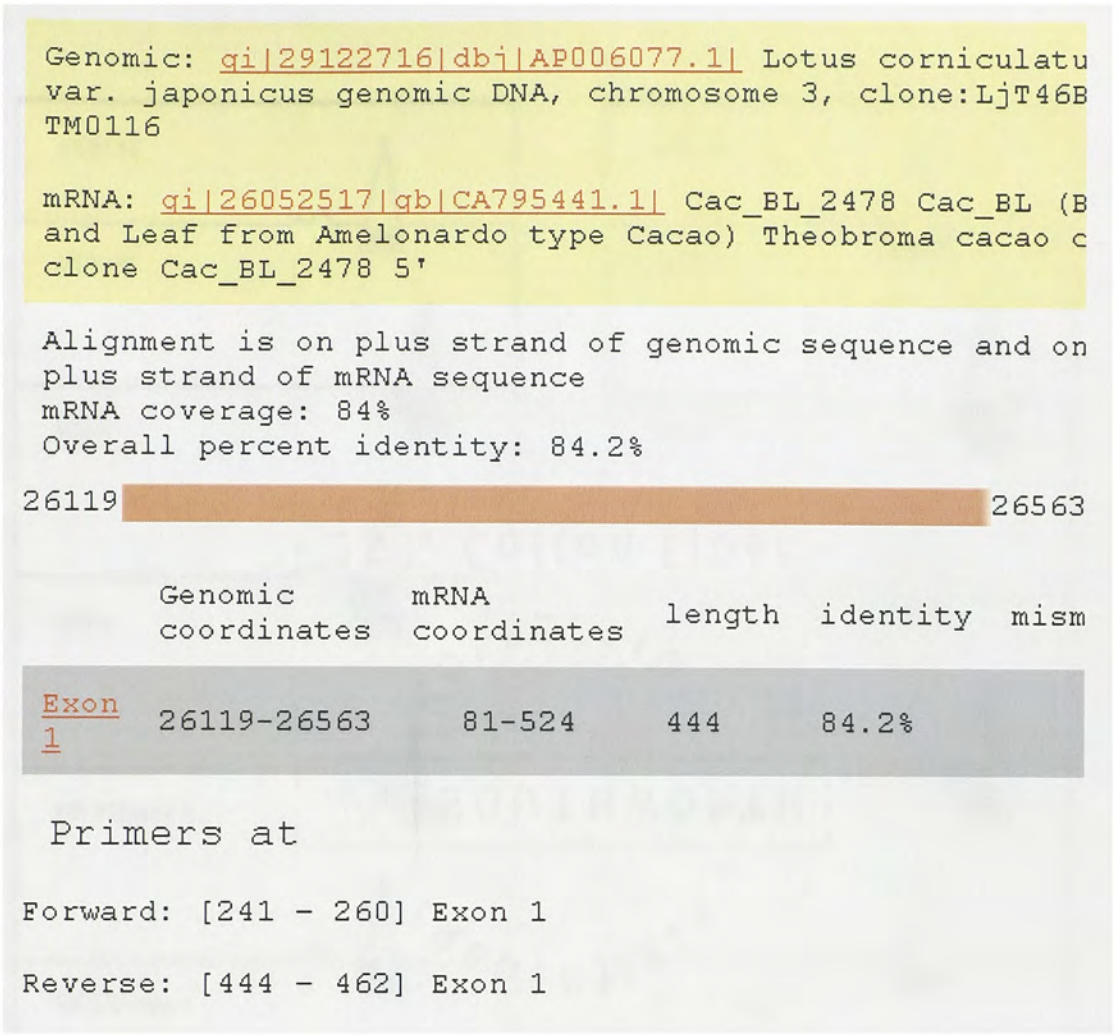

Figure 15. SPIDEY data for primer 3. Both forward and reverse primers are located in exon 1.

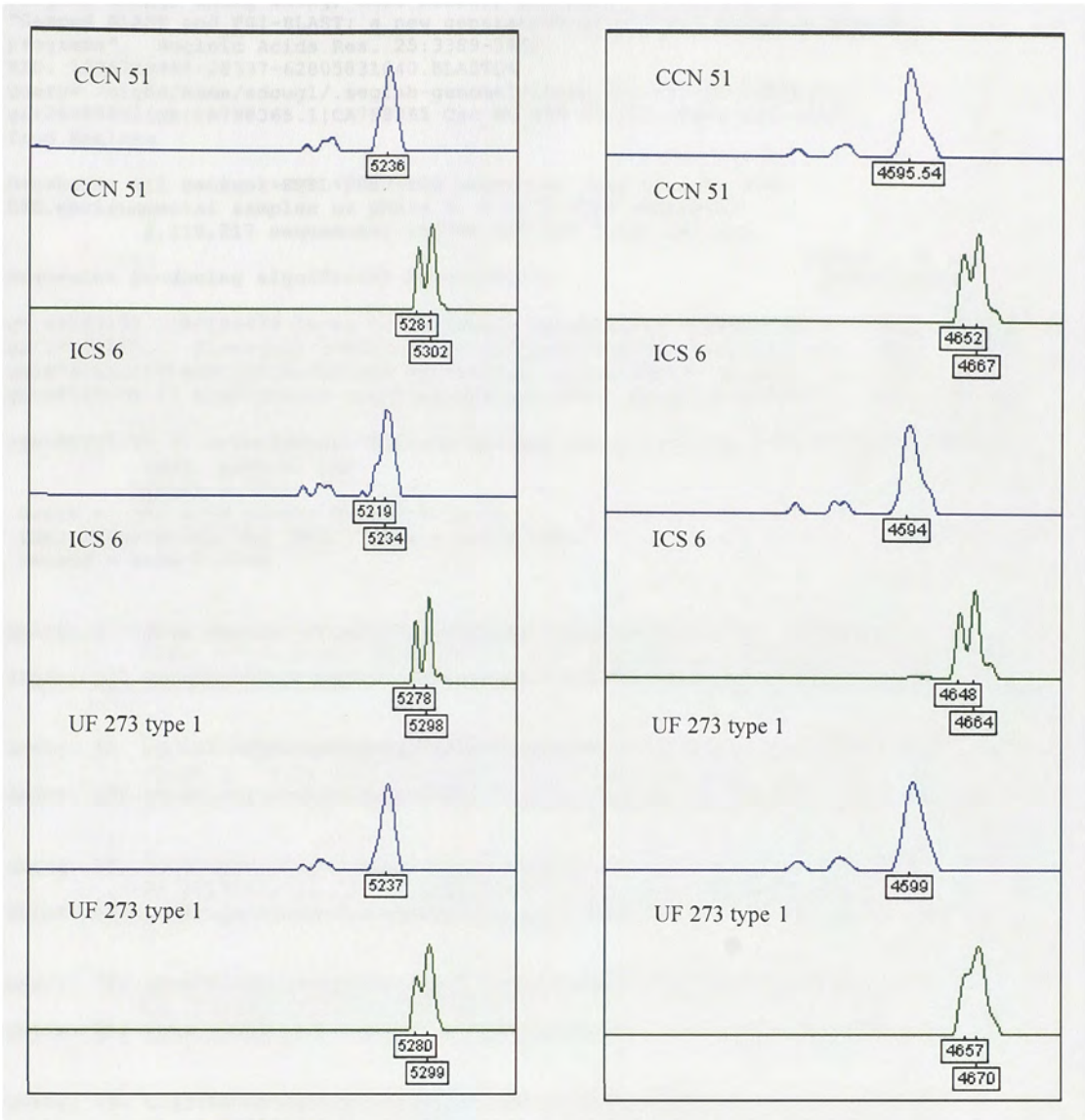

Figure 16. SSCP data for T. cacao for primer 3 at 22˚C and 28˚C respectively. Data does not indicate polymorphism for this primer. Cultivar names located in top left corner of each electropherogram.

```
BLASTN 2.2.12 [Aug-07-2005]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.
RID: 1086719965-28337-62805831040.BLASTQ4
Query= /bighd/home/sdougl/.seqlab-genome1/input_282.rsf{gi-26055451}
gi|26055451|gb|CA798365.1|CA798365 Cac_BL_696 Cac_BL (Bean and Leaf
from Amelona

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           2,219,217 sequences; 10,709,859,898 total letters
                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gb|AF193438.1|AF193438 Hevea brasiliensis ubiquitin precursor (u...   412   e-112
gb|AF061807.1| Elaeagnus umbellata polyubiquitin mRNA, complete cds   377   e-101
gb|M74101.1|TOBUBI11A Nicotiana sylvestris (clone 6PU11) hexamer...   303   2e-79
gb|AF528709.1| Arabidopsis thaliana ecotype Chi-1 putative polyu...   301   7e-79

>gb|AF528709.1| Arabidopsis thaliana ecotype Chi-1 putative polyubiquitin (UBQ13)
           gene, partial cds
          Length = 1011
 Score =  301 bits (152), Expect = 7e-79
 Identities = 363/432 (84%), Gaps = 1/432 (0%)
 Strand = Plus / Plus


Query: 1   gaggtggaaagctctgacactattgataatgtcaaagctaagattcaagacaaggaaggg 60
           ||||| ||||| ||||| |||||||| ||||| || |||||||| || || ||||||||
Sbjct: 133 gaggttgaaagttctgatactattgacaatgttaaggctaagatccaggataaggaagga 192


Query: 61  atcccaccagacCAACAGAGGCTTATCTTTGCtggtaaacagctagaggatggccgcacc 120
           ||||| || || || ||||||||||||||||| ||||| |||||||||||||| |||||
Sbjct: 193 atccctccggatcagcagaggcttatctttgccggtaagcagctagaggatggtcgcaca 252


Query: 121 cttgctgattacaatatccagaaggagtccaccctccaccttgttcttcgacttaggggt 180
           || |||||||||| ||||| ||||||||||||||||  | || ||||| || || |||
Sbjct: 253 ctcgctgattacaacatccaaaaggagtccaccctccatttggtgcttcgtctcagaggt 312


Query: 181 ggcatgcagatatttgtgaaaactctcactgggaaaaccatcactttggaggtggagagt 240
           || |||||||||||| |||| || |||||||| || || |||||||||||||| |||||
Sbjct: 313 ggtatgcagatatttttgaagaccctcactggaaagacaatcactttggaggttgagagc 372


Query: 241 tcggataccatcgataatgtgaaggccaaGATTCAGGACAAGGAGGGTatcccaccagac 300
           || || |||||||||||| |||||||| |||||||||||||||| ||||||||||||||
Sbjct: 373 tcagacaccatcgataacgtgaaggctaagattcaggacaaggaaggtatcccaccagac 432
```

Figure 17. NETBLAST report showing the highest scoring matches of the *T. cacao* sequence for primer 4 against the non-redundant database at NCBI. The sequence alignment shows the nucleotide match between *T.cacao* and *A. thaliana*. Forward and reverse primers are labeled in bold type.

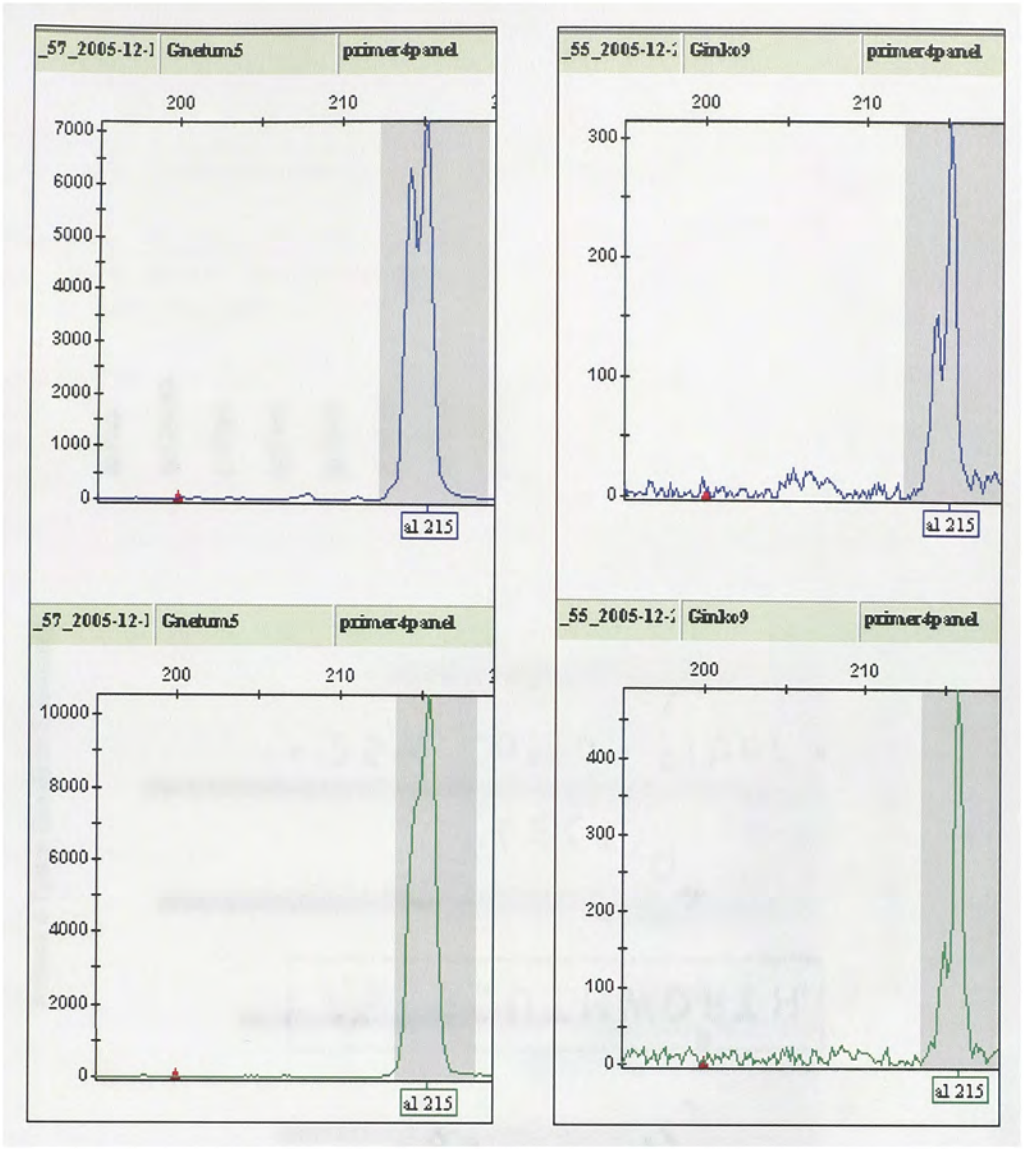

Figure 18. Electropherogram of amplification data for *Gnetum gnemon* and *Ginkgo biloba* for primer 4. Both plant species show a product of 215 base pairs.

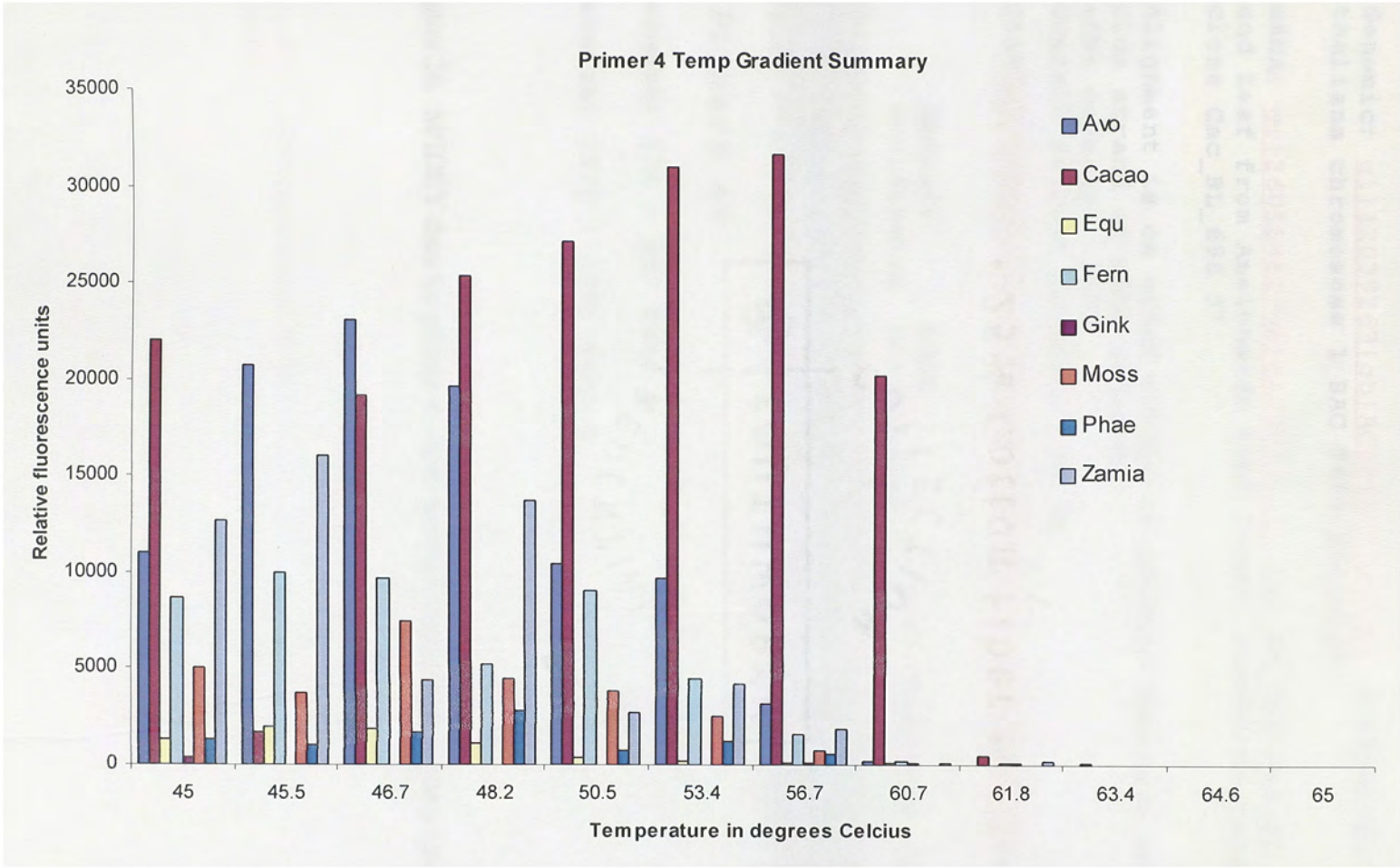

Figure 19. Temperature gradient summary for primer 4.

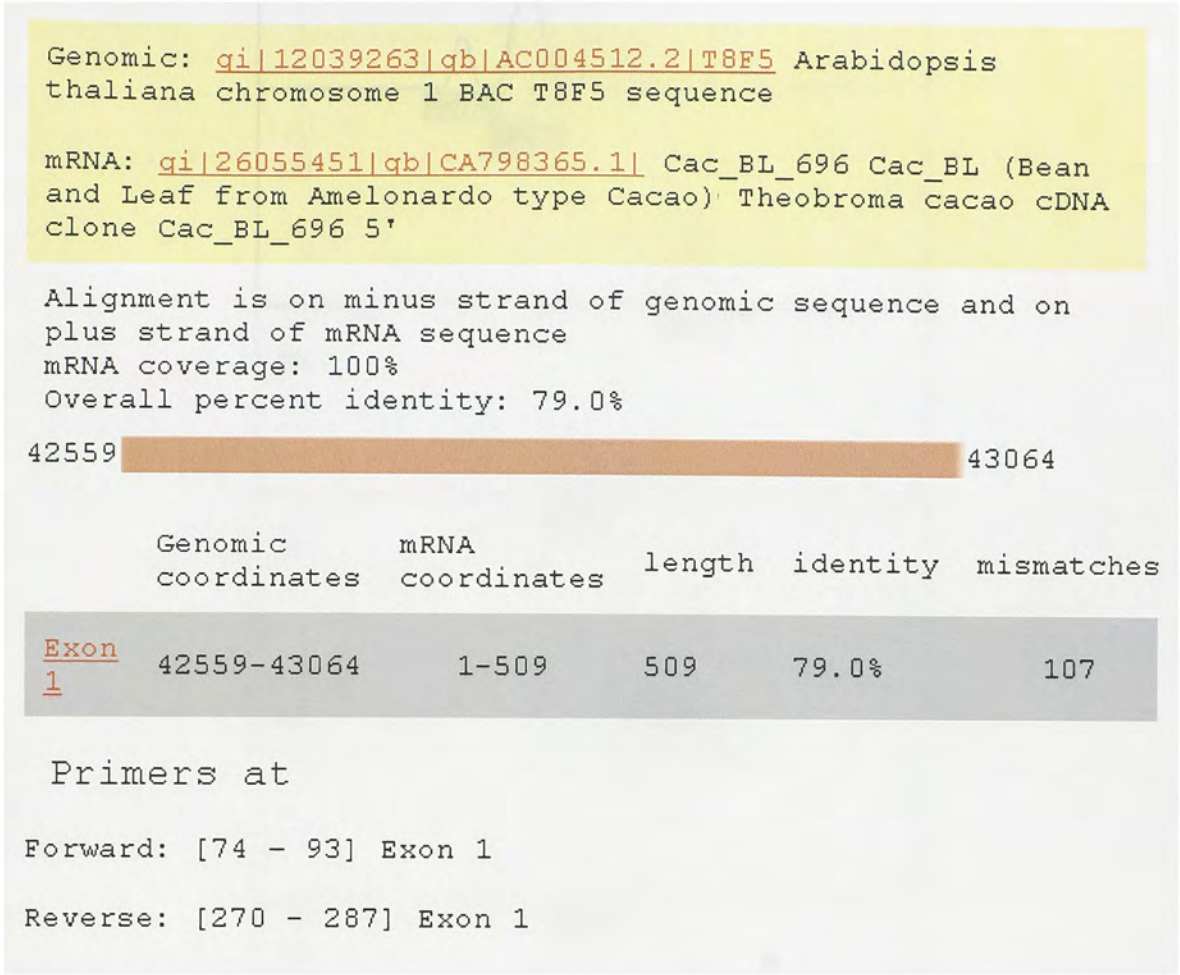

Figure 20. SPIDEY data for primer 4. Both forward and reverse primers located in exon 1.

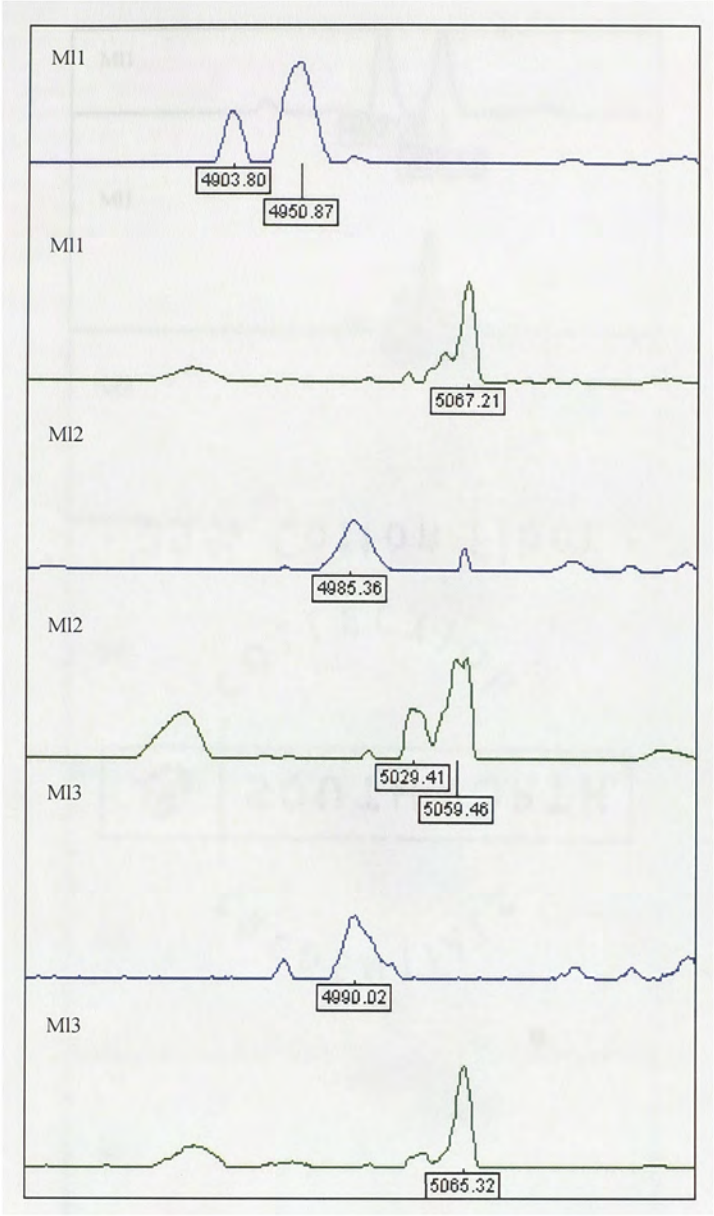

Figure 21. Electropherogram of SSCP data of *Myodocarpus lanceolata* at 22˚C for primer 4.

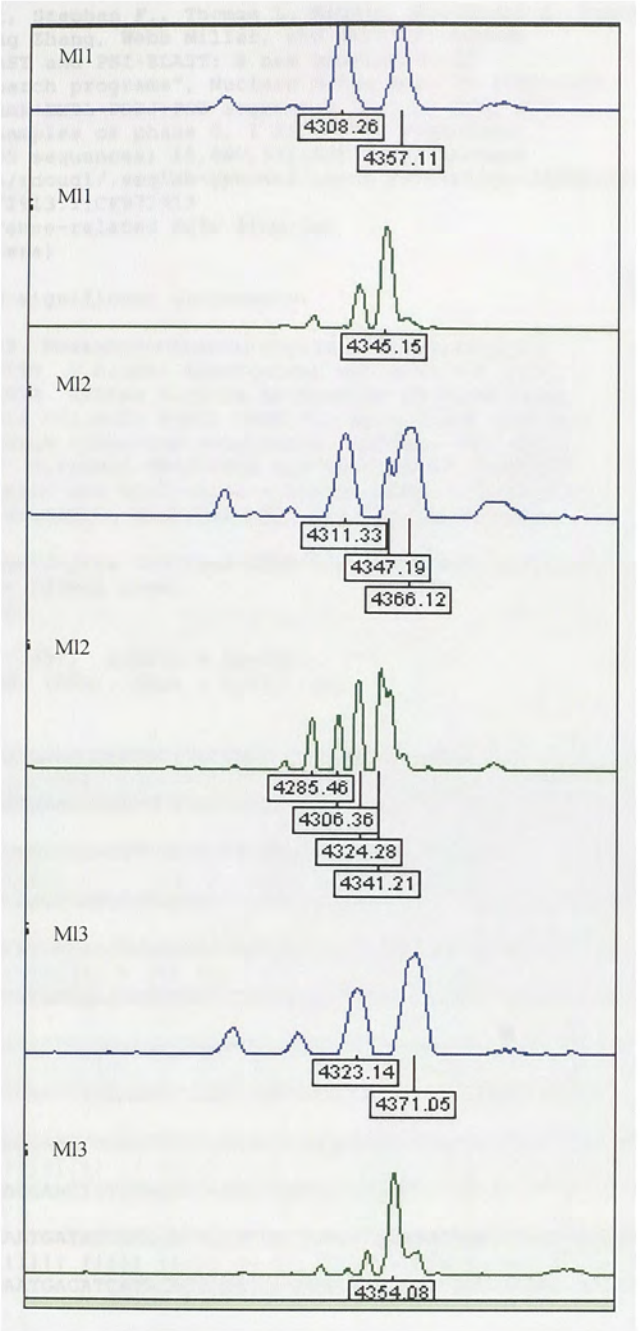

Figure 22. Electropherogram of SSCP data for *Myodocarpus lanceolata* at 28˚C for primer 4.

```
BLASTN 2.2.12 [Aug-07-2005]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schdffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.
Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           3,559,705 sequences; 15,880,532,508 total letters
Query=  /bighd/home/sdougl/.seqlab-genome1/input_247.rsf{gi-38498673}
gi|38498673|gb|CF972913.1|CF972913
PSU_2blonup1A01 Defense-related ESTs from Cac
          (630 letters)
                                                                  Score     E
Sequences producing significant alignments:                       (Bits)  Value

gb|U84889.1|MCU84889  Mesembryanthemum crystallinum methionine...   543    2e-151
emb|Z49150.1|CBKPMETGN  C.blumei kinetoplast met gene for coba...   505    5e-140
gb|AF220054.1|AF220054  Coffea arabica methionine synthase mRNA,    488    1e-134
dbj|AB221011.1|  Beta vulgaris BvMS1 mRNA for methionine synthase   466    4e-128
gb|AF082893.1|  Solanum tuberosum methionine synthase (MS) mRNA,    454    2e-124
emb|X83499.1|CRMETS  C.roseus MetE mRNA for methionine synthase     444    1e-121
gb|AF518566.1|  Glycine max methionine synthase mRNA, complete cd   408    8e-111
emb|AJ608673.1|  Arabidopsis thaliana mRNA for cobalamin-indep...   383    5e-103

>emb|AJ608673.1| Arabidopsis thaliana mRNA for cobalamin-independent methionine
            synthase (atms1 gene)
          Length=2326

 Score =  383 bits (193),  Expect = 5e-103
 Identities = 398/467 (85%), Gaps = 0/467 (0%)
 Strand=Plus/Plus

Query  61    CGCCCAATGAAGGGAATGCTTACTGGCCCTGTAACCATCCTCAACTGGTCCTTTGTCAGA  120
             ||||||||||||| |||||||| || || || ||||| ||||||||||||||||||||
Sbjct  1659  CGCCCAATGAAGGGTATGCTTACCGGTCCCGTCACCATTCTCAACTGGTCCTTTGTCAGG  1718

Query  121   AATGATCAGCCTAGATTTGAGACATGCTACCAGATTGCTTTGGCCATCAAGGACGAAGTG  180
             || || ||||| ||    || || || ||||||||||||||||||||||||||||||||
Sbjct  1719  AACGACCAGCCCAGGCACGAAACCTGTTACCAGATTGCTTTGGCCATCAAGGACGAAGTC  1778

Query  181   GAGGATCTTGAGAAGGCTGGTATCAATGTTATCCAAATTGATGAGGCTGCTTTGAGAGAG  240
             |||||||||||||| | ||| |||  ||| || || ||||||||||||  | |||||
Sbjct  1779  GAGGATCTTGAGAAAGGTGGAATCGGTGTCATTCAGATTGATGAGGCTGCACTTAGAGAA  1838

Query  241   GGGTTACCTCTCAGGAAGTCTGAACATGCCTTCTACTTGAAATGGGCTGTCCACTCCTTC  300
             ||  |||| |||||||| || || ||||| ||||||||| | ||||| ||||||||||||
Sbjct  1839  GGACTACCACTCAGGAAATCCGAGCATGCTTTCTACTTGGACTGGGCCGTCCACTCCTTC  1898

Query  301   AGGATCACCAACTGCGGCGTCCAGGACACTACCCAGATCCACACCCACATGTGCTACTCC  360
             || |||||||||| || ||||| ||||  |||||||||||||| ||||||||||||||||
Sbjct  1899  AGAATCACCAACTGTGGAGTCCAAGACAGCACCCAGATCCACACTCACATGTGCTACTCC  1958

Query  361   AACTTCAATGATATCATCCACTCTATTATTGACATGGATGCTGATGTCATCACCATTGAA  420
              |||||||||| ||||| ||||| || || ||||||||||||||||||||||||||||||
Sbjct  1959  CACTTCAATGACATCATACACTCCATCATCGACATGGATGCTGATGTCATCACCATTGAG  2018
```

Figure 23. NETBLAST report showing the highest scoring matches of the *T. cacao* sequence for primer 5 against the non-redundant database at NCBI. The sequence alignment shows the nucleotide match between *T.cacao* and *A. thaliana*. Forward and reverse primers are labeled in bold type.

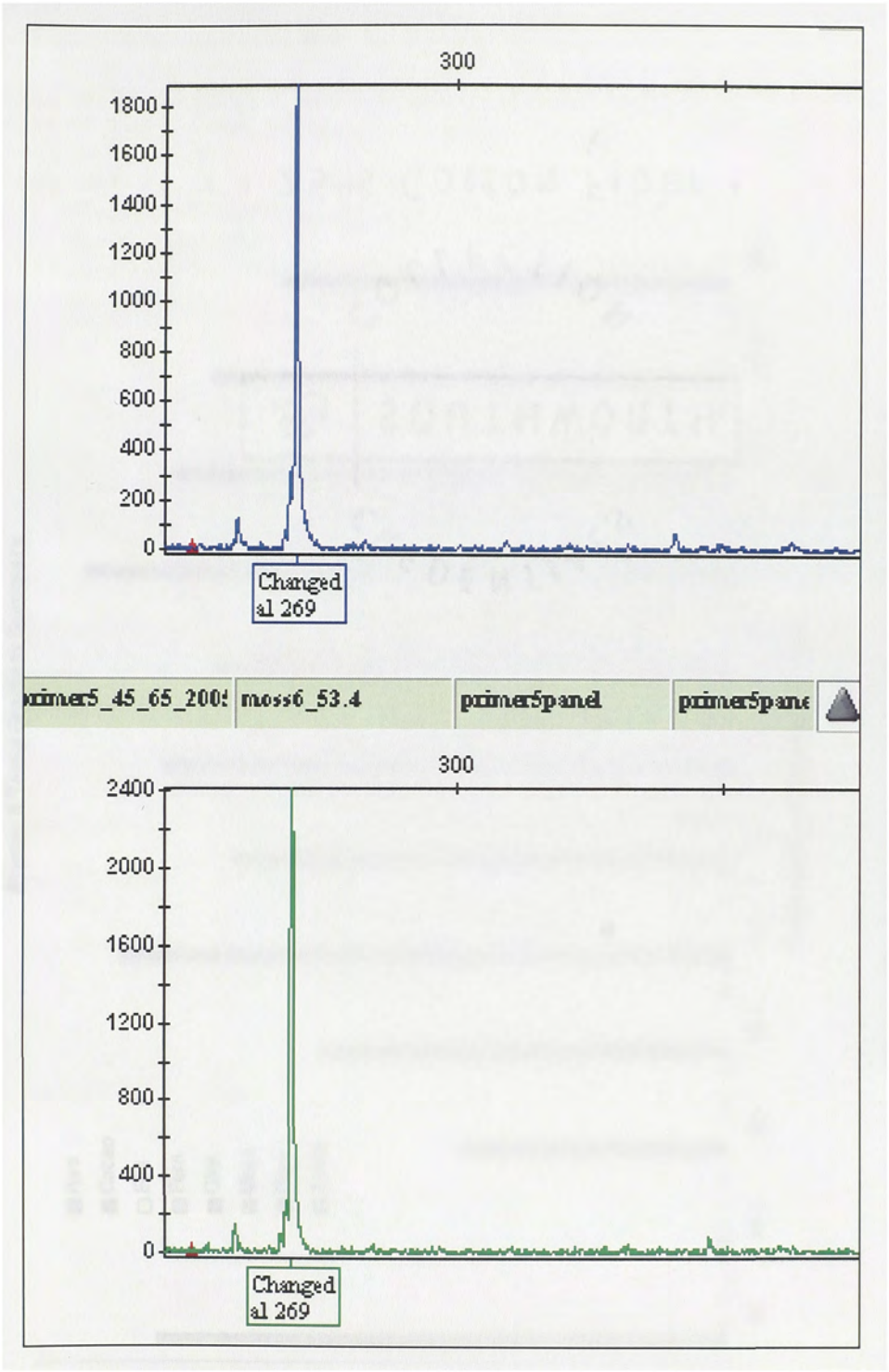

Figure 24. Electropherogram of amplification data for a moss species for primer 5. Amplification product size is 269 base pairs.

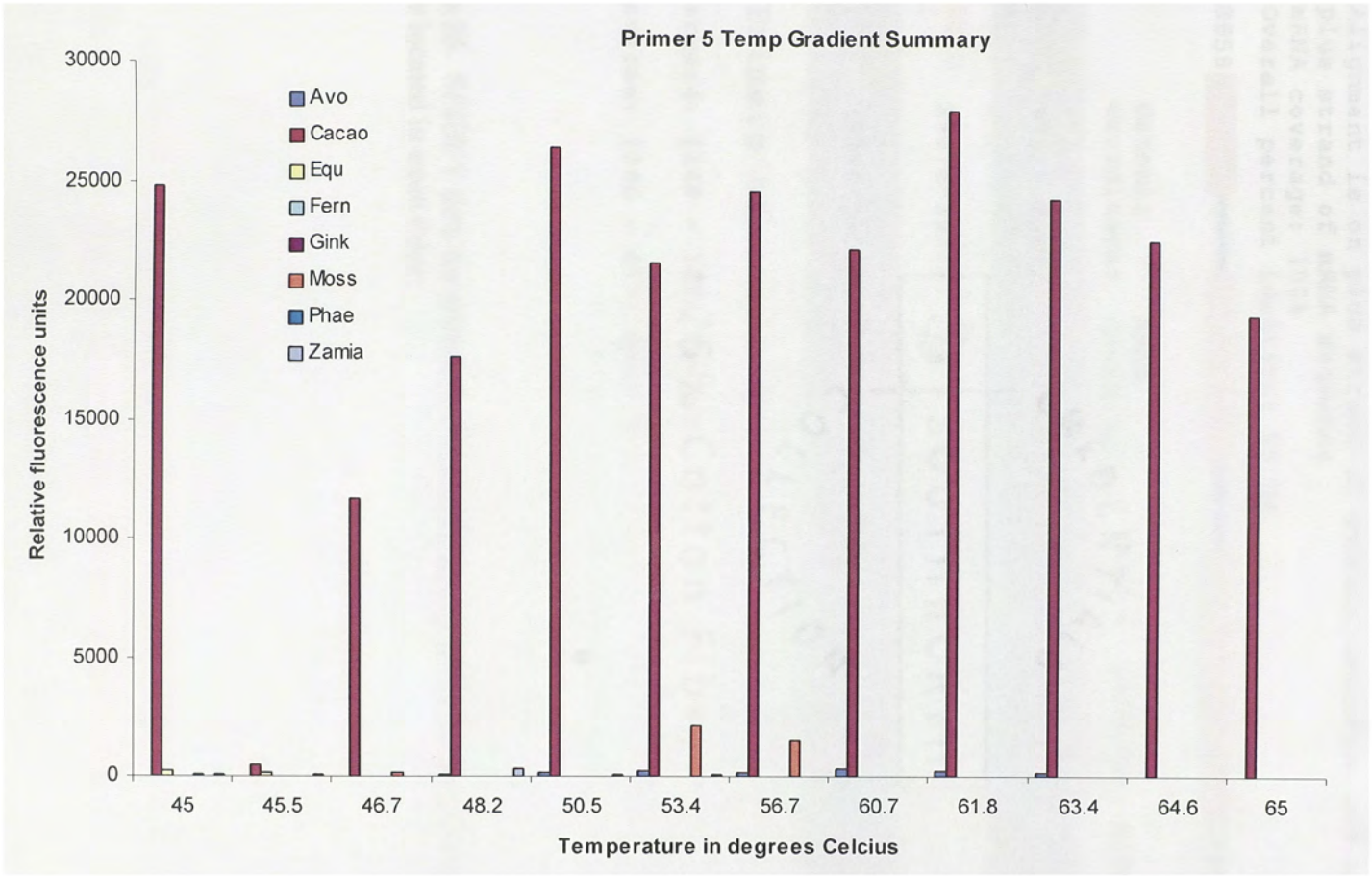

Figure 25. Temperature gradient summary for primer 5.

Genomic: qi|2924730|dbj|AB011480.1| Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MPI7

mRNA: qi|38498673|gb|CF972913.1| PSU_2blonup1A01 Defense-related ESTs from Cacao Leaves Theobroma cacao cDNA

Alignment is on plus strand of genomic sequence and on plus strand of mRNA sequence
mRNA coverage: 100%
Overall percent identity: 81.0%

	Genomic coordinates	mRNA coordinates	length	identity	mismatches
Exon 1	38858-38991	1-134	134	80.6%	26
Exon 2	39076-39277	135-336	202	81.7%	37
Exon 3	39376-39666	337-630	294	80.6%	57

Primers at

Forward: [148 - 169] Exon 2

Reverse: [396 - 417] Exon 3

Figure 26. SPIDEY data for primer 5. Forward primer located in exon two and reverse primer located in exon three.

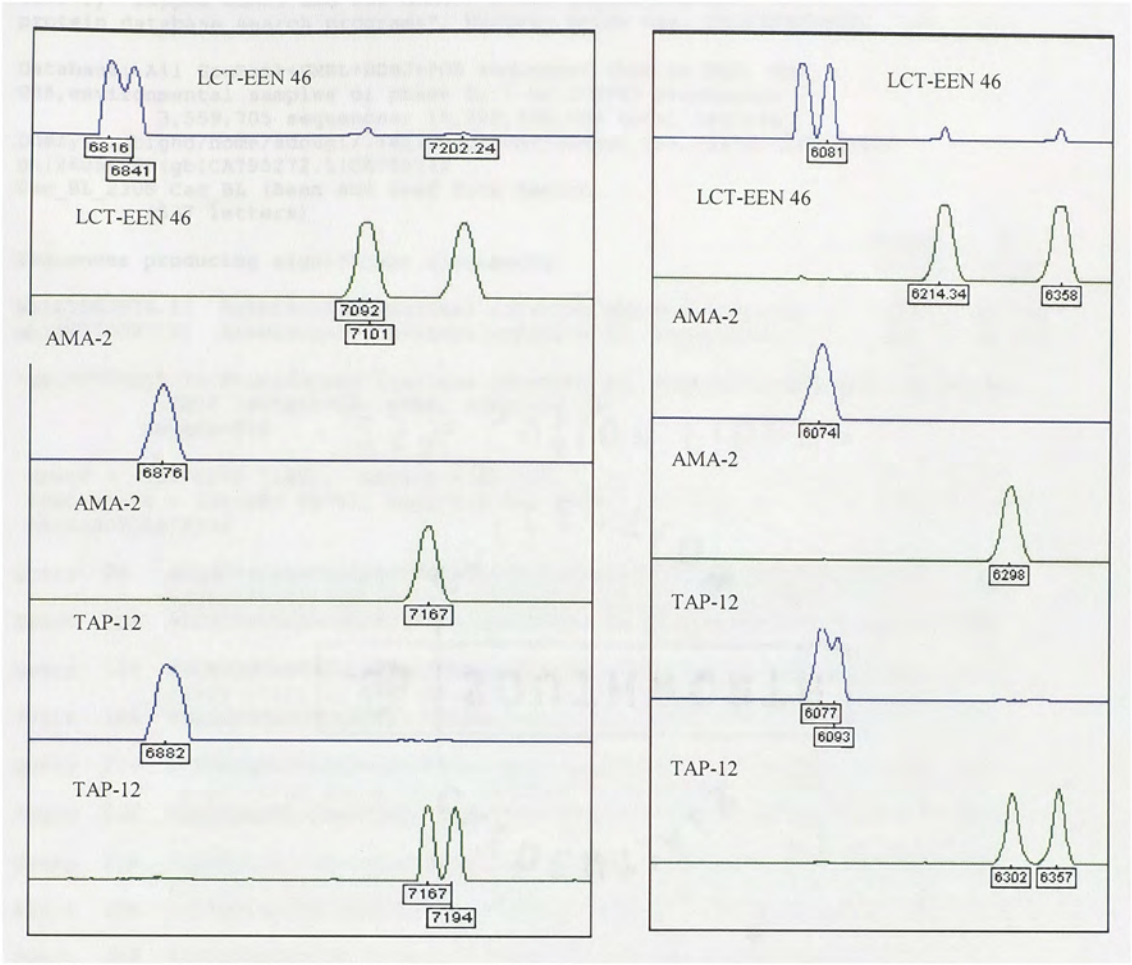

Figure 27. Electropherogram of *T. cacao* for primer 5 at 22˚C and 28˚C respectively. Polymorphism is clear at both temperatures for this primer.

```
BLASTN 2.2.12 [Aug-07-2005]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schdffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           3,559,705 sequences; 15,880,532,508 total letters
Query=  /bighd/home/sdougl/.seqlab-genome1/input_254.rsf{gi-26052348}
gi|26052348|gb|CA795272.1|CA795272
Cac_BL_2308 Cac_BL (Bean and Leaf from Amelon
          (537 letters)
                                                                  Score     E
Sequences producing significant alignments:                       (Bits)  Value

gb|AY063074.1|  Arabidopsis thaliana putative ubiquitin-conjug...   381    2e-102
gb|AF370257.1|  Arabidopsis thaliana putative E2, ubiquitin-co...   381    2e-102

>gb|AF370257.1| Arabidopsis thaliana putative E2, ubiquitin-conjugating enzyme
           UBC8 (At5g41700) mRNA, complete cds
         Length=810

 Score =  381 bits (192),  Expect = 2e-102
 Identities = 335/383 (87%), Gaps = 0/383 (0%)
 Strand=Plus/Plus

Query  94   ATGGCTTCAAAGCGGATCTTGAAGGAGCTCAAGGATCTCCAGAAAGATCCTCCTACCTCT  153
            ||||||||  ||  ||||||||||||||||  ||||||||||||||||  |||||  |||||
Sbjct  126  ATGGCTTCGAAACGGATCTTGAAGGAGCTGAAGGATCTCCAGAAAGACCCTCCAACCTCC  185

Query  154  TGCAGCGCAGGCCCT[GTTGCTGAAGACATGTTTCATTGG]CAAGCAACAATTATGGGTCCT  213
            |||||  |||||  ||  ||||||||||||||||||||||||||||||  ||||||||||||||
Sbjct  186  TGCAGTGCAGGTCCAGTTGCTGAAGACATGTTTCATTGGCAAGCTACAATTATGGGTCCT  245

Query  214  CCTGACAGTCCATATGCCGGTGGAGTGTTTCTAGTCACAATTCATTTCCCTCCGGACTAT  273
             |  ||  |||||  |||  |  ||  ||  ||||||||  ||  ||  |||||  ||||||||||||||
Sbjct  246  GCAGAGAGTCCGTATTCAGGCGGTGTGTTTCTCGTTACCATTCACTTCCCTCCGGACTAT  305

Query  274  CCATTCAAACCACCCAAGGTTGCATTCAGGACAAAGGTCTTTCATCCCAATATTAACAGC  333
            ||||||||||||||  ||||||||||  |||||  |||||  |||||  ||  |||||  ||||||
Sbjct  306  CCATTCAAACCACCAAAGGTTGCATTTAGGACGAAGGTGTTTCACCCTAATATCAACAGC  365

Query  334  AACGGTAGCATTTGCCTCGATATTTTGAAGGAGCAGTGG[AGCCCTGCCCTCACCATA]TCC  393
            |||||  ||||||||||  ||  ||||||||  ||  ||  ||||||||||||||||||||  |||
Sbjct  366  AACGGAAGCATTTGCCTTGACATTTTGAAAGAACAATGGAGCCCTGCCCTCACCATTTCC  425

Query  394  AAGGTGTTGCTCTCAATCTGCTCACTCTTGACGGATCCAAATCCTGATGACCCGTTGNTA  453
            ||||||||||||||  ||  ||  ||  ||  ||  ||||||||||  ||||||||  |||  ||
Sbjct  426  AAGGTGTTGCTCTCGATATGTTCGCTGTTAACAGATCCAAATCCAGATGACCCTTTGGTA  485
```

Figure 28. NETBLAST report showing the highest scoring matches of the *T. cacao* sequence for primer 6 against the non-redundant database at NCBI. The sequence alignment shows the nucleotide match between *T.cacao* and *A. thaliana*. Forward and reverse primers are labeled in bold type.

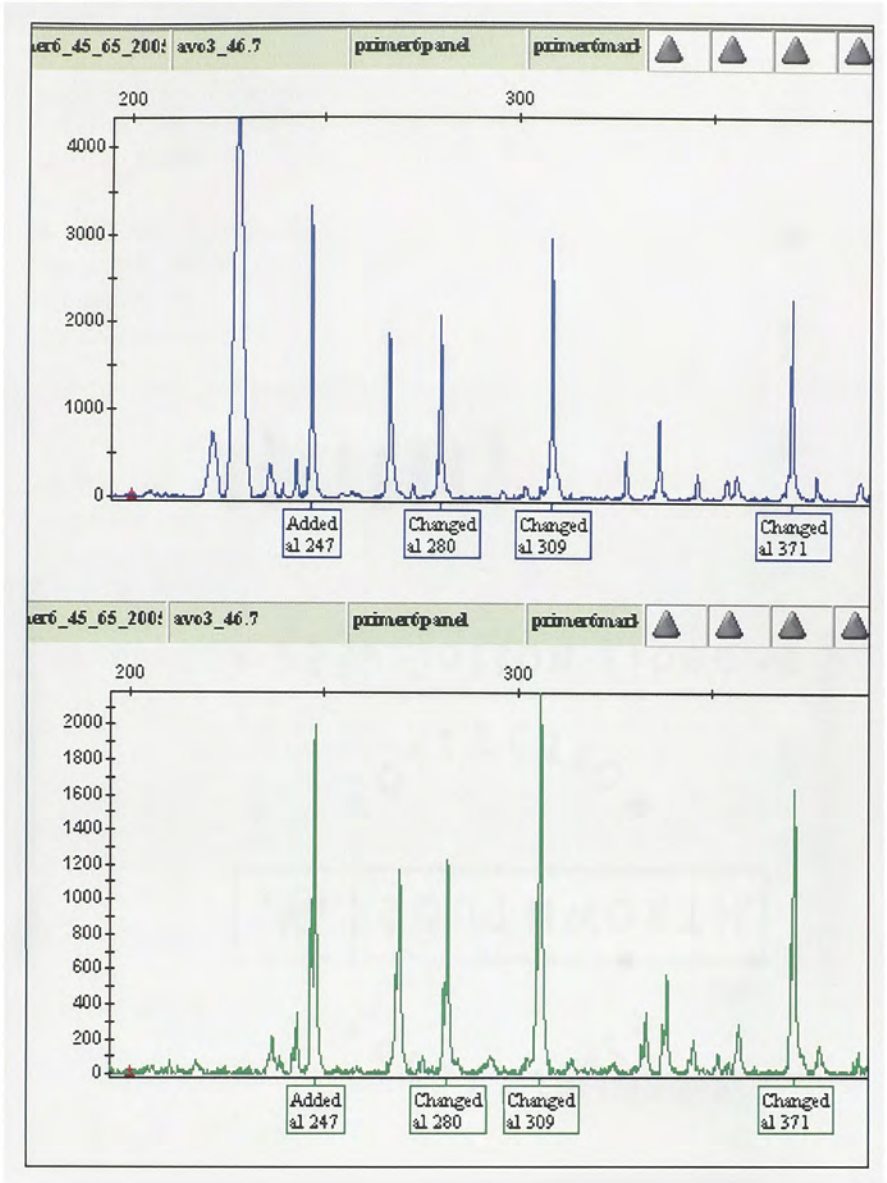

Figure 29. Electropherogram of several amplification products for *Persea americana* for primer 6 at 46.7˚C. Cultivar name shown is, Pa 17 Irwing 78.

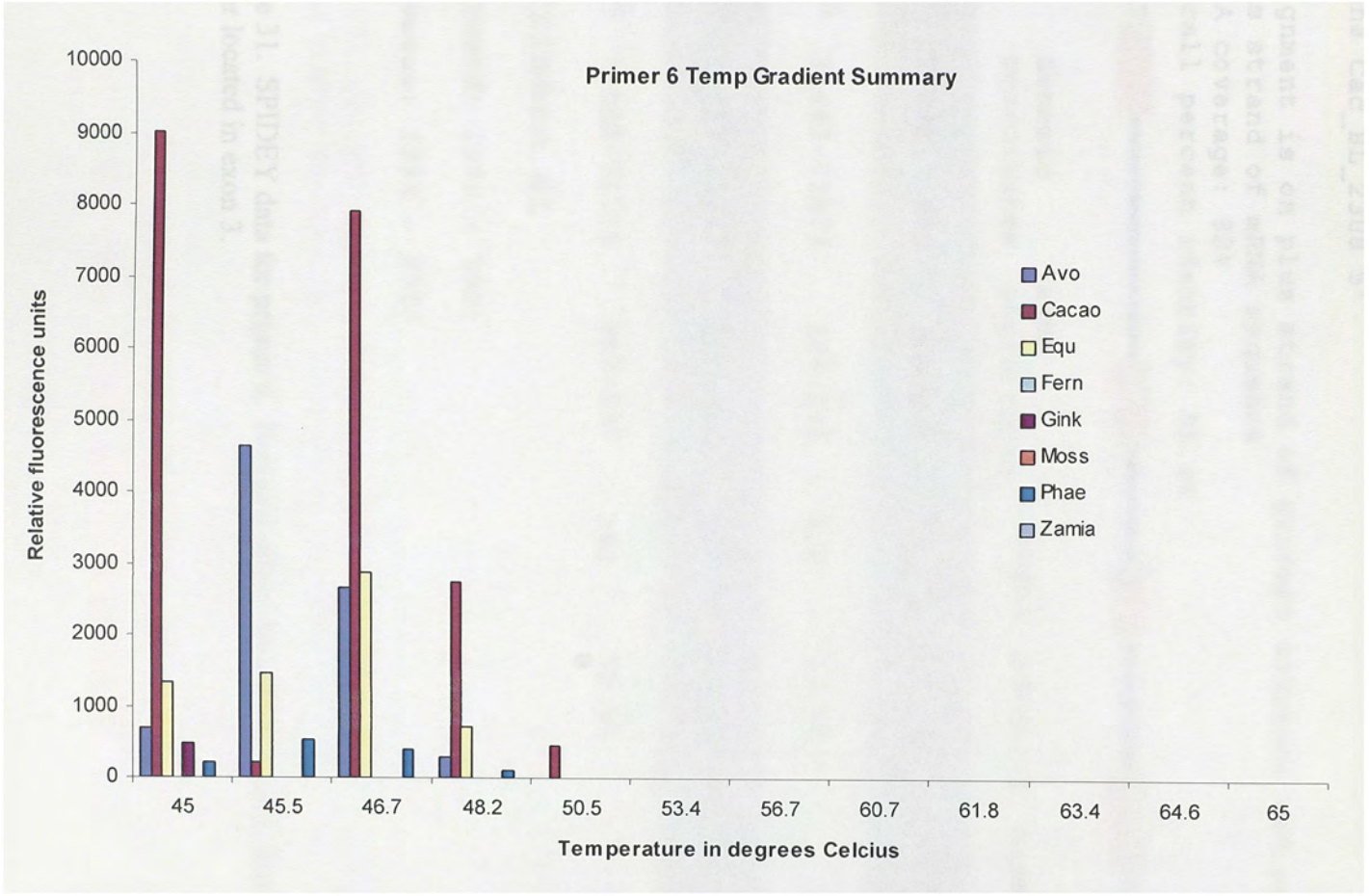

Figure 30. Temperature gradient summary for primer 6.

Genomic: gi|2264305|dbj|AB005233.1| Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MBK23

mRNA: gi|26052348|gb|CA795272.1| Cac_BL_2308 Cac_BL (Bean and Leaf from Amelonardo type Cacao) Theobroma cacao cDNA clone Cac_BL_2308 5'

Alignment is on plus strand of genomic sequence and on plus strand of mRNA sequence
mRNA coverage: 82%
Overall percent identity: 85.4%

	Genomic coordinates	mRNA coordinates	length	identity	mismatches
Exon 1	71027-71096	94-163	70	90.0%	7
Exon 2	71447-71574	164-291	128	87.5%	16
Exon 3	71737-71841	292-396	105	87.6%	13
Exon 4	72035-72175	397-537	141	79.4%	29

Primers at

Forward: [170 - 193]

Reverse: [390 - 373]

Figure 31. SPIDEY data for primer 6. Forward primer located in exon 2, and reverse primer located in exon 3.

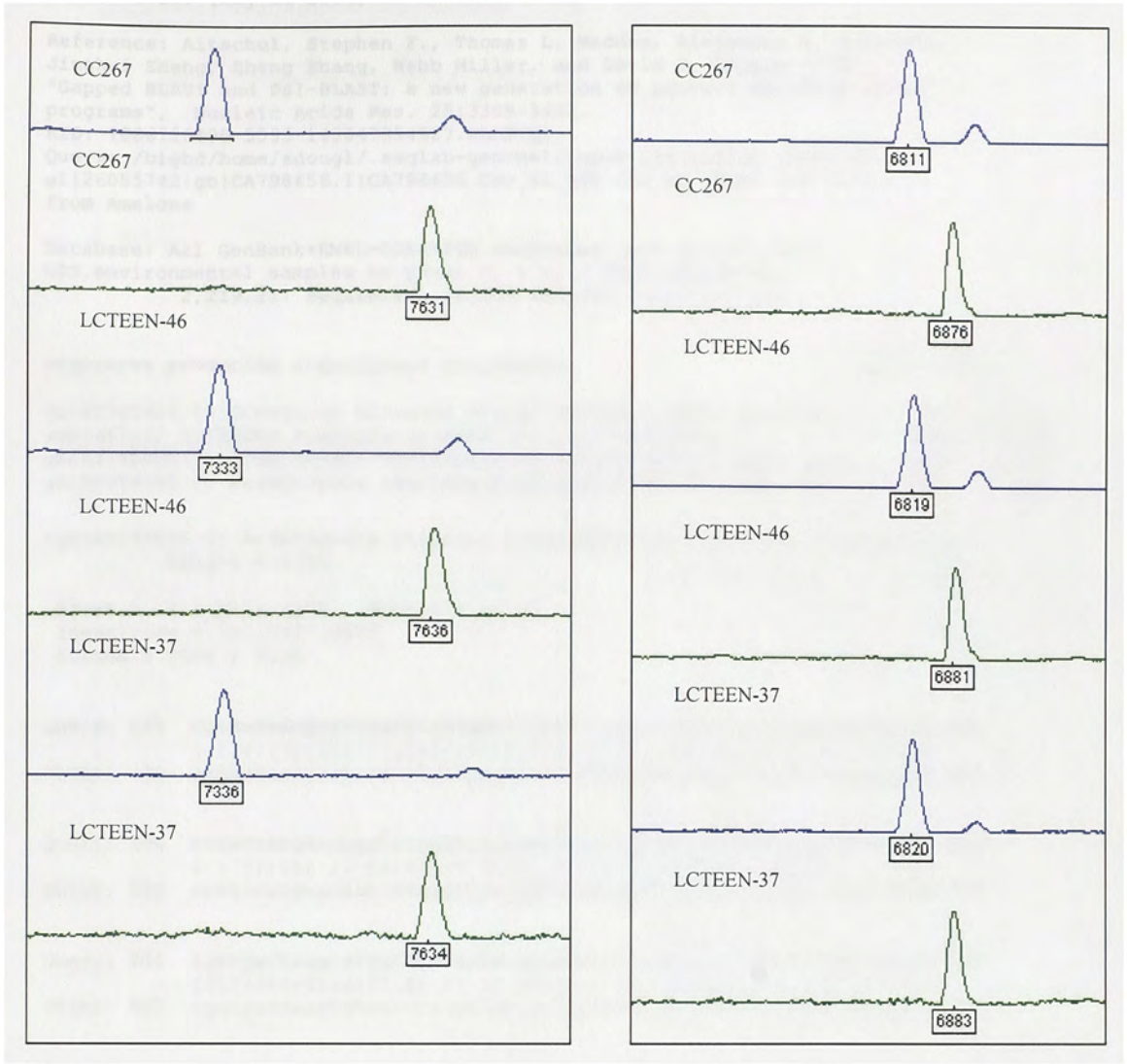

Figure 32. Electropherogram of SSCP data of *T. cacao* for primer 6. Cultivar names located in upper left corner of each electropherogram slide.

```
BLASTN 2.2.9 [May-01-2004]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.
RID: 1086720490-5533-143047954627.BLASTQ4
Query= /bighd/home/sdougl/.seqlab-genome1/input_296.rsf{gi-26055742}
gi|26055742|gb|CA798656.1|CA798656 Cac_BL_995 Cac_BL (Bean and Leaf
from Amelona

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           2,219,217 sequences; 10,709,859,898 total letters

                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

gb|AY345604.1| Gossypium hirsutum alpha-tubulin 2 mRNA, complete...   547   e-153
emb|X67162.1|PAATUB P.amygdalus mRNA for alpha-tubulin               500   e-138
gb|AY142005.1| Arabidopsis thaliana At1g04820/F13M7_26 mRNA, com...   414   e-113
gb|AY058199.1| Arabidopsis thaliana At1g04820/F13M7_26 mRNA, com...   414   e-113

>gb|AY142005.1| Arabidopsis thaliana At1g04820/F13M7_26 mRNA, complete cds
          Length = 1353

 Score =  414 bits (209), Expect = e-113
 Identities = 350/397 (88%)
 Strand = Plus / Plus


Query: 124  cctacccaagaatccacttcatgctttcctcttatgctccagtcatctctgccgagaaag 183
            | ||||||||||||||||||||||||||||| ||||| |||||||||||| || |||||||
Sbjct: 782  catacccaagaatccacttcatgctttcctcctatgccccagtcatctccgcagagaaag 841


Query: 184  cttaccatgaacagctctcagtcgctgaaatcaccaacagtgcctttgagccctcatcta 243
            | | |||||| || |||||||| ||||| ||||| |||||||| |||||||   | || |
Sbjct: 842  ccttccatgagcaactctcagttgctgagatcacaaacagtgcttttgagccagcttcca 901


Query: 244  tgatggctaagtgtgatcctcgccatggaaagtatatggcttgctgtttgatgtaccgtg 303
            ||||||||||||||| || || || |||||||| ||||||||||||||||||||||||
Sbjct: 902  tgatggctaagtgtgacccacgtcacggaaagtacatggcttgctgtttgatgtaccgtg 961


Query: 304  gtgatgttgtacctaaggatgtcaatgctgcagttgctaccatcaagaccaagcgcacca 363
            |||||||||| || |||||||| || || || ||||  |||||||||||||||||| |
Sbjct: 962  gtgatgttgtccccaaggatgtaaacgcagctgttggcaccatcaagaccaagcgcacta 1021


Query: 364  ttcagtttgtggactggtgccaactggattcaagtg 400
            |||||||||| |||||||| || ||||||||||||||
Sbjct: 1022 ttcagtttgttgactggtgtcctactggattcaagtg 1058
```

Figure 33. NETBLAST report showing the highest scoring matches of the *T. cacao* sequence for primer 7 against the non-redundant database at NCBI. The sequence alignment shows the nucleotide match between *T.cacao* and *A. thaliana*. Forward and reverse primers are labeled in bold type.

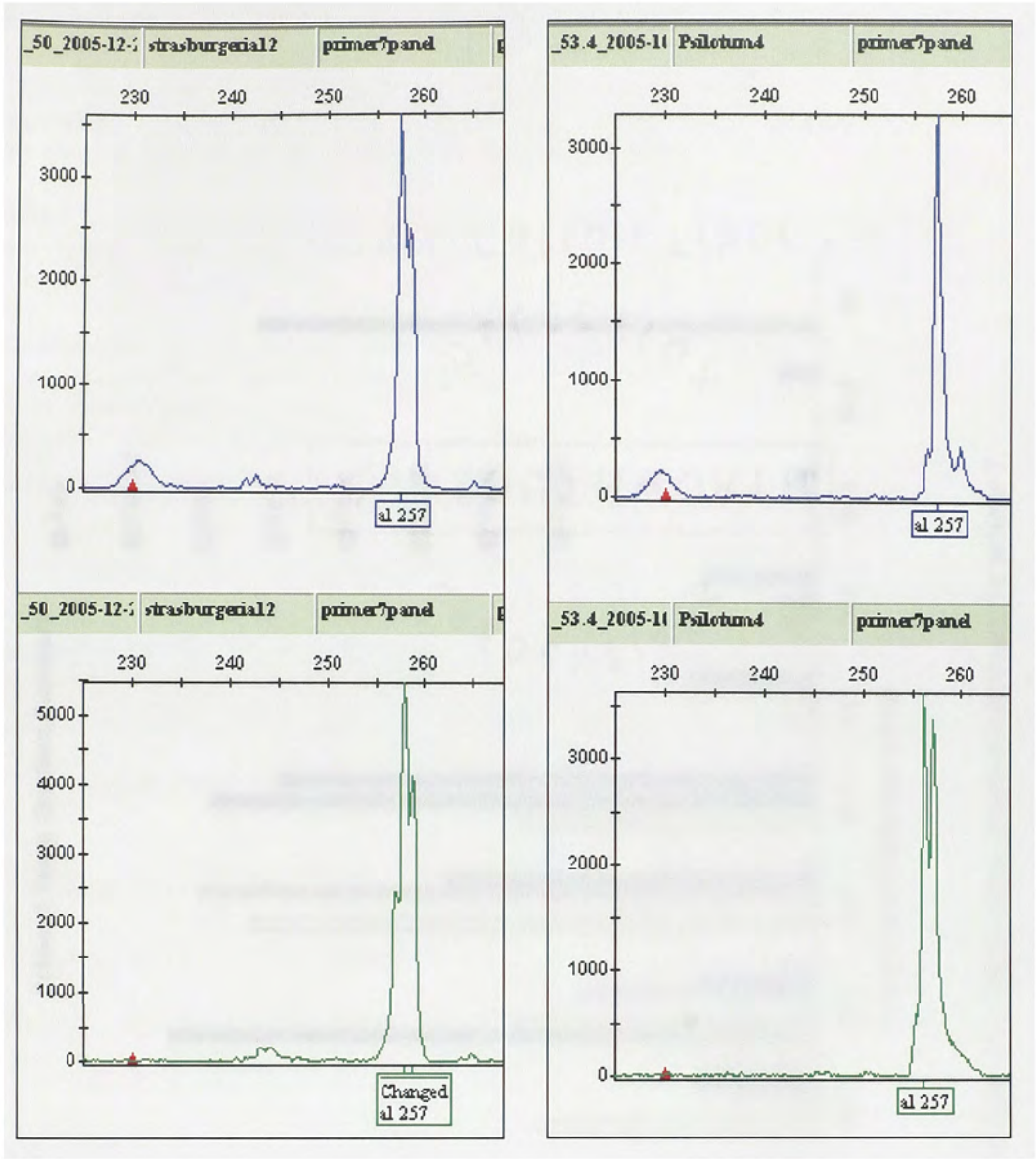

Figure 34. Electropherogram slides of amplification products for *Strasburgeria robusta* and *Psilotum nudum*. Both plant species amplified fragments of 257 base pairs.

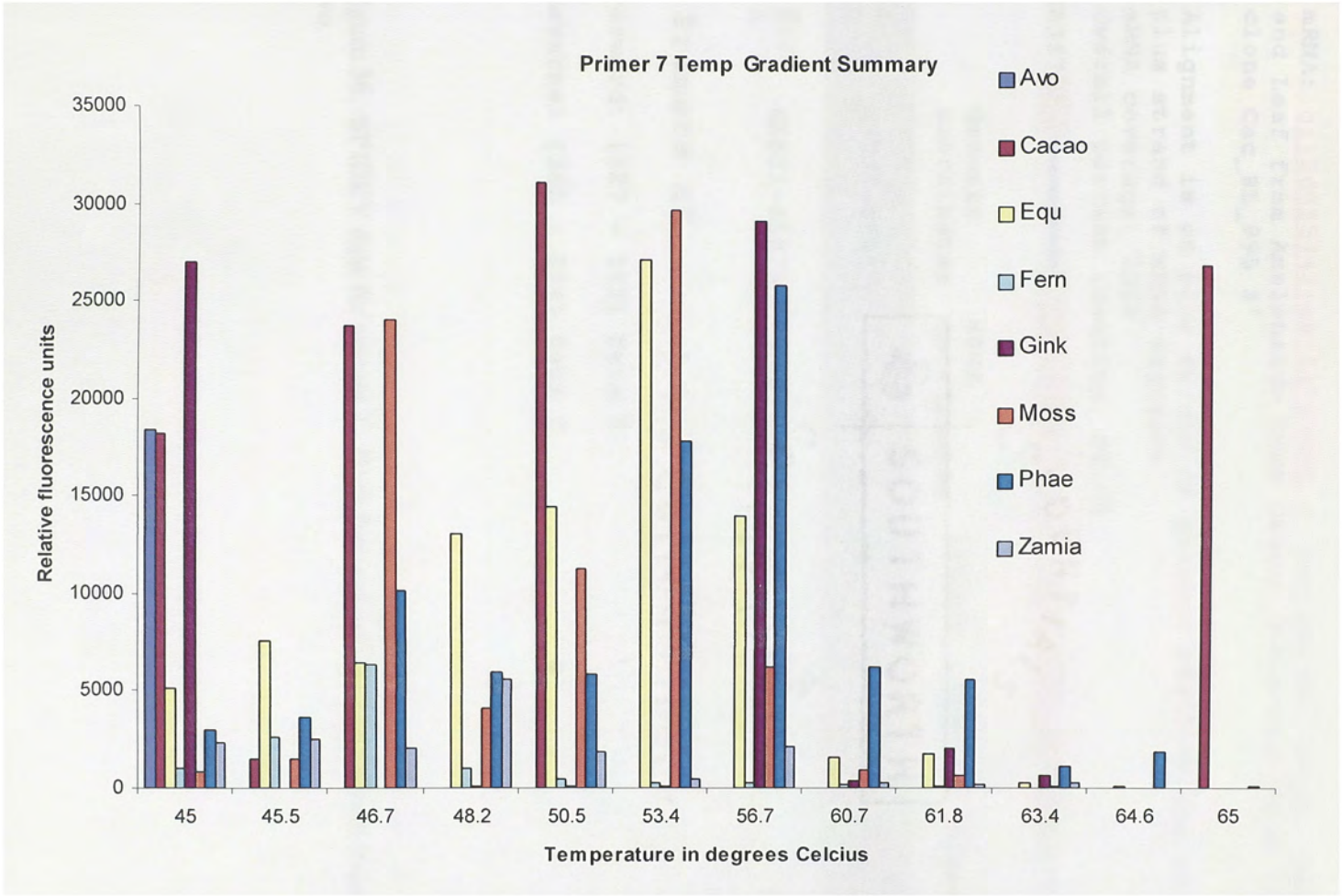

Figure 35. Temperature gradient summary for primer 7.

Genomic: gi|3399678|gb|AC004809.1|F13M7 Arabidopsis thaliana chromosome 1 BAC F13M7 sequence

mRNA: gi|26055742|gb|CA798656.1| Cac_BL_995 Cac_BL (Bean and Leaf from Amelonardo type Cacao) Theobroma cacao cDNA clone Cac_BL_995 5'

Alignment is on plus strand of genomic sequence and on plus strand of mRNA sequence
mRNA coverage: 100%
Overall percent identity: 88.0%

65397  65879

	Genomic coordinates	mRNA coordinates	length	identity	mismatches
Exon 1	65397-65437	1-41	41	87.8%	5
Exon 2	65521-65879	42-400	359	88.0%	43

Primers at

Forward: [127 - 147] Exon 2

Reverse: [365 - 384] Exon 2

Figure 36. SPIDEY data for primer 7. Both forward and reverse primers located in exon two.

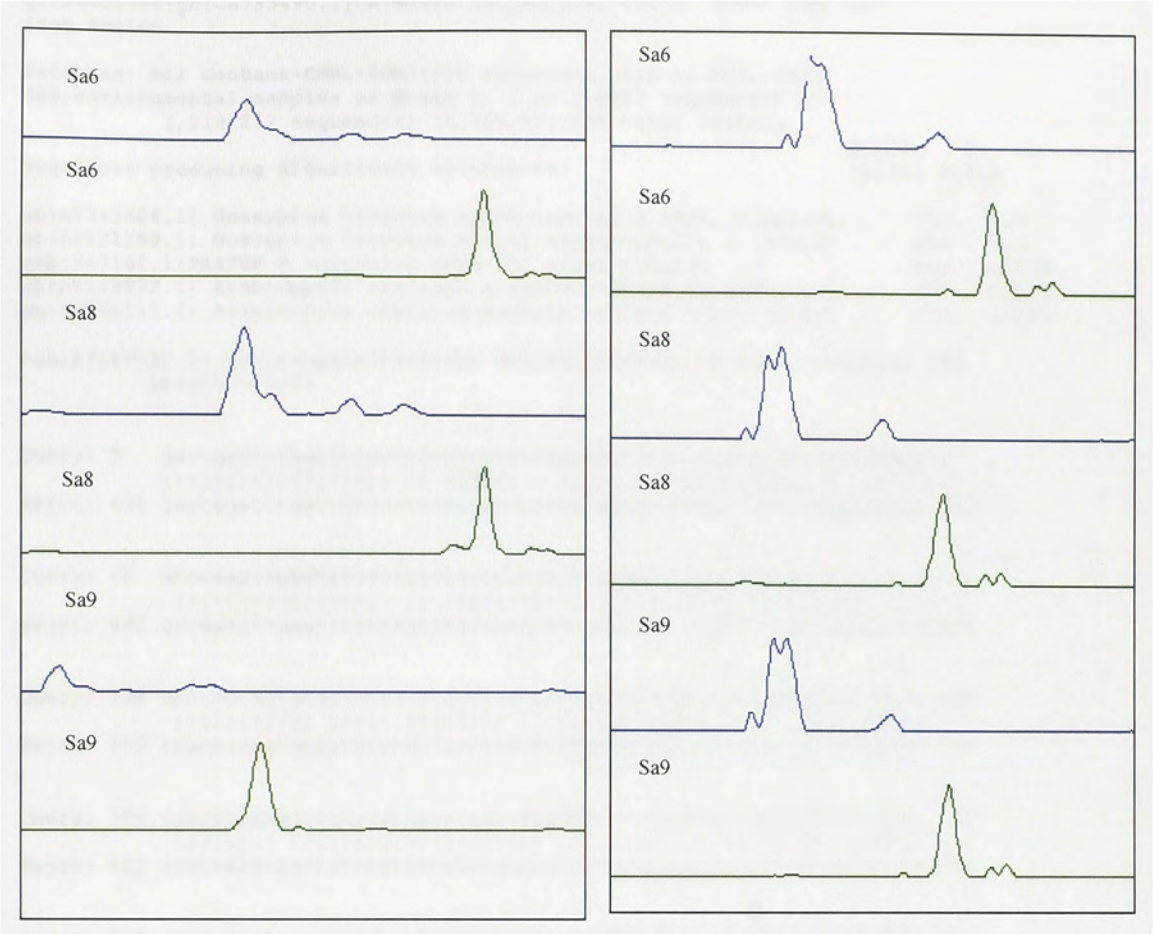

Figure 37. Electropherogram of SSCP data for *Selaginella aramta* var. *etonii* for primer 7 at 22˚C and 28˚C respectively.

```
BLASTN 2.2.12 [Aug-07-2005
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.
Query= /bighd/home/sdougl/.seqlab-genome1/input_124.rsf{gi-26052566}
gi|26052566|gb|CA795490.1|CA795490 Cac_BL_2523 Cac_BL (Bean and Leaf
from Amelon

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS,
GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
           2,219,217 sequences; 10,709,859,898 total letters
                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

gb|AY345604.1| Gossypium hirsutum alpha-tubulin 2 mRNA, complete...   724   0.0
gb|AF521250.1| Gossypium hirsutum Xu-142 alpha-tubulin 2 (ATub2)...   684   0.0
emb|X67162.1|PAATUB P.amygdalus mRNA for alpha-tubulin                504   e-139
gb|AY149922.1| Arabidopsis thaliana At1g50010/F2J10_12 mRNA, com...   450   e-123
gb|AY065117.1| Arabidopsis thaliana Tubulin Alpha-6 Chain (At1g5...   450   e-123

>gb|AY149922.1| Arabidopsis thaliana At1g50010/F2J10_12 mRNA, complete cds
          Length = 1353


Query: 8   gactggatctggtcttggttccctccttttggagcgtttgtctgttgactatggcaagaa 67
            |||||||||||||||||| || |||||  | ||| |  | ||||||||||| || || ||
Sbjct: 432 gactggatctggtcttggatctctcctccttgagagactttctgttgactacgggaaaaa 491


Query: 68  atccaagttgggtttcactgtctacccatctccccaggtctccacgtcagttgttgagcc 127
            |||||||||||||||| || |||||||||| |||||||||| || ||||||||||
Sbjct: 492 gtccaagttgggtttcacagtttacccatctccacaggtctccacctctgttgttgagcc 551


Query: 128 ctacaacagtgttctctcaactcactccctttggaacacactgatgtggctgttctcct 187
            |||||||||| ||||| |||||||| || ||||||||||||||||  || | |||||
Sbjct: 552 ttacaacagtgtcctctccactcactcactcttggaacacactgatgtctctatcctcct 611


Query: 188 tgacaatgaggctatctatgacatctgcaggcgttctcttgacattgagcgacccaccta 247
            |||||||| ||||||||||||||||||||| ||||| ||   ||||||| ||||||||||
Sbjct: 612 cgacaatgaagctatctatgacatctgcagacgttccctaagcattgagagacccaccta 671


Query: 248 cactaaccttaaccgccttgtctctcaggttatttcctccttgactgcctcacttaggtt 307
            ||| ||||| ||||| || |||||||||||||| || |||||||||| || || |||||
Sbjct: 672 caccaacctcaaccgtctcgtctctcaggttatctcatccttgactgcttctctgaggtt 731


Query: 308 tgatggtgccttgaatgtggatgtgactgaattccagaccaacttggtcccctacccaag 367
            |||||||||||||||| |||||||||| ||||| |||||||||||||| ||||||||
Sbjct: 732 cgatggtgccttgaatgttgatgtgactgagttccaaaccaacttggtcccatacccaag 791
```

Figure 38. NETBLAST report showing the highest scoring matches of the *T. cacao* sequence for primer 8 against the non-redundant database at NCBI. The sequence alignment shows the nucleotide match between *T.cacao* and *A. thaliana*. Forward and reverse primers are labeled in bold type.

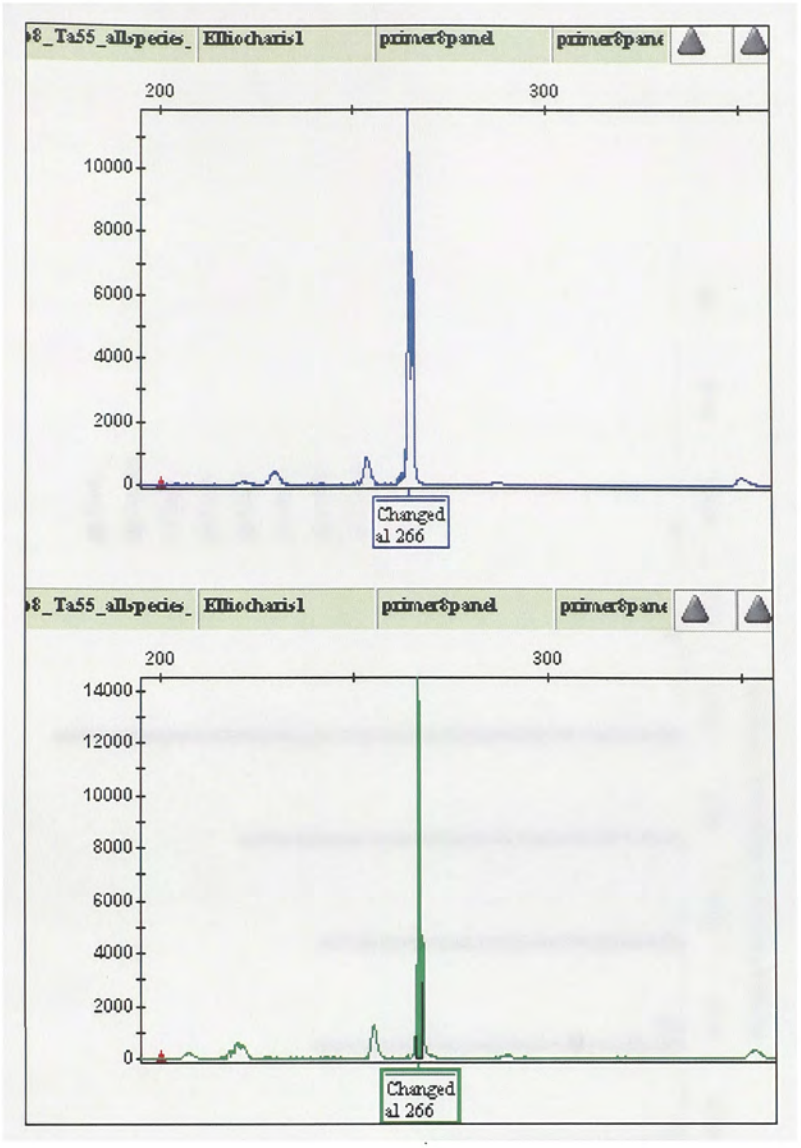

Figure 39. Electropherogram for amplification of *Eleocharis cellulosa* for primer 8. Product size is 266 base pairs.

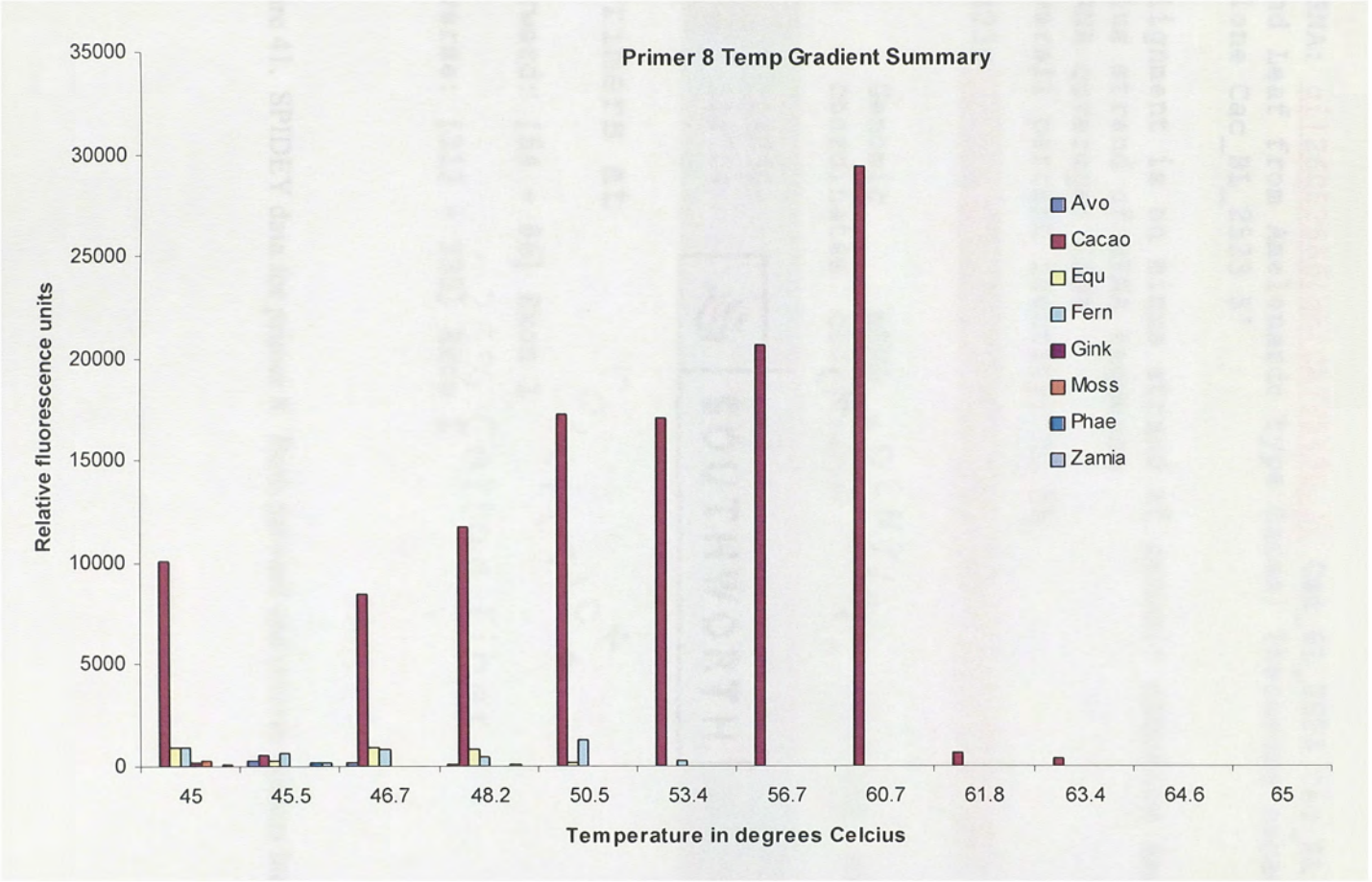

Figure 40. Temperature gradient summary for primer 8.

Genomic: gi|5302774|emb|Z97337.2|ATFCA2 Arabidopsis thaliana DNA chromosome 4, ESSA I FCA contig fragment No. 2

mRNA: gi|26052566|gb|CA795490.1| Cac_BL_2523 Cac_BL (Bean and Leaf from Amelonardo type Cacao) Theobroma cacao cDNA clone Cac_BL_2523 5'

Alignment is on minus strand of genomic sequence and on plus strand of mRNA sequence
mRNA coverage: 100%
Overall percent identity: 83.5%

	Genomic coordinates	mRNA coordinates	length	identity	mismatches
Exon 1	114230-114764	1-538	538	83.5%	89

Primers at

Forward: [64 - 86] Exon 1

Reverse: [312 - 330] Exon 1

Figure 41. SPIDEY data for primer 8. Both forward and reverse primers located in exon one.

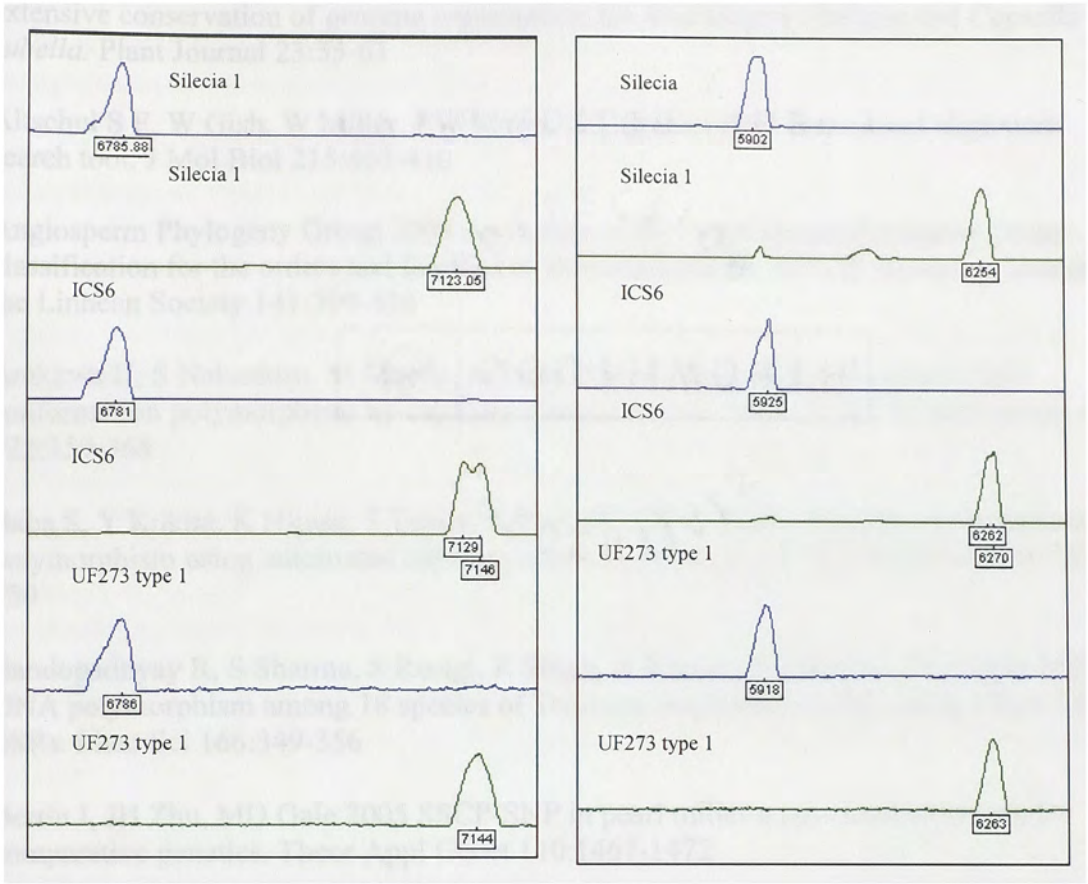

Figure 42. Electropherogram of SSCP data of *T. cacao* for primer 8 at 22˚C and 28˚C respectively. Cultivar names listed in upper left corner of each slide.

LIST OF REFERENCES

Acarkan A,M Roβberg, M Koch, R Schmidt, 2000 Comparative genome analysis reveals extensive conservation of genome organisation for *Arabidopsis thaliana* and *Capsella rubella.* Plant Journal 23:55-63

Altschul S F, W Gish, W Miller, EW Myers, DJ Lipman 1990 Basic local alignment search tool. J Mol Biol 215:403-410

Angiosperm Phylogeny Group 2003 An undate of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Botanical Journal of the Linnean Society 141:399-436

Arakawa H, S Nakashiro, M Maeda, A Tsuji 1996 Analysis of single-strand DNA conformation polymorphism by capillary electrophoresis. Journal of Chromotograpy A 722:359-368

Baba S, Y Kukita, K Hijasa, T Tahira, K Hayashi 2003 Single-stranded conformational polymorphism using automated capillary array electrophoresis. BioTechniques 34:746-750

Bandopadhyay R, S Sharma, S Rustgi, R Singh, A Kumar, HS Balyan, PK Gupta 2003 DNA polymorphism among 18 species of Triticum-Aegilops complex using wheat ESTs-SSRs. Plant Sci 166:349-356

Bertin I, JH Zhu, MD Gale 2005 SSCP-SNP in pearl millet-a new marker system for comparative genetics. Theor Appl Genet 110:1467-1472

Borrone J, D Kuhn, R Schnell 2004 Isolation, characterization, and development of WRKY genes as useful genetic markers in *Theobroma cacao*. Theor Appl Genet 109:495-507

Bowe ML, G Coat, CW dePamphilis 2000 Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc Natl Aca Sci 97:4092-4097

Brenner ED, MS Katart, DW Stevenson, SA Rudd, AW Douglas, WN Moss, RW Twigg, SJ Runko, GM Stellari, WR MnCombie, GM Coruzzi 2005 EST analysis in *Ginko biloba*: an assessment of conserved developmental regulators and gymnosperm specific genes. BMC Genomics 6:143-153

Bryja J, M Galan, N CHarbonnel, JF Cosson 2005 Analysis of major histocompatibility complex class II gene in water voles using capillary electrophoresis-single stranded conformation polymorphism. Molecular Ecology Notes 5:173-176

Bundock PC, MJ Cross, FM Shapter, RJ Henry 2006 Robust allele-specific polymerase chain reaction markers developed for single nucleotide polymorphisms in expressed barley sequences. Theor Appl Genet 112:358-365

Canavas JM, A Morse, B West 2004 PCR primer selection tool optimized for high-thoughput proteomics and structural genomics. BioTechniques 36:1040-1042

Chaw S M, CL Parkinson, Y Cheng, TM Vincent, JD Palmer 2000 Seed plant phylogeny inferred from all three plant genomes: Monophyly of extant gymnosperms and origin of Gnetales from conifers. Proceedings of the National Academy of Science of the United States of America 97:4086-4091

Chen X, YG Cho, SR McCouch 2002 Sequence divergence of rice microsatellites in Oryza and other plant species. Mol Genet Genomics 268:331-343

Cordeiro GM, R Casu , CL McIntyre, JM Manners, RJ Henry 2001 Microsatellite markers from sugarcane (*Saccharum spp.*) ESTs cross transferable to erianthus and sorghum. Plant Sci 160:1115-1123

Decroocq V, MG Fave, L Hagen, L Bordenave, S Decroocq 2003 Development and transferability of apricot and grape EST microsatellite markers across taxa. Theor Appl Genet 106:912-922

Devos KM, MD Gale 1997 Comparative genetics in grasses. Plant Molecular Biology 35:3-15

Dirlewanger E, E Graziano, T Joobeur, F Garriga-Caldere, P Cosson, W Howad, P Arus 2004 Comparative mapping and marker-assisted selection in Rosaceae fruit crops. Proc Natl Acad Sci U S A 101:9891-9896

Doyle JA 1998 Molecules, morphology, fossils and the relationship of angiosperms and Gnetales. Molecular Phylogenetics and Evolution 9:448-462

Erdemoglu N, B Sener 2000 Taxol sources. Fabad Journal of Pharmaceutical Sciences 25:67-74

Fazio G, SM Chung, JE Staub, 2003 Comparative analysis of response to phenotypic and marker-assisted selection for multiple lateral branching in cucumber (*Cucumis sativus* L.). Theor Appl Genet 107:875-883

Fraser LG, MA McNeilage, GK Tsang, CF Harvey, HN De Silva 2005 Cross-species amplification of microsatellite loci within the dioecious polyploidy genus *Actinidia* (Actinidiaceae). Theor Appl Genet 112:149-157

Gallego JF, MA Perez, Y Nunez, P Hidalgo 2005 Comparison of RAPDs, AFLPs and SSR markers for the genetic analysis of yeast strains of *Saccharomyces cerevisiae.* Food Microbiology 22:561-568

Garcia AAF, LL Benchimol, AMM Barbosa, IO Geraldi, CL Souza, AP de Souza 2004 Comparison of RAPD, RFLP,AFLP, and SSR markers for diversity studies in tropical maize inbred lines. Genetics and Molecular Biology 27:579-588

Glaszmann JC, P Dufour, L Grivet, A D'Hont, M Deu, F Paulet, P Hamon 1997 Comparative genome analysis between several tropical grasses. Euphytica 96:13-21

Goremykin, VV, FH Hellwig 2005 Evidence for the most basal split in land plants dividing bryophyte and tracheophyte lineages. Pl. Syst. Evol. 254:93-103

Grant D, P Cregan, RC Shoemaker 2000 Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. Proc Natl Acad Sci 97:4168-4173

Gualtieri G, O Kulikova, E Limpens, D-J Kim, DR Cook, T Bisseling, R Geurts 2002 Microsynteny between pea and *Medicago truncatula* in the SYM2 region. Plant Molecular Biology 50:225-235

Gupta PK, RK Varshney 2000 The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. Euphytica 113:163-185

Harry DE, B Temesgen, DB Neale 1998 Codominant PCR-based markers for *Pinus taeda* developed form mapped cDNA clones. Theo Appl Genet 97:327-336

Hartl DL, EW Jones 2002 Essential genetics a genomics perspective. Jones and Bartlett, Sudbury Massachusetts.

Jordanova A, L Kalaydijeva, A Savov, M Clautre, M Schwarz, X Estivill, D Angelicheva, A Haworth, T Casals, I Kremensky 1997 SSCP analysis: A blind sensitivity trial. Human Mutation 10:65-70

Judd W,CS Campbell, EA Kellogg, PF Stevens, MJ Donoghue 2002 Plant systematics a phylogenetic approach. Sinauer Associates Inc, Sunderland Massachusetts.

Karol, KG, RM McCourt, MC Cimino, CF Delwiche 2001 The closest living relatives of land plants. Science 294:2351-2353

Kato, M., H. Akiyama 2005 Interpolation hypothesis for origin of the vegetative sporophyte of land plants. Taxon 54:443-450

Kenrick, P, PR Crane 1997 The origin and early evolution of plants on land. Nature 389:33-39

Kirk BW, M Feinsod, R Favis, RM Kliman, F Barany 2002 Single nucleotide polymorphism seeking long term association with complex disease. Nucleic Acids research 30:3295-3311

Kong P, C Hong, PA Richardson, ME Gallegly 2003 Single-strand-conformational polymorphism of ribosomal DNA for rapid species determination in genus *Phytophthora*. Fungal Genetics and Biology 39:238-249

Kress JW, KJ Wurdack, EA Zimmer, LA Weigt, DH Janzen 2005 Use of DNA barcodes to identify flowering plants. Proc Natl Acac Sci 102:8369-8374

Kuhn DN, M Heath , RJ Wisser, A Meerow A, JS Brown, U Lopes, RJ Schnell 2003 Resistance gene homologues in *Theobroma cacao* as useful genetic markers. Theor Appl Genet 107:191-202

Kuhn, DN., J Borrone, AW Meerow, JC Motamayor, JS Brown, RJ Schnell 2005 Single-strand conformational polymorphism analysis of candidate genes for reliable identificaton of alleles by capillary array electrophoresis. Electrophoresis 26:112-125

Kwok P-Y 2001 methods for genotyping single nucleotide polymorphisms. Annual Review of Genomics and Human Genetics 2:235-258

Lande R, R Thompson 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743-756

Liewlaksaneeyanawin C, CE Ritland, YA El-Kassaby, K Ritland 2004 Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theor Appl Genet 109:361-369

McCallum J, D Leite, M Pither-Joyce, MJ Havey 2001 Expressed sequence markers for genetic analysis of bulb onion. Theor Appl Genet 103:979-991

Milbourne D, Meyer R, Bradshaw JE, Baird E, Bonar N, Provan J, Powell W, Waugh R 1997 Comparison of PCR-based marker systems for the analysis of genetic relationships in cultivated potato Molecular Breeding 3:127-136

Mishler, BD 2000 Deep phylogenetic relationships among 'plants' and their implications for classification. Taxon 49:661-683

Mohan M, S Nair, A Bhagwat, TG Krishna, M Yano, CR Bhatia, T Sasaki 1997 Genome mapping, molecular markers and marker-assisted selection in crop plants. Mol Breed 3:87-103

Mukhtar MS, M Rahman, Y Zafar 2002 Assesment of genetic diversity among wheat (*Triticum aestivum* L.) cultivars from a range of localities across Pakistan using random amplified polymorphic DNA (RAPD) analysis. Euphytica 128:417-425

Mullis K, F Faloona, S Scharf 1986 Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. Cold Spring Harbor Symposia on Quantitative Biology 51:263-273

Mullis KB 1990 The unusual origin of the polymerase chain reaction. Scientific American 262:56-65

Murray CG, TP Larsson, T Hill, R Bjorklind, R Fredriksson, HB Schioth 2005 Evaluation of EST-data using the genome assembly. Biochemical and Biophysical Research Communications 331:1566-1576

Nair S, JS Bentur, RU Prasada, M Mohan 1995 DNA markers tightly linked to a gall midge resistant gene (*Gm2*) are potentially useful for marker-aided selection in rice breeding. Theor Appl Genet 91:68-73

Orita M, Iwahana, H Kanazawa, K Hayashi, T Sekiya 1989 Detection of polymorphism of human DNA by gel electrophoresis as single-strand conformational polymorphism. Proc Natl Acac Sci 86:2766-2770

Orti G, MP Hare, JC Avise 1997 Detection and isolation of nuclear haplotypes by PCR-SSCP. Molecular Ecology 6:575-580

Pappas GJ Jr, K Benabdellah, B Zingales, A Gonzalez 2005 Expressed sequence tags from the plant trypanosomatid *Phytomonas serpens*. Molecular & Biochemical Parasitology 142:149-157

Paterson AH, JE Bowers, MD Burow, X Drayne, CG Elsik, C-X Jiang, CS Katsar, T-H Lan, Y-R Lin, R Ming, RJ Wright 2000 Comparative genomics of plant chromosomes. The Plant Cell 12:1523-1539

Plomion C, P Hurme, J-M Frigerio, M Ridolfi, D Pot, C Pinneau, C Avila, F Gallardo, H David, G Neutelings, M Campbell, F M Canovas, O Savolainen, C Bodenes, A Kremer 1999 Developing SSCP markers in two *Pinus* species. Molecular Breeding 5:21-31

Pryer KM, H Schneider, AR Smith, R Cranfill, PG Wolf, JS Hunt, SD Snipes 2001 Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. Nature 409:618-621

Qiu Y-L, J Lee, F Berasconi-Quadroni, DE Soltis, PS Soltis, M Zanis, EA Zimmer, Z Chen, V Savolainen, M Chase 1999 The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402:404-407

Rafalski A 2002 Applications of single nucleotide polymorphisms in crop genetics. Curr Opin in Plant Biol 5:94-100

Saha MC, MA Mian, I Eujayl, JC Zwonitzer, L Wang, GD May 2004 Tall fescue EST-SSR markers with transferability across several grass species. Theor Appl Genet 109:783-791

Salmaso M, G Faes, C Segala, M Stefanini, I Salakhutdinov, E Zyprian, R Toepfer, MS Grando, R Velasco 2004 Genome diversity and gene haplotypes in the grapevine (*Vitis vinifera* L.), as revealed by single nucleotide polymorphisms. Mol Breed 14:385-395

Salse J, B Piegu, R Cooke, M Delseny 2002 Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. Nucleic Acids Research 11:2316-2328

Sato Y, T Nishio 2003 Mutation in rice waxy mutants by PCR-RF-SSCP. Theor Appl Genet 107:1282-1284

Schmidt R 2000 Synteny: recent advances and future prospects. Current Opinion in Plant Biology 3:97-102

Slabaugh MB, GM Huestis, J Leonard, JL Holloway, C Rosato, V Hongtrakul, N Martini, R Toepfer, J Schnell, SJ Knapp 1997 Sequence-based genetic markers for genes and gene families: single-strand conformational polymorphism for the fatty acid synthesis genes of *Cuphea*. Theor Appl Genet 94:400-408

Soltis PS, DE Soltis, MJ Zanis, S Kim 2000 Basal lineages of angiosperms: Relationships and implications for floral evolution. International Journal of Plant Sciences 161:S97-S107

Soltis DE, PS Soltis, MJ Zanis 2002 Phylogeny of seed plants based on evidence from eight genes. American Journal of Botany 89:1670-1681

Steele KA, AH Price, HE Shashidhar, JR Witcombe 2006 Marker-assisted selection to introgress rice QTLs controlling root traits into an Indian upland rice variety. Theor Appl Genet 112:208-221

Stern KR 2000 Introductory Plant Biology. McGraw-Hill Companies Inc.,California State University-Chico

Sunnucks P, ACC Wilson, B Beheregaray, K Zenger, J French, AC Taylor 2000 SSCP is not so difficult: the application and utility of single-stranded conformational polymorphism in evolutionary biology and molecular ecology. Molecular Ecology 9:1699-1710

Torjek O, D Berger, RC Meyer, C Mussig, KJ Schmidt, T Rosleff Sorensen, B Weisshaar, T Mitchell-Olds, T Altmann 2003 Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis*

Vidal JR, P Delavault, M Coarer, A Defontaine 2000 Design of grapevine (*Vitis vinifera* L.) cultivar-specific SCAR primers for PCR fingerprinting. Theor Appl Genet 101:1194-1201

Vos P, R Hogers, M Bleeker, M Reijans, T Van de Lee, M Hornes, A Frijters, J Pot , J Peleman, M Kuiper, M Zabeau 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acid Research **23**: 4407-4414

Wolf PG, KG Karol, DF Mandoli, J Kuehl, K Arumuganathan, MW Ellis, BD Mishler, DG Kelch, RD Olmstead, JL Boore 2005 The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). Gene 350:117-128

Yan HH, J Mudge, D-J Kim, RC Shoemaker, SR Cook ND Young 2004 Comparative physical mapping reveals features of microsynteny between *Glycine max*, *Medicago truncatula*, and *Arabidopsis thaliana*. Genome 47:141-155

Yu JK, M la Rota, RV Kantety, ME Sorrells 2004 EST derived SSR markers for comparative mapping in wheat and rice. Mol Gen Genomics 271: 742-751

Zimmer EA, Y-L Qiu, PK Endress, EM Friis 2000 Current Perspectives on Basal Angiosperms: Introduction. Int J.Plant Sci 161:S1-S2