2006-07-20

# Identification, Sequencing, Expression and Evolutionary Relationships of the 11S Seed Storage Protein Gene in Chenopodium quinoa Willd.

Marie Renee Barrett Balzotti
*Brigham Young University - Provo*

# IDENTIFICATION, SEQUENCING, EXPRESSION AND EVOLUTIONARY RELATIONSHIPS OF THE 11S SEED STORAGE PROTEIN GENE IN *CHENOPODIUM QUINOA* WILLD.

by

Marie Renee Barrett Balzotti

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Plant and Animal Sciences

Brigham Young University

August 2006

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Marie R. B. Balzotti

This thesis has been read by each member of the following graduate committee and by majority vote has been found satisfactory.

| | |
|---|---|
| _____ | _____ |
| Date | Craig E. Coleman, Chair |
| _____ | _____ |
| Date | P. Jeff Maughan |
| _____ | _____ |
| Date | David A. McClellan |

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Marie R. B. Balzotti in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative material including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

| | |
|---|---|
| _____ | _____ |
| Date | Craig E. Coleman<br>Chair, Graduate Committee |

Accepted for the Department

| | |
|---|---|
| _____ | _____ |
| Date | Von D. Jolley<br>Graduate Coordinator |

Accepted for the College

| | |
|---|---|
| _____ | _____ |
| Date | Rodney J. Brown<br>Dean, Biology and Agriculture |

ABSTRACT

IDENTIFICATION, SEQUENCING, EXPRESSION AND EVOLUTIONARY RELATIONSHIPS OF THE 11S SEED STORAGE PROTEIN GENE IN *CHENOPODIUM QUINOA* WILLD.

Marie R.B. Balzotti

Department of Plant and Animal Sciences

Master of Science

Quinoa (*Chenopodium quinoa*) is an Andean crop adapted to harsh environmental conditions and containing a high and well-balanced protein composition. Two seed storage proteins, the 2S albumin and 11S globulin, are the major amino acid reservoirs for the developing seedling. An in-depth study of the genes encoding these proteins is necessary to understand the source of quinoa's protein quality. Our objectives include identification and sequencing of the 11S genomic and cDNA clones, analysis of 11S expression profiles in different quinoa accessions and evaluation of evolutionary relationships between the sequence of the 11S gene in quinoa and homologous genes in other species.

Clones containing 11S cDNA and genomic DNA were identified and sequenced. Two copies of the gene were found to be present at two different loci in the quinoa genome. The amino acid composition of the 11S gene was also analyzed. Results show that the 11S gene contains a well-balanced assortment of essential amino acids with

relatively high levels of glutamate/glutamine, aspartate/asparagine, arginine, serine, leucine and glycine, typical of other 11S seed storage proteins.

Total RNA and globulin was extracted from seed collected at different developmental intervals from nine quinoa accessions. Expression profiles were determined by evaluating 11S transcript levels using relative quantification real time RT-PCR and comparing relative 11S globulin accumulation using sodium SDS-PAGE. The 11S gene was found to be expressed during late-maturation regardless of differences in maturation rate.

A portion of the amino acid sequence of the 11S and homologous genes was chosen for phylogenetic analysis. Fifty five such sequences from 50 different plant species were assembled and aligned. Two phylogenetic reconstructions, one using the parsimony method and the other using Bayesian analysis, were generated in order to analyze evolutionary relationships between the 11S gene in quinoa and homologous genes in other species. Relationships shown by both reconstructions for sequences from closely-related species were consistent with taxonomic clustering. Both reconstructions showed less resolution involving relationships between distantly-related angiosperm taxa indicating a wide divergence of sequence at the angiosperm level and a need for additional angiosperm sequence data for finer resolution.

# ACKNOWLEDGEMENTS

I was extremely fortunate to have such a patient, supportive and talented graduate committee. I would like to thank Dr. Craig Coleman, my advisor, for his guidance and expertise throughout this project and for providing me with so many valuable learning opportunities. I enjoyed working with Dr. Jeff Maughan as a teaching assistant and am so grateful for his encouragement and mentoring. I would like to thank Dr. David McClellan for being extremely accommodating and patient while teaching me phylogenetic principles and techniques. I am also grateful to Drs. Eric Jellen, Mikel Stevens, Von Jolley and Daniel Fairbanks for sharing their knowledge and providing useful counsel, Jenny Thornton for her assistance and time and Dorine Jesperson for her friendship and kindness. I am especially grateful to my husband, Chris, for his love, time and support and for all the encouragement and advice I received from my family and friends.

## TABLE OF CONTENTS

**Appendix**

# Isolation of BAC clones containing the 11S gene from a BAC library for *Chenopodium quinoa* Willd.


## ABSTRACT

Quinoa is a crop from the Andean region of South America. Its seed has an unusually well-balanced amino acid composition and the plant is well-adapted to harsh environmental conditions including frost, drought and saline soils. The objective of this study was to identify BAC clones containing the 11S seed storage protein, one of two major storage proteins found in quinoa seed using a quinoa BAC library. Seven positive clones were identified from the library using an 11S cDNA probe. Southern blotting was performed using BAC DNA and genomic DNA from four accessions, all digested with *EcoRI*. Two copies of the 11S gene were identified which represent two distinct loci within the quinoa genome.


## INTRODUCTION

Quinoa (*Chenopodium quinoa)* is a putative allotetraploid ($2n = 4x = 36$) and a member of the Amaranthaceae (or Chenopodiaceae) family. Other members of this family having some commercial importance are amaranth (*Amaranthus* spp.), spinach (*Spinacea oleracea* L.) and sugar beet (*Beta vulgaris* L.). Quinoa inhabits a variety of environments from the high Andean region to valleys and coastal areas and was an important crop of the Incan empire (Galwey, 1995). After the Spanish conquest, quinoa cultivation was discouraged due to its affiliation with cultural and religious practices and

replaced by introduced species such as wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.). Quinoa cultivation was preserved throughout the years only by poor or isolated indigenous communities and small farms where it was able to maintain high genetic diversity (Castillo, 1995). Recently, there has been a revival of quinoa production and renewed interest has spread outside the Andean region. Commercially grown quinoa seed has been successfully promoted not only in South America but also in Europe, Australia, North America and Japan. Quinoa's ability to adapt to drought, frost and alkaline or salty soil conditions as well as its high protein content and favorable assortment of amino acids has made it an attractive candidate for crop improvement and attracted interest in the fields of genetic and molecular research.

There have been several studies involved in the characterization of quinoa's seed storage proteins. Quinoa seed proteins were electrophoretically characterized in order to create genetic markers for identification and classification of quinoa accessions as well as a resource for phylogenetic and genetic variability studies (Fairbanks et al., 1990). An analysis of quinoa seed fractions including protein analysis was conducted by Chauhan et al. (1992). Brinegar and Goundan (1993) specifically characterized individual seed storage proteins with the isolation and characterization of the 11S seed storage protein, which they call chenopodin. In their study, protein structure, amino acid compositions, and the N-terminal sequence for the basic subunit were determined. In addition, the 2S seed storage albumin, the other major protein component of quinoa, was isolated and characterized (Brinegar et al., 1996).

Since the discovery of quinoa's specific seed storage proteins, recent studies have provided genetic tools for further characterization of the genes encoding these proteins.

Over 400 simple sequence repeats were characterized by Mason et al. (2004) in order to provide groundwork for the development of a genetic linkage map containing 230 amplified length polymorphism (AFLP), 19 simple sequence repeat (SSR), and 6 randomly amplified polymorphic DNA (RAPD) markers (Maughan et al., 2004). Coles et al. (2004) reported the construction of a quinoa cDNA library from seed and floral tissues and its utility in the identification of 51 single nucleotide polymorphisms (SNPs) and 424 expressed sequence tags (ESTs) from which the clones containing 11S cDNA were identified and sequenced. And finally, a BAC library for quinoa was developed and reported by Stevens et al. (2006) containing 74,496 clones with approximately 9x coverage of the quinoa genome.

One utility of the BAC library is sequence detection and isolation. Here we report the identification of clones containing putative 11S seed storage protein genomic DNA from the quinoa BAC library.

**RESULTS**

Identification and characterization of seed storage proteins in quinoa was an important first step in understanding how protein is stored in the seed in such high quantities and with such a well-balanced assortment of essential amino acids. In order to identify clones of the 11S gene, double spotted high-density blots representing the quinoa BAC library were hybridized with a radiolabeled fragment of the 11S gene. Ten clones hybridized with the 11S probe and seven of these were identified and selected from the BAC library.

Figure 1 shows the results of Southern blotting of *EcoRI*-digested BAC and genomic DNA probed with a fragment of the 11S gene. The radiolabeled probe was developed using primers specific to the 11S cDNA clone and amplifying a portion of it from BAC DNA extracted from one of the positive clones, 77L9. Lanes 1-7 include the seven BAC clones: 77L9, 13H24, 36L11, 82M10, 119J6, 155I19, and 164F2, respectively. Lanes 8-11 include genomic DNA from four homozygous lines: Chucapaca, KU-2, NL-6, and 0654, respectively. These homozygous are parentals of populations used for genetic mapping (Maughan et al., 2004). Two different bands, 4 and 4.5kb, were observed in lanes representing the seven BAC clones, similar in size to the bands of Chucapaca and 0654, two highland ecotypes. In addition, Southern blotting with 'Real', a highland variety and the quinoa accession used to construct the BAC library (Stevens et al., 2006), showed similar banding to Chucapaca and 0654 (data not shown). The 4.5kb band appears to be monomorphic across the four accessions whereas the 4kb band in the highland varieties and the 5.5kb band in the lowland varieties indicate a polymorphism in the *EcoRI* restriction site between the highland and lowland ecotypes.

**DISCUSSION**

Following construction of a BAC library, clones containing the 11S gene were identified and characterized. Restriction enzyme digestion followed by Southern blotting revealed two distinct bands evident in four parental accessions and a single band either 4kb or 4.5kb in size in the BAC clones. The relatively small size of the fragments suggests that the 11S gene may be present as a single copy at two separate loci within the genome. This notion is further supported by the fact that the two copies of the gene show

up in separate BAC clones and are located within separate linkage groups on the developing quinoa map (Jarvis, 2006). The restriction fragment length polymorphism identified between the highland and lowland ecotypes will be useful for genetic marker development. In addition, because quinoa is a putative allotetraploid and thought to have arisen from two diploid ancestors, it is possible that future mapping studies may show these two genes to be on homeologous chromosomes providing valuable information for studies involving quinoa ancestry. Quinoa's extraordinarily nutritional quality and ability to tolerate a wide variety of environmental conditions make it a valuable commodity. Understanding the proteins responsible for quinoa's unique amino acid composition is essential to future improvements in the quantity and quality of its seed storage proteins.

## MATERIALS AND METHODS

### Plant materials

Tissue for use in DNA extractions were collected from the first two true leaves of two-to-four-week-old plants grown under greenhouse conditions at Brigham Young University (BYU) of four inbred lines: NL-6, KU-2, 0654, and Chucapaca. Harvested tissue was immediately frozen and stored at -80˚C.

### Hybridization of the quinoa BAC library

An 11S gene probe was developed by amplification of a fragment of the gene with the following primers: forward 5'- GCACTAGTGTTGTTGAATGGGTGC -3' and reverse 5'- GCGTTCACCCCTCTGGGATT -3' derived from the cDNA sequences (GenBank accession AY562550). Hybridization of the BAC library was carried out according to the method described by Stevens et al. (2006).

**DNA extraction and Southern blotting**

Positive clones from the BAC library screening were grown in LB media containing 12.5 μg/ml chloramphenicol and BAC DNA was extracted using a BACMAX™ DNA purification kit (Epicentre, Madison, WI). Genomic DNA was extracted from NL-6, KU-2, 0654, and Chucapaca according to the protocol reported by Saghai-Maroof et al. (1984). Genomic and BAC DNA was digested using *EcoRI* and fragments were separated on a 0.8% agarose gel by electrophoresis. The gel was blotted using downward capillary transfer to a Hybond™-N+ positively charged nylon membrane (Amersham, Piscataway, NJ) and hybridized with a $^{32}$P-labeled 11S probe at 60˚C overnight. The membrane was washed and visualized the same as described by Stevens et al. (2006).

## LITERATURE CITED

**Brinegar C, Goundan S** (1993) Isolation and characterization of chenopodin, the 11S seed storage protein of quinoa (*Chenopodium quinoa*). J Agric Food Chem **41:** 182-185

**Brinegar C, Sine B, Nwokocha L** (1996) High-cysteine 2S seed storage proteins from quinoa (*Chenopodium quinoa*). J Agric Food Chem **44:** 1621-1622

**Castillo RO** (1995) Plant genetic resources in the Andes: impact, conservation, and management. Crop Sci **35:** 355-360

**Chauhan GS, Eskin NAM, Tkachuk R** (1992) Nutrients and antinutrients in quinoa seed. Cereal Chem **69(1):** 85-88

**Coles ND, Coleman CE, Christensen SA, Jellen EN, Stevens MR, Bonifacio A, Rojas-Beltran JA, Fairbanks DJ, Maughan PJ** (2004) Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. Plant Science **168:** 439-447

**Fairbanks DJ, Burgener KW, Robison LR, Andersen WR, Ballon E** (1990) Electrophortetic characterization of quinoa seed proteins. Plant Breeding **104:** 190-195

**Galwey NW** (1995) Quinoa and relatives Chenopodium spp. (Chenopodiacea). In: Evolution of Crop Plants Longman Scientific Technical, Harlow Essex. J Smartt and NW Simmonds eds. Pp. 41-46

**Jarvis DE** (2006) Simple sequence repeat development, polymorphism, and genetic mapping in quinoa (*Chenopodium quinoa* Willd.). Unpublished masters thesis, Brigham Young University, Provo, Utah, USA

**Mason SL, Stevens MR, Jellen EN, Bonifacio A, Fairbanks DJ, Coleman CE, McCarty RR, Rasmussen AG, Maughan PJ** (2005) Development and use of microsatellite markers for germplasm characterization in quinoa (*Chenopodium quinoa* Willd.) Crop Science **45:** 1618-1630

**Maughan PJ, Bonifacio A, Jellen EN, Stevens MR, Coleman CE, Ricks M, Mason SL, Jarvis DE, Gardunia BW, Fairbanks DJ** (2004) A genetic linkage map of quinoa (*Chenopodium quinoa*) based on AFLP, RAPD, and SSR markers. Theor Appl Genet **109(6):** 1188-1195

**Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW** (1984) Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location and population dynamics. Proc Natl Acad Sci USA **81:** 8014-8018

**Stevens MR, Coleman CE, Parkinson SE, Maughan PJ, Zhang H-B, Balzotti MR, Kooyman DL, Arumuganathan K, Bonifacio A, Fairbanks DJ, Jellen EN, Stevens JJ** (2006) Construction of a quinoa (*Chenopodium quinoa* Willd.) BAC library and its use in identifying genes encoding seed storage proteins. Theor Appl Gen **112:** 1593-1600

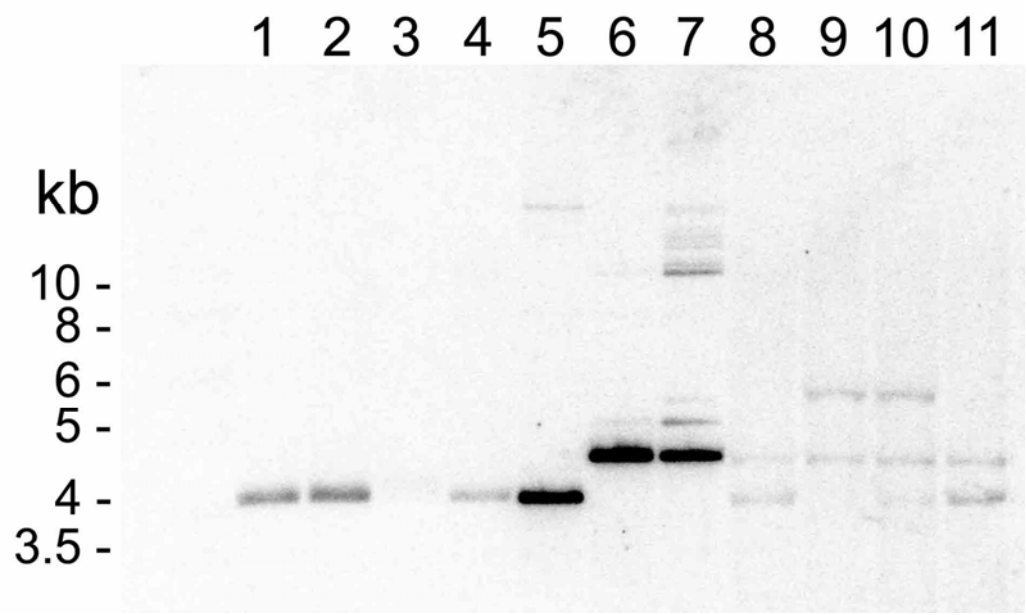


**Figure 1.** Southern blot of quinoa BAC DNA containing the 11S gene and genomic DNA from four quinoa accessions from inbred lines; lanes 1-7: 36L11, 82M10, 119J6, 155I19, and 164F2, respectively; lanes 8-11: genomic DNA from Chucapaca, KU-2, NL-6, and 0654 respectively. All DNA was digested with *Eco*RI and probed with a fragment of the 11S seed storage protein gene. The sizes of DNA markers are given in kilobases (kb) shown on the left. Lane 3 containing 119J6 BAC DNA shows a faint band at 4kb due to poor DNA concentration. Additional bands are shown in lanes 5-7 which is most likely due to incomplete digestion.

# Sequencing, expression and evolutionary relationships of the 11S seed storage protein gene in *Chenopodium quinoa* Willd.


**ABSTRACT**

Quinoa (*Chenopodium quinoa* Willd.) is a pseudocereal that provides nutritional value for high Andean plateau farmers and commercial value for many South American countries. Present interest in quinoa is due to its tolerance of harsh environments, its well-balanced amino acid composition and nutritional value. The seed storage proteins of quinoa, particularly the 11S globulins and 2S albumins are responsible for providing a reservoir of nitrogen and amino acids for the developing seedling upon germination. The genomic and cDNA sequences for two copies of the 11S gene at two different loci were determined. Transcription initiation and termination sites, coding regions, and introns were determined by comparing genomic to cDNA sequences. Gene expression and protein accumulation data were obtained using real time RT-PCR and separation by SDS-PAGE, respectively. Using the coding DNA sequence for the well-conserved 11S basic subunit, phylogentic relationships were analyzed between quinoa and other species for which 11S and homologous sequences have been reported.

## INTRODUCTION

Quinoa (*Chenopodium quinoa* Willd.) is an allotetraploid dicotyledonous pseudocereal whose seed is used as a source of nutrition for people in the Andean region to which it is indigenous. It has been an important cultural and dietary component for people of the Andean region since 3000 B.C. and continues to be a major resource for the Quechua and Aymara people (Brinegar, 1997). Quinoa is able to tolerate a wide variety of environments from valleys and coastal foothills to the semi-arid and cool Altiplano. In addition, approximately 12.6-13.7% of the dry weight of quinoa seed is protein (Cardoza and Tapia, 1979; Chauhan et al., 1992) which is high compared to most important cereal crops such as rice, corn and millet (Charalampopoulos, 2002). In addition, its amino acid composition exceeds the FAO recommendations for human protein consumption (Ruas et al. 1999) with relatively high quantities of lysine and sulfur-containing amino acids which are usually limiting in the seeds of cereals and legumes, respectively. Seed storage proteins play a vital role in the protein composition of the seed. It is therefore important to understand the genes encoding these proteins, their expression profiles and how they relate to analogous genes in other species in order to understand how protein is deposited in the seed and why quinoa, specifically, contains a relatively high quantity and well balanced assortment of amino acids.

Large quantities of seed storage proteins accumulate in developing seeds, as a source of nitrogen, carbon, sulfur and amino acids for use in germination and growth of the developing seedling. These protein reserves are stored in the cells of the endosperm and embryo in protein storage vacuoles or specialized aggregates called protein bodies, assembled within the endoplasmic reticulum (Prego et al., 1998). Seed storage proteins

are generally classified according to their solubilities: albumins, globulins, prolamins, and glutelins which are soluble in water, saline solutions, alcohol, and alkali solutions respectively (Osborne, 1924). Storage proteins are further classified based on their sedimentation coefficients ($S_{20,W}$). Globulins generally fall into two major groups based on these coefficients: the 7-8S vicilin-type and the 11-12S legumin-type. Since legumin-type seed storage proteins vary in size, we will refer to the 11-12S globulins collectively in other species as legumins. The 11S globulin, therefore, is a sub-type of the legumin seed storage proteins.

Recently, efforts have been made to understand the legumin gene families and the structures of the proteins they encode. Legumin sequences have been reported for economically important seed crops such as rice (Okita et al., 1989; Wen et al, 1989; Takaiwa et al., 1991), oat (Shotwell et al, 1988; Schubert et al., 1990; Tanchak et al., 1995), maize (Woo et al., 2001; Yamagata et al., 2003), and soybean (Momma et al., 1985; Utsumi et al., 1987; Nielsen et al, 1989; Xue et al., 1992). The crystal structure of an 11S globulin from soybean was determined, providing new insights into conserved binding domains, quaternary structure and how structure relates to packaging, storage and degradation (Adachi et al., 2003).

The 11S globulin is a hexameric protein consisting of six pairs of acidic and basic subunits with each subunit pair connected by a disulfide bond. It is translated as a single precursor polypeptide containing acidic and basic subunits as well as a signal peptide responsible for translocation of the precursor into the endoplasmic reticulum (ER). This signal peptide is cleaved upon entry into the lumen of the ER (Herman and Larkins, 1999). The 11S protein assembles as a trimer in the ER which is transported to the

vacuoles through the Golgi apparatus. Further processing occurs at a well-conserved cleavage site by an asparaginyl endopeptidase which divides the subunit into its acidic and basic components and hexamer formation finally takes place. The assembly of the hexamer is a direct result of this second cleavage event (Dickinson et al., 1989). At the onset of germination, degradation results due to pH-influenced flexibility and dissociation of the 11S hexamer as well as the actions of vacuolar-targeted proteinases formed by germination-related regulatory mechanisms (Adachi et al., 2003; Shutov et al., 2003).

Conserved regions within the 11S amino acid sequence are necessary for proper folding, packaging, storage, and proteolysis. Temporally controlled protein protection and degradation is determined by specific structural features. Distance-based phylogenetic analysis of legumin-encoding sequences as well as a review of the major groups of legumins was reported by Häger and Fischer (1999). The origin and early evolution of legumin structures have been reported by Shutov and Baumlein (1999) who describe their evolution from a germin-like ancestor. Following divergence of seed storage proteins into vicilin and legumin classes, each acquired independently their storage-related properties. Seed storage and analogous proteins seem to have an ancient history, present in the early progenitors of plants including mosses and fungi (Shutov et al., 2003). Thus, these gene sequences may indeed prove useful as phylogenetic markers in molecular evolution studies as well as provide insight into evolutionary modifications under the selective constrains of either better storage capabilities or degradation. In addition, the study of conserved and variable regions and how they relate to structure and

metabolism may provide valuable information into how these protein sequences may be altered and improved for nutritional purposes in quinoa and other species.

The 11S globulin and the 2S albumin are the two major storage proteins in quinoa seed (Brinegar, 1997). The 11S globulin in quinoa has a typical 11S quaternary structure with an estimated native molecular weight of 320 kDa. The acidic and basic polypeptides have average relative molecular weights of 35,500 Da and 22,500 Da, respectively (Brinegar and Goundan, 1993).

Here we report the genomic and cDNA sequences from the quinoa 11S gene. We present data on 11S gene expression during seed development and the corresponding accumulation of 11S globulin within the seed. Coding DNA sequences were used to conduct a phylogenetic study between the 11S sequence in quinoa and homologous seed storage proteins of diverse species. The results of this analysis may answer questions concerning legumin molecular evolution, why quinoa contains such an extraordinary protein content compared to important cereal crops and how the 11S sequence of quinoa is different from or similar to those of other species.

## RESULTS

### Isolation and Characterization of the Quinoa 11S Gene

The cDNA sequences for 17B7 (figure 1) and 8B14 (figure 2) were determined by primer walking off clones reported by Coles et al. (2004). Isolation of BAC clone 77L9 was previously reported by Stevens et al. (2006) and the same protocol was used in the recent identification of BAC clone 164F2. These two clones contain putative 11S genes from two separate loci within the quinoa genome. The nucleotide sequence of the entire

gene and flanking regions were obtained directly from the BAC clones using a primer walking strategy. The nucleotide sequence of the 11S gene region from clones 77L9 and 164F2 are shown in figures 3 and 4, respectively. Coding and flanking regions of the genomic nucleotide sequences of 77L9 and 164F2 had 99.5% and 99.8% sequence identity with that of the cDNA sequences of 17B7 and 8B14, respectively.

Initiation and termination codons and intron positions were determined by comparing genomic and cDNA sequences. Three introns were identified within the 11S gene and are shown in Figs. 3 and 4. The locus corresponding to BAC clone 164F2 and cDNA clone 8B14 contain a 1437 base pair open-reading frame that encodes a polypeptide 479 amino acids in length, whereas the locus corresponding to BAC clone 77L9 and cDNA clone 17B7 contain a 1440 base pair reading frame coding for a polypeptide with 480 amino acids. The three-base difference between the two clones results in an arginine insertion at position 285. Although the coding and flanking regions of genomic nucleotide sequences 77L9 and 164F2 share approximately 97% and 92% sequence identity, respectively, intron sequence identity between the two was only 57% on average.

The 11S amino acid sequence in quinoa is similar to 11S seed storage proteins found in other species. Amino acid compositions of polypeptides from the two loci are shown in Table 1. The first 25 hydrophobic amino acid residues are typical of a signal peptide found in other legumin-like seed storage proteins that is removed once entering the lumen of the ER (Utsumi, 1992). An additional processing site (between asparagine 293 and glycine 294 in 17B7 and asparagine 291 and glycine 292 in 8B14) in which the precursor polypeptide is separated into acidic and basic subunits is shown in figures 1-4. This

cleavage event would result in acidic and basic subunits 267-268 and 188 amino acids in length respectively. Other conserved sequences include regions surrounding amino acid residues that participate in covalent and hydrogen binding. Four of these regions contain cysteine residues which participate in disulfide bond formation in other species (Adachi et al., 2003).

**11S Gene Expression and Protein Accumulation**

To determine the 11S gene expression profile in quinoa, seeds were collected at eight- and ten-day intervals after anthesis until dessication from nine quinoa accessions: NL-6, KU-2, Mocko, Maniqueña, Ollague, Ratuqui, Sayaña, Chucapaca, and 0654. Accessions were chosen based on geographic location and maturation rate. Plants were selected from coastal regions and the Altiplano (high Andean plateaus) and are shown in Table 2. Altiplano types are subdivided into Salares (areas containing salt flats) and Valley. Maturity was determined at the time 90% of the seeds were fully dessicated.

Following seed collection, total RNA was extracted and quantified. Relative quantification of 11S mRNA in seeds from different developmental stages was determined by one-step reverse transcriptase real time PCR using the GAPDH gene as an endogenous control. In order to confirm expression at the protein level, relative 11S globulin quantity was determined by SDS-PAGE analysis. Figure 5 shows the results of 11S mRNA and globulin quantification at specific days after anthesis (DAA). Few 11S mRNA transcripts were detected during early seed development in all accessions. In general, 11S expression was detected followed by a substantial accumulation of 11S protein. As seeds became mature, expression decreased and protein accumulation remained high.

Based on our observations of plants grown under the same greenhouse conditions, five types of maturation rates were identified according to days to dessication, 11S RNA and protein accumulation patterns and are listed together with accession names and locations in Table 2. Type I consists of accessions KU-2 and NL-6 from coastal regions showing peak gene expression and 11S globulin accumulation at 24 DAA and maturity between 32 and 40 DAA. Type II, including Maniqueña and Mocko, two varieties located in the Salares, had peak gene expression at 24 DAA, detectable 11S globulin quantity occurring 32 DAA and maturity achieved between 40 and 50 DAA. Ollague (also located in the Salares), Ratuqui and Sayana (two accessions found in the Altiplano) were included as type III accessions and showed peak 11S transcript levels and substantial 11S globulin accumulation at 40 DAA and reached maturity between 50 and 60 DAA. Chucapaca, also located in the Altiplano, and 0654 from valleys, were included in type IV and type V, respectively, showed detectable 11S protein quantities at 50 DAA. However, Chucapaca reached maturity at approximately 60 DAA and 11S transcript levels reached a maximum amount 40 DAA while those of 0654 did so at 50 DAA and reached maturity between 60 and 70 DAA. As shown in Figure 5, the increases in 11S globulin levels in these different varieties appear to be correlated with the expression profiles for the 11S gene.

**Phylogenetic Analysis of 11S Seed Storage Proteins**

The coding DNA sequences of complete legumin sequences from 50 different plant species were translated, aligned and evaluated for phylogenetic signal quality. Common name, species, families, and orders for the plant species utilized in the phylogenetic analysis are presented in Table 3. Alignments were created using whole or portions of

the legumin sequence. Using distance-based and maximum parsimony methods for tree reconstruction in MEGA3 (Kumar et al., 2004) preliminary trees were evaluated to assess the quality of the alignments. Amino acid alignments generated from only the basic subunit sequence of the legumin gene (partial alignment shown in Figure 6) resulted in detectable amino acid conservation, consistency between trees generated by the aforementioned methods and taxonomic clustering. It was therefore selected for further analysis.

Figures 7 and 8 show parsimony and Bayesian tree reconstructions, respectively. In both reconstructions, with posterior probabilities and bootstrap values of 100, the two quinoa 11S sequences formed a monophyletic group with that of a *Amaranthus hypochondriacus* sequence (Genbank accession X82121) , the only other available sequence from a species within the Amaranthaceae family. Using amino acid alignments so identities and positives were maximized, the 11S basic subunit showed over 74% sequence identity between amaranth and quinoa. Other monophyletic and well-resolved taxonomic groups represented in both trees include taxa from the Poaceae, Brassicaceae, Cupressaceae, and Pinaceae families, the Fagales order and the clade for gymnosperms. As expected, the sequences from members of the Fabaceae family formed two monophyletic groups, each containing 11S sequences derived from different subfamilies that diverged early in angiosperm evolution (Yagasaki, 1997). The phylogenetic relationships between the gymnosperm outgroup taxa are well-resolved in both trees and reveal a clear bifurcation of angiosperm and gymnosperm legumin genes. Although both reconstructions show similar relationships between taxa, the Bayesian consensus tree has better support, indicated by high posterior probability values, than the parsimony tree for

unresolved portions of the tree including relationships between taxa from species of different families and orders.

## DISCUSSION

### 11S Gene Sequences

Two BAC clones, 77L9 and 164F2, containing 11S genomic DNA and two cDNA clones, 17B7 and 8B14, were sequenced and analyzed. The coding sequences for 77L9 and 164F2 highly correlated with 99.7% and 99.9% identity to 17B7 and 8B14, respectively. The quinoa cultivar 'Real' was used in construction of both cDNA and BAC libraries. However, it is not truebreeding. Therefore, the slight difference in DNA sequence between the BAC and cDNA sequences is likely due to heterogeneity in the 'Real' line. Although the coding and flanking regions were well conserved between the two loci (approximately 96% identity overall), the introns only correlated with approximately 57% identity.

The amino acid sequence of the 11S gene in quinoa is very similar to that of other species in domains necessary for binding and hexamer formation. Sequence similarities of the first 25 amino acid residues between quinoa and other species suggest that this region is a signal peptide that localizes the precursor to the ER where it is subsequently removed (Utsumi, 1992). Brinegar (1993) reported the N-terminal sequence of the basic subunit indicating that the polypeptide is indeed cleaved at a well-conserved site reported in other species (Utsumi, 1992). The sequence similarities of six binding regions suggest that the quinoa 11S hexamer has a structure similar to glycinin, for which the crystal structure has been determined (Adachi et al., 2003). Four of these binding regions

contain cysteine residues that participate in disulfide bond formation, one intrachain disulfide bond which occurs within the acidic polypeptide between the first two cysteine residues and one interchain disulfide bond which occurs between the last cysteine residue within the acidic subunit and the first cysteine in the basic subunit (Utsumi, 1992). (See Figures 1-4.)

The 11S amino acid sequence indicates a well-balanced composition and a high level of essential amino acids. Typical of 11S storage proteins, chenopodin contains high levels of glutamate/glutamine, aspartate/asparagine, arginine, serine, leucine and glycine (Brinegar, 1993). However, we note that the 11S gene is only one of two major seed storage proteins in the quinoa seed and that it is the combined contribution of the 11S and the 2S seed proteins that produce quinoa's unique seed protein quantity and composition. Thus, cloning and sequencing of the 2S albumin will be necessary to gain a complete understanding of the relative contributions of each storage protein within the developing seed and how quinoa seed accumulate such a high quantity of protein and essential amino acids.

**Expression Patterns of 11S Genes in Different Quinoa Accessions**

The gene expression and protein data presented here suggests that the accumulation of seed storage protein is associated with maturation rate. Early maturing varieties KU-2 and NL-6, for example had the first substantial accumulation of 11S mRNA (16 DAA) and protein (24 DAA) while late maturing variety 0654 did not show increased transcript levels and protein accumulation until 50 DAA. These data and those from the other quinoa accessions suggest that high expression of 11S mRNA occurs, on average, during late-maturation, consistent with the pattern of storage protein

accumulation during seed development and those of other 11S seed storage proteins (Nakamura, 2004). In this analysis, we have observed relative increases in 11S transcript levels and globulin. However, a quantitative study into the total protein content per seed of these and other quinoa accessions will be needed to determine if the timing of gene expression has an effect on the overall quantity of seed protein. Also, while analyzing plants grown in the greenhouse is important and informative, this study should be repeated in the field to see if our observations are consistent with those of field-grown plants.

**Phylogenetic Relationships between Legumins of Various Species**

The legumin coding DNA sequence of the basic subunit was used in order to analyze phylogenetic relationships between the 11S gene of quinoa and other species. The basic subunit of legumin is well-conserved and contains sequence elements that are necessary for its proper assembly and packaging. These include the first N-terminal amino acids involved in the recognition of the peptide cleavage site between the acidic and basic subunits followed by an amino acid sequence which forms a β-barrel structure. Also included are conserved regions surrounding several amino acids involved in hydrogen bonding, hydrogen-bonded salt bridges and a cysteine residue involved in disulfide bond formation (Adachi et al., 2003). Thus the entire legumin basic subunit sequence or portions of the basic subunit are suitable phylogenetic sequences for studies involving the molecular evolution of seed storage proteins (Fischer et al., 1996; Häger and Wind, 1997; Shutov and Bäumlein, 1999 ).

Two phylogenetic trees, one based on maximum parsimony and another on Bayesian analysis, are shown in figures 7 and 8, respectively. These two methods, which

utilize very different assumptions, were selected in order to validate the relationships shown in the reconstructions. Maximum parsimony assumes that the best tree is one in which evolutionary steps are minimized while Bayesian analysis implements a model of evolution in addition to the sequence data in order to maximize the probability of a tree. Because one method lacks outside parameters assuming that one evolutionary event is just as likely as another and the other uses established parameters in order to predict associations, we assumed that relationships established by both reconstructions would be well-supported. Indeed, all monophyletic groups defined by the maximum parsimony tree were also supported by the Bayesian tree.

The data suggest that the Bayesian analysis was better suited for resolving relationships between more distantly-related taxa than the parsimony method. Indeed, most unresolved portions of the Bayesian reconstruction angiosperm backbone had posterior probabilities just under 90 and are therefore not shown in Figure 8. A third and fourth node with posterior probabilities of 83 and 88, respectively was present following the bifurcation of lower angiosperm legumin sequences in ginger, mint and yam (Fischer et al., 1996). The third node represented the divergence in a sequence from arrowhead. The fourth node separated two clades from the remaining angiosperms, one including sequences from members of the Poaceae family, magnolia, african oil palm, almond and coffee and the other composed of sequences from sesame, pumpkin, orange and buckwheat (data not shown). Although these relationships are based on posterior probabilities under 90, they are still supported by the majority of the data.

Following the divergence between gymnosperms and angiosperms, there appears to be an explosion of sequence divergence in the legumin genes as evident by the

presence of multigene families and the lack of resolution in the phylogenetic tree reconstruction. Multigene families for the legumin-like genes have been identified in many species including those of legumes (Domeney et al., 1986; Heim et al., 1994; Nielsen et al., 1989), brassicas (Breen and Crouch, 1992; Depigny-This et al., 1992; Pang et al., 1988) and cereals (Okita et al., 1989; Shotwell et al., 1988). Multigene families have also been identified in gymnosperms (Hager et al., 1995; Wind and Hager, 1996; Hager and Wind, 1997). However, these sequences have a much higher degree of homology with each other than those of angiosperms. There does not appear to be an 11S multigene family in quinoa (Stevens et al., 2006), nor have there been any identified in amaranth (Barba de la Rosa et al., 1996). However, other 11S genes or pseudogenes may be present in the quinoa genome but have not yet been identified due sequence dissimilarity. If additional 11S genes are present but nonfunctional, it would be interesting to understand what type of selection may have resulted in such an inactivation.

Both phylogenetic trees were well-resolved between closely-related taxa. Quinoa's 11S sequence placement with amaranth exhibited posterior probabilities and bootstrap values of 100. The observation that the quinoa sequence was located where it would be expected to fall (with the only other sequence from a member of the Amaranthacae family) and the high degree of homology between it and amaranth suggests that selection on this sequence is absent or subtle or that opposing selective forces do not allow the sequence to change substantially. Indeed, Shutov et al. (2003) suggested that seed storage protein structure evolution is under the constraints of two antagonistic selective forces: selection for better storage capabilities and selection for improved degradation. Quinoa is a putative allotetraploid and we postulate that the two

copies of the 11S gene may be located on homeologous chromosomes and represent genes present in the diploid ancestors of quinoa (Stevens et al., 2006). Although the sequences for both loci are similar, transcript levels from each may not be equal. As an alternative, it is possible that the unusually high protein content of quinoa is a direct result of tetraploidization and the unaltered expression of both genes since polyploidy generally increases gene expression levels overall (Osborn et al., 2003). Thus, seeds containing higher protein reserves for germination and the developing seedling may have been at a selective advantage.

The ability to resolve relationships between sequences from gymnosperms and other ancient taxa is shown by both parsimony and Bayesian trees. We note, however, that the radiation at the angiosperm level as well as the lack in sufficient sequence data to match the broad range of angiosperm species creates some challenges when using the legumin sequence as a phylogenetic tool. Indeed, without adequate sequence data and the added complication of multigene families that may have arisen at different occasions during evolutionary time, it is difficult to know if homologous genes are being compared and taxa may be grouped inappropriately. However, it is evident from this study that there is sufficient phylogenetic signal even from diverse species to conduct future studies into molecular evolution of legumin genes in angiosperms as new sequences are discovered.

## MATERIAL AND METHODS

### Plant materials

Quinoa accession Real was grown under greenhouse conditions at Brigham Young University. Tissue for use in DNA extractions were collected from the first two true leaves of two-to-four-week-old plants. Harvested tissue was immediately frozen and stored at -80˚C.

Nine quinoa accessions were planted simultaneously and grown under greenhouse conditions: NL-6, KU-2, Maniqueña, Mocko, Ollague, Ratuqui, Sayaña, Chucapaca, and 0654. Seed was collected at eight day intervals after anthesis until maturity from each accession and immediately frozen and stored at -80˚C.

### cDNA and genomic sequencing

Genomic DNA was extracted from quinoa accession Real according to the protocol reported by Saghai-Maroof et al. (1984). An initial fragment of the quinoa 11S gene was obtained by amplification of genomic DNA using degenerate PCR primers designed from a partial quinoa 11S amino acid sequence (Brinegar and Goundan, 1993) as well as an alignment of a cDNA sequence from *Amaranthus hypochondriacus* (Barba de la Rosa et al., 1996). The forward primer sequence was 5'-AATGGKGTGGARGARACYATTTGC -3' and the reverse 5'-TGTKKGCGTTKAGGTTSYAGTG -3'. This gene fragment was used as a probe to screen the quinoa developing seed cDNA library reported by Coles et al. (2004). Positive clones were selected and their identities confirmed by Southern blotting. Clones 17B7 and 8B14 were selected for sequencing. Two BAC clones containing 11S genomic DNA, 77L9 previously isolated by Stevens et al. (2006) and 164F2 subsequently

identified and representing a second locus were identified by screening of a quinoa BAC library. The 11S genomic inserts within clones 77L9 and 164F2 and the cDNA inserts within clones 17B7 and 8B14 were sequenced by primer walking at the Brigham Young University DNA Sequencing Center (Provo, UT).

**RNA extraction and relative quantification**

Nine quinoa accessions were germinated and grown under the same greenhouse conditions. Temperatures were 20-25°C during the day and 15-18°C at night and plants were exposed to 12h days. Seed collected at eight- and ten-day intervals after anthesis from these accessions were ground to a fine powder in liquid nitrogen. Total RNA was extracted using an RNeasy® Plant Mini Kit (Qiagen, Valencia, CA) and quantified using a NanoDrop® ND-1000 Spectrophotometer (NanoDrop Technologies, Inc., Wilmington, DE). Primers and probes were designed for the 11S gene and the GAPDH gene: 11S forward 5'- GGCGGTCGCTTCCAAGA -3', reverse 5'- TTGCGAAAATGTGGCCTTGAC -3', probe 5'- CCAACACCAGAAGATCA -3'; GAPDH forward 5'- GGTTACAGTCATTCAGACACCATCA -3', reverse 5'- AACAAAGGGAGCCAAGCAGTT -3', probe 5'- CGCTTCCTGTACCAC -3'. Using the GAPDH gene as an endogenous control, multiplexed RNA samples were quantified using a TaqMan® One-Step RT-PCR Master Mix Reagents Kit (Applied Biosystems, Foster City, CA) on a 7300 Real Time PCR System (Applied Biosystems, Foster City, CA). Each sample contained 0.5µg of total RNA. Thermocycler conditions included the following: samples were held at 48°C for 30min, heated to 95° and held for 10min, followed by forty repetitions of temperatures at 95°C for 15s and 60°C for 1min.

**Globulin extraction, quantification and separation by SDS-PAGE**

Fifty milligrams of seed collected at eight day intervals after anthesis from the eight quinoa accessions were ground to a fine powder and 500µl of water was added to each sample. The samples were shaken at 4˚C for over an hour, centrifuged at 13,000xg, and the supernatant containing the albumin fraction removed. The pellet was washed twice using the same protocol and dried. To each sample, 400µl 0.5M NaCl/50mM Tris-HCl, pH 8.0 was added. The samples were shaken at 4˚C for over an hour, centrifuged at 13,000xg and the supernatant containing the globulin fraction was collected into separate tubes. Globulin was quantified using a BCA™ Protein Assay Kit (Pierce, Rockford, IL) on a NanoDrop® ND-1000 Spectrophotometer (NanoDrop Technologies, Inc., Wilmington, DE). Globulin samples were diluted 1:2 with Laemmli Sample Buffer (BioRad, Hercules, CA) and separated with SDS-PAGE on Ready Gels (Biorad, Hercules, CA). Gels were fixed overnight according to the manufacturer's instructions and visualized using Flamingo™ Fluorescent Gel Stain (Biorad, Hercules, CA) under UV light.

**Taxonomic sampling and sequence alignment**

Legumin sequences from 50 different species of plants were assembled using MEGA3 software (Kumar et al., 2004). These species and their accession numbers are listed in Table 3. Coding DNA sequences were downloaded from GenBank. Using the amino acid sequence of the 11S seed storage protein from *Chenopodium quinoa* (GenBank accession AAS67037) as a query sequence, a BLASTP search was performed (Altschul et al., 1990). Only legumin sequences from different plant species were chosen with E-values less than 1e-10 as to include sequences substantially similar and likely

related by descent to the 11S gene in quinoa. Incomplete sequences, less than 75% of the average length of the legumin gene, and redundant sequences were eliminated because the C-terminal basic amino acid sequence was used for analysis. One coding DNA legumin sequence from each species, except for quinoa and species from the Fabaceae family was selected based on its similarity to the 11S sequence from quinoa. Due to a greater amount of sequence data available for members of the Fabaceae family, legumin sequences from two major gene types (Yagasaki, 1997) were selected if available. The two quinoa sequences included in the analysis corresponded to the coding DNA sequences of cDNA clones 17B7 and 8B14 (GenBank accessions AY562549 and AY562550, respectively). The legumin basic subunit, a well-conserved sequence region, was chosen to improve the alignment because distantly related sequences were being compared. The selected coding DNA legumin sequences were translated into amino acid sequences using MEGA3. Amino acid sequences were aligned with ClustalW using the BLOSUM matrix and used as a template for the nucleotide alignment. ClustalW performs multiple sequence alignments of nucleotide or protein sequences using a progressive alignment algorithm including a gap cost model which penalizes insertions and deletions according to a linear function and sequence similarity for construction of a guide tree (Thompson et al., 1994).

**Phylogenetic Reconstruction**

Trees were constructed using distance, maximum parsimony and Bayesian methods. Branch support of parsimony and Bayesian reconstructions was tested using bootstrapping and posterior probabilities, respectively. ModelTest (Posada and Crandall,

1998) was utilized in order to select the most appropriate nucleotide substitution model for molecular evolution.

Distance, maximum parsimony and Bayesan analysis are three kinds of tree reconstruction methods reliant on different assumptions. Neighbor-joining is a distance method that expresses distance as a fraction of sites that differ between sequences. Maximum parsimony assumes that the tree generated by the fewest number of sequence changes between taxa is the most likely one. Bayesian analysis relies upon the concept of posterior probabilities, based on a postulated model of evolution that represents the probability of the relationship given the data. Trees are generated based on the model and the sequence alignment, sampled periodically and used to compute a consensus tree.

ModelTest (Posada and Crandall, 1998) is a program designed to assign the most statistically appropriate model of evolution to the alignment under evaluation. It employs two statistical approaches to model selection, the likelihood ratio test (LRT) and the Akaike information criterion (AIC). The LRT is generated by pairwise comparisons between the likelihood score of nested alternative hypotheses under a complex model. The AIC compares several models concurrently to determine the model that best fits the data while penalizing increasing numbers of parameters in the model (increasing parameters will always generate a better fit for the data but doesn't always represent the best model). The general time-reversible (GTR) model assumes that the probability of changing from one base into another is the same in one direction as it is in the reverse direction. Therefore, among four nucleotides there are six possible substitution rates. Additional parameters allow probabilities to be assigned to each site for designation to a specific rate category.

Measures of branch support for parsimony and Bayesian reconstructions can be tested by using bootstrapping and posterior probabilities, respectively. Bootstrap values are expressed as probabilities that taxa within a clade are always grouped within that clade. Posterior probabilities express the probability of relationships given the data and the model of evolution.

The phylogenetic signal of the data was explored using distance tree reconstruction in MEGA3. Parsimony reconstruction was accomplished using PAUP* version 4.0b10 (Swofford, 2001) and a bootstrap analysis with 1000 replicates was conducted with bootstrap values over 70 representing supported relationships. A parameter-rich GTR+I+G model to be implemented into the Bayesian analysis was selected from 56 alternative models with ModelTest as the most appropriate nucleotide substitution model for the aligned sequences. Bayesian consensus tree reconstruction was accomplished using MrBayes version 3.0b4 (Ronquist and Huelsenbeck, 2003) using default values as the starting parameters. Five million tree generations were conducted and trees were saved every $1000^{th}$ generation. The consensus tree was generated with a burn-in at 20,000 tree generations and posterior probabilities above 90 represented supported relationships. Members of the Coniferales, Ginkgoales and Gnetales were assigned as a monophyletic outgroup based on putative sequence similarity to ancient taxa (Shutov et al., 1998)

## LITERATURE CITED

**Adachi M, Kanamori J, Masuda T, Yagasaki K, Kitamura K, Mikami B, Utsumi S** (2003) Crystal structure of soybean 11S globulin: glycinin A3B4 homohexamer. PNAS **100:** 7395-7400

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

**Barba de la Rosa AP, Herrera-Estrella A, Utsumi S, Paredes-López O** (1996) Molecular characterization, cloning and structural analysis of a cDNA encoding an amaranth globulin. J Plant Physiol **149:** 527-532

**Breen JP, Crouch ML** (1992) Molecular analysis of a cruciferin storage protein gene family in *Brassica napus*. Plant Mol Biol **19:** 1049-1055

**Brinegar C** (1997) The seed storage proteins of quinoa. In: Food Proteins and Lipids (ed. Damodaran) Plenum Press, New York. pp. 109-115

**Brinegar C, Goundan S** (1993) Isolation and characterization of chenopodin, the 11S seed storage protein of quinoa (*Chenopodium quinoa*). J Agric Food Chem **41:** 182 185

**Cardoza A, Tapia M** (1979) Valor nutricio. In: Quinoa y Kaniwa. M. Tapia ed. Serie libros y materials educativos No. 49CIID-IICA. Bogota

**Charalampopoulos D, Wang R, Pandiella SS, Webb C** (2002) Application of cereals and cereal components in functional foods: a review. International Journal of Food Microbiology **79:** 131-141

**Chauhan GS, Eskin NAM, Tkachuk R** (1992) Nutrients and antinutrients in quinoa seed. Cereal Chem **69:** 85-88

**Coles ND, Coleman CE, Christensen SA, Jellen EN, Stevens MR, Bonifacio A, Rojas-Beltran JA, Fairbanks DJ, Maughan PJ** (2004) Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. Plant Science **168:** 439-447

**Depigny-This D, Raynal M, Aspart L, Delseny M, Grellet F** (1992) The cruciferin gene family in radish. Plant Mol Biol **20:** 467-479

**Dickinson CD, Hussein EHA, Nielsen NC** (1989) Role of posttranslational cleavage in glycinin assembly. The Plant Cell **1:** 459-469

**Domeney C, Barker D, Casey R** (1986) The complete deduced amino acid sequence of legumin Beta-polypeptides from different genetic loci in Pisum. Plant Mol Biol **7:** 467-474

**Fairbanks DJ, Burgener KW, Robison LR, Andersen WR, Ballon E** (1990) Electrophortetic characterization of quinoa seed proteins. Plant Breeding **104:** 190-195

**FAO/WHO/UNU** (1985) Energy and protein requirements. Technical Report Series 724. WHO, Geneva

**Fischer H, Chen L, Wallisch S** (1996) The evolution of angiosperm seed proteins: a methionine-rich legumin subfamily in lower angiosperm clades. J Mol Evol **43:** 399 404

**Häger KP, Braun H, Czihal A, Muller B, Baumlein H** (1995) Evolution of seed storage protein genes: legumin genes of *Ginkgo biloba*. J Mol Evol **41:** 457-466

**Häger KP, Fischer H** (1999) Molecular phylogenies and structural diversification of gymnosperm and angiosperm storage globulins. In: Shewry PR, Casey R, eds. Seed proteins. The Netherlands: Kluwer Academic Publishers, 543-561

**Häger KP, Wind C** (1997) Two ways of legumin-precursor processing in conifers. Characterization and evolutionary relationships of *Metasequoia* cDNAs representing two divergent legumin gene subfamilies. Eur J Biochem **246:** 763 771

**Heim** (1994) The legumin gene family: a reconstructed *Vicia faba* legumin gene encoding a high-molecular-weight subunit is related to type B genes. Plant Mol Biol **25:** 131-135

**Herman EM, Larkins BA** (1999) Protein storage bodies and vacuoles. The Plant Cell **11:** 601-613

**Kumar S, Tamura K, Nei M** (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform **5:** 150-163

**McClellan DA, Palfreyman EJ, Smith MJ, Moss JL, Christensen RG, Sailsbery JK** (2004) Physicochemical evolution and molecular adaptation of the cetacean and artiodactyls cytochrome *b* proteins. Molecular Biology and Evolution **22:** 437-455

**Momma T, Negoro T, Udaka K, Fukazawa C** (1985) A complete cDNA coding for the sequence of glycinin A2B1a subunit precursor. FEBS **188(1):** 117-122

**Nielsen NC, Dickinson CD, Cho T-J, Thanh VH, Scallon BJ, Fischer RL, Sims TL, Drews GN, Goldberg RB** (1989) Characterization of the glycinin gene family in soybean. The Plant Cell **1:** 313-320

**Okita TW, Hwang YS, Hnilo J, Kim WT, Aryan AP, Larson R, Krishnan HB** (1989) Structure and expression of the rice glutelin multigene family. Journal of Biological Chemistry **264:** 12573-12581

**Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee H-S, Comai L, Madlung A, Doerge RW, Colot V, Martienssen RA** (2003) Understanding mechanisms of novel gene expression in polyploids. Trends in Genetics **19:** 141-147

**Osborne TB** (1924) The vegetable proteins. 2nd edn. Longmans, Green and Co. London pp. 154

**Pang PP, Pruitt RE, Meyerowitz EM** (1988) Molecular cloning, genomic organization, expression and evolution of 12S seed storage protein genes of *Arabidopsis thaliana*. Plant Mol Biol **11:** 805-820

**Posada D, Crandall KA** (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics **14:** 817-818

**Prego I, Maldonado S, Otegui M** (1998) Seed structure and localization of reserves in *Chenopodium quinoa*. Annals of Botany **82:** 481-488

**Ronquist F, Huelsenbeck JP** (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:** 1572-1574

**Ruas PM, Bonifacio A, Ruas CF, Fairbanks DJ, Anderson WR** (1999) Genetic relationship among 19 accessions of six species of *Chenopodium* L. by random amplified polymorphic DNA fragments (RAPD). Euphytica **105:** 25-32

**Schubert R, Baumlein H, Czihal A, Wobus U** (1990) Genomic sequence of a 12S seed storage protein gene from oat (*Avena sativa* L. cv. 'Solidor'). Nucleic Acids Research **18:** 377

**Shotwell MA, Afonso C, Davies E, Chesnut RS, Larkins BA** (1988) Molecular characterization of oat seed globulins. Plant Physiol **87:** 698-704

**Shutov AD, Braun H, Chesnokov YV, Horstmann C, Kakhovskaya IA, Bäumlein H** (1998) Sequence peculiarity of gnetalean legumin-like seed storage proteins. J Mol Evol **47:** 486-492

**Shutov AD, Bäumlein H** (1999) Origin and evolution of seed storage globulins. In: Shewry PR, Casey R, eds. Seed proteins. The Netherlands: Kluwer Academic Publishers, 543-561

**Shutov AD, Bäumlein H, Blattner FR, Müntz K** (2003) Storage and mobilization as antagonistic functional constraints on seed storage globulin evolution. Journal of Experimental Botany **54:** 1645-1654

**Stevens MR, Coleman CE, Parkinson SE, Maughan PJ, Zhang H-B, Balzotti MR, Kooyman DL, Arumuganathan K, Bonifacio A, Fairbanks DJ, Jellen EN, Stevens JJ** (2006) Construction of a quinoa (*Chenopodium quinoa* Willd.) BAC library and its use in identifying genes encoding seed storage proteins. Theor Appl Gen **112:** 1593-1600

**Swofford D** (2001) PAUP*: Phylogentic analysis using parsimony (*and other methods). Version 4.0b10. Sinauer Associates, Sunderland, Mass.

**Takaiwa F, Oono K, Wing D, Kato A** (1991) Sequence of three members and expression of a new major subfamily of glutelin genes from rice. Plant Mol Biol **17:** 875-885

**Tanchak MA, Giband M, Potier B, Schernthaner JP, Dukiandjiev S, Altosaar I** (1995) Genomic clones encoding 11S globulins in oats (*Avena sativa*). Genome **38:** 627-634

**Thompson JD, Higgins DG, Gibson TJ** (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Research **22:** 673-680

**Utsumi S** (1992) Plant food protein engineering. Adv Food Nutr Research **36:** 89-207

**Utsumi S, Kim CS, Kohno M, Kito M** (1987) Polymorphism and expression of cDNAs encoding glycinin subunits. Agric Biol Chem **51:** 3267-3273

**Wen L, Huang J-K, Johnson BH, Reeck GR** (1989) Nucleotide sequence of a cDNA that encodes a rice glutelin. Nucleic Acids Research **17:** 9490

**Wind C, Hager KP** (1996) Legumin encoding sequences from the redwood family (Taxodiaceae) reveal precursors laking the conserved Asn-Gly processing site. FEBS Lett **383:** 46-50

**Woo Y-M, Hu DW-N, Larkins BA, Jung R** (2001) Genomics analysis of genes expressed in maize endosperm identifies novel seed proteins and clarifies patterns of zein gene expression. The Plant Cell **13:** 2297-2317

**Xue ZT, Xu ML, Shen W, Zhuang NL, Hu WM, Shen SC** (1992) Characterization of a Gy4 glycinin gene from soybean *Glycine max* cv forrest. Plant Mol Biol **18:** 897-908

**Yagasaki K, Takagi T, Sakai M, Kitamura K** (1997) Biochemical characterization of soybean protein consisting of different subunits of glycinin. J Agric Food Chem **45:** 656-660

**Yamagata T, Kato H, Kuroda S, Abe S, Davies E** (2003) Uncleaved legumin in develping maize endosperm: identification, accumulation and putative subcellular localization. Journal of Experimental Botany **54:** 913-922

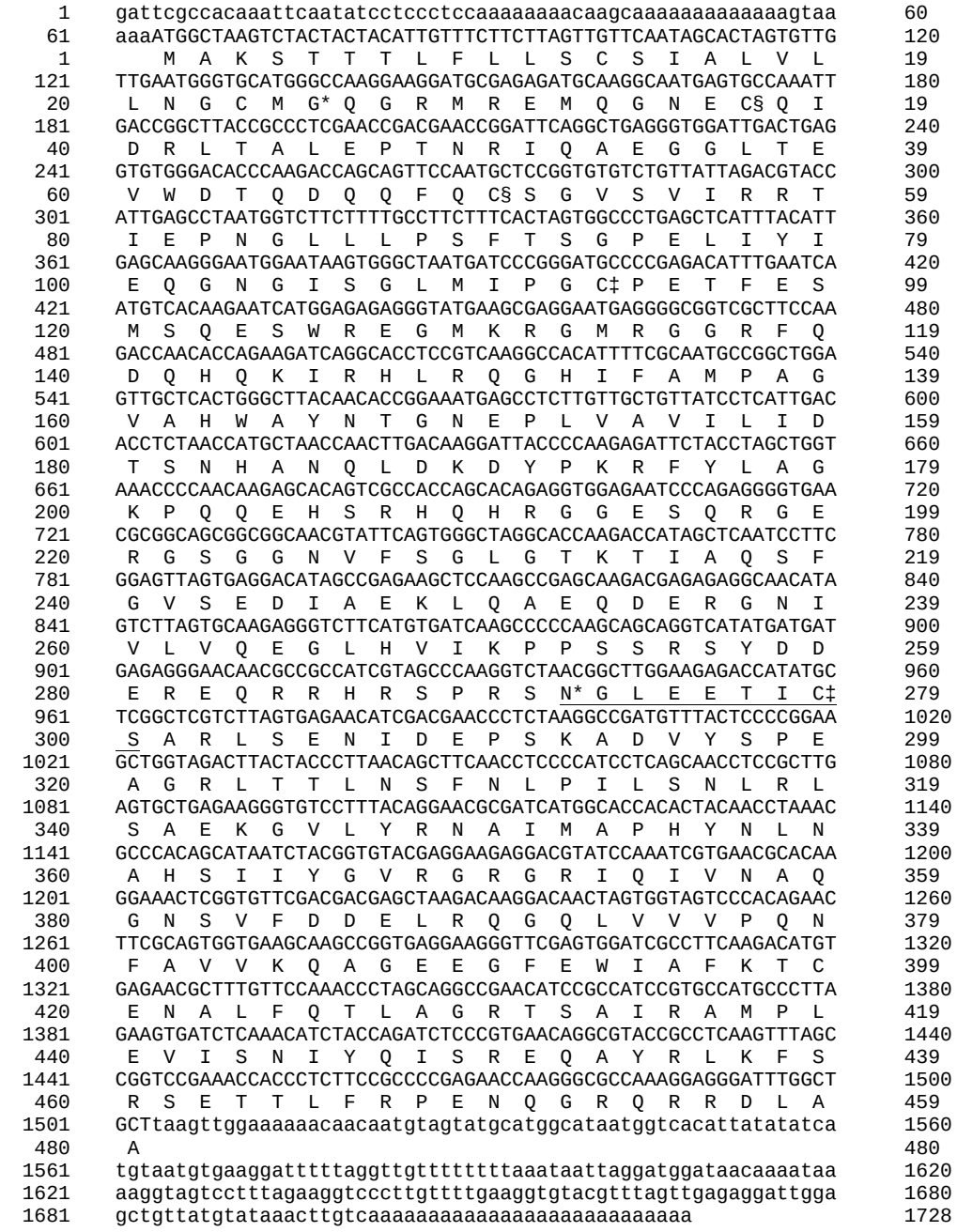


**Figure 1.** The nucleotide and predicted amino acid sequence of cDNA clone 17B7. The * indicates putative cleavage sites that divide the precursor polypeptide into a 25-amino acid signal peptide, 267-residue acidic and 118-residue basic subunits. The ‡ and § indicate cysteine residues predicted to participate in intrachain and interchain disulfide bonds, respectively. The 9 underlined amino acids indicate a well-conserved processing site involved in peptidase recognition. Untranslated portions are indicated in lowercase.

```
   1    gattcgccacaaattaaatatccttctccaaaaatacaagaaaaaaaaagagagtaaaaA      60
  61    TGGCTAAGTCTACTACTACATTGTTTCTTCTTAGTTGTTCAATAGCACTAGTGTTGTTGA      120
   1    M  A  K  S  T  T  T  L  F  L  L  S  C  S  I  A  L  V  L  L       20
 121    ATGGGTGCATGGGCCAAGGAAGGATGCGAGAAATGCAAGGCAATGAGTGCCAAATCGACC      180
  21    N  G  C  M  G* Q  G  R  M  R  E  M  Q  G  N  E  C§ Q  I  D       40
 181    GGCTTACCGCCCTCGAACCGACGTATCGGATTCAGGCTGAGGGTGGGTTGACTGAGGTGT      240
  41    R  L  T  A  L  E  P  T  Y  R  I  Q  A  E  G  G  L  T  E  V       60
 241    GGGACACCCAGGACCAGCAGTTCCAATGCTCCGGTGTGTCTGTTATTAGGCGTACCATTG      300
  61    W  D  T  Q  D  Q  Q  F  Q  C§ S  G  V  S  V  I  R  R  T  I       80
 301    AGCCTAATGGTCTTCTTTTGCCTTCTTTCACTAGTGGCCCTGAGCTCATTTACATTGAGC      360
  81    E  P  N  G  L  L  L  P  S  F  T  S  G  P  E  L  I  Y  I  E      100
 361    AAGGGAATGGAATAAGTGGGCTGATGATCCCGGGGTGCCCCGAGACATTTGAATCAATGT      420
 101    Q  G  N  G  I  S  G  L  M  I  P  G  C‡ P  E  T  F  E  S  M      120
 421    CACAAGAATCATGGAGAGAGGGTATGGAGCGAGGAATGAGGGGCGGTCGCTTCCAAGACC      480
 121    S  Q  E  S  W  R  E  G  M  E  R  G  M  R  G  G  R  F  Q  D      140
 481    AACACCAGAAGATCAGGCACCTCCGTCAAGGCCACATTTTCGCAATGCCGGCCGGAGTTG      540
 141    Q  H  Q  K  I  R  H  L  R  Q  G  H  I  F  A  M  P  A  G  V      160
 541    CTCACTGGGCTTACAACTCCGGAAATGAGCCACTTGTTGCTGTTATCCTCATTGACACCT      600
 161    A  H  W  A  Y  N  S  G  N  E  P  L  V  A  V  I  L  I  D  T      180
 601    CTAACCATGCTAACCAACTTGACAAGGATTACCCCAAGAGATTCTACCTAGCTGGTAAAC      660
 181    S  N  H  A  N  Q  L  D  K  D  Y  P  K  R  F  Y  L  A  G  K      200
 661    CCCAACAAGAGCACAGTCGCCACCATCACAGGGGTGGAGAATCCCAGAGGGGTGAACACG      720
 201    P  Q  Q  E  H  S  R  H  H  H  R  G  G  E  S  Q  R  G  E  H      220
 721    GCAGCGACGGCAACGTATTCAGTGGCCTAGACACCAAGTCCGTAGTTCAATCCTTCGGAG      780
 221    G  S  D  G  N  V  F  S  G  L  D  T  K  S  V  V  Q  S  F  G      240
 781    TTAGTGAGGACATAGCGGAGAAGCTCCAAGCCAAGCAGGACGAGAGAGGCAACATAGTCT      840
 241    V  S  E  D  I  A  E  K  L  Q  A  K  Q  D  E  R  G  N  I  V      260
 841    TAGTGCAAGAGGGTCTTCATGTGATCAAGCCCCCAAGCAGCAGGTCATATGATGATGAGA      900
 261    L  V  Q  E  G  L  H  V  I  K  P  P  S  S  R  S  Y  D  D  E      280
 901    GGGAACAGCGCCATCGTAGCCCAAGGTCTAACGGCTTGGAAGAGACCATATGCTCGGCTC      960
 281    R  E  Q  R  H  R  S  P  R  S  N* G  L  E  E  T  I  C‡ S  A      300
 961    GCCTTAGTGAGAACATCGACGATCCCTCTAAGGCCGATGTTTACTCCCCGGAAGCTGGTA     1020
 301    R  L  S  E  N  I  D  D  P  S  K  A  D  V  Y  S  P  E  A  G      320
1021    GACTTACCACCCTTAACAGCTTCAACCTCCCCATCCTCAGCAACCTCCGCTTGAGTGCTG     1080
 321    R  L  T  T  L  N  S  F  N  L  P  I  L  S  N  L  R  L  S  A      340
1081    AGAAGGGTGTCCTCTACAGGAACGCAATCATGGCACCACACTACAACCTAAACGCCCACA     1140
 341    E  K  G  V  L  Y  R  N  A  I  M  A  P  H  Y  N  L  N  A  H      360
1141    GCATAATCTATGGTGTACGAGGACGAGGACGTATCCAAATCGTGAACGCACAAGGAAACT     1200
 361    S  I  I  Y  G  V  R  G  R  G  R  I  Q  I  V  N  A  Q  G  N      380
1201    CAGTGTTTGACGATGAGCTAAGACAAGGACAACTAGTGGTAGTCCCACAGAACTTCGCAG     1260
 381    S  V  F  D  D  E  L  R  Q  G  Q  L  V  V  V  P  Q  N  F  A      400
1261    TGGTGAAGCAAGCCGGTGAGGAAGGATTCGAGTGGATCGCCTTCAAGACATGTGAGAACG     1320
 401    V  V  K  Q  A  G  E  E  G  F  E  W  I  A  F  K  T  C  E  N      420
1321    CTTTGTTCCAAACCCTAGCAGGCCGAACATCCGCCATACGTGCCATGCCCGTAGAAGTAA     1380
 421    A  L  F  Q  T  L  A  G  R  T  S  A  I  R  A  M  P  V  E  V      440
1381    TCTCAAACATCTACCAGATCTCCCGCGAACAGGCGTACCGCCTCAAGTTTAGCCGGTCCG     1440
 441    I  S  N  I  Y  Q  I  S  R  E  Q  A  Y  R  L  K  F  S  R  S      460
1441    AAACCACCCTCTTCCGCCCCGAGAACCAAGGGCGCCAAAGGAGGGAAATGGCTGCTtaag     1500
 461    E  T  T  L  F  R  P  E  N  Q  G  R  Q  R  R  E  M  A  A        479
1501    ttgaaaaaacaacaacgtagtaagcatggcataatgatcacgttatatatcatgtaatgt      1560
1561    gaaggagttttaggttgttttttttaaaataattaggatggataacaaaataaaaggta      1620
1621    gtcctttagaaggtcccttgttttgaaggtgttttgttgagaggattggagctgttatgt      1680
1681    ataaacttgtcaatgatcaaatgcaaagttccaacatcaaaaaaaaaaaaaaaaaaaaa      1740
1741    aa
```

**Figure 2**. The nucleotide and predicted amino acid sequence of cDNA clone 8B14. Designations are the same as those listed in Figure 1.

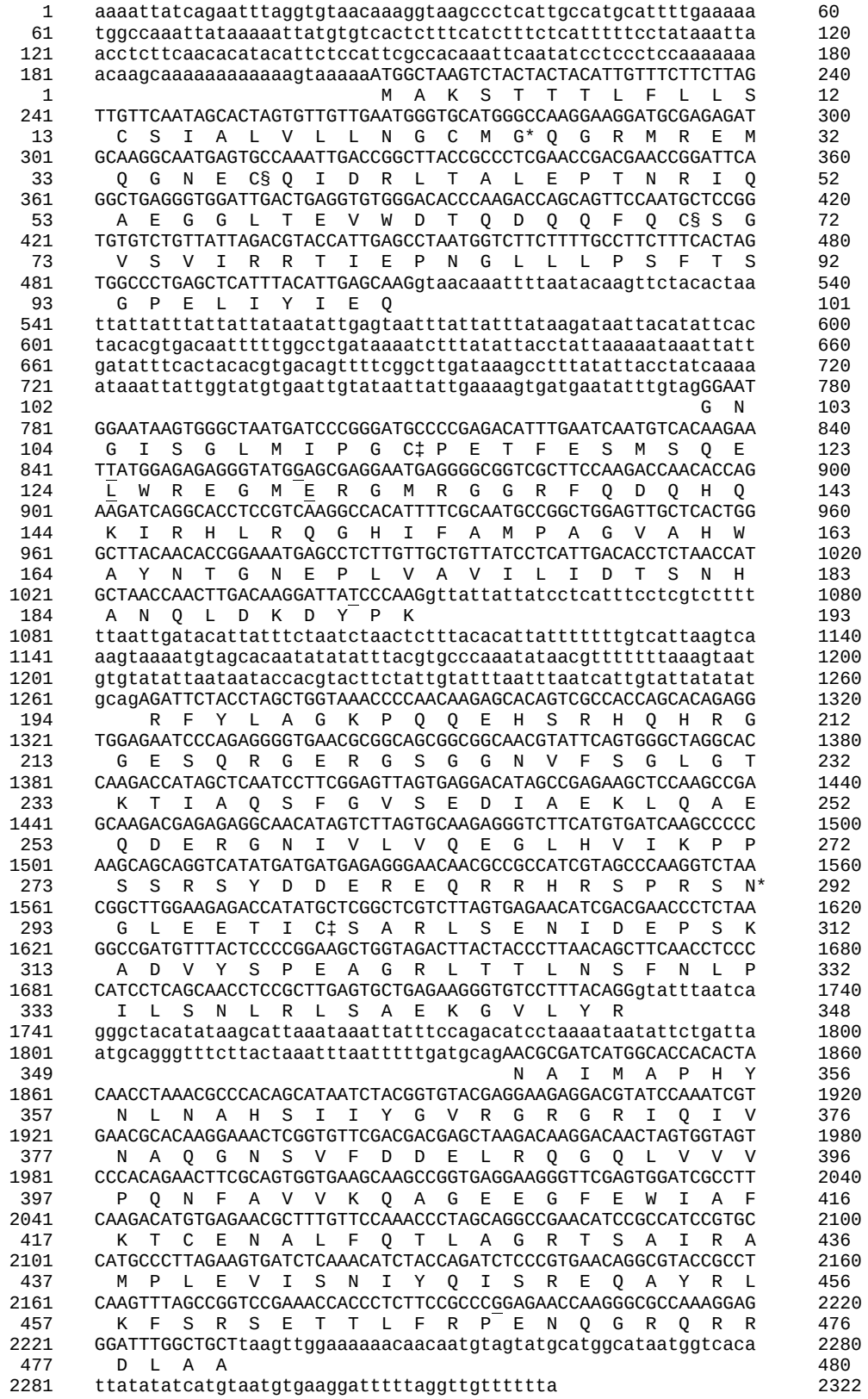

**Figure 3.** The nucleotide and predicted amino acid sequence of the genomic 11S gene from BAC clone 77L9. The *, § and ‡ indicate the same as those listed in Figure 1. Underlined nucleotides and amino acids indicate differences in the coding regions between sequences 77L9 and 17B7. Introns and untranslated regions are indicated in lowercase.

```
   1   ggtaagccctcattgccatgcattttgaagaatggccaaactatcaaaattatgtgtcac     60
  61   tctttcatctttctcatttttcctataaattcacctcttcaacacatacattttccattc     120
 121   gccacaaattaaataccttctccaaaaatacaagaaaaaaaaagagagtaaaaATGGCT     180
   1                                                          M  A       2
 181   AAGTCTACTACTACATTGTTTCTTCTTAGTTGTTCAATAGCACTAGTGTTGTTGAATGGG     240
   3    K  S  T  T  T  L  F  L  L  S  C  S  I  A  L  V  L  L  N  G      22
 241   TGCATGGGCCAAGGAAGGATGCGAGAAATGCAAGGCAATGAGTGCCAAATCGACCGGCTT     300
  23    C  M  G* Q  G  R  M  R  E  M  Q  G  N  E  C§ Q  I  D  R  L      42
 301   ACCGCCCTCGAACCGACGCATCGGATTCAGGCTGAGGGTGGGTTGACTGAGGTGTGGGAC     360
  43    T  A  L  E  P  T  H  R  I  Q  A  E  G  G  L  T  E  V  W  D      62
 361   ACCCAGGACCAGCAGTTCCAATGCTCCGGTGTGTCTGTTATTAGGCGTACCATTGAGCCT     420
  63    T  Q  D  Q  Q  F  Q  C§ S  G  V  S  V  I  R  R  T  I  E  P      82
 421   AATGGTCTTCTTTTGCCTTCTTTCACTAGTGGCCCTGAGCTCATTTACATTGAGCAAGgt     480
  83    N  G  L  L  L  P  S  F  T  S  G  P  E  L  I  Y  I  E  Q       101
 481   aacaaatattaatacaagttctacaaataatcattataataatacggagtaatttataat     540
 541   ttatattataaggtaattaattacgtattgagtacttgtgagatatttcggtctaaaaat     600
 601   aacattcatatgttatattgtgtagattattgaaaagtgatagaatttttgtagGGAATG     660
 102                                                              G  N      103
 661   GAATAAGTGGGCTGATGATCCCGGGGTGCCCCGAGACATTTGAATCAATGTCACAAGAAT     720
 104   G  I  S  G  L  M  I  P  G  C‡ P  E  T  F  E  S  M  S  Q  E      123
 721   CATGGAGAGAGGGTATGGAGCGAGGAATGAGGGGCGGTCGCTTCCAAGACCAACACCAGA     780
 124   S  W  R  E  G  M  E  R  G  M  R  G  G  R  F  Q  D  Q  H  Q      143
 781   AGATCAGGCACCTCCGTCAAGGCCACATTTTCGCAATGCCGGCCGGAGTTGCTCACTGGG     840
 144   K  I  R  H  L  R  Q  G  H  I  F  A  M  P  A  G  V  A  H  W      163
 841   CTTACAACTCCGGAAATGAGCCACTTGTTGCTGTTATCCTCATTGACACCTCTAACCATG     900
 164   A  Y  N  S  G  N  E  P  L  V  A  V  I  L  I  D  T  S  N  H      183
 901   CTAACCAACTTGACAAGGATTACCCCAAGgtacgttacacttcaataatattccttatta     960
 184   A  N  Q  L  D  K  D  Y  P  K                                    193
 961   cctcgtctttttgttacatatttcttatcttatagtctttttattcgttacaaatttaca    1020
1021   attattcttggcaagcgctttctccaatattagactcaaataatgtgtcaaaattaaatg    1080
1081   tagcaaacaataaaagactagagaaaataatgtaacgtttttttaaaatgatgcttacgta    1140
1141   ccaacttcaacaataattaataacaccacttgtgcatttgatcaatgatttatatgcagA    1200
1201   GATTCTACCTAGCTGGTAAACCCCAACAAGAGCACAGTCGCCACCATCACAGGGGTGGAG    1260
 194   R  F  Y  L  A  G  K  P  Q  Q  E  H  S  R  H  H  H  R  G  G      213
1261   AATCCCAGAGGGGTGAACACGGCAGCGACGGCAACGTATTCAGTGGCCTAGACACCAAGT    1320
 214   E  S  Q  R  G  E  H  G  S  D  G  N  V  F  S  G  L  D  T  K      233
1321   CCGTAGCTCAATCCTTCGGAGTTAGTGAGGACATAGCGGAGAAGCTCCAAGCCAAGCAGG    1380
 234   S  V  A  Q  S  F  G  V  S  E  D  I  A  E  K  L  Q  A  K  Q      253
1381   ACGAGAGAGGCAACATAGTCTTAGTGCAAGAGGGTCTTCATGTGATCAAGCCCCCAAGCA    1440
 254   D  E  R  G  N  I  V  L  V  Q  E  G  L  H  V  I  K  P  P  S      273
1441   GCAGGTCATATGATGATGAGAGGGAACAGCGCCATCGTAGCCCAAGGTCTAACGGCTTGG    1500
 274   S  R  S  Y  D  D  E  R  E  Q  R  H  R  S  P  R  S  N* G  L      293
1501   AAGAGACCATATGCTCGGCTCGCCTTAGTGAGAACATCGACGATCCCTCTAAGGCCGATG    1560
 294   E  E  T  I  C‡ S  A  R  L  S  E  N  I  D  D  P  S  K  A  D      313
1561   TTTACTCCCCGGAAGCTGGTAGACTTACCACCCTTAACAGCTTCAACCTCCCCATCCTCA    1620
 314   V  Y  S  P  E  A  G  R  L  T  T  L  N  S  F  N  L  P  I  L      333
1621   GCAACCTCCGCTTGAGTGCTGAGAAGGGTGTCCTCTACAGGgtaattaatcagggctata    1680
 334   S  N  L  R  L  S  A  E  K  G  V  L  Y  R                       347
1681   tataagtattaaatatatatttttccgtatatcctaaaatagtaatctgtcagtagaact    1740
1741   ttaactaaattatttttgatgcagAACGCAATCATGGCACCACACTACAACCTAAACGCC    1800
 348                           N  A  I  M  A  P  H  Y  N  L  N  A      359
1801   CACAGCATAATCTATGGTGTACGAGGACGAGGACGTATCCAAATCGTGAACGCACAAGGA    1860
 360    H  S  I  I  Y  G  V  R  G  R  G  R  I  Q  I  V  N  A  Q  G      379
1861   AACTCAGTGTTTGACGATGAGCTAAGACAAGGACAACTAGTGGTAGTCCCACAGAACTTC    1920
 380    N  S  V  F  D  D  E  L  R  Q  G  Q  L  V  V  V  P  Q  N  F      399
1921   GCAGTGGTGAAGCAAGCCGGTGAGGAAGGATTCGAGTGGATCGCCTTCAAGACATGTGAG    1980
 400    A  V  V  K  Q  A  G  E  E  G  F  E  W  I  A  F  K  T  C  E      419
1981   AACGCTTTGTTCCAAACCCTAGCAGGCCGAACATCCGCCATACGTGCCATGCCCGTAGAA    2040
 420    N  A  L  F  Q  T  L  A  G  R  T  S  A  I  R  A  M  P  V  E      439
2041   GTAATCTCAAACATCTACCAGATCTCCCGCGAACAGGCGTACCGCCTCAAGTTTAGCCGG    2100
 440    V  I  S  N  I  Y  Q  I  S  R  E  Q  A  Y  R  L  K  F  S  R      459
2101   TCCGAAACCACCCTCTTCCGCCCCGAGAACCAAGGGCGCCAAAGGAGGGAAATGGCTGCT    2160
 460    S  E  T  T  L  F  R  P  E  N  Q  G  R  Q  R  R  E  M  A  A      479
2161   taagttgaaaaaacaacaacgtagtaagcatggcataatgatcacgttatatatcatgta    2220
2221   atgtgaaggagttttaggttgtttttttttaaaataattaggatggataacaaaataaaa    2280
2281   ggtagtcctttagaaggtcccttgttttgaaggtgttttgttgagaggattggagctgtt    2340
2341   atgta
```

**Figure 4.** The nucleotide and predicted amino acid sequence of the genomic 11S gene from BAC clone 164F2. Designations are the same as those listed in Figure 3. Underlined nucleotides and amino acids indicate differences between the 164F2 and 8B14 sequences.

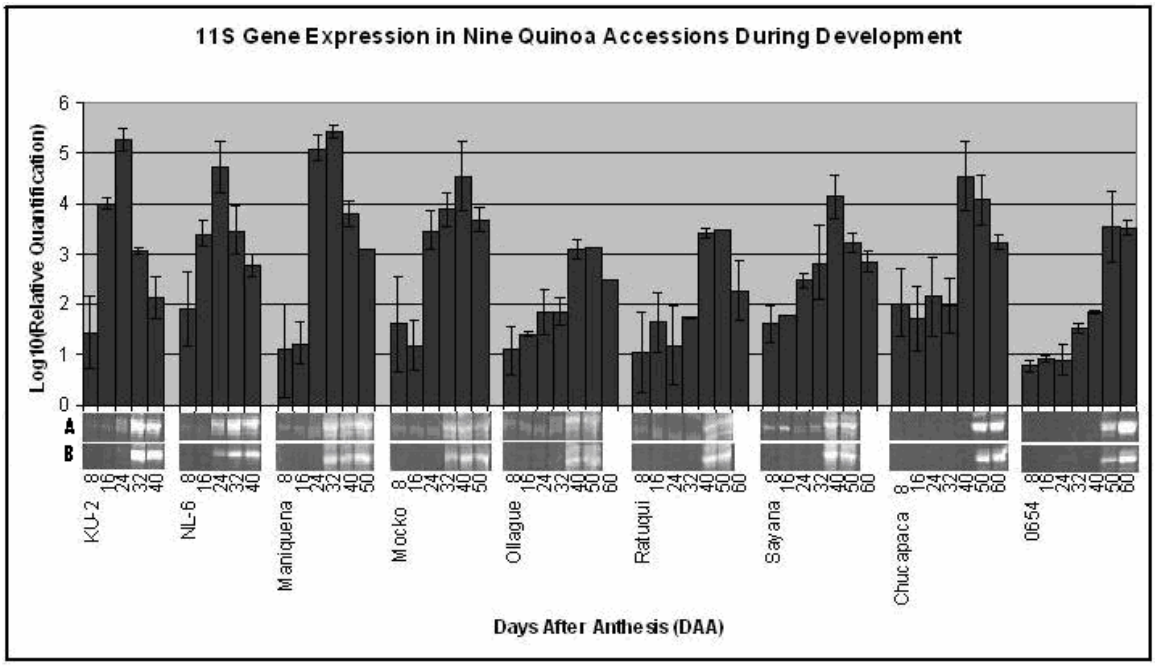


**Figure 5.** 11S gene expression and protein accumulation during the development of quinoa seeds. Quinoa accessions and days after anthesis (DAA) are listed on the horizontal axis. Gene expression was measured in $Log_{10}(RQ)$ units using the GAPDH gene as an endogenous control. The results of the SDS-PAGE analysis are shown below the graph. A represents the acidic subunit of the 11S polypeptide and B represents the basic subunit of the 11S polypeptide.

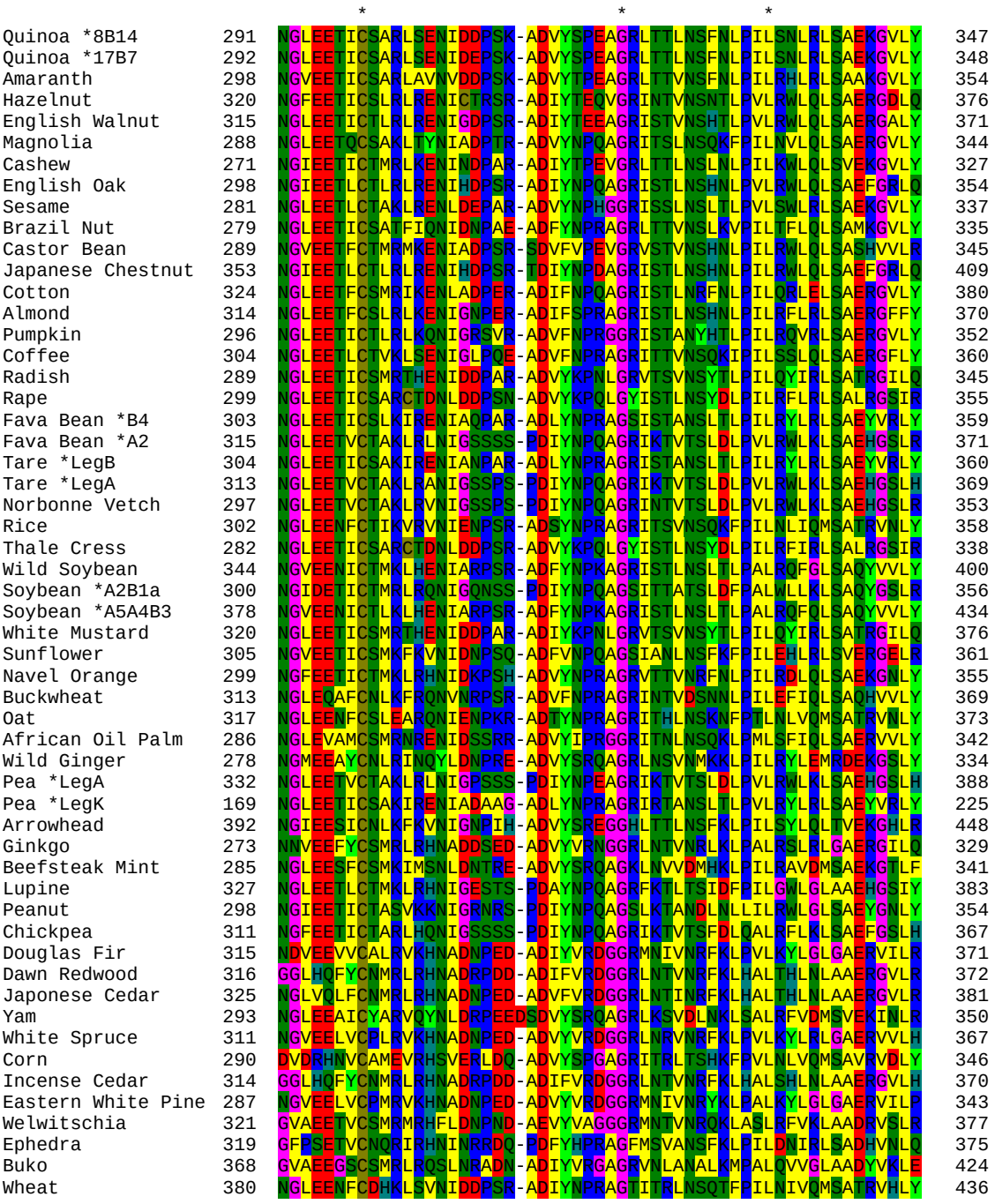


**Figure 6.** Alignment using ClustalW of 55 legumin amino acid sequences from 50 different species. The amino acid alignment was used as a template for the CDS alignment. Only approximately 56 amino acid residues are shown which make up the N-terminal sequence of the basic subunit and in most cases, the C-terminal-asparagine residue of the acidic subunit. The * indicates amino acid residues that are the same in all sequences. The numbers on either side of the alignment indicate amino acid position.

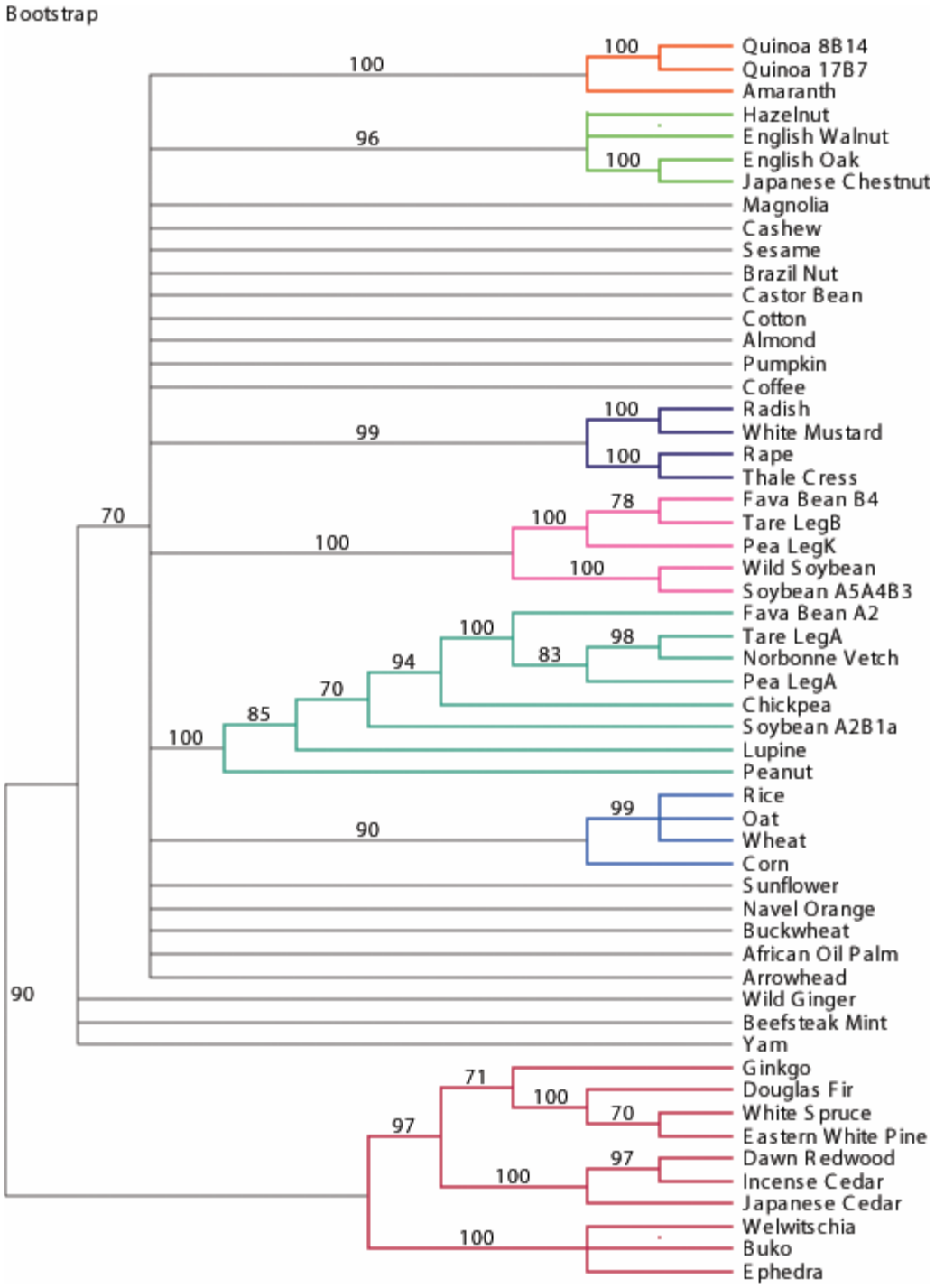


**Figure 7.** Parsimony tree reconstruction of legumin basic subunit coding DNA sequences. Unresolved portions of the tree represent bootstrap values under 70. Colored portions represent monophyletic groups identified by both Bayesian and Parsimony tree reconstructions.

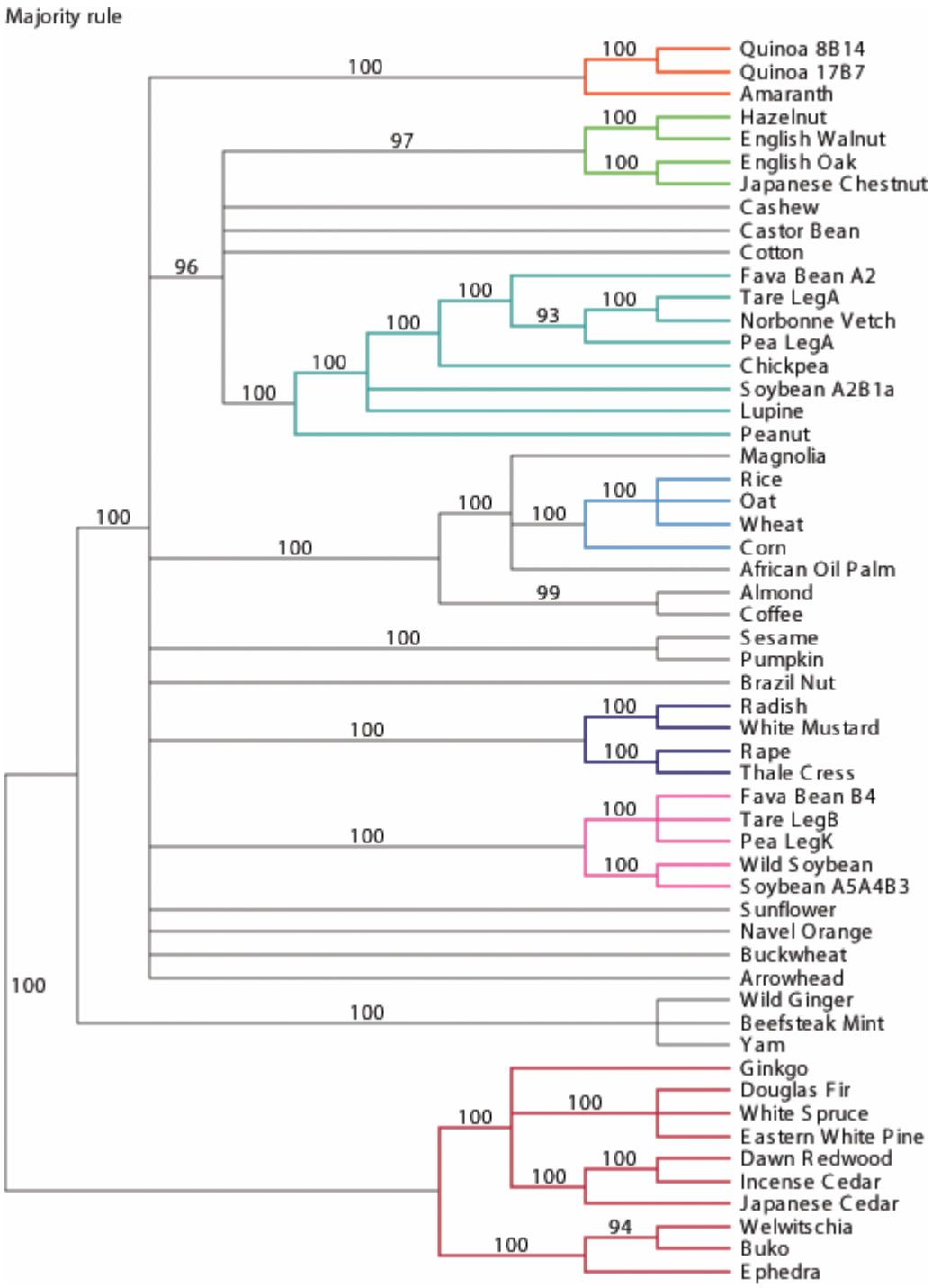


**Figure 8.** Bayesion consensus tree reconstruction of coding DNA sequences of the legumin basic subunit. Unresolved portions of the tree represent posterior probabilities below 90. Colored portions represent monophyletic groups identified by both Bayesian and Parsimony tree reconstructions.

**Table 1.** Amino acid composition of the two loci of the 11S gene in quinoa 8B14 and 17B4 compared to analogous genes in rice, corn and soybean

| Amino Acid Composition | | | | | |
|---|---|---|---|---|---|
| Amino acid | *C. quinoa* 8B14 | *C. quinoa* 17B7 | Amino acid | *C. quinoa* 8B14 | *C. quinoa* 17B7 |
| Ala | 6.39 | 6.59 | Leu* | 7.27 | 7.69 |
| Arg | 8.59 | 9.01 | Lys* | 2.86 | 2.86 |
| Asn | 5.07 | 5.27 | Met* | 2.20 | 1.98 |
| Asp | 3.96 | 3.52 | Phe* | 3.52 | 3.52 |
| Cys* | 1.10 | 1.10 | Pro | 4.41 | 4.40 |
| Glu | 8.37 | 8.35 | Ser | 7.93 | 7.47 |
| Gln | 7.49 | 7.69 | Thr* | 3.74 | 4.18 |
| Gly | 8.81 | 9.23 | Trp* | 0.88 | 0.88 |
| His* | 3.08 | 2.64 | Tyr* | 2.64 | 2.42 |
| Ile* | 6.17 | 6.37 | Val | 5.51 | 4.84 |

*Essential amino acids

**Table 2.** Quinoa accession, location and maturation type for which 11S gene and protein expression data were obtained

| **Accession** | **Location*** | **Maturation Type** |
|---|---|---|
| KU-2 | Coastal | Type I |
| NL-6 | Coastal | Type I |
| Maniqueña | Salares | Type II |
| Mocko | Salares | Type II |
| Ollague | Salares | Type III |
| Ratuqui | Altiplano | Type III |
| Sayaña | Altiplano | Type III |
| Chucapaca | Altiplano | Type IV |
| 0654 | Valley | Type V |

*Valley and Salares are subdivisions of the Altiplano type

**Table 3.** Organisms' common name, species, family, order and GenBank accession number from which the 11S gene cDNA sequences were obtained

| Common Name | Species | Family | Order | Accession Number |
|---|---|---|---|---|
| Arrowhead | Sagittaria sagittifolia | Alismataceae | Alismatales | Y09116. |
| Aftrican Oil Palm | Elaeis guineensis | Arecaceae | Arecales | AF261691. |
| Sunflower | Helianthus annuus | Asteraceae | Asterales | M28832. |
| Thale Cress | Arabidopsis thaliana | Brassicaceae | Brassicales | NM_123779. |
| Rape | Brassica napus | Brassicaceae | Brassicales | M16860. |
| Radish | Raphanus sativus | Brassicaceae | Brassicales | X59808. |
| White Mustard | Sinapsis alba | Brassicaceae | Brassicales | AY846388. |
| Quinoa *17B7 | Chenopodium quinoa | Amaranthaceae | Caryophyllales | AY562549. |
| Quinoa *8B14 | Chenopodium quinoa | Amaranthaceae | Caryophyllales | AY562550. |
| Amaranth | Amaranthus hypochondriacus | Amaranthaceae | Caryophyllales | X82121. |
| Buckwheat | Fagopyrum esculentum | Polygonaceae | Caryophyllales | AF152003. |
| Incense Cedar | Calocedrus decurrens | Cupressaceae | Coniferales | X95540. |
| Japanese Cedar | Cryptomeria japonica | Cupressaceae | Coniferales | X95542. |
| Dawn Redwood | Metasequoia glyptostroboides | Cupressaceae | Coniferales | X95544. |
| White Spruce | Picea glauca | Pinaceae | Coniferales | X63192. |
| Eastern White Pine | Pinus strobus | Pinaceae | Coniferales | Z11486. |
| Douglas Fir | Pseudotsuga menziesii | Pinaceae | Coniferales | L07484. |
| Pumpkin | Cucurbita pepo | Cucurbitaceae | Cucurbitales | M36407. |
| Yam | Dioscorea caucasica | Dioscoreaceae | Dioscoreales | X95510. |
| Ephedra | Ephedra gerardiana | Ephedraceae | Ephedrales | Z50777. |
| Brazil Nut | Bertholletia excelsa | Lecythidaceae | Ericales | AY221641. |
| Peanut | Arachis hypogaea | Fabaceae | Fabales | AF125192. |
| Chickpea | Cicer arietinum | Fabaceae | Fabales | Y15527. |
| Soybean *A2B1a | Glycine max | Fabaceae | Fabales | D00216. |
| Soybean *A5A4B3 | Glycine max | Fabaceae | Fabales | AB195712. |
| Wild Soybean | Glycine soja | Fabaceae | Fabales | X79467. |
| Lupine | Lupinus albus | Fabaceae | Fabales | AJ938034. |
| Pea *LegA | Pisum sativum | Fabaceae | Fabales | AJ132614. |
| Pea *LegK | Pisum sativum | Fabaceae | Fabales | X07015. |
| Fava Bean *A2 | Vicia faba | Fabaceae | Fabales | X55014. |
| Fava Bean *B4 | Vicia faba | Fabaceae | Fabales | X14237. |
| Norbonne Vetch | Vicia narbonensis | Fabaceae | Fabales | Z46803. |
| Tare *LegA | Vicia sativa | Fabaceae | Fabales | Z32835. |
| Tare *LegB | Vicia sativa | Fabaceae | Fabales | Z32796. |
| Hazelnut | Corylus avellana | Betulaceae | Fagales | AF449424. |
| Japanese Chestnut | Castanea crenata | Fagaceae | Fagales | AF525749. |
| English Oak | Quercus robur | Fagaceae | Fagales | X99539. |
| English Walnut | Juglans regia | Juglandaceae | Fagales | AY692446. |
| Coffee | Coffea arabica | Rubiaceae | Gentianales | AF054895. |
| Ginkgo | Ginkgo biloba | Ginkgoaceae | Ginkgoales | Z50778. |
| Buko | Gnetum gnemon | Gnetaceae | Gnetales | Z50779. |
| Beefsteak Mint | Perilla frutescens | Lamiaceae | Lamiales | AF180392. |
| Sesame | Sesamum indicum | Pedaliaceae | Lamiales | AF240004. |
| Magnolia | Magnolia salicifolia | Magnoliaceae | Magnoliales | X82464. |

| | | | | |
|---|---|---|---|---|
| Castor Bean | Ricinus communis | Euphorbiaceae | Malpighiales | AF262998. |
| Cotton | Gossypium hirsutum | Malvaceae | Malvales | M16905. |
| Wild Ginger | Asarum europaeum | Aristolochiaceae | Piperales | X95508. |
| Oat | Avena sativa | Poaceae | Poales | X17637. |
| Rice | Oryza sativa | Poaceae | Poales | XM_464834. |
| Wheat | Triticum aestivum | Poaceae | Poales | S62630. |
| Corn | Zea mays | Poaceae | Poales | AF371279. |
| Almond | Prunus dulcis | Rosaceae | Rosales | X78119. |
| Cashew | Anacardium occidentale | Anacardiaceae | Sapindales | AF453947. |
| Navel Orange | Citrus sinensis | Rutaceae | Sapindales | U38914. |
| Welwitschia | Welwitschia mirabilis | Welwitschiaceae | Welwitschiales | Z50780. |

* Indicates the type of subunit

# Appendix

## Relative quantification of mRNA using one-step real time RT-PCR

Quantitative real-time PCR is a technique capable of recording the progress of the PCR throughout the process and therefore determine quantities of specified starting materials. Reverse transcriptase (RT) real-time PCR is a method used to quantify RNA transcripts and so is useful in gene expression assays. Relative quantification (RQ) determines the change in expression of a target gene sequence in reference to an appropriate endogenous control gene. RQ is capable of determining the initial quantities of specific transcripts without knowing the total initial RNA concentration and without the use of a standard curve that is necessary for absolute quantification.

There are several components that are necessary for an RQ experiment. The target is the sequence under study. An endogenous control is a gene with a constant level of expression in order to normalize the quantification of the target. Housekeeping genes such as glyceraldehyde-3-phosphate dehydrogenase (GAPDH), ribosomal RNA (rRNA) and β-actin are commonly used as endogenous controls due to their stable expression levels. Each sample must include a target and an endogenous control. A calibrator is a sample which serves as a basis for comparison for all the other samples within the experiment. In addition, sample replication is necessary to ensure statistical significance.

## General Outline:

A. RNA Preparation

B. Primers and Probe Selection

C. Multiplexing vs. Singleplexing

D. RQ RTPCR cocktail for multiplexing

E. RQ RTPCR cocktail for singleplexing

F. Creating a RQ Plate Document using Sequence Detection Systems Software

G. Thermocycler conditions

H. Creating an RQ Study

### A. RNA Preparation

1. Collect plant material and immediately freeze in liquid nitrogen.
2. Store plant material at -80°C.
3. Extract RNA using a RNeasy® Plant Mini Kit (Qiagen, Valencia, CA) or appropriate protocol and follow the manufacturer's instructions.
4. Quantify total RNA.
5. Adjust RNA concentration to 50 µg/ml using RNase free water.

    Note: The total RNA concentration may need to be increased or decreased based on the quantity of target mRNA present in the total RNA.

**B. Primers and Probe Selection**

There are several different methods for selecting primers and probes for the target sequence and the endogenous control. Assays-by-Design$^{SM}$ is a service provided by Applied Biosystems (Foster City, CA) in which only a sequence is provided and primer and probe sets are designed, synthesized, and delivered in the appropriate concentrations. Primer Express® Software is designed to assist in selecting primers and probes from a given sequence. Primer and probe sets are necessary for the target and endogenous control sequences.

Note: Multiplexing requires the probe for the target sequence and the probe for the endogenous control to have two different dyes. In singleplexing experiments, the probe for the target sequence and the probe for the endogenous control may have the same dye. In multiplexing, FAM and VIC dyes are commonly used. In singleplexing, the FAM dye may be used for both

**C. Multiplexing vs. Singleplexing**

A PCR experiment can either be conducted as a singleplex in which only a single primer set is present or a multiplex reaction in which two or more primer sets are present within a sample. In a singleplex experiment, the primers and probes specific to the target sequence and those for the endogenous control are in separate wells or tubes while multiple primer sets for single or multiple targets and endogenous controls may be in a single well or tube in a multiplex experiment. Multiplexing requires less sample, less reagents and minimizing pipetting error. However, factors such as the number of targets or the abundance of the endogenous control must be taken into consideration. As the

number of targets increase, the multiplexing method becomes less effective. In addition, the endogenous control must be more abundant than the targets.

*Materials:*

RNA samples

96-well optical reaction plate

Optical adhesive cover

TaqMan® Universal PCR Master Mix with ROX (2X)

20X Assay Mix for target sequence

| | |
|---|---|
| Forward Primer Concentration (μM): | 18μM |
| Reverse Primer Concentration (μM): | 18μM |
| Probe Concentratin (μM): | 5 μM |

20X Assay Mix for endogenous control

| | |
|---|---|
| Forward Primer Concentration (μM): | 18μM |
| Reverse Primer Concentration (μM): | 18μM |
| Probe Concentratin (μM): | 5 μM |

Multiscribe™ Reverse Transcriptase

RNase-free water

*Instructions*

1. In a 96-well optical reaction plate, apply 10μl RNA sample (50μg/ml) to each well in duplicates or triplicates for statistical significance.
2. Prepare cocktail sufficient for the number of wells to be used on ice.

3. Add 15µl of cocktail (see instructions below) to each well.
4. Cover with an optical adhesive cover and ensure wells are sealed to avoid evaporation during the PCR.
5. Centrifuge briefly to ensure sample is at the bottom of each well.
6. Keep plate on ice until ready for use.

RQ RT-PCR cocktail for multiplexing

| One Well Cocktail | Volume (µl) |
|---|---|
| 1. TaqMan® Universal PCR Master Mix with ROX (2X) | 12.500 |
| 2. 20X Assay Mix for target sequence | 0.625 |
| 3. 20X Assay Mix for endogenous control | 0.625 |
| 4. Multiscribe™ Reverse Transcriptase (on ice) | 0.500 |
| 5. RNase-free water | 0.750 |
| Total | 15.000 |

RQ RT-PCR cocktail for singleplexing

Note: Singleplexing experiments require two cocktails to be prepared, one containing the assay mix for the target sequence and the other containing the assay mix for the endogenous control

| One Well Cocktail | Volume (µl) |
|---|---|
| 1. TaqMan® Universal PCR Master Mix with ROX (2X) | 12.50 |
| 2. 20X Assay Mix for target sequence or endogenous control | 1.25 |
| 3. Multiscribe™ Reverse Transcriptase (on ice) | 0.50 |
| 4. RNase-free water | 0.75 |
| Total | 15.00 |

**D. Creating a RQ Plate Document using Sequence Detection Systems Software**

These steps are also illustrated in the Relative Quantification Getting Started Guide provided by Applied Biosystems. Refer to the guide for further details.

1. Create a new RQ Plate document:
    a. Assay: Relative Quantification (ddCt) Plate
    b. Container: 96-Well Clear
    c. Template: Bland Document
    d. Plate Name: Enter plate name
2. Select detectors: Add any dyes that have been used to label probes that will be used in the experiment.
    a. Multiplexing: generally the probe for the target is labeled with FAM dye and the probe for the endogenous control is labeled with VIC® dye
    b. Singleplexing: generally only FAM dye is used to label the probes of both the target and endogenous control.
3. Determine the detectors and tasks of each well.
    a. The task will be either to detect target or endogenous control or both (if multiplexing).
4. Enter sample names using the well inspector (detectors and tasks can also be determined at this step).

**E. Thermocycler Conditions for One-Step RT-PCR**

1. Stage 1 (1 Rep): 48.0°C for 30 minutes

   Note: This stage may be lengthened or shortened depending on the type of reverse transcriptase being used.

2. Stage 2 (1 Rep): 95.0°C for 10 minutes
3. Stage 3 (40 Reps): 95.0°C for 15 seconds

   60.0°C for 1 minute

   Note: More reps may be used, especially if the quantity of target mRNA is small

4. Sample volume: 25µl

**F. Creating an RQ Study**

1. Create a new RQ Study document:
   a. Assay: Relative Quantification (ddCt) Study
   b. Container: 96-Well Clear
   c. Plate Name: Enter plate name
2. Add plates to the study.

   Note: Only plates that have been run under the same conditions will be allowed to be compared in an RQ Study. Also, samples with the same name will be averaged and shown under a single sample name.

3. RQ Study Document:
   a. RQ Detector Grid: shows detectors used in the experiment.
   b. RQ Sample Grid: shows samples associated with selected detectors.

i. Sample summary: includes information about detectors, tasks, and RQ calculations.

ii. Well information: includes information about the location of samples on the plate.

c. RQ Results Panel: contains information about the plates included in the study, the amplification plot, and relative gene expression.

4. Analysis Settings

a. Detector: All

b. Auto Ct

c. Determine which sample will be used as the calibrator.

Note: A calibrator is a sample which serves as a basis for comparison for all the other samples within the experiment.

d. Select the endogenous control detector.

e. Control type: only an option if plates contain multiplexed and singleplexed reactions.

f. RQ Min/Max Confidence (for error bar calculations): 95.00%

g. Select 'Remove Outlier'

h. Click 'OK & Reanalyze'

5. Verify the accuracy of the baseline and threshold.

a. The amplification curve should include a:

i. Plateau phase

ii. Linear phase

iii. Geometric phase

iv. Background

v. Baseline

b. The maximum baseline should be set before the amplification curve.

c. The threshold should be set in the geometric phase of the amplification curve.

d. Adjust the baseline and threshold manually if necessary by selecting Manual Ct under the Analysis Settings.

e. Always reanalyze after making any changes in the analysis settings.

6. Export the RQ Study Data

    a. Under file, select export, then results, then sample summary.

    b. Data can be further analyzed in Microsoft® Office Excel.