2008-02-01

Expression and Evolutionary Relationships of the Chenopodium Quinoa 11S Seed Storage Protein Gene

Mikel Stevens
mikel_stevens@byu.edu

Peter J. Maughan

Daniel J. Fairbanks

Marie R. B. Balzotti

Jennifer N. Thornton

See next page for additional authors

Original Publication Citation

International Journal of Plant Sciences 169.2(Feb 28): 281-291.

Int. J. Plant Sci. 169(2):281–291. 2008.

1058-5893/2008/16902-0007$15.00 DOI: 10.1086/523874

EXPRESSION AND EVOLUTIONARY RELATIONSHIPS OF THE *CHENOPODIUM QUINOA* 11S SEED STORAGE PROTEIN GENE

Marie R. B. Balzotti,* Jennifer N. Thornton,* Peter J. Maughan,* David A. McClellan,† Mikel R. Stevens,* Eric N. Jellen,* Daniel J. Fairbanks,* and Craig E. Coleman[1,]*

*Department of Plant and Wildlife Sciences, Brigham Young University, Provo, Utah 84602, U.S.A.; and †Department of Biology, Brigham Young University, Provo, Utah 84602, U.S.A.

Quinoa (*Chenopodium quinoa* Willd.) is a food crop cultivated by subsistence farmers and commercial growers on the high Andean plateau, primarily in Bolivia, Peru, and Chile. Present interest in quinoa is due to its tolerance of harsh environments and its nutritional value. It is thought that the seed storage proteins of quinoa, particularly the 11S globulins and 2S albumins, are responsible for the relatively high protein content and ideal amino acid balance of the quinoa seed. Here we report the genomic and cDNA sequences for two 11S genes representing two orthologous loci from the quinoa genome. Important features of the genes and the proteins they encode are described on the basis of a comparison with homologous 11S sequences from other plant species. Gene expression and protein accumulation determined via reverse transcriptase real-time PCR and SDS-PAGE analyses are described. Additionally, we report the phylogenetic relationships between quinoa and 49 other species by using the coding DNA sequence for the well-conserved 11S basic subunit.

Keywords: quinoa, RT-PCR, 11S globulin, seed storage protein, gene expression.

Introduction

Quinoa (*Chenopodium quinoa* Willd.) is a dicotyledonous pseudocereal native to the Andean Altiplano (high plains). It is an important cultural and dietary component of the Quechua and Aymara people, who have lived on the Altiplano for centuries (Risi and Galwey 1984). Quinoa thrives in a wide variety of environments extending from mountain valleys and coastal foothills to the semiarid and cool Andean Altiplano. Ca. 12%–14% of the dry mass of quinoa seed is protein (Cardoza and Tapia 1979; Chauhan et al. 1992), and its amino acid composition exceeds the Food and Agriculture Organization's recommendations for human protein consumption (Ruas et al. 1999). The percent protein in quinoa seed is high compared with that in the seeds of cereal crops such as rice, corn, and millet (Charalampopoulos et al. 2002). Seed storage proteins are thought to play a vital role in determining the nutritional value of quinoa seed (Brinegar 1997).

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, which are assembled within the endoplasmic reticulum (Herman and Larkins 1999). Seed storage proteins are generally classified according to their solubility: albumins, globulins, prolamins, and glutelins, which are soluble in water, saline solutions, alcohol, and alkali solutions, respectively (Osborne 1924). They are further classified on the basis of 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. Because legumin-type seed storage proteins vary in size, the 11–12S globulins are referred to collectively in other species as legumins.

The nucleotide sequence of genes encoding legumins has 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), and the crystal structure of an 11S globulin from soybean has been determined (Adachi et al. 2003). The 11S globulin is a hexamer 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 containing acidic and basic subunits as well as a signal peptide responsible for translocation of the precursor into the endoplasmic reticulum (ER). The 11S proprotein assembles as a trimer in the ER and is transported to the vacuole via the Golgi apparatus. In the vacuole, the proprotein is cleaved by asparaginyl endopeptidase to yield the acidic and basic subunits connected by a disulfide linkage. The hexamer assembles directly after this cleavage event (Dickinson et al. 1989).

The origin and early evolution of legumin structure has been discussed by Shutov and Bäumlein (1999), who describe legumin evolution from a germinlike ancestor. After divergence of seed storage proteins into vicilin and legumin classes, each independently acquired their storage-related properties. Seed storage proteins and analogous proteins seem to have an ancient history, as they are present in the early progenitors of plants, including mosses and fungi (Shutov et al. 2003). Thus, these gene sequences may prove useful as phylogenetic markers in molecular evolution studies and may provide insight into

[1] Author for correspondence; e-mail: craig_coleman@byu.edu.

Manuscript received July 2007; revised manuscript received September 2007.

evolutionary modifications under the selective constraints of either better storage capabilities or degradation. The study of conserved and variable regions and how they relate to structure and metabolism may provide valuable information about 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 mass of 320 kDa. The acidic and basic polypeptides have average relative molecular masses of 35.5 and 22.5 kDa, respectively (Brinegar and Goundan 1993). Here we report the genomic and cDNA sequences of two orthologous 11S genes from quinoa. 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 the homologous seed storage proteins of diverse species. We discuss legumin molecular evolution in light of the results of our analysis.

Material and Methods

Plant Materials

Quinoa Real was grown under greenhouse conditions at Brigham Young University. Tissue for use in DNA extractions was collected from the first two true leaves of 2- to 4-wk-old plants. Harvested tissue was immediately frozen and stored at −80°C. Quinoa accessions NL-6, KU-2, Maniqueña, Mocko, Ollague, Ratuqui, Sayaña, Chucapaca, and 0654 were planted simultaneously and grown under greenhouse conditions. Temperatures were 20°–25°C during the day and 15°–18°C at night, and plants were exposed to 12 h of daylight. Seed was collected at 8-d intervals after anthesis until maturity from each accession and was immediately frozen and stored at −80°C.

cDNA and Genomic Sequencing

Genomic DNA was extracted from quinoa 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 was 5′-TGTKKGCGTTKAGGTTSYAGTG-3′. The gene fragment was used as a probe to screen the quinoa developing seed cDNA library reported by Coles et al. (2005). Positive clones were selected and their identities confirmed by Southern blotting. Two clones homologous to 11S genes from other species were isolated from the cDNA library and have been assigned GenBank accession numbers AY562549 and AY562550. A quinoa bacterial artificial chromosome (BAC) library was screened as previously described, and a second nonoverlapping clone was identified in addition to the clone already reported by Stevens et al. (2006). The 11S genomic and cDNA inserts within positively identified clones were sequenced by primer walking at the Brigham Young University DNA Sequencing Center (Provo, UT).

RNA Extraction and Relative Quantification

Seeds were ground to a fine powder in liquid nitrogen, and total RNA was extracted using an RNeasy Plant Mini Kit (Qiagen, Valencia, CA). The RNA was quantified using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, 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, we quantified multiplexed RNA samples by using a TaqMan One-Step RT-PCR Master Mix Reagents Kit on a 7300 Real-Time PCR System (Applied Biosystems, Foster City, CA). Each sample contained 0.5 μg of total RNA. The samples were held at 48°C for 30 min, heated to 95°C and held for 10 min, followed by 40 repetitions of temperatures at 95°C for 15 s and 60°C for 60 s.

Globulin Extraction, Quantification, and Separation by SDS-PAGE

Fifty milligrams of seed collected at 8-d intervals after anthesis was ground to a fine powder, and 500 μL of water was added to each sample. The samples were shaken at 4°C for more than 1 h and centrifuged at 13,000 *g*, and the supernatant containing the albumin fraction was removed. The pellet was washed twice, using the same procedure described above, and dried. To each pellet, 400 μL 0.5 M NaCl/50 mM Tris-HCl, pH 8.0, was added. The suspension was shaken at 4°C for more than 1 h and was centrifuged at 13,000 *g*, and the supernatant containing the globulin fraction was collected into separate tubes. Total protein in the globulin fraction was quantified using a BCA Protein Assay Kit (Pierce, Rockford, IL) on a NanoDrop ND-1000 spectrophotometer. The samples were diluted 1 : 2 with Laemmli Sample Buffer 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) 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 the GenBank accession numbers for their legumin sequences are listed in table 2. Coding DNA sequences were downloaded from GenBank. Using the amino acid sequence of the 11S seed storage protein from *Chenopodium quinoa* (GenBank accession no. AAS67037) as a query sequence, we preformed a BLASTp search (Altschul et al. 1990). Only legumin sequences from different plant species with E values less than 1e-10 were chosen, as to include sequences substantially similar and likely related by descent to the 11S gene in quinoa. One coding DNA legumin sequence from each species, except for quinoa and species from the Fabaceae family, was selected

on the basis of similarity to the quinoa 11S sequence. Two major legumin genes are present in many species in the Fabaceae family (Yagasaki et al. 1997), and sequence information for both genes from these species, if available, was included in the analyses. The two quinoa sequences included in the analysis corresponded to the coding DNA sequences of cDNA clones 17B7 and 8B14 (GenBank accession nos. 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 by using MEGA3. Amino acid sequences were aligned with CLUSTAL W using the BLOSUM matrix and used as a template for the nucleotide alignment. CLUSTAL W performs multiple sequence alignments of nucleotide or protein sequences by 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 used to select the most appropriate likelihood-based nucleotide substitution model.

Distance, maximum parsimony, and Bayesian analyses are three kinds of tree reconstruction methods that rely 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 on 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 on the basis of the model and the sequence alignment. Periodic sampling, in this case every 1000 tree generations, allowed us to compute a consensus tree by using the software package PAUP* (Swofford 2001).

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 does not 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. Bayesian consensus tree reconstruction was accomplished using MrBayes, version 3.0b4 (Ronquist and Huelsenbeck 2003), with default values as the starting parameters. Five million tree generations were conducted, and trees were saved every one-thousandth 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).

Results

Isolation and Characterization of the Quinoa 11S Genes

Two BAC clones, representing distinct 11S seed storage globulins from quinoa, have been identified and characterized. The nucleotide sequences of the coding and flanking regions of both genes were obtained directly from the BAC clones by using a primer walking strategy, and these have been assigned GenBank accession numbers DQ917482 and DQ917483. Coding and flanking regions of BAC clone DQ917482 are 99.5% identical to those of cDNA clone AY562549; likewise, the coding and flanking regions of BAC clone DQ917483 are 99.8% identical to those of cDNA clone AY562550, suggesting that the cDNAs represent transcripts of the two 11S gene loci. We have named the two loci 11SA and 11SB, respectively. A comparison between 11SA and 11SB shows 97% homology in the coding regions and 92% homology in the untranslated regions. Each gene includes three introns whose positions are conserved between the two genes. The 11SA gene encodes a polypeptide of 480 amino acids, whereas the 11SB gene encodes a polypeptide of 479 amino acids.

An alignment of the amino acid sequences of the two quinoa 11S proteins clearly shows that the quinoa 11S proteins are similar to the 11S seed storage proteins found in other species (fig. 1). The first 25 hydrophobic amino acid residues are typical of a signal peptide found in other legumin-like seed storage proteins. The conserved asparaginyl endopeptidase recognition site (between asparagine 292 and glycine 293 in 11SA and between asparagine 291 and glycine 292 in 11SB) at which the precursor polypeptide is separated into acidic and basic subunits is highlighted in figure 1. Cleavage at this conserved site yields acidic subunits with 266 or 267 amino acids and basic subunits with 188 amino acids. As shown in figure 1, there are four conserved cysteine residues involved in disulfide bond formation in 11S proteins of other species.

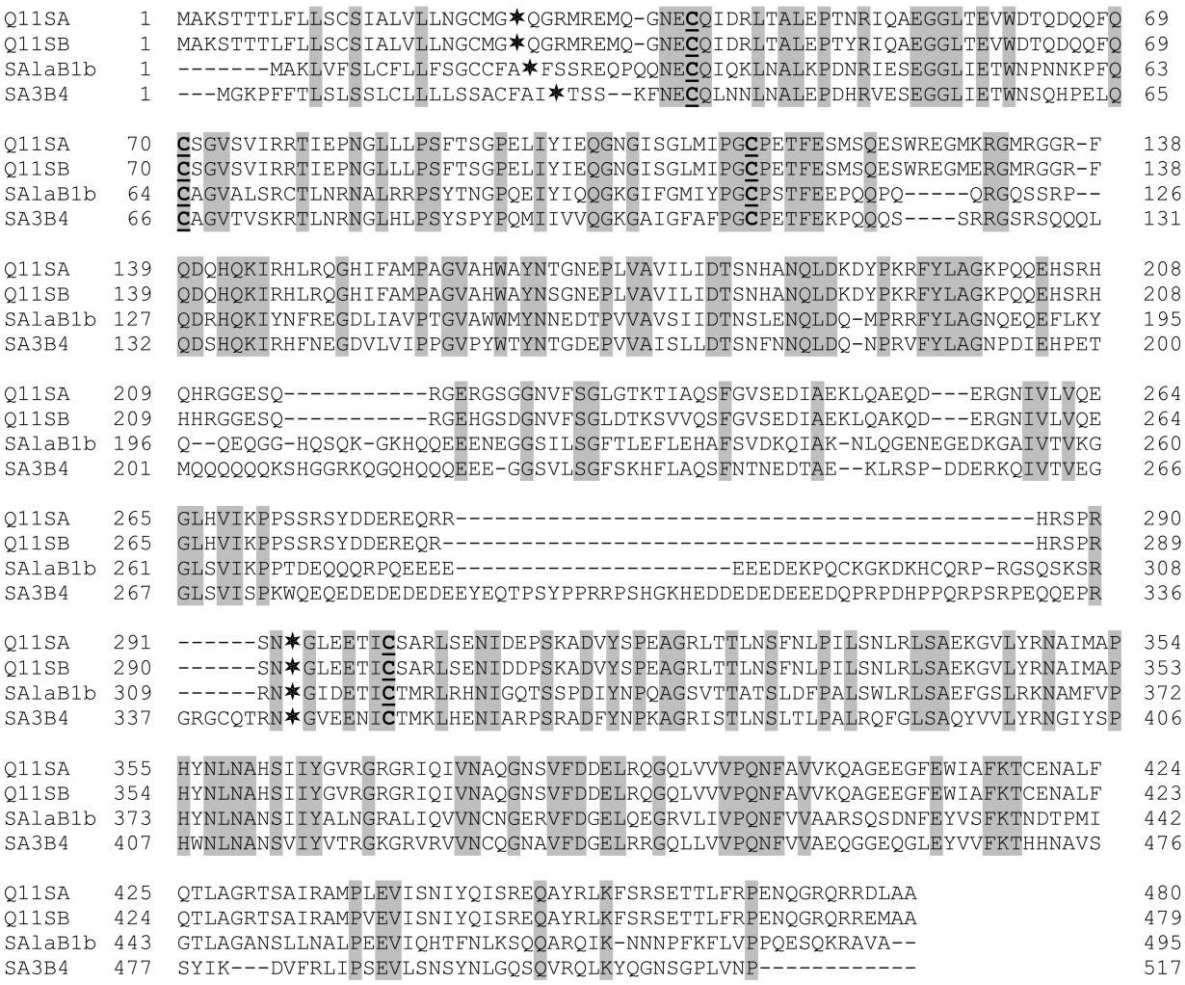

Fig. 1 Alignment between the 11SA (AAS67036) and 11SB (AAS67037) amino acid sequences from quinoa and between the AlaBlb (BAC78522) and A3B4 (BAB15802) amino acid sequences from soybean. Shaded portions represent identical amino acids between the four sequences. The asterisk represents cleavage sites. Underlined cysteine residues represent those involved with disulfide bond formation (Adachi et al. 2003).

11S Gene Expression and Protein Accumulation

To determine the 11S gene expression profile, we collected quinoa seeds at 8- and 10-d intervals after anthesis until desiccation from nine accessions: NL-6, KU-2, Mocko, Maniqueña, Ollague, Ratuqui, Sayaña, Chucapaca, and 0654. Accessions were chosen on the basis of geographic origin (ecotype) and maturation rate (table 1). Altiplano ecotypes originate either in the Salares (areas containing salt flats) or in valley regions, while coastal ecotypes are native to the Chilean lowlands. Total RNA was extracted and quantified from each accession, and 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. Relative amounts of 11S globulin protein were assayed via SDS-PAGE. Figure 2 shows the results of the 11S mRNA and globulin quantification at specific days after anthesis (DAA). The bars in the graph (fig. 2) depict the relative quantity of globulin mRNA at each developmental time point, whereas the bands below the bars are a qualitative representation of globulin protein as seen on SDS-polyacrylamide gels. In general, only minor amounts of 11S mRNA transcripts were detected during early seed development, followed by a substantial accumulation of 11S protein concurrent with peak gene expression levels. As seeds became mature, expression decreased and protein accumulation remained high.

On the basis of days to desiccation and 11S RNA and protein accumulation patterns, five classes of maturation rates were identified and are listed together with accession names and geographic locations in table 1. Type I consists of two Chilean lowland accessions (KU-2 and NL-6) that showed peak gene expression and 11S globulin accumulation at 24 DAA and maturity between 32 and 40 DAA. Type II consists of Maniqueña and Mocko, two Salares Altiplano ecotypes that showed peak gene expression at 24 DAA, detectable 11S globulin quantity occurring at 32 DAA, and maturity achieved between 40 and 50 DAA. Ollague (Salares Altiplano), Ratuqui, and Sayaña (Northern Altiplano) were included as type III accessions and showed peak 11S transcript levels and substantial

Table 1

Quinoa Accession, Location, and Maturation Type for Which 11S Gene and Protein Expression Data Were Obtained

Accession	Ecotype[a]	Maturation type
KU-2	Coastal	I
NL-6	Coastal	I
Maniqueña	Salares	II
Mocko	Salares	II
Ollague	Salares	III
Ratuqui	Altiplano	III
Sayaña	Altiplano	III
Chucapaca	Altiplano	IV
0654	Valley	V

[a] Valley and Salares are subdivisions of the Altiplano type.

11S globulin accumulation at 40 DAA and reached maturity between 50 and 60 DAA. Chucapaca (Salares Altiplano ecotype) and 0654 (Peruvian valley ecotype) were included as types IV and V, respectively. Both showed detectable 11S protein quantities at 50 DAA; however, Chucapaca reached maturity at ca. 60 DAA, and 11S transcript levels reached a maximum of 40 DAA, while 0654 reached maturity significantly later (between 60 and 70 DAA) and showed maximal 11S transcript levels at 50 DAA. As shown in figure 2, 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

A comparison of 11S globulins from quinoa and 49 other species was performed to determine the phylogenetic relationship of the quinoa protein with homologous proteins from other plant species. 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 used in the phylogenetic analysis are presented in table 2. Alignments were created from the whole or portions of the legumin sequence. Using distance-based and maximum parsimony methods for tree reconstruction in MEGA3, we evaluated preliminary trees to assess the quality of the alignments. An amino acid alignment generated from the basic subunit sequence of the legumin gene resulted in detectable amino acid conservation, consistency between trees generated by the aforementioned methods, and taxonomic clustering. It was therefore selected for further analysis.

A second phylogenetic analysis of the basic subunits was performed using the Bayesian algorithm in MrBayes. In order to do this appropriately, we used MODELTEST to determine the likelihood model with the fewest number of parameters that failed to differ from the most general model. This analysis resulted in the identification of the GTR model with measures of invariable sites (*I*) and rate heterogeneity (*G*) as the most appropriate model for these data. Bayesian approaches for phylogeny reconstruction do not initially assume specific parameter values but iterate through the process of calculating trees and then parameter values from those trees until parameters converge on values that are emergent from the data. For this analysis, we let the algorithm iterate through 5 million tree generations.

Figures 3 and 4 show the parsimony and Bayesian tree reconstructions, respectively. In both reconstructions, with posterior probabilities and bootstrap values of 100, the two

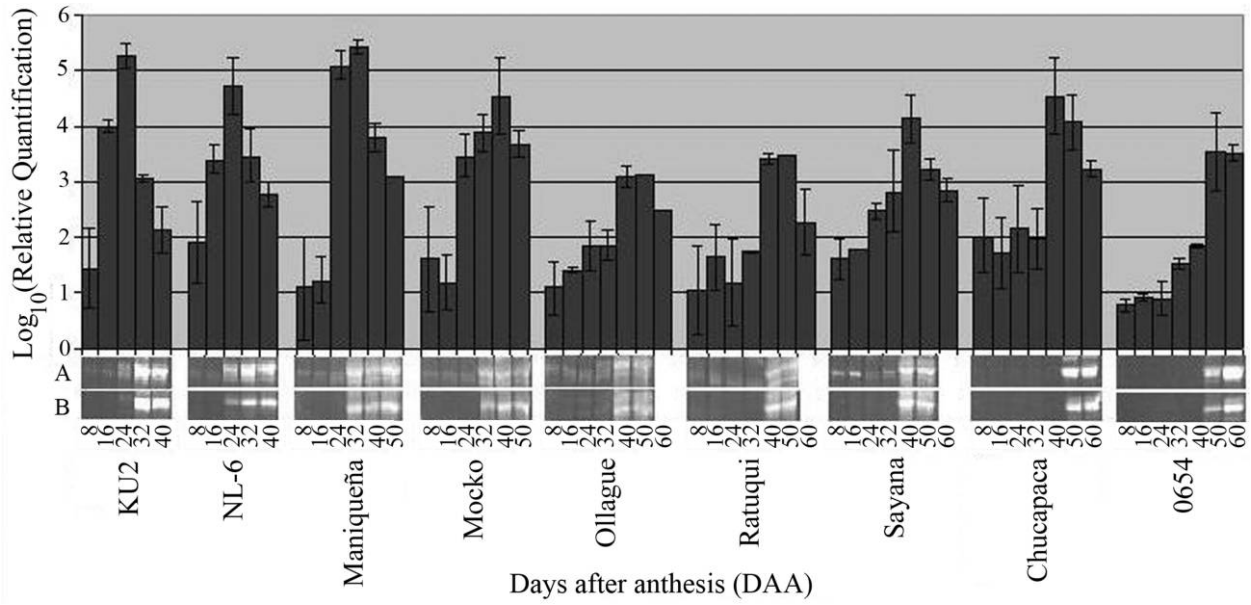

Fig. 2 11S gene expression and protein accumulation during the development of quinoa seeds. Quinoa accessions and days after anthesis are listed on the horizontal axis. Gene expression was measured in $\log_{10}$(relative quantification) units using the *GAPDH* gene as an endogenous control. The results of the SDS-PAGE analysis are shown below the graph. *A*, Acidic subunit of the 11S polypeptide; *B*, basic subunit of the 11S polypeptide.

Table 2

Organism Common Name, Species, Family, Order, and GenBank Accession Number from Which the 11S Gene cDNA Sequences Were Obtained

Common name	Species	Family	Order	Accession no.
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[a]	*Chenopodium quinoa*	Amaranthaceae	Caryophyllales	AY562549
Quinoa 8B14[a]	*C. 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[a]	*Glycine max*	Fabaceae	Fabales	D00216
Soybean A5A4B3[a]	*G. max*	Fabaceae	Fabales	AB195712
Wild soybean	*Glycine soja*	Fabaceae	Fabales	X79467
Lupine	*Lupinus albus*	Fabaceae	Fabales	AJ938034
Pea LegA[a]	*Pisum sativum*	Fabaceae	Fabales	AJ132614
Pea LegK[a]	*P. sativum*	Fabaceae	Fabales	X07015
Fava bean A2[a]	*Vicia faba*	Fabaceae	Fabales	X55014
Fava bean B4[a]	*V. faba*	Fabaceae	Fabales	X14237
Norbonne vetch	*Vicia narbonensis*	Fabaceae	Fabales	Z46803
Tare LegA[a]	*Vicia sativa*	Fabaceae	Fabales	Z32835
Tare LegB[a]	*V. 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

[a] Type of subunit.

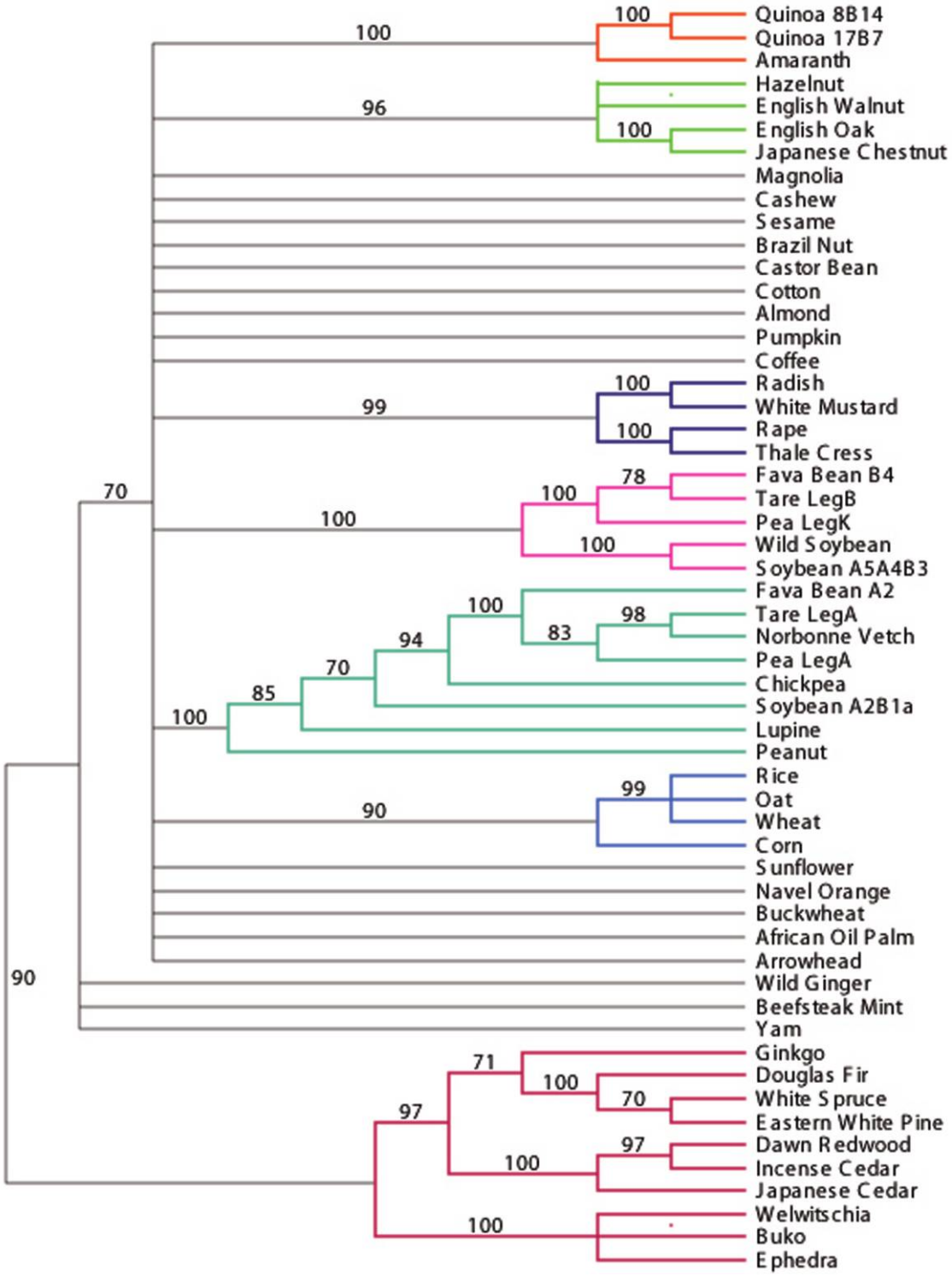

Fig. 3 Parsimony tree reconstruction of legumin basic subunit coding DNA sequences. Unresolved portions of the tree represent bootstrap values under 70. Colored portions represent monophyletic groups identified by both Bayesian and parsimony tree reconstructions.

quinoa 11S sequences form a monophyletic group with an *Amaranthus hypochondriacus* 11S sequence (GenBank accession X82121), the only other reported sequence from a species within the Amaranthaceae family. On the basis of amino acid alignments, the 11S basic subunit shows more than 74% sequence identity between amaranth and quinoa. Other monophyletic and well-resolved taxonomic groups represented in both trees include taxa from the Poaceae, Brassicaceae,

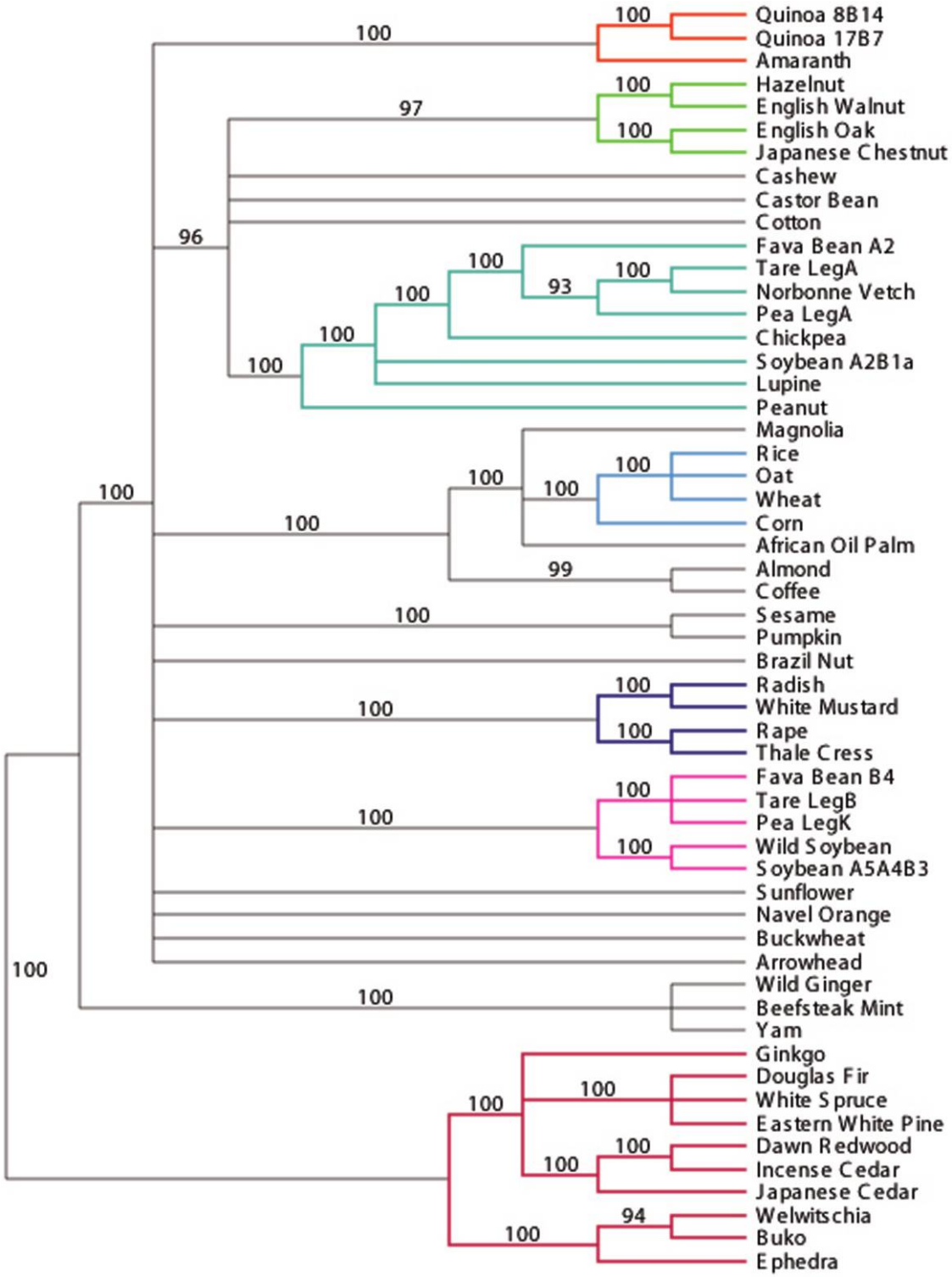

Fig. 4 Bayesian consensus tree reconstruction of coding DNA sequences of the legumin basic subunit. Unresolved portions of the tree represent posterior probabilities below 90. Colored portions represent monophyletic groups identified by both Bayesian and parsimony tree reconstructions.

Cupressaceae, and Pinaceae families, the Fagales order, and the clade for gymnosperms. The sequences from members of the Fabaceae family form two monophyletic groups, each containing 11S sequences derived from different subfamilies. The phylogenetic relationships between the gymnosperm outgroup taxa are well resolved in both trees and reveal a clear bifurcation of angiosperm and gymnosperm legumin gene lineages.

Discussion

11S Gene Sequences

BAC clones and cDNAs representing two quinoa genetic loci, 11SA and 11SB, that encode an 11S legumin-like seed storage protein have been identified and characterized. The cDNAs for 11SA and 11SB have 99.7% and 99.9% sequence identity with their corresponding BAC clone sequences. The small sequence divergence observed between the cDNAs and the BACs is most likely due to sequence heterogeneity in quinoa Real, which was used to construct both the BAC and the cDNA libraries.

On the basis of the deduced amino acid sequence analysis, the domains and residues required for processing and assembly are conserved in the quinoa 11S preproprotein. The first 25 amino acid residues constitute a putative signal peptide responsible for localizing the proprotein to the lumen of the ER. As reported by Brinegar and Goundan (1993), the first 21 amino acids of the N-terminus of the quinoa 11S basic subunit are GLEETICSARLSENIDDPSKA, suggesting that the asparaginyl endopeptidase cleavage site is between Asp292 and Gly293 in 11SA and between Asp291 and Gly292 in 11SB. The position of the cleavage site and the N-terminal sequence of the basic subunit are highly conserved among 11S proteins (Utsumi 1992). Glycinin, the legumin-like protein from soybean, has four major protein binding domains, as revealed by its crystal structure (Adachi et al. 2003). These domains are also found in the quinoa 11S protein, suggesting that it has a hexamer quaternary structure similar to that of glycinin. Four cysteine residues that participate in disulfide bond formation in other species are found in the quinoa sequences (fig. 1). The first two cysteine residues (Cys37 and Cys70) form an intrachain disulfide bridge within the acidic subunit, and the second two cysteine residues (Cys113 and Cys299/300) form an interchain disulfide bridge between the acidic and basic subunits (Utsumi 1992).

The quinoa 11S amino acid sequence is well balanced and shows a high level of essential amino acids with respect to human nutrition. Typical of 11S storage proteins, the quinoa protein is rich in glutamate/glutamine, aspartate/asparagine, arginine, serine, leucine, and glycine (Brinegar and Goundan 1993). We note, however, that the 11S gene is only one of two major seed storage proteins in quinoa seed and that it is the combined contribution of the 11S and 2S seed proteins that produce quinoa's unique seed protein quantity and composition. Thus, the cloning, sequencing, and expression characterization of the 2S albumin remains an important goal to complete our understanding of how quinoa seeds 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 indicate that the accumulation of seed storage protein is correlated to maturation rate. For example, 11S mRNA begins peaking at 16 DAA and 11S protein at 24 DAA in the early-maturing varieties KU-2 and NL-6, whereas similar increases in mRNA and protein accumulation are not seen in late-maturing varieties until at least 24 and 32 DAA, respectively. In extreme cases (i.e., Chucapaca and 0654), the 11S protein is not detected until 50 DAA. These data suggest that high expression of 11S mRNA occurs, on average, during late maturation, consistent with the pattern of storage protein (including 11S) accumulation during seed development in other plant species (Nakamura et al. 2004). While analyzing plants grown in the greenhouse is important and informative, we recognize that future field-based experiments are needed to confirm our observations. Furthermore, a quantitative study of total protein content per seed is needed to determine whether the timing of gene expression has an effect on the overall quantity of seed protein in quinoa.

Phylogenetic Relationships between Legumins of Various Species

The legumin coding DNA sequence of the basic subunit was used 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 amino acids involved in the recognition of the peptide cleavage site between the acidic and basic subunits and a cysteine residue involved in disulfide bond formation (Adachi et al. 2003). Indeed, several researchers have used the legumin basic subunit sequence or portions of the basic subunit for studies involving the molecular evolution of seed storage proteins (Fischer et al. 1996; Häger and Wind 1997; Shutov and Bäumlein 1999).

The maximum parsimony and Bayesian analyses used in the phylogenetic reconstructions reported here (figs. 3 and 4, respectively) use very different assumptions. 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. Both phylogenetic trees were well resolved between closely related taxa, and the placement of the quinoa 11S sequences with amaranth exhibited posterior probabilities and bootstrap values of 100.

In both analyses, all monophyletic groups were equally supported; however, the data suggest that the Bayesian analysis was better suited for resolving relationships between more distantly related taxa than was the parsimony method. Indeed, most unresolved portions of the Bayesian reconstruction angiosperm backbone had posterior probabilities just under 90. Third and fourth nodes with posterior probabilities of 83 and 88, respectively, were present following the bifurcation of lower angiosperm legumin sequences in ginger, mint, and yam. The third node represented the divergence in a sequence from arrowhead, while 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 one composed of sequences from sesame, pumpkin, orange, and buckwheat. Although these relationships are

based on posterior probabilities under 90, they are still well supported by the majority of the data.

Following the divergence between gymnosperms and angiosperms, there appears to have been a sudden increase in sequence divergence among the legumin genes as evidenced 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 (Domoney et al. 1986; Nielsen et al. 1989; Heim et al. 1994), brassicas (Pang et al. 1988; Breen and Crouch 1992; Depigny-This et al. 1992), and cereals (Shotwell et al. 1988; Okita et al. 1989). Multigene families have also been identified in gymnosperms (Häger et al. 1995; Wind and Häger 1996; Häger and Wind 1997). Interestingly, these gymnosperm genes have a much higher degree of sequence homology with each other than do the legumin gene sequences of the angiosperms. We also note that 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 because of sequence dissimilarity.

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 of 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 evolutionary points, it is difficult to know whether homologous genes are being compared, and thus, taxa may be grouped inappropriately. However, it appears that there is sufficient phylogenetic signal, even from diverse species, to support future studies investigating the molecular evolution of legumin genes in angiosperms.

Quinoa is a putative allotetraploid, and we postulate that the two copies of the 11S gene may be located on homeologous chromosomes and may represent genes present in the diploid ancestors of quinoa. Although the sequences for both loci are similar, transcript levels from each may not be equal. Alternatively, it is also possible that the unusually high protein content of quinoa is a direct result of tetraploidization and the unaltered expression of both genes because polyploidy generally increases gene expression levels overall (Osborn et al. 2003). Indeed, seeds containing higher protein reserves for germination and the developing seedling may have been a selective force in the tetraploidization of quinoa.

Acknowledgments

This research was supported by the McKnight Foundation Collaborative Crop Research Program and the Holmes Family Foundation.

Literature Cited

Adachi M, J Kanamori, T Masuda, K Yagasaki, K Kitamura, B Mikami, S Utsumi 2003 Crystal structure of soybean 11S globulin: glycinin A3B4 homohexamer. Proc Natl Acad Sci USA 100:7395–7400.

Altschul SF, W Gish, W Miller, EW Meyers, DJ Lipman 1990 Basic local alignment search tool. J Mol Biol 215:403–410.

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

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

Brinegar C 1997 The seed storage proteins of quinoa. Pages 109–115 *in* S Damodaran, ed. Food proteins and lipids. Vol 415. Advances in experimental medicine and biology. Plenum, New York.

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

Cardoza A, ME Tapia 1979 Valor nutritivo. Pages 149–192 *in* ME Tapia, ed. Quinua y kañiwa cultivos Andinos: serie libros y materials educativos. No 49. Instituto Interamericano de Ciencias Agrícolas, Bogotá.

Charalampopoulos D, R Wang, SS Pandiella, C Webb 2002 Application of cereals and cereal components in functional foods: a review. Int J Food Microbiol 79:131–141.

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

Coles ND, CE Coleman, SA Christensen, EN Jellen, MR Stevens, A Bonifacio, JA Rojas-Beltran, DJ Fairbanks, PJ Maughan 2005 Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. Plant Sci 168:439–447.

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

Dickinson CD, EHA Hussein, NC Nielsen 1989 Role of posttranslational cleavage in glycinin assembly. Plant Cell 1:459–469.

Domoney C, D Barker, R Casey 1986 The complete deduced amino acid sequences of legumin β-polypeptides from different genetic loci in *Pisum*. Plant Mol Biol 7:467–474.

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

Häger K-P, H Braun, A Czihal, B Muller, H Bäumlein 1995 Evolution of seed storage protein genes: legumin genes of *Ginkgo biloba*. J Mol Evol 41:457–466.

Häger K-P, C Wind 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 U, H Bäumlein, U Wobus 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, BA Larkins 1999 Protein storage bodies and vacuoles. Plant Cell 11:601–614.

Kumar S, K Tamura, M Nei 2004 MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5:150–163.

Momma T, T Negoro, K Udaka, C Fukazawa 1985 A complete cDNA coding for the sequence of glycinin A_2B_{1a} subunit precursor. FEBS Lett 188:117–122.

Nakamura S, A Ikegami, M Mizuno, F Yagi, K Nomura 2004 The expression profile of lectin differs from that of seed storage proteins in *Castanea crenata* trees. Biosci Biotechnol Biochem 68:1698–1705.

Nielsen NC, CD Dickinson, TJ Cho, VH Thanh, BJ Scallon, RL Fischer, TL Sims, GN Drews, RB Goldberg 1989 Characterization of the glycinin gene gamily in soybean. Plant Cell 1:313–328.

Okita TW, YS Hwang, J Hnilo, WT Kim, AP Aryan, R Larson, HB Krishnan 1989 Structure and expression of the rice glutelin multigene family. J Biol Chem 264:12573–12581.

Osborn TC, JC Pires, JA Birchler, DL Auger, ZJ Chen, H-S Lee, L Comai, et al 2003 Understanding mechanisms of novel gene expression in polyploids. Trends Genet 19:141–147.

Osborne TB 1924 The vegetable proteins. Longmans, Green, London.

Pang PP, RE Pruitt, EM Meyerowitz 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, KA Crandall 1998 MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818.

Risi JC, NW Galwey 1984 The *Chenopodium* grains of the Andes: Inca crops for modern agriculture. Adv Appl Biol 10:145–216.

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

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

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

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

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

Shutov AD, H Bäumlein 1999 Origin and evolution of seed storage globulins. Pages 543–561 *in* PR Shewry, R Casey, eds. Seed proteins. Kluwer Academic, Dordrecht.

Shutov AD, H Bäumlein, FR Blattner, K Muntz 2003 Storage and mobilization as antagonistic functional constraints on seed storage globulin evolution. J Exp Bot 54:1645–1654.

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

Stevens MR, CE Coleman, S Parkinson, PJ Maughan, HB Zhang, MRB Balzotti, D Kooyman, et al 2006 Construction of a quinoa (*Chenopodium quinoa* Willd.) BAC library and its use in identifying genes encoding seed storage proteins. Theor Appl Genet 112:1593–1600.

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

Takaiwa F, K Oono, A Kato 1991 Analysis of the 5′ flanking region responsible for the endosperm-specific expression of a rice glutelin chimeric gene in transgenic tobacco. Plant Mol Biol 15:49–58.

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

Thompson JD, DG Higgins, TJ Gibson 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

Utsumi S 1992 Plant and food protein engineering. Adv Food Nutr Res 36:89–208.

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

Wen L, JK Huang, BH Johnson, GR Reeck 1989 Nucleotide sequence of a cDNA that encodes a rice glutelin. Nucleic Acids Res 17:9490.

Wind C, K-P Häger 1996 Legumin encoding sequences from the redwood family (*Taxodiaceae*) reveal precursors lacking the conserved Asn-Gly processing site. FEBS Lett 383:46–50.

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

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

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

Yamagata T, H Kato, S Kuroda, S Abe, E Davies 2003 Uncleaved legumin in developing maize endosperm: identification, accumulation and putative subcellular localization. J Exp Bot 54:913–922.