Three Genes from Solanum Chacoense Coding for Squalene Synthase

William Herring Wadlington

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

in

Horticulture

James G. Tokuhisa, Chair

Richard E. Veilleux

Eva Čolláková

Sakiko Okumoto

May 4, 2011

Blacksburg, VA

Keywords: *Solanum chacoense,* squalene synthases, gene family

Three Genes from Solanum Chacoense Coding for Squalene Synthase

William Herring Wadlington

Abstract

Squalene synthase (EC 2.5.1.2.1; SQS) is located at a branch point in the isoprenoid pathway and catalyzes the condensation of two molecules of farnesyl diphosphate to form squalene. SQS activity contributes to the formation of triterpenes and sterols, including phytosterols, brassinosteroids, cholesterol, and in potato plants, steroidal glycoalkaloids (SGAs). These compounds have diverse functions in the plant. SGAs are defense compounds that deter feeding by potato pests. The wild potato *Solanum chacoense* accumulates higher amounts of SGAs than cultivated potato and some of its accessions produce leptines, a rare class of SGAs that is toxic to Colorado potato beetle. Unlike most eukaryotes, higher plants have more than one gene coding for SQS. Three *sqs* gene homologs were isolated from *S. chacoense*, $sqs1_{Sc}$, $sqs2_{Sc}$, and $sqs4_{Sc}$, that have 74 to 83% identity at the amino acid level. Some of the amino acid differences between *sqs* isoforms are likely to affect enzyme activity. Each of the three genes contained an intron in the 3'UTR. This feature may have a role in the nonsense-mediated decay of incomplete *sqs* mRNAs. A partial SQS polypeptide retaining catalytic activity but lacking the membrane anchoring domain could adversely affect a cell with the randomly distributed accumulation of squalene. The mRNA of $sqs1_{Sc}$ and $sqs2_{Sc}$ was detected in all tissues whereas $sqs4_{Sc}$ transcript was limited to bud tissue. The $sqs2_{Sc}$ transcript was less uniformly distributed in the plant than $sqs1_{Sc}$ and accumulated most abundantly in floral tissue. The results demonstrate that the three *sqs* genes have different patterns of gene expression and encode proteins with different primary structures indicating distinct roles in plant squalene metabolism.

Table of Contents

List of Figures

List of Tables

Chapter 1

A General Introduction

The Importance of Potato

Potato is the most important vegetable crop in the world according to Food and Agriculture Organization (2008). In the last twenty years, the potato market has been changing along with global economic trends. Although potatoes are grown mostly in North America, Europe and the former the Soviet Union, potato cultivation is decreasing in these regions. Between 1991 and 2007, potato production in the developed world decreased from 183.13 million imperial tons to 159.89 million imperial tons. While in the same time frame, the developing countries in Asia, Africa, and Latin America have increased production from 84.86 to 165.41 million imperial tons (FAO, 2008). The need for potato cultivars that are more resistant to pests will grow as organic cultivation becomes more popular and as cultivation of potato shifts to countries that have limited infrastructure to support regulated chemical pesticide use. By understanding the metabolism of potato defense compounds, such as steroidal glycoalkaloids (SGAs), researchers can implement breeding programs that will develop the naturally pest resistant cultivars for the future.

SGA accumulation in cultivated potato

The concentration of SGAs in the tuber is very important for potato culinary quality and human health. SGA content below 10 mg 100 g^{-1} fresh wt makes a positive contribution to potato flavor (Valkonen et al., 1996). In a taste evaluation, potatoes containing SGA concentrations greater than 14 mg 100 g^{-1} fresh wt were noticeably bitter tasting and with content over 22 mg 100 g^{-1} fresh wt, panelists felt a burning sensation in the mouth and throat (Sinden et al., 1976). SGA concentrations over 20 mg 100 g^{-1} fresh wt can be toxic to humans (Valkonen et al., 1996). By contrast, high concentrations of SGAs in the above ground tissues are of benefit as toxins or antifeedants against a broad range of pests, including Colorado potato beetle, (CPB, *Leptinotarsa decemlineata*; (Rangarajan et al., 2000). Different SGA structures confer different levels of resistance to pests (Sinden et al., 1980). For instance, leptine glycoalkaloids confer resistance to CPB herbivory at concentrations as low as 120 mg 100 g^{-1} fresh wt of leaf tissue (Sinden et al., 1986). The high content of foliar SGA in *Solanum chacoense* line 8380-1 (chc 80-1) is effective as a CPB resistance mechanism in Bt-cyt3a transgenic lines of *S. tuberosum* cv. Yukon Gold (Coombs et al., 2002, 2003) An ideal cultivated potato would accumulate enough

SGAs in the aerial tissues to provide adequate defense against herbivory, but low enough levels in the tuber tissues to maintain culinary quality and health safety. To create a potato with an ideal SGA profile, researchers have investigated the metabolic pathway that is responsible for SGA biosynthesis (Krits et al., 2007; Mweetwa, 2009).

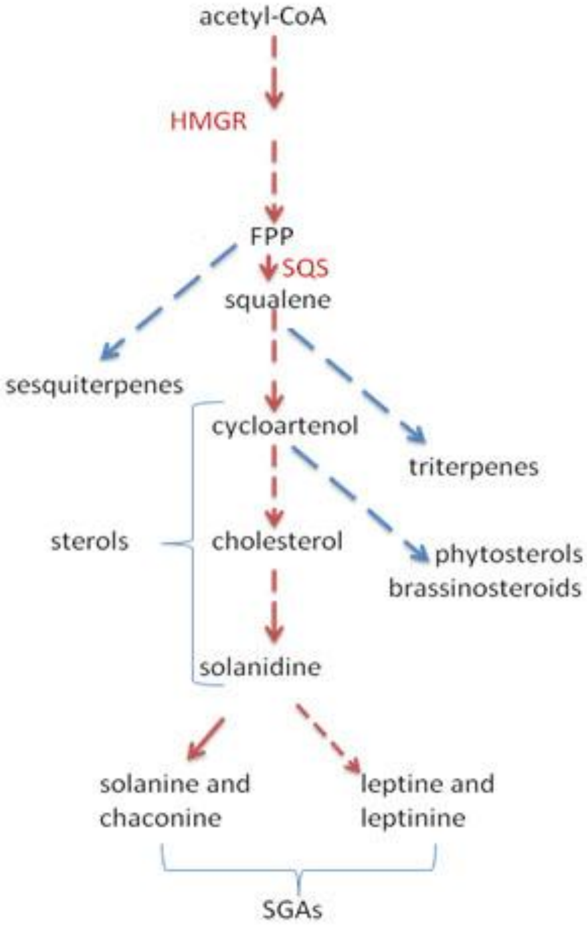


Figure 1.1 Pathways for the biosynthesis of SGAs and related metabolites. Acetyl-CoA is the initial substrate of the terpene pathway. 3-Hydroxy-3-methylglutaryl coenzyme A reductase (HMGR) commits carbon into the pathway which results in the formation of farnesyl diphosphate (FPP). FPP is the direct precursor of sequiterpenes or is converted to squalene by squalene synthase (SQS). Squalene is the substrate for triterpene biosynthesis or is converted to cycloartenol and cholesterol as precursors for sterol and SGA formation. Solanidine can be converted to the SGAs α-solanine, α-chaconine, leptines, and leptinines. Red dashed arrows indicate SGA biosynthesis pathways. Blue dashed arrows represent pathways competing with SGA biosynthesis. (alternative sounds like there's a different way of making SGAs)Enzymes are labeled red. Metabolites are labeled black.

SGA biosynthesis

Many species of the Solanaceae and Melanthiaceae use the sterol biosynthetic pathway to produce plant natural products called SGAs (**Figure 1.1**). The first committed step of this pathway is the condensation of two molecules of farnesyl diphosphate (FPP) by squalene synthase (SQS) to form squalene.

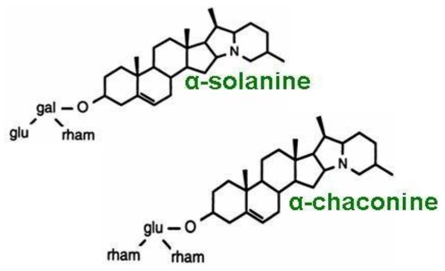


Figure 1.2 α-Solanine and α-chaconine. SGAs are composed of a sterol backbone with an alkaloid group and a sugar moiety. The two major SGAs of potato are shown. The trisaccharide of α-solanine is composed of one molecule each of galactose (gal), glucose (glu) and rhamnose (rham). The trisaccharide of α-chaconine is composed of one glucose and two rhamnoses.

Potato accumulates three vital classes of compounds made from squalene, namely brassinosteroids, phytosterols, and SGAs. Although they are structurally and biochemically related, they have very different functions in the plant. The abundance of individual or classes of compounds varies during the growth and development of the plant and changes with the various biotic and abiotic stresses the plant encounters (Newman and Chappell, 1999). Brassinosteroids are plant growth regulators derived from the phytosterol campesterol and they accumulate to pmol g^{-1} fr. wt levels (Bajguz and Tretyn, 2003). Based on mutants that are deficient or insensitive to brassinosteroids, they are involved in regulating cell proliferation, cell elongation, and floral organ development and are associated with etiolation and general abiotic stress tolerance (Clouse 2011). Phytosterols are derived from the cyclization of 2,3-oxidosqualene and incorporated into the cell membranes that comprise the endomembrane system, primarily the plasma membrane and endoplasmic reticulum (Benveniste, 2004). They contribute to membrane fluidity and are principal constituents of lipid rafts, specialized regions of cell membranes that have a distinct lipid composition and cellular functions (Piironen et al. 2000, Tanner et al. 2011). In potato, phytosterol content in leaves and tubers is 0.14 μmol g^{-1} fr. wt. and 0.09 μmol g^{-1} fr. wt. (Normen et al., 1999), respectively.

These levels are among the lowest observed in plants and the tuber content is the lowest among commonly consumed vegetables. Phytosterols are important nutraceutical compounds because they lower cholesterol uptake in humans by competing with cholesterol for uptake in the gut (L. Calpe-Berdiel, 2009).

Squalene epoxidase catalyzes the epoxidation at the C2-C3 double bond of squalene using oxygen and NADPH resulting in the formation of 2,3-oxidosqualene. 2,3-Oxidosqualene is cyclized to form triterpenes and sterols. Cycloartenol and lanosterol are precursors for other sterols, including cholesterol (Xu et al., 2004a). SGAs have four rings (A, B, C, and D) derived from cholesterol and an additional two rings (E and F) formed from the addition of a nitrogen and ring closures of the aliphatic tail of cholesterol (Valkonen et al., 1996). The activities of glucosyltransferases lead to the formation of a trisaccharide or tetrasaccharide moiety at C3 to complete the SGA structure (**Figure 1.2**) (Valkonen et al., 1996).

The degradation of SGAs is not as well characterized as the biosynthesis. Enzymes have been isolated that hydrolyze the sugar groups from SGAs. An enzyme isolated from potato peels hydrolyzes linkages of the rhamnose sugar groups of α-chaconine (Bushway et al., 1990). Hydrolytic enzymes that cleave bonds between the sugars of the trisaccharide of α-solanine and α-chaconine have been isolated from fungal pathogens of *Solanum* such as *Septoria lycopersici* and *Plectosphaerella cucumerina* (Sandrock et al., 1995; Oda et al., 2002). In field trials, three different soil types treated with SGAs had breakdown products of the compounds indicating catabolic activity in the rhizosphere (Jensen et al., 2009). The hydrolytic enzymes used by the pathogens to break down SGAs contribute to the pathogenicity of the fungi. However, no plant enzymes, except as described above, have been isolated that degrade the aglycones of SGAs.

The concentration and structures of SGAs in different species of potato are quite variable. Cultivated lines have mostly α-chaconine and α-solanine, whereas SGAs in wild potatoes are structurally more diverse. A recent study identified 56 glycoalkaloids in just seven genotypes of potato (Shakya and Navarre, 2008). *S. chacoense* is described as having particularly high SGA content with variation between accessions. Some accessions of *S. chacoense* accumulate leptines in addition to α-chaconine and α-solanine. Leptines differ from solanine and chaconine only in that leptines are hydroxylated and then acetylated on carbon 23 (Ronning et al., 1999). Leptines are of particular interest because of their CPB antifeedant properties (Sinden et al., 1986).

Gene families in SGA biosynthesis

This study is part of a project to define the as yet unknown metabolic pathway of SGA formation in potato, with a particular focus on leptine formation. The current model of SGA biosynthesis was deduced from the structures of SGA-like compounds extracted from *Veratrum* (Kaneko et al., 1977). Although these novel structures are the products of specialized and currently unknown enzymes, enzymes in the earlier steps of the pathway, which are used in the formation of other essential metabolites (Figure 1.1), are known and often are products of gene families. Gene families arise from gene and genome duplications and the subsequent specialization of individual genes for expression and encoded enzyme activity can lead to pathways with metabolic activities that are adapted to the developmental program and environmental stress responses of a plant.

Completely redundant duplicated genes can coexist in the genome when the additional RNA/protein resulting from the multiple genes is either beneficial or not detrimental to an organism. Also, functionally redundant genes can act as back up genes to protect against mutation. Often though, genes in families are thought to have diverged from a single progenitor gene with each duplicated gene member having a more limited function (subfunctionalized) or acquiring novel function (neofunctionalized; (Cooke et al., 1997; Zhang, 2003). Gene expression patterns can be altered by mutations in the *cis*-promoter elements and other noncoding regions of a gene (Force et al., 1999) whereas mutations in the coding region can alter the activity of the encoded enzymes by expanding or restricting the substrate specificity relative to the progenitor enzyme.

One extensively characterized example of gene duplication and diversification in potato is the *hmg* gene family. *hmg* codes for 3-hydroxy-3-methyglutaryl-CoA reductase (HMGR), which commits carbon to the terpene pathway localized in the cytosol (**Figure 1.1**). Three *hmg* genes have been identified. Under normal development, *hmg1* is expressed in expanding leaf tissue and in the tubers, whereas transcript of *hmg2* accumulates in expanding leaves and additionally in the budding flowers, ovaries and sepals. The transcript for *hmg3* has more limited expression accumulating mostly in mature petals and anthers (Korth et al., 1997).

Each member of the gene family is also regulated differently in response to environmental stress (Yang et al., 1991; Choi et al., 1992). Upon wounding, transcripts of *hmg1, hmg2* and *hmg3* increase, whereas following pathogen attack, levels of *hmg2* and *hmg3* transcript increase while *hmg1* transcript levels are reduced (Yang et al., 1991). The wound-induced expression of *hmg1* is correlated with SGA accumulation. On the other hand, pathogen attack or arachidonate elicitation lead to the accumulation of the sesquiterpene phytoalexins, rishitin, and lubimin (Choi et al., 1992). As for amino acid differences between the three isozymes, HMGR2 contains a unique tyrosine kinase phosphorylation site that may be a distinct regulatory mechanism in potato (Korth et al., 1997). These results demonstrate that differential expression of *hmg* gene family members influences the abundance of C_{15} and C_{30} terpene products. The possibility that a similar mode of regulation has been employed at the branch point for sterol and triterpene biosynthesis led us to focus on SQS.

A biochemical rationale for a *sqs* gene family

Whole genome sequencing and previous characterizations of SQS indicate that the enzyme may be represented in potato by a gene family. Two *sqs* genes have been isolated from *Nicotiana tabacum* and *Glycyrrhiza glabra* (Hayashi et al., 1999). *Panax ginseng* is the source of three *sqs* genes (Kim et al., 2011a). Duplicated *sqs* genes may have specialized functions just as observed with the *hmg* gene family: different tissue specificities or responses to various environmental elicitations. For example, the *sqs* gene family members in the above-mentioned species have different patterns of tissue-specific expression or catalytic efficiency converting FPP to squalene (Hayashi et al., 1999; Devarenne et al., 2002; Kim et al., 2011a).

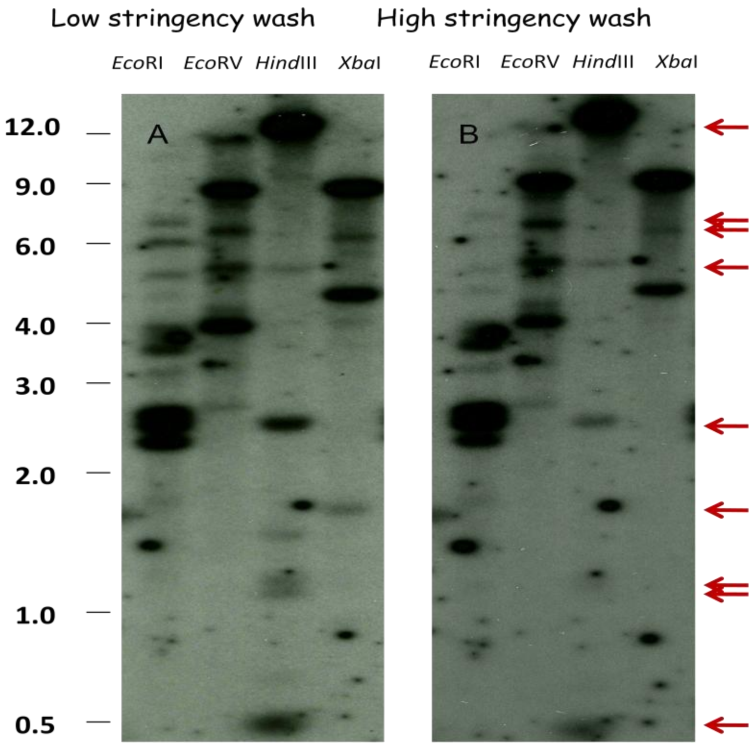


Figure 1.3 DNA blot of *S. chacoense* chc 80-1 using a partial *sqs1$_{Sc}$* cDNA as a probe, prepared by Mweetwa (2009). Low (left) and high (right) stringency washes of a DNA-blot are compared. The restriction enzymes used are listed at the top. DNA marker sizes (kbp) are listed on the left. On the right, red arrows indicated bands that are of variable intensity.

To date, one *sqs* gene (*sqs1$_{St}$*) has been described in tetraploid *S. tuberosum* though the existence of additional *sqs* genes was not discounted (Yoshioka et al., 1999). Orthologs of *sqs1$_{St}$* have been described in at least three species of potato (Krits et al., 2007). A DNA blot of *S. chacoense* chc 80-1 was probed with a radiolabeled portion of *sqs1$_{Sc}$* cDNA to detect homologous *sqs* genes (**Figure 1.3**). The pattern and relative intensities of the bands between the high and low stringency washes of the blot indicate the probe is binding to several loci in the *S. chacoense* genome. The different band intensities indicate relative differences in homology amongst the loci binding to the probe. This indicates a small family of *sqs* in potato. My study sought to isolate the gene homologs detected by the DNA blot analysis by Mweetwa (2009), to characterize their transcript profile in different tissues, and to examine their deduced amino acid sequences for differences that could change the protein synthesis profile or enzyme activity.

Chapter 2

Three *squalene synthase* homologs in *Solanum chacoense*

Abstract

Squalene synthase (EC 2.5.1.21; SQS) catalyzes the head-to-head dimerization of farnesyl diphosphate at a branch point in the isoprenoid biosynthetic pathway. Squalene synthase activity contributes to pathways leading to the formation of triterpenes and sterols, which in plants includes cholesterol, phytosterols, triterpenes, brassinosteroids and defense compounds such as latex and steroidal glycoalkaloids (SGAs), the latter compounds being natural products found in specific plant taxa. The enzyme is encoded by a gene family in many plant species. The *sqs* gene family members may have differential expression patterns and their encoded enzymes may have different effects on metabolism, as has been demonstrated with the *hmg* gene family. To identify *sqs* genes in the wild potato species *Solanum chacoense*, the Basic Local Alignment Search Tool (BLAST) was used to screen a provisional assembly of the genome sequence of *Solanum phureja* with the amino acid sequence of the *sqs* gene from cultivated potato. Four homologous gene structures were identified. Reverse transcriptase-PCR using oligonucleotide primers based on the *S. phureja sqs* sequences and RNA extracted from flower and bud tissue allowed for the isolation of three partial cDNAs. Comparisons of the three coding regions revealed more than 76% nucleotide identity between each other and therefore the sequences were designated *sqs1*, *sqs2,* and *sqs4*. An alignment of the genomic and cDNA sequences of *sqs1* identified the intron-exon boundaries and highlighted the presence of an intron in the 3'-UTR of the gene. An identical gene structure has been found in *sqs2* from *Nicotiana tabacum,* demonstrating conservation of the *sqs* gene structure within plants of the Solanaceae. Alignment of the *S. chacoense* cDNA sequences of *sqs1, sqs2* and *sqs4* with the orthologous genes of *S. phureja* indicated more than 97% DNA sequence identity and the presence of an intron in the 3'-UTR in each of the *S. phureja* genes. Based on the close phylogenetic similarity between *S. phureja* and *S. chacoense,* we predict that *sqs2* and *sqs4* of *S. chacoense* each have an intron in the 3'-UTR.

Semiquantitative RT-PCR of RNA extracted from various tissues of *S. chacoense* showed tissue-specific patterns of steady-state transcript abundance for each of the three genes.

Transcripts of *sqs1*$_{Sc}$ and *sqs2*$_{Sc}$ were detected in all tissues screened. Transcript abundance of *sqs2*$_{Sc}$ was greatest in the flowers. Transcript levels of *sqs4*$_{Sc}$ were highest in the buds. Tissue-specific profiles of *sqs* transcript abundance were compared with SGA content to assess the contribution of individual gene family members to SGA biosynthesis. The results indicate that the transcript profile pattern for *sqs2*$_{Sc}$ alone or in combination with *sqs1*$_{Sc}$ is more similar to the pattern of SGA abundance than the pattern for *sqs1*$_{Sc}$ alone.

Introduction

The isoprenoid pathway begins with the serial condensation of three acetyl-coenzyme A (Ac-CoA) molecules to produce 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) which is reduced to mevalonate by HMG-CoA reductase (HMGR). Additional steps including the decarboxylation of diphosphomevalonate are necessary to yield isopentenyl diphosphate (IPP). An IPP isomerase converts IPP into dimethylallyl diphosphate (DMAPP), which combines with two IPPs to form farnesyl diphosphate (FPP) (Bouvier et al., 2005). FPP can be dimerized in a head-to-head condensation catalyzed by SQS to produce squalene or cyclized by terpene synthases to form sesquiterpenes (**Figure 2.**1). Squalene is converted into 2,3-oxidosqualene by squalene monooxygenase (epoxidase) and then cyclized by several oxidosqualene cyclases that produce the different carbocation intermediates that lead to the formation of various triterpenes and sterols (Xu et al., 2004b).

Metabolites derived from squalene are structurally and functionally diverse across the plant kingdom. Triterpenes are the principal medicinal compounds in *Bupleurum falcatum* (Siberian ginseng), *Panax ginseng*, *Lotus japonica*, *Glycyrrhiza glabra* (licorice) (Xu et al., 2004a). Triterpenes are major component of wax and latex present in *Euphorbia tirucalli* (petroleum plant). Phytosterols, such as sitosterol and campesterol are major structural components in plant cell membranes (Gallova et al., 2011). The brassinosteroids are major class of plant hormones that are also derived from squalene (Choe, 2010).

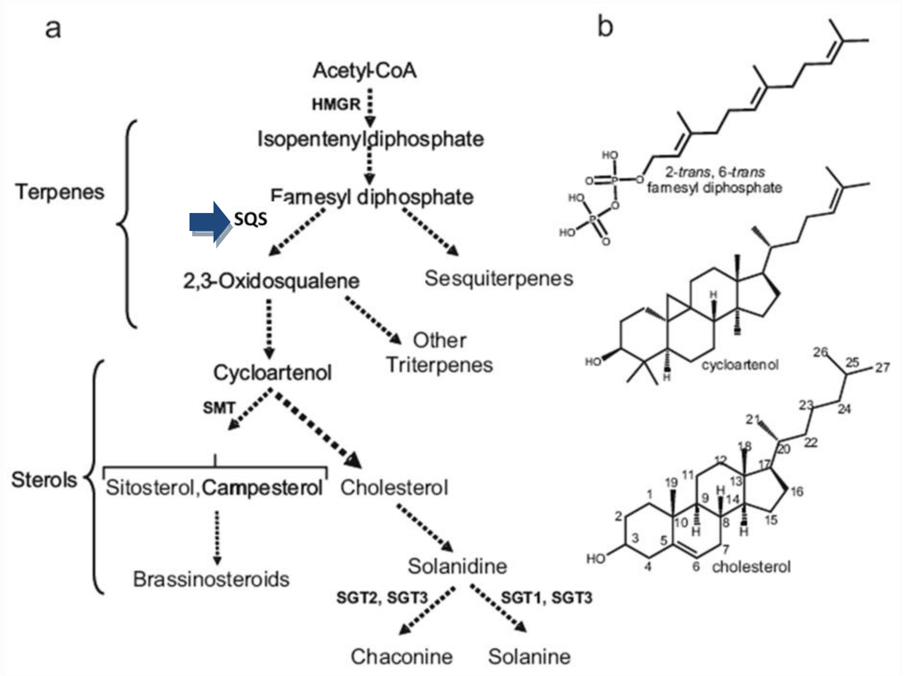


Figure 2.1. Biosynthetic pathways associated with squalene synthase. a. The initial substrate of the isoprenoid pathway is acetyl-CoA. Three of these molecules are used to make 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA), which is then reduced by HMG-CoA reductase (HMGR). Two molecules of isopentenyldiphosphate and one molecule of its isomer, are condensed to form farnesyl diphosphate (FPP). Squalene synthase (SQS; blue arrow) dimerizes FPP to squalene which is epoxidated to form squalene epoxide which leads to the formation of triterpenes and sterols. FPP is a substrate for sesquiterpene formation. The arrows indicate multiple enzymatic steps. Metabolites are indicated between the arrows. Enzyme names are abbreviated beside the arrows: sterol C24-methyltransferase (SMT), solanidine glycosyltransferase (SGT). b. The chemical structures of FPP, cycloartenol, and cholesterol.

Overexpression of *sqs* in plants has been shown to result in the altered accumulation of other metabolites derived from squalene. Overexpression of *sqs1* isolated from *P. ginseng* in *Eleutherococcus senticosus* increased the accumulation of phytosterols and triterpene saponins 2 to 2.5 times that of the wild type (Seo et al., 2005). Upregulation of *sqs* expression in roots of *B. falcatum* led to an increased accumulation of squalene, sterols, and triterpenes and also the upregulated expression of genes in the pathways downstream of *sqs* (Kim et al., 2011b).

The tissue-specific accumulation of steady-state *sqs* transcript and triterpene and sterol metabolites has been described in the literature. Studies with *E. tirucalli* revealed that the level of *sqs* transcript is highest in the leaf and stem cambial tissues which exude triterpene saponins in the form of latex (Uchida et al., 2009). The two *sqs* genes isolated from *N. tabacum* are transcribed in the apical meristem, which requires phytosterols for cell membrane biogenesis (Devarenne et al., 2002).

Mweetwa (2009) concluded that in potato there was a correlation between total SGA content and transcript abundance of *sqs1*. In other studies, plant defense responses elicited by chemical treatment or wounding show altered steady-state levels of *sqs* gene expression with corresponding changes in SGA levels. Potato tubers that were wounded elicited the accumulation of SGAs simultaneously with higher levels of *sqs* transcript relative to unwounded tubers (Yoshioka et al., 1999). Extracts of the potato pathogen *Phytophthora infestans* applied to tuber disks were shown to inhibit accumulation of SGAs. This inhibition occurred concurrent with an increase in sesquiterpene and sesquiterpene cyclase accumulation and also with a reduction in *sqs* transcript abundance (Tjamos and Kuc, 1982; Yoshioka et al., 1999). A similar response of *sqs* transcript and sesquiterpene biosynthesis was observed when sesquiterpene biosynthesis was elicited in tobacco cell culture with arachidonic acid (Threlfall and Whitehead, 1988).

In *Solanum tuberosum,* SQS protein, *sqs1* transcript and SGA accumulations were shown to increase in tubers after wounding (Yoshioka et al., 1999). Amongst accessions of potato, six potato genotypes that accumulated different amounts of SGA in the tissues showed a correlation between *sqs* transcript and SGA levels. A correlation between *sqs* transcript and SGA levels was documented in the leaves and also the phelloderm and parenchymous tuber tissues of *S. chacoense* line chc 80-1 and *S. tuberosum* cv. Desirée (Krits et al., 2007). However, when the tissue specificity of *$sqs1_{Sc}$* transcription was compared to the SGA content in the organs of *S. chacoense* line chc 80-1 (**Figure 2.2**), limited correlation was observed (Mweetwa, 2009). Transcript levels of *$sqs1_{Sc}$* were high in the bud and low in the tuber tissue, similar to the SGA profile, but similar correlations were not observed with floral and stem tissues. All three of the aforementioned studies were based on the notion of a single *sqs* gene.

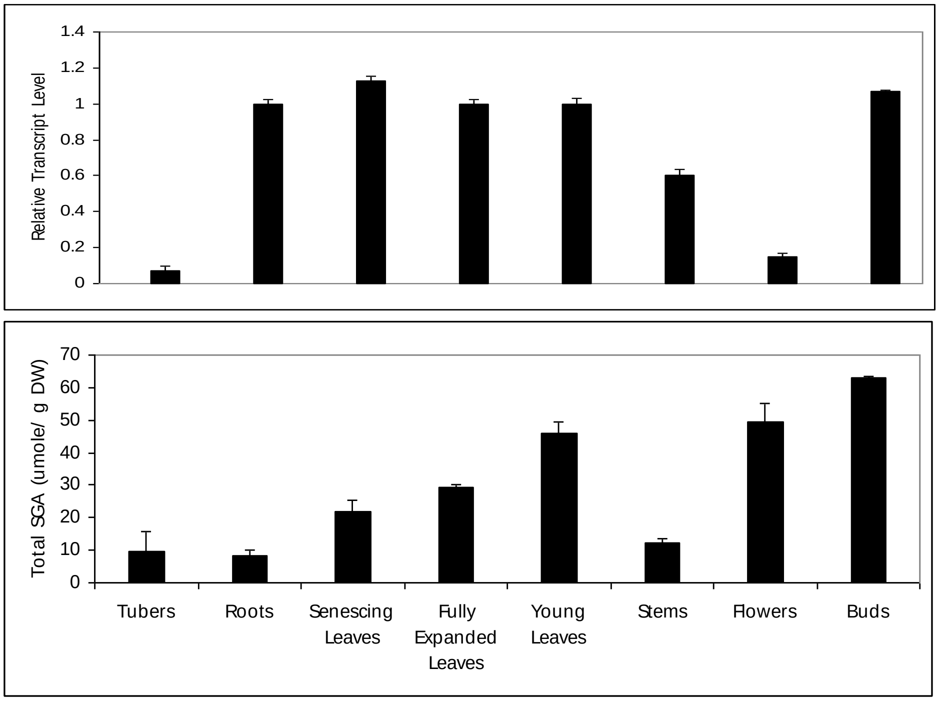


Figure 2.2 s*qs1*$_{Sc}$ transcript (top) and SGA levels (bottom) in different tissues of chc 80-1. Relative transcript abundance of *sqs1*$_{Sc}$ and total SGA levels in tissues of *S. chacoense*. The indicated tissues were collected at anthesis from plants grown under controlled conditions (Mweetwa, 2009). The figure is from Mweetwa (2009).

However, two genes of *sqs* have been identified in many plant species, including *N. tabacum* (Devarenne et al., 2002), and *G. glabra* (Hayashi et al., 1999). *P. ginseng* has been characterized as having three *sqs* genes (Kim et al., 2011a). In the tetraploid *S. tuberosum,* a single *sqs* (*sqs1*$_{St}$) has been reported although multiple genes were not precluded (Yoshioka et al., 1999). To identify genomic DNA homologous to *sqs* in *S. chacoense,* Mweetwa (2009) probed a genomic DNA blot with a radiolabeled partial cDNA of *sqs1*$_{Sc}$ and showed differential intensities of bands following washes of low and high stringency (**Figure 1.3**) (Mweetwa, 2009). The different levels of hybridization to the probe indicated a small *sqs* gene family in *S. chacoense*.

Here I report the isolation and cloning of cDNAs of three *sqs* gene homologs from *S. chacoense*. The three genes showed tissue specific patterns of transcript accumulation with *sqs1*$_{Sc}$ and *sqs2*$_{Sc}$ transcribed in all tissues, and with *sqs4*$_{Sc}$ transcript accumulating primarily in the buds. I identified *sqs* gene families in other species of the genus *Solanum,* including *S.*

tuberosum, *S. lycopersicum*, and *S. phureja*. An intron in the 3’ UTR of each of the *sqs*$_{Sc}$ gene was identified. This intron may function as a proofreading mechanism to ensure protein synthesis of only full-length transcripts of the *sqs* gene.

Results

Genomic sequence of $sqs1_{Sc}$

In order to complete the genomic sequence of $sqs1_{Sc}$ started by Mweetwa (2009), I amplified genomic regions of introns 10 and 13. Partial genomic sequence of $sqs1_{Sc}$ had been obtained from PCR products using genomic DNA extracted from *S. chacoense* chc 80-1 as a template (Mweetwa, 2009). The sequences assembled into two contigs (**Figure 2.4**). One contig began 165 bp upstream of the start codon and continued into intron 10. The other contig began in intron 10 and continued to exon 13. Preliminary results indicated that intron 10 spanned 2.5 kb but it was not cloned or sequenced entirely.

To generate a single contig of the $sqs1_{Sc}$ gene, a PCR spanning the two contigs was optimized to produce sufficient copies of intron 10 for PCR sequencing. Intron 10 harbored both a polyT and dinucleotide repeat rich regions toward the middle of the intron (**Figure 2.4)**. This 65-bp-region spanned the two original contigs resulting in a revised intron 10 length of 1620 bp (**Figure 2.3**).

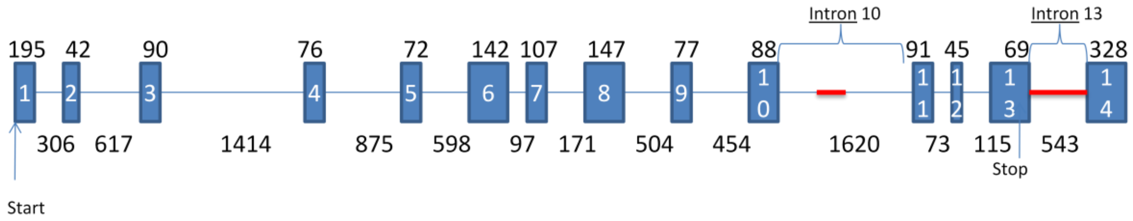


Figure 2.3. Intron-exon organization of $sqs1_{Sc}$. Exons (blue rectangles) and introns (lines) are represented proportionately with their sizes (bp) indicated above or below. Exons are numbered. The start codon is indicated by a vertical blue arrow, and stop codon is indicated by a vertical blue line. The parts of the sequence found previously (Mweetwa, 2009) are indicated in blue. The genomic regions in intron 10 and intron 13 sequenced in this study are highlighted in red.

To identify the 3' end of the *sqs1* gene, oligonucleotide primers hybridizing to the 3' end of exon 13 and the nucleotide sequence just upstream of the polyA tail of the corresponding cDNA were used to amplify the 3' untranslated region extending from exon 13. Alignment of the cDNA and the resulting genomic fragment sequences revealed an additional 540 bp in the genomic sequence that was lacking in the cDNA. The insert in the genomic sequence began 6 bp downstream of the stop codon of the cDNA and had the canonical intron slice sites (GT-AG). Together, these results indicate that the gap represents an intron in the 3’ UTR (**Figure 2.3**). As a result of these studies, the genomic structure of *sqs1*$_{Sc}$ can be defined from 165 bp upstream of the start codon to the polyadenylation site (**Appendix A**).

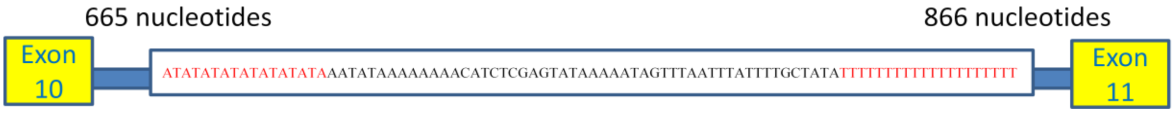

665 nucleotides
866 nucleotides
Exon 10
ATATATATATATATATAAATATAAAAAAAACATCTCGAGTATAAAAATAGTTTAATTTATTTTGCTATATTTTTTTTTTTTTTTTTTTT
Exon 11

Figure 2.4. Internal sequence of intron 10 flanked by repetitive regions. The area between exon 10 and exon 11 represents intron 10 and includes 665 nucleotides at th 5’end and 866 nucleotides at at the 3’end sequenced previously (blue bars, Mweetwa 2009) and the 65 bp sequenced in this project. The red letters represent the repetitive sequences.

The complete genomic sequence of *sqs1* was used to assess a genomic DNA blot generated previously of *S. chacoense* chc 80-1 probed with a partial *sqs1*$_{Sc}$ cDNA (Mweetwa, 2009). Fragment sizes generated by the restriction enzyme recognition sites predicted in the genomic sequence (**Appendix B**) map to hybridizing bands observed in both high and low stringency washes of the DNA-blot for all four enzyme digests.

Coding region for four *sqs*$_{Sp}$ and isolation of three *sqs*$_{Sc}$ partial cDNAs

The doubled monoploid accession, DM1-3 516R44 (CIP801092), of the diploid wild potato species *S. phureja* generated in the Veilleux lab at Virginia Tech has been sequenced (Xu et al., 2011) and preliminary versions of the genome assembly are accessible through the Potato Genome Sequencing Consortium (PGSC; http://potatogenomics.plantbiology.msu.edu/). In version 3, the genome is represented by scaffolds with an average length of 1.3 Mbp representing portions of the chromosomes. A BLAST search of version 3 was conducted using the coding region of *sqs1* from *S. tuberosum* (*sqs1*$_{St}$, AB022599; (Yoshioka et al., 1999). Four genomic domains on three scaffolds were identified with e-values in the range of 1.0e-187 to 3.1e-45 indicating the presence of four homologs of *sqs1*$_{St}$ in *S. phureja* (*sqs*$_{Sp}$) (**Table 2.4**). The homolog with the lowest e-value was designated *sqs1*$_{Sp}$ because of the 99% similarity to *sqs1*$_{Sc}$ (Mweetwa, 2009). The homolog with the second lowest e-value was called *sqs2*$_{Sp}$ because of the similarity to *sqs1*$_{Sc}$ in genomic size between the start and stop codons (8.7 kbp for *sqs1*, 5.7 kbp for *sqs2*) and 83% sequence identity. The third scaffold identified by the BLAST search contained two *sqs*$_{Sp}$ genes of approximately 3 kb length each that are in an inverted tandem array separated by 10 kbp. These two genes were both 76% identical to *sqs1*$_{Sc}$ and arbitrarily named *sqs3*$_{Sp}$ and *sqs4*$_{Sp}$. Together, these results identified four *sqs* gene homologs in the *S. phureja* provisional genome sequence (**Figure 2.6**), and raises the possibility that other diploid potato species also have four *sqs* genes.

The coding region of plant *sqs* (Devarenne et al., 2002; Mweetwa, 2009) was used to predict *sqs* gene structure from the genomic scaffold sequences. The *sqs*$_{Sp}$ genes had 13 exons all of the same lengths as in *sqs1*$_{Sc}$, *sqs1*$_{Nt}$ and *sqs2*$_{Nt}$ with introns having both expected and non-canonical intron-exon dinucleotide sequences. The deduced amino acid sequences of all four *sqs*$_{Sp}$ cDNAs have the six amino acid peptide domains (domains 1 through 6;) previously defined as highly conserved in all *sqs* genes (**Appendix C)** (Robinson et al., 1993). The predicted cDNAs of *sqs1*$_{Sp}$, *sqs2*$_{Sp}$, and *sqs4*$_{Sp}$ had coding regions of 1236 bp, identical in length to *sqs1*$_{Sc}$, *sqs1*$_{Nt}$, and *sqs2*$_{Nt}$, (Devarenne et al., 2002; Mweetwa, 2009). The coding region of *sqs3*$_{Sp}$ had a length of 1197 bp because of a 9-bp deletion in exon 1 and a 30-bp deletion in exon 8. To isolate the *S. chacoense* homologs of the four *sqs*$_{Sp}$ genes (*sqs*$_{Sc}$), RT-PCR of RNA from *S. chacoense* was undertaken using oligonucleotide primers based on unique sequence domains in the predicted cDNA of each *S. phureja* homolog (**Table 2.3**). PCR products were obtained for *sqs1*$_{Sc}$ and *sqs4*$_{Sc}$ from RNA of bud tissue whereas *sqs2*$_{Sc}$ was isolated from RNA of flower tissue. No product homologous to *sqs3*$_{Sp}$ was ever detected and the isolation of *sqs3*$_{Sc}$ was not pursued further. The predicted exon structure of each of the *sqs*$_{Sc}$ was compared to identify differences **(Figure 2.5)**. Exons 1 to 13 of *sqs1*$_{Sc}$ had been described by (Mweetwa, 2009), but the gene structure was completed in this study (**Figure 2.3**). The coding sequences of *sqs1*$_{Sc}$, *sqs2*$_{Sc}$, and *sqs4*$_{Sc}$ indicate a gene composed of 13 exons. The exon structures of *sqs2*$_{Sc}$ and *sqs1*$_{Sc}$ are identical but exon 1 of *sqs4*$_{Sc}$ was 9 bp longer and the stop codon in exon 13 was 6 bp further downstream compared to those in the other *sqs* genes, making the coding region of *sqs4*$_{Sc}$ 1239 bp rather than the 1236 bp of *sqs1*$_{Sc}$ and *sqs2*$_{Sc}$.

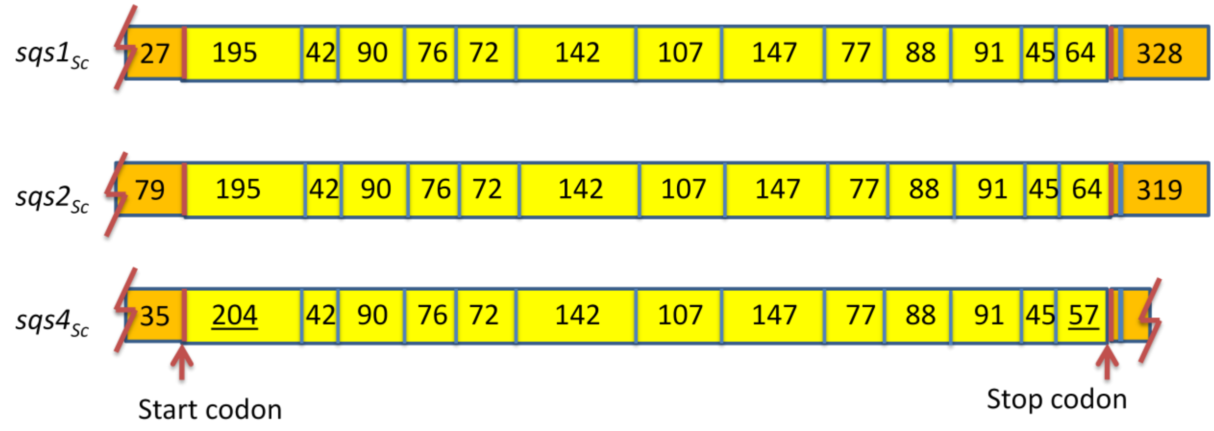


Figure 2.5. Comparison of exon length of *sqs*$_{Sc}$ partial cDNAs. The coding region (yellow) and additional 5' and 3' sequences (brown) defined by exon length (bp) are compared for *sqs1*, *sqs2,* and *sqs4*. Splice junctions are represented by blue bars. The underlined numbers in *sqs4*$_{Sc}$ indicate exon length differences. Lengths (bp) of the known sequence of the 5' UTR are listed. Length of *sqs4*$_{Sc}$ exon 1 and the coding portion of exon 13 are underlined to highlight the sequence differences. Jagged red lines indicated where the cDNA sequence is incomplete in the UTR.

Comparison of *sqs*$_{Sc}$ to other *sqs* genes

To determine the relationship between *sqs* genes from different plants of the Solanaceae, the genes identified in the *S. phureja* genomic assembly, those isolated from *S. chacoense,* and other publically available sequences were assembled in a phylogenetic tree generated using Clustal W (DNASTAR Lasergene program version 8.0, Megalign; **Figure 2.6**). The *sqs1* coding region of *S. chacoense* clustered with genes from *S. phureja*, *S. tuberosum*, *S. lycopersicum* and *Capsicum annuum* with at least 94% identity, suggesting that the *sqs1* gene type is abundant in five species (**Figure 2.6, Figure 2.7**). The coding region sequence of *sqs2*$_{Sc}$ and *sqs2*$_{Sp}$ clustered with 97% identity, indicating that *sqs2*$_{Sc}$ is the homolog of *sqs2*$_{Sp}$. The two genes *sqs3*$_{Sp}$ and *sqs4*$_{Sp}$ are more similar (90% identity) to each other than to the other *sqs*$_{Sc}$ sequences. The sequences of *sqs4*$_{Sp}$ and *sqs4*$_{Sc}$ have 97% identity, indicating they are homologs. No homologs of *sqs3*$_{Sp}$ were isolated from *S. chacoense*, but one was identified by BLAST in a *S. lycopersicum* BAC library (http://solgenomics.net, SL2.40ch10: 60,010,501 - 60,015,400 bp; **Figure 2.6**). Thus, three distinct *sqs* gene homologs (out of four) predicted by the *S. phureja* genome sequence were isolated from *S. chacoense*.

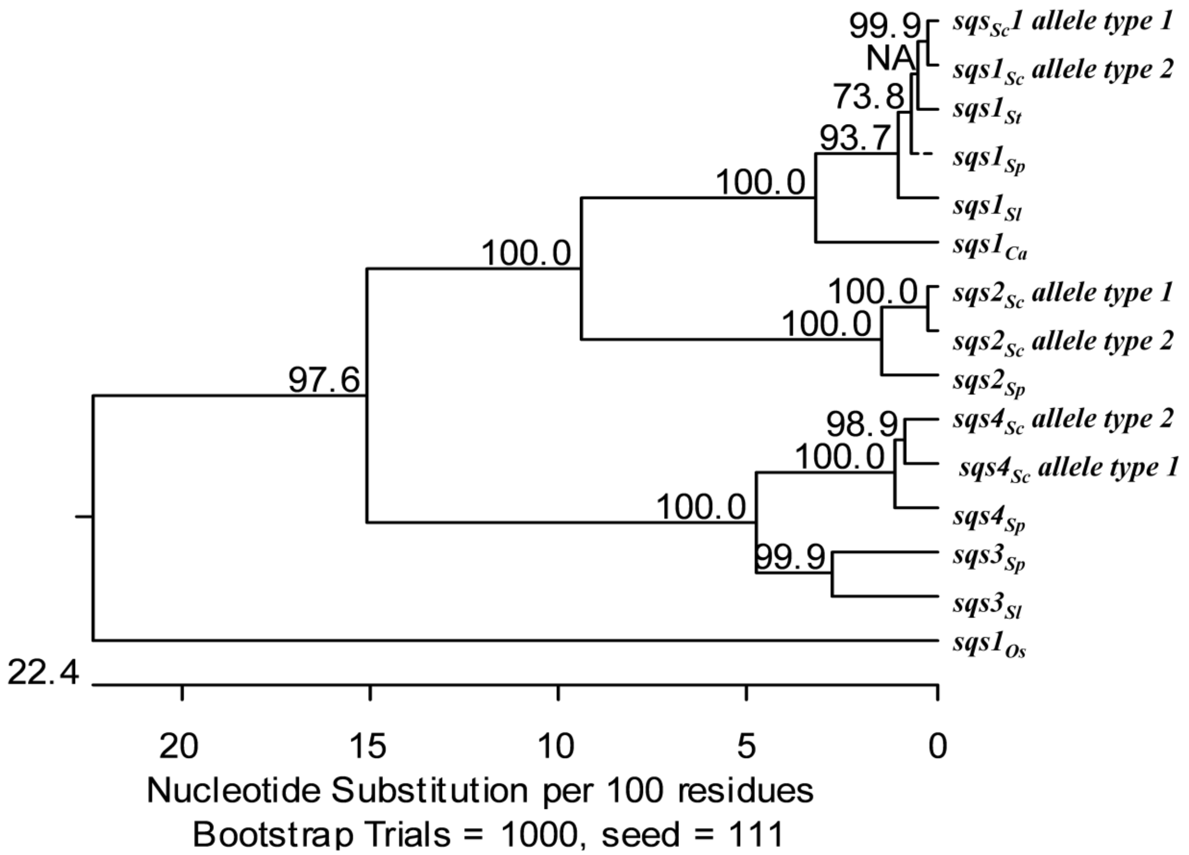

99.9
NA
73.8
93.7
100.0
100.0
97.6
100.0
100.0
100.0
98.9
100.0
100.0
99.9
22.4
$sqs_{Sc}1$ allele type 1
$sqs1_{Sc}$ allele type 2
$sqs1_{St}$
$sqs1_{Sp}$
$sqs1_{Sl}$
$sqs1_{Ca}$
$sqs2_{Sc}$ allele type 1
$sqs2_{Sc}$ allele type 2
$sqs2_{Sp}$
$sqs4_{Sc}$ allele type 2
$sqs4_{Sc}$ allele type 1
$sqs4_{Sp}$
$sqs3_{Sp}$
$sqs3_{Sl}$
$sqs1_{Os}$
20
15
10
5
0
Nucleotide Substitution per 100 residues
Bootstrap Trials = 1000, seed = 111

Figure 2.6. Phylogenetic alignment of *Solanum sqs* coding region. Clustal W was used to create a neighbor joining tree of plant *sqs*. Two alleles of three *sqs*$_{Sc}$ are compared to the previously identified *sqs* from *S. tuberosum* (AB022599; (Yoshioka et al., 1999)), *Solanum lycopersicum*; GU075687), *Capsicum annuum* (AF124842; (Lee et al., 2002), an *sqs3*$_{Sp}$ homolog found in a *S. lycopersicum* BAC library, the four predicted *sqs*$_{Sp}$ coding regions from the genome, and *Oryza sativa sqs1* (BAA22558; (Hata et al., 1997). The bootstrap values (%) are from 1000 trials.

Most diploid potato species, including *S. chacoense*, are self-incompatible. Therefore, genes isolated from a plant are likely to have two distinct allelic sequences. To distinguish allelic differences from PCR-generated errors or recombination between allelic templates during PCR, I defined a wild-type sequence as the same nucleotide present in two out of at least four sequences, with a minimum of three sequences each from at least three independent PCRs of a cDNA preparation. Two alleles were identified for each the three *sqs* genes isolated from *S. chacoense* chc 80-1 (**Appendix C**). The allelic differences were all single nucleotide substitutions leading to both synonymous and non-synonymous changes. There were no variations in the length of any of the exons between alleles.

3’ UTR introns in *sqs*$_{Sp}$ genes

Nucleotide sequence comparison of the genomic sequence of *sqs1*$_{Sc}$ and the corresponding cDNA indicates the presence of an intron (intron 13) in the 3’UTR. A region of sequence homologous to intron 13 was observed in the genomic sequence of *sqs1*$_{Sp}$. To determine if *sqs2*$_{Sc}$ and *sqs4*$_{Sc}$ had introns in the 3’UTR (in the absence of cloning the genomic regions from *S. chacoense*), the cDNA sequence at the 3’end of the *sqs2*$_{Sc}$ and *sqs4*$_{Sc}$ was aligned with their respective homologs in the *S. phureja* genomic scaffolds (**Figure 2.8**). The cDNAs aligned in the 3’UTR region to two regions on the genomic sequence separated by 403 bp. The gap starts 12 bp after the stop codon. Similarly, the 3’end of *sqs4*$_{Sc}$ cDNA aligned with the *sqs4*$_{Sp}$ genomic sequence with the introduction of a 436 bp gap starting 9 bp after the stop codon. The alignment gaps begin and end with the conventional dinucleotide sequences of intron-exon boundaries (GT-AG) and therefore are consistent with an intron in the 3’ UTR of each of the *S. phureja sqs* genes. By homology, this intron is likely to be present in each of the *S. chacoense sqs* genes.

sqs	***sqs1*$_{Sc}$**	***sqs2*$_{Sc}$**	***sqs4*$_{Sc}$**
***sqs1*$_{Sc}$**	**-**	**83%**	**76%**
***sqs1*$_{Sp}$**	**99%**	**84%**	**77%**
***sqs1*$_{Sl}$**	**98%**	**83%**	**76%**
***sqs1*$_{Ca}$**	**94%**	**84%**	**77%**
***sqs2*$_{Sc}$**	**83%**	**-**	**74%**
***sqs2*$_{Sp}$**	**84%**	**97%**	**75%**
***sqs4*$_{Sc}$**	**76%**	**74%**	**-**
***sqs4*$_{Sp}$**	**77%**	**74%**	**98%**

Figure 2.7. Percent identity of *Solanum sqs* coding regions from *S. phureja, S. chacoense, S. lycopersicum*, and *C. annuum*. Identity was generated in the same alignment produced for **Figure 2.6.**

To determine if the intron 13 sequence was conserved across different species, the sequence was used as a query to screen GenBank, the Sol Genomics Network (http://solgenomics.net/) and the *S. phureja* genomic assembly using BLAST (NCBI; http://www.ncbi.nlm.nih.gov/). No significant homologies were found to any of the intron 13 sequences.

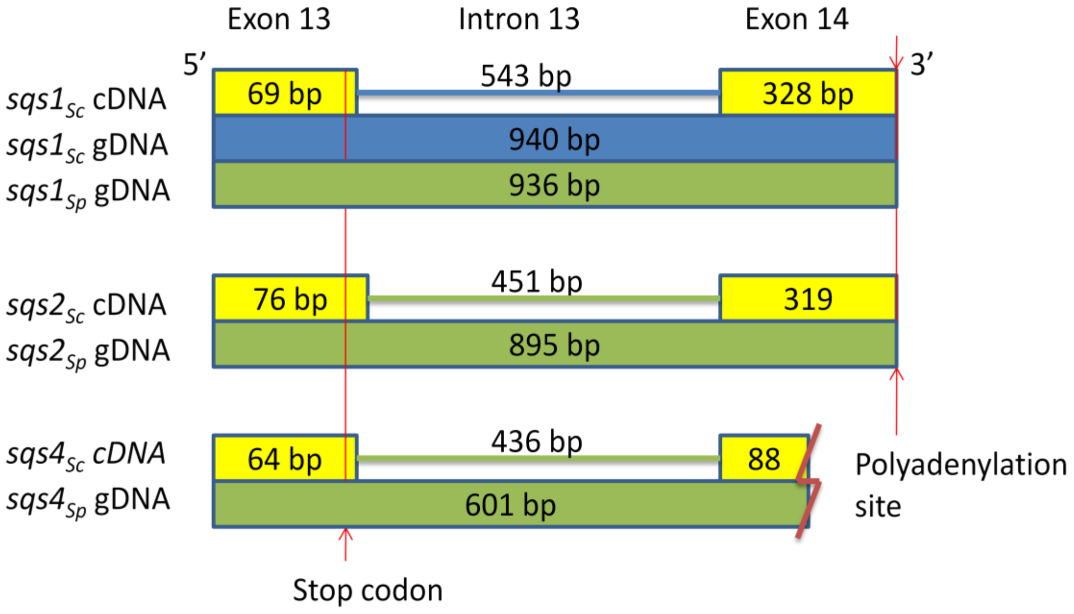


Figure 2.8 Intron placement at the 3'end of *sqs* cDNAs and genomic DNA. The cDNA of *sqs*$_{Sc}$ (yellow rectangles), the genomic sequence of *sqs1*$_{Sc}$ (blue rectangle), and genomic sequence from *sqs*$_{Sp}$*1*, *sqs*$_{Sp}$*2*, and *sqs*$_{Sp}$*4*, (green rectangles) were compared to determine the lengths of exon 13, intron 13 and exon 14 of *sqs1*, *sqs2*, and *sqs4*. The cDNA gaps are represented by lines connecting the exons in the cDNA. The approximate position of the stop codon and the approximate site where the poly A-tails are added to the RNA are annotated with vertical red lines. Sequence for the 3' end of the UTR of *sqs4*$_{Sc}$ is incomplete and the 3'extent is indicated by a jagged line.

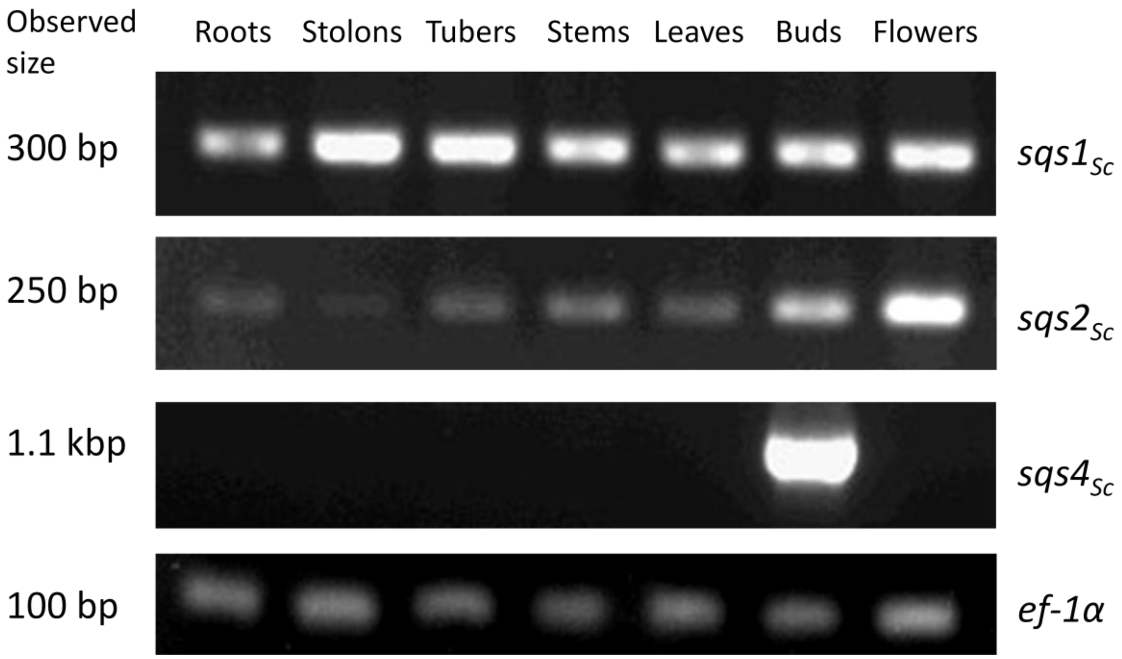


Figure 2.9 Tissue specific abundance of steady-state transcripts for *$sqs1_{Sc}$*, *$sqs2_{Sc}$*, and *$sqs4_{Sc}$*. The *sqs_{Sc}* products were generated by gene specific oligonucleotide primers for the three *sqs_{Sc}* using reverse-transcribed RNA purified from the indicated tissues. The *sqs_{Sc}* products were fractionated on a single agarose gel and imaged at a non-saturating exposure. Potato *elongation factor 1-α* (*ef1-α*; AB061263), a housekeeping gene involved in protein synthesis, was used as a constitutively transcribed gene to control for the amount of cDNA template. Observed sizes (left margin) are recorded.

Tissue specificity of *sqs_{Sc}* gene expression

The levels of *sqs* transcript and metabolites derived from SQS have been compared in potato without the recognition that *sqs* is a gene family. To assess the distribution of *sqs* transcript in different parts of the plant, RT-PCR was performed with gene-specific oligonucleotide primers to detect *sqs_{Sc}* transcript levels in tissues of *S. chacoense* and compare them with the SGA content determined previously by Mweetwa (2009) **(Table 2.3)**. The primers flanked intron regions so that genomic DNA contamination could be identified as products longer than expected. PCR products of *$sqs2_{Sc}$*, and *$sqs4_{Sc}$* were incubated with restriction enzymes and fragmentation patterns noted or sequenced to confirm that the PCR product was generated from the expected *sqs* cDNA template. *$sqs1_{Sc}$* appeared to be transcribed in all tissues as shown previously (Mweetwa, 2009). Transcript for *sqs2*$_{Sc}$ was detected in all tissues screened but was most abundant in floral tissue followed by bud tissue **(Figure 2.9).** *sqs4*$_{Sc}$ transcript was detected in bud tissue only and was detected faintly after 40 cycles of PCR in leaves, flowers, and stems. While *sqs1*$_{Sc}$ seemed to be transcribed uniformly across all tissues and *sqs2*$_{Sc}$ transcript had a more varied but ubiquitous profile, *sqs4*$_{Sc}$ appeared to be transcribed specifically in floral buds.

Discussion

Squalene synthase occupies a critical position in the biosynthetic pathway leading to the formation of phytosterols, sterols, brassinosteroids and triterpenes. This enzyme catalyzes the conversion of a soluble substrate FPP into squalene, a hydrophobic compound. Plants of the Solanaceae family use this pathway to biosynthesize SGAs, which can accumulate in the plant to similar levels as the phytosterols. These different requirements for squalene raise the possibility of a family of squalene synthases in solanaceous plants to accommodate the diverse metabolic requirements for squalene. Here, we present the identification of three *sqs* gene homologs, the genomic organization of *sqs1*, the unusual feature of an intron in the 3'UTR of all three sqs_{Sc} genes, and the transcript accumulation patterns of members of the *sqs* gene family found in *S. chacoense*, a diploid wild relative of the cultivated potato.

The coding regions of the four *sqs* homologs identified in the *S. phureja* genome assembly (Version 3) were used to design gene-specific oligonucleotide primers to isolate the *sqs* orthologs from *S. chacoense*. The objective was to clone the four coding regions so that the genes could be expressed heterologously and the gene products characterized for their enzyme kinetics. All angiosperms have orthologs of $sqs1_{Sc}$ and most if not all sequenced genomes have an additional *sqs* gene (data not shown). The cDNAs for three out of the four *S. phureja* genes were obtained from *S. chacoense* (**Figure 2.6**). No homolog of $sqs3_{Sp}$ was detected by RT-PCR of RNA from *S. chacoense* **(Figure 2.6)**. However, a nucleotide sequence of a *S. lycopersicum* BAC (http://solgenomics.net, SL2.40ch10: 60,010,501 - 60,015,400 bp) clustered with the $sqs3_{Sp}$ sequence on a phylogenetic tree with 94.6% identity (**Figure 2.6)** and had the same shortened length of exon 8 (120 bp rather than 147 bp) as does $sqs3_{Sp}$ indicating that the *S. lycopersicum* sequence is most closely related to *sqs3*. The related species *C. annuum* has 2 to 5 *sqs* homologs based on a DNA blot using $sqs1_{Ca}$ as a probe (Lee et al., 2002). These results indicate that *sqs* exists as a gene family in higher plants, in contrast to the single gene in yeast and mammals, and that in certain taxa such as *Capsicum* and *Solanum*, which have similar genomic structures, 3 to 4 *sqs* homologs per haploid genome can be expected.

The similarity of the *sqs* gene family members and the conservation of the intron-exon gene structure including the intron in the 3' UTR indicate that the gene family members in *Solanum* are products of gene duplication. Gene duplication is a common occurrence, and there are several evolutionary outcomes predicted for duplicated genes that can contribute to enhanced plant fitness (Moore and Purugganan, 2005). The majority of duplicated genes are eliminated from the genome (Cooke et al., 1997), but those that are preserved can have a dose-dependent effect or subfunctionalized/neofunctionalized properties relative to the parental gene (Flagel and Wendel, 2009; Hahn, 2009). In *S. chacoense*, transcripts of *sqs1*$_{Sc}$ and *sqs2*$_{Sc}$ are detected in all tissues, but the levels vary (Force et al., 1999). Furthermore, as described in Chapter 3, the coding region of each gene has amino acid sequence differences that may alter SQS catalytic activity in the case of SQS2$_{Sc}$ or substrate preference leading to neofunctionalization as may be the case with SQS4$_{Sc}$. The *sqs4*$_{Sc}$ gene was expressed almost exclusively in bud tissue and had only 76.1% identity at the nucleotide level and 68.5% at the amino acid level to *sqs1*$_{Sc}$. These distinct patterns of transcript accumulation and nucleotide sequence indicate distinct phenotypes for each member of the *sqs*$_{Sc}$ gene family.

The *sqs*$_{Sc}$ gene family in *S. chacoense* has introns in the 3'UTR

Alignment of the three cDNA sequences of *sqs*$_{Sc}$ genes with their orthologous genomic sequences of the *S. phureja* genome assembly revealed gaps in the 3'UTRs defined at the beginning and end by intron splice junction sequences (GT-AG).   These gaps began approximately at the same position in all three genes, 6-12 bp after the stop codon (**Figure 2.3, Figure 2.8**). These properties indicate the presence of an intron, which is typically absent in the 3' UTR of genes. For instance, only 1.7% of *A. thaliana* genes have them, and their occurrence is often linked to posttranscriptional control of mRNA fidelity (Kertesz et al., 2006). Despite major gene size differences between *sqs4*$_{Sp}$ and the other *sqs* genes due to shorter introns in *sqs4* than in *sqs1*, the 3' UTR intron of *sqs4*$_{Sp}$ was conserved suggesting an essential role for the intron.

Only one of the two *sqs* genes in *N. tabacum, sqs2*$_{Nt}$, has an intron in the 3' UTR (Devarenne et al., 2002). Comparison of the cDNA and genomic 3'UTR sequence of *sqs1* from *C. annuum* also indicated the lack of an intron in the 3' UTR (data not shown).   Comparison of the *C. annuum* and *S. chacoense sqs1*$_{Sc}$ genomic sequences revealed a gap in the *C. annuum* sequence equivalent to intron 13 of *sqs1*$_{Sc}$. The corresponding cDNAs (Lee et al., 2002) for both species align 73% in the 3'UTR, while the coding region is 94% identical (**Figure 2.7**) (**Appendix E**). Introns were not detected in the 3'UTR of *sqs* genes isolated from *A. thaliana*

(Kribii et al., 1997), *T. cuspidata* (Huang et al., 2007), *S. cerevisiae* (Zhang, 1993) or *R. norvegicus* (Gu et al., 1998).

The conservation of intron 13 in *sqs2*$_{Nt}$ and other *Solanum* species yet the absence of the intron in *sqs1*$_{Ca}$ and *sqs1*$_{Nt}$ suggests a functional role for the intron that is critical in specific but not all taxa. Splicing differences in the 3'UTR would not affect the translated gene products but may have a posttranscriptional effect on the rates of messenger transcription or transcript stability (Rose, 2004; Mignone et al., 2005; Nyiko et al., 2009). Silhavy and colleagues (Kertesz et al., 2006) have observed that 3'UTR introns contain *cis*-acting elements that regulate transcription and can be a component of a nonsense-mediated decay mechanism that ensures that the mRNA is full-length and capable of producing a full-length polypeptide. A SQS polypeptide lacking the C-terminal domain is likely to have catalytic activity but will not be associated with the endoplasmic reticulum and the other enzymes of sterol/triterpene biosynthesis due to the loss of the membrane-binding domain (see Chapter 3). This intron may be an important feature of *sqs* in potato because of the large amounts of squalene required for phytosterol and SGA biosynthesis.

Transcript abundance of *sqs*$_{Sc}$ gene family in floral tissues

To assess whether an *sqs*$_{Sc}$ gene family member participates in SGA biosynthesis , we have compared the transcript abundance of each gene in different tissues of *S. chacoense* (**Figure 2.9**) with the accumulation of SGAs in various tissues (**Figure 1.3**) (Mweetwa, 2009). Although previous studies have shown a lack of correlation between *sqs* transcript levels and metabolite accumulation (Devarenne et al., 2002; Wentzinger et al., 2002), the comparison can be informative, as in the complete absence of a gene transcript. Clearly, the transcript profile for *sqs4*$_{Sc}$ does not correlate with SGA levels. For *sqs1* transcript, the most accumulation was in the tubers and stolons, whereas SGA levels in tubers and stolons were the second and third lowest, 8 and 10 μmol g^{-1} dry wt, respectively, after roots (2 μmol g^{-1} dry wt). For *sqs2*, transcript levels were highest in floral and bud tissue, which is consistent with the SGA accumulation pattern at anthesis where flowers and floral buds have the highest SGA concentration of 50 and 65 μmol g^{-1} dry wt, respectively. Aside from the exceptions noted, both *sqs1* and *sqs2* have transcript profiles that approximate the pattern of SGA abundance. Together the results indicate that either *sqs1*, *sqs2*, or both could contribute to SGA biosynthesis.

We confirmed the transcript levels detected for *sqs4*$_{Sc}$ in floral buds with transcript profiles of the doubled monoploid accession of *S. phureja* made accessible though the PGSC (http://www.potatogenome.net/). Based on RNA-seq profiles of different tissues, as well as published reports, elevated transcript levels have been detected for at least one member of the *sqs*, *hmg*, and *squalene epoxidase* gene families in stamen tissue (Korth et al., 1997) **(Appendix F).** It is unknown what the basis is for the high steady-state abundance of *hmg* family members in the floral tissue, but genes induced under plant defense are commonly constitutively expressed during floral development (Choi et al., 1992; Korth et al., 1997). In plant species including *Lotus japonicus, N. tabacum,* and *E. tirucalli sqs* transcript is more abundant in tissues that are proliferating or accumulating sterols and triterpenes (Hayashi et al., 1999; Devarenne et al., 2002; Uchida et al., 2009). In potato, genes controlling the carbon flow into the sterol or triterpene pathways are increased in stamen tissue. These genes may be supporting increased synthesis of triterpenes, phytosterols, or SGAs.

The *in planta* role of *sqs*$_{Sc}$ genes can be addressed by inducing or suppressing *sqs*$_{Sc}$ gene expression in a transgenic plant and monitoring the resulting metabolite profile. The general phenotype and profiles of squalene-derived compounds in plants with RNAi-mediated suppression of *sqs1*$_{Sc}$, *sqs2*$_{Sc}$ and *sqs4*$_{Sc}$ gene expression would help to determine if the *sqs* genes are functionally redundant. Stable transgenic potato lines that overexpress *sqs*$_{Sc}$ genes may accumulate higher levels of sterols, triterpenes, or other terpenes than potato lines transformed with empty vector. Transgenic lines could be screened for genes that are upregulated coordinately with the overexpression of *sqs* to identify genes that are functionally related to SQS activity.

Materials and Methods

RNA extraction, genomic DNA extraction, and cDNA preparation

DNA and RNA used in this study were isolated by Mweetwa (2009) as described. RNA was quantified spectrophotometrically and visualized on denaturing agarose gels. All cDNA preparations were generated from 2.0 µg of RNA using the Superscript II kit from Invitrogen (Carlsbad, CA) according to the manufacturer's instructions.

Table 2.1 Primer sequence and function

Primer (gene)	Sequence	Purpose
1SQS1utr5	GGAACAGTGTTTGAATTTGTTG	cDNA isolation
3'utrR (*sqs1*)	CTGGGAAAACCTCTGAAACTGT	cDNA isolation
1SQS2utr5b	AACATTCCCTCCAACGCTTC	cDNA isolation
1SQS2utr3	CCAAATGACTCCTAAGTTACAG	cDNA isolation
1SQS4utr5	TAAATTGACACTCCTTAATTAAAC	cDNA isolation
SQS4utrc	GTAGAAAGGATATTATGGGTAC	cDNA isolation
1SQS1ex6ex7	TATGTAGCTGGGCTTGTTGG	Used in Fig. 2.7
1SQS2ex10ex9	TCTTGGTCCGGTCAATGACC	Used in Fig. 2.7
2SQS2e	TGCCTCTGGGAAAGAAGATG	Used in Fig. 2.7
2SQS2d	GCAGGATCCCGCAGTGTAG	Used in Fig. 2.7
1SQS4utr5	TAAATTGACACTCCTTAATTAAAC	Used in Fig. 2.7
SQS4ex13R	GATAAAAGAATAGCCATCATGATG	Used in Fig. 2.7
Ef 1α F	ATTGGAAACGGATATGCTCCA	Used in Fig. 2.7
Ef 1α R	TCCTTACCTGAACGCCTGTCA	Used in Fig. 2.7
10F-internal (*sqs1*)	ACCTTGCATTTGGTGGTATTAC	Genomic DNA
10 R-3 (*sqs1*)	GTTTCTCGTAGCATGATGCA	Genomic DNA
1SQS1.B F (*sqs1*)	TGTCATCTTCATCATACTGGCT	Genomic DNA

Isolation of *sqs1*$_{Sc}$ genomic sequence

Intron 10 and intron 13 were amplified by PCR from 50 ng of genomic DNA as described (**Table 2.1**). Genomic DNA was isolated from leaf tissue by Alice Mweetwa from plants grown under controlled conditions as described (Mweetwa, 2009). For intron 10, the Accuprime enzyme kit was used with the following modifications. The reaction had 2.5 times the suggested amount of reaction mixture and enzyme, and had 3 times the suggested concentration of dNTP in a 20 µl reaction mixture. PCR fragments were purified (NucleoSpin Extract II Kit, Macherey-Nagel, Bethlehem, PA). Intron 13 was amplified using the conditions specified (**Table 2.2).** The intron 13 genomic DNA was cloned into the vector pGEM T Easy (Promega, Madison, WI) and the resulting vector was introduced into chemically competent *E. coli* cells (Top10, Invitrogen, Carlsbad, CA). Plasmids were purified from *E. coli* using the NucleoSpin Plasmid kit (Macherey-Nagel). DNA was sequenced by Quintara Biosciences (Berkeley, CA).

Isolating sqs_{Sc} cDNA

The three cDNAs were amplified by RT-PCR as described in **Table 2.2**. Each cDNA was amplified using the indicated enzyme kit according to the manufacturer instructions. Both $sqs1_{Sc}$ and $sqs2_{Sc}$ were isolated from RNA of flower tissue, while $sqs4_{Sc}$ was isolated from bud tissue RNA. RNA was purified by Mweetwa (2009) from plants grown under controlled conditions. The cDNA fragments were cloned, prepared, and sequenced as described for intron 13.

Table 2.2 PCR conditions for sqs_{Sc} cDNA and intron 10 and intron 13

Amplicon	PCR Kit	Tissue source	Forward Primer	Reverse Primer	PCR Conditions for 30 cycles
$sqs1_{Sc}$	Accuprime	Flower	1SQS1utr5	3'utrR	95°C: 15 s, 55°C: 30 s, 68°C : 90 s
$sqs2_{Sc}$	Accuprime	Flower	1SQS2utr5b	1SQS2utr3	95°C: 15 s, 55°C: 30 s, 68°C : 90 s
$sqs4_{Sc}$	NEB Taq	Bud	1SQS4utr5	SQS4utrc	95°C: 15 s, 55°C: 30 s, 72°C : 90 s
intron 13	NEB Taq	Leaf	1SQS1.BF	3'utrR	95°C: 15 s, 55°C: 30 s, 72°C : 60 s
intron 10	Accuprime	Leaf	10F-internal	10 R-3	95°C: 15 s, 55°C: 30 s, 72°C : 60 s

Sequence and phylogenetic analysis

The Megalign program (DNASTAR Lasergene program version 8.0) was used to align nucleic acid sequences for phylogenetic analysis. The Clustal W algorithm was used with default parameters to produce a neighbor joining tree with bootstraps values indicated.

Tissue specificity screen

RNA used for quantifying *sqs* transcript abundance in specific tissues was isolated from *S. chacoense* chc 80-1 by Mweetwa (2009). The cDNA for the RT-PCR was prepared immediately before PCR. The PCR products were separated by agarose gel electrophoresis and photographed with a sub-saturated exposure. PCR cycle conditions are described in **Table 2.3.**

Table 2.3 PCR conditions for semiquantitative RT-PCR

Gene	Forward primer	Reverse Primer	PCR conditions for 30 cycles
sqs1$_{Sc}$	1SQS1ex6ex7	1SQS2ex10ex9	95°C : 15 s, 55°C : 30 s, 72°C : 20 s
sqs2$_{Sc}$	2SQS2e	2SQS2d	95°C : 15 s, 55°C : 30 s, 72°C : 20 s
sqs4$_{Sc}$	1SQS4utr5	SQS4ex13R	95°C : 15 s, 50°C : 30 s, 72°C : 60 s
ef-1α	Ef 1α F	Ef 1α R	95°C : 15 s, 55°C : 30 s, 72°C : 20 s

Table 2.4. BLAST results of *sqs1*$_{st}$ (AB022599) queried in Version 3 of the genomic assembly of *S. phureja.*

Gene	PGSC Scaffold	e-value	*S. chacoense* name	Scaffold location
$sqs1_{Sp}$	PGSC0003DMS000002689	1.0e-187	$sqs1_{Sc}$	210815-219262
$sqs2_{Sp}$	PGSC0003DMS000002188	2.0e-82	$sqs2_{Sc}$	107238-113092
$sqs3_{Sp}$	PGSC0003DMS000003447	3.1e-45	$sqs3_{Sc}$	436353-432085
$sqs4_{Sp}$	PGSC0003DMS000003447	3.1e-45	$sqs4_{Sc}$	419044-422624

Appendices:

Appendix A Genomic Sequence of $sqs1_{Sc}$ and DNA-blot verified

Appendix B Predicted cDNA coding region of sqs_{Sp} genes

Appendix C Wild type alleles of sqs_{Sc} cDNAs

Appendix D 3’UTR of $sqs1_{Sc}$ and $sqs1_{Ca}$ aligned

Appendix F RNA-seq for *hmg*, *sqs* and *spe* genes

Chapter 3

Squalene synthase from *Solanum chacoense*

Abstract

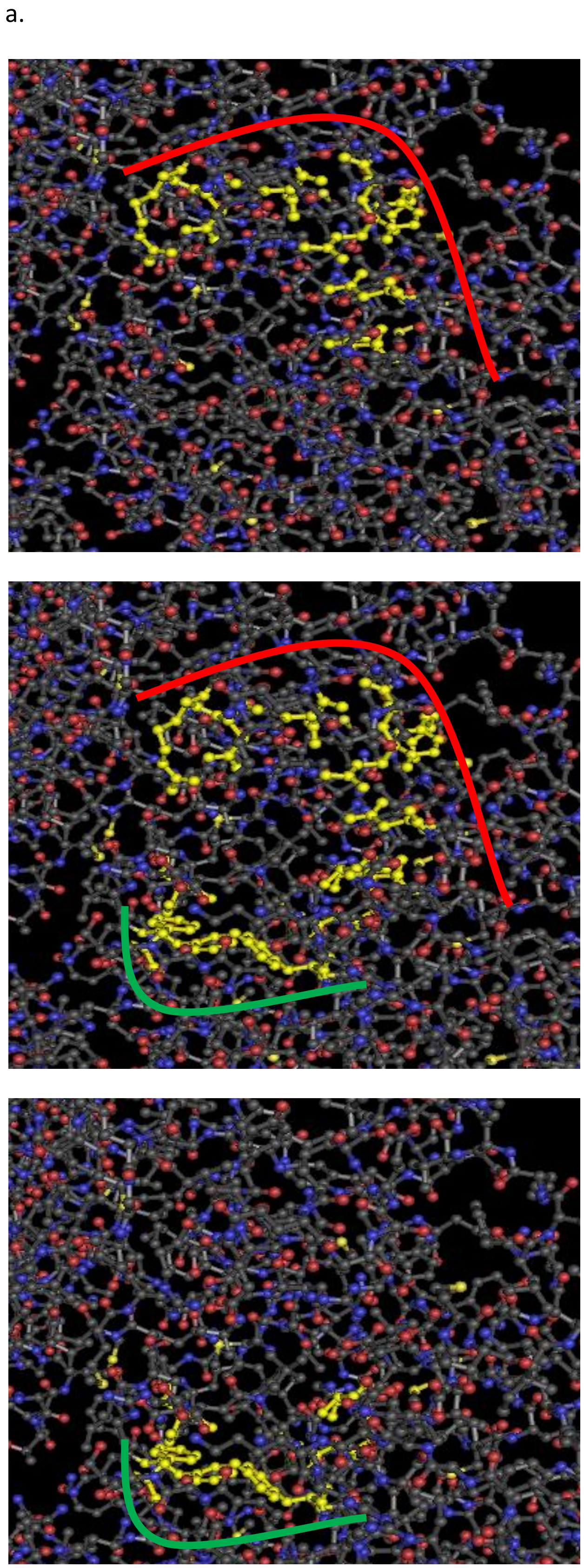
Squalene synthase (EC 2.5.1.21; SQS) catalyzes two separate reactions in the condensation of farnesyl diphosphate to squalene. The reaction steps and the amino acid residues critical for the reactions were used to predict the SQS activity of the uncharacterized enzymes of the wild potato *Solanum chacoense* chc 8380-1. Six different transcripts representing three *sqs*$_{Sc}$ gene homologs were isolated by reverse transcription-PCR of RNA preparations from floral and bud tissues of *S. chacoense*. The six isoforms had between 38% and 42% amino acid identity to the mammalian enzymes from *Homo sapiens* and *Rattus norvegicus* where secondary structure and amino acid residues critical for enzyme activity have been determined. Amino acid residues in SQS$_{Sc}$ that correspond to the catalytic residues in mammalian SQS sequences were identified by sequence alignment. Allele 1 of SQS1$_{Sc}$ and both alleles of SQS2$_{Sc}$ have all of the conserved domains and amino acid residues found in a functional SQS enzyme. Allele 2 of SQS1$_{Sc}$ has an Arg225Cys substitution in domain IV that is at a critical residue for enzyme activity. SQS4$_{Sc}$ has conserved amino acid residues in the catalytic domains required for the first half reaction involved in the condensation of two farnesyl diphosphates (FPPs) to presqualenepyrophosphate, but lacks the conserved residues found in the active site for the second half reaction, indicating that SQS4$_{Sc}$ may only have partial SQS activity. Whereas SQS1 and SQS2 have peptide motifs indicating targeting to the endoplasmic reticulum, SQS4$_{Sc}$ lacks such a motif. Together, these results suggest that these genes encode enzymes with altered activities, subcellular localization, and regulation.

Introduction

The SQS reaction involves two separate half-reactions (**Figure 3.1**). In the first half reaction, two FPPs are combined to produce presqualene pyrophosphate (PSPP) and PP_i and in a second half reaction, PSPP is reduced by NADPH and dephosphorylated resulting in the formation of squalene and PP_i (Gu et al., 1998).

SQS is characterized by six domains (domains I-VI), defined by their sequence conservation in the fungal, plant, and mammalian enzymes (**Figure 3.2**). Domains I-V participate in enzyme catalysis (**Figure 3.1**) (Pandit et al., 2000), whereas domain VI, located at the C-terminal end of the protein, is a transmembrane domain composed of non-polar amino acids with no catalytic function (Busquets et al., 2008). Certain amino acid residues in domains I, III, and IV of SQS from *Rattus norvegicus* were modified to assess the role of individual amino acids in squalene biosynthesis (Gu et al., 1998). Changes in conserved amino acids (Tyr171, Asp219 and Glu222 Asp223, Glu226; numbering position in SQS_{Rn}) in domains III and IV resulted in enzymes that were not able to use FPP as a substrate. Enzymes with changes at nonconserved residues such as Gln283, Phe286, Phe288, Gln293 in SQS_{Rn} domain V were able to catalyze formation of presqualene pyrophosphate PSPP and inorganic diphosphate PP_i, but not into the final product squalene (Gu et al., 1998). The two outcomes of mutagenesis indicated SQS has two active sites that catalyze two reactions (**Figure 3.1**).

A model of the SQS reaction center was generated following X-ray crystallography of human SQS at a 2.15-Ǻngström resolution (**Figure 3.1**) (Pandit et al., 2000). The predicted structure was compared to results of mutagenesis studies that determined which conserved residues were involved in each reaction (Gu et al., 1998). The combined data were sufficient to predict a reaction mechanism that can be applied to other SQSs. The amino acids composing the active site of the second half-reaction were defined, but no mechanism was proposed (Pandit et al., 2000). The amino acids directly involved in the two half-reactions of the enzyme are presented in **Figure 3.1** and **Figure 3.2.**



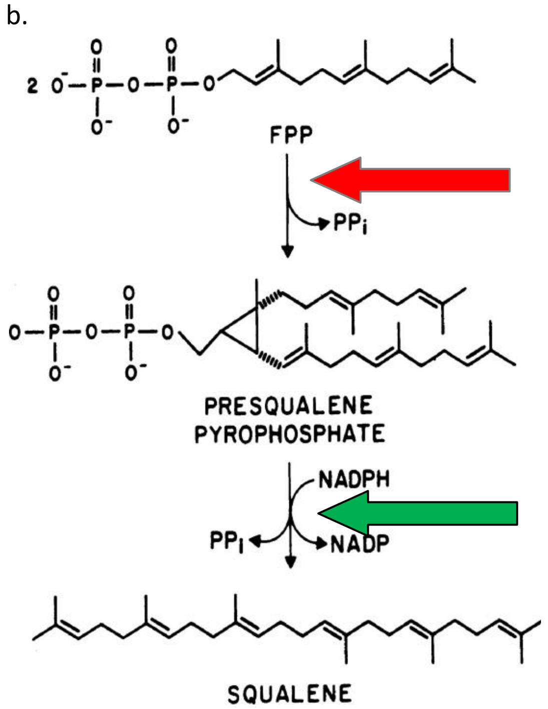


Figure 3.1 Organization of the SQS reaction center (a) and reaction steps (b). (a) The active site of SQS from *Homo sapiens* is shown with the residues involved in catalysis highlighted in yellow. Red curved lines indicate the residues associated with the first half-reaction. Green curved lines indicate the second half-reaction. In the top image the first –half reaction center is highlighted. In the middle image both reaction centers are highlighted. In the bottom image, the second half-reaction site is highlighted. The rest of the protein is colored coded by element: carbon (gray), nitrogen (blue), oxygen (red), and sulfur (yellow). (b) The two reaction steps of squalene synthase, the condensation of two farnesyl diphosphate (FPP) into presqualene pyrophosphate and its conversion to squalene using NADPH as a cofactor.

Unlike fungi and animals, plants have multiple *sqs* genes. Two *sqs* genes have been isolated from many plant species including *Nicotiana tabacum* (Devarenne et al., 2002) and *Glycyrrhiza glabra* (Hayashi et al., 1999). *P. ginseng* has three *sqs* genes (Kim et al., 2011a). A comparison of the first two SQS enzymes isolated from *P. ginseng* found SQS2$_{Pg}$ to be only 35% as efficient as SQS1$_{Pg}$ at converting FPP to squalene (Lee et al., 2004).

In the previous chapter, I described how *S. chacoense* produces three different transcripts coding for proteins that are homologous to previously identified SQS enzymes. Here I describe the deduced amino acid sequences of the three coding regions isolated from *S. chacoense* and compare them with known SQS enzymes.

Results

Comparison of SQS$_{Sc}$ with previously characterized SQS

Using Clustal W (Megalign, DNASTAR Lasergene Version 8), the amino acid sequences of SQS1$_{Sc}$, SQS2$_{Sc}$, and SQS4$_{Sc}$ were aligned with those of *H. sapiens* (Pandit et al., 2000), *R. norvegicus* (Gu et al., 1998), and *N. tabacum* (Devarenne et al., 2002) to determine if the deduced amino acid sequences of the *sqs* genes isolated from *S. chacoense* resemble those of previously described SQS enzymes. As shown in **Figure 3.2**, the SQS$_{Sc}$ proteins contain the five catalytic domains and the carboxy-terminal transmembrane domain that are characteristic of SQS proteins (Pandit et al., 2000). Three gaps corresponding to 3 amino acids each were introduced to optimize the alignment. The six sequences consist of 410-417 amino acids. The predicted molecular masses range from 46.9 kDa to 48.1 kDa. SQS1$_{Sc}$ and SQS4$_{Sc}$ have neutral isoelectric points whereas the isoelectric point of SQS2$_{Sc}$ is high (**Table 3.1**). Because of these structural similarities, predictions about enzyme properties of the sequences isolated from *S. chacoense* can be made based on the previous studies on SQS enzymes.

To assess the probability of SQS enzyme activity for each SQS$_{Sc}$ sequence, I compared the amino acids that participate in the reaction center to other SQS proteins. There were non-synonymous nucleotide polymorphisms in the coding regions of the two alleles of each *sqs* (**Figure 3.3**). Comparisons of the predicted protein sequences between each allele of SQS1$_{Sc}$, SQS2$_{Sc}$, and SQS4$_{Sc}$ revealed 99.0%, 98.5%, and 97.3% identity respectively. SQS4$_{Sc}$ has more polymorphisms between its alleles than do SQS1$_{Sc}$ and SQS2$_{Sc}$. The amino acid sequence coded by each allele of the 3 isoforms

In SQS1$_{Sc}$, one unexpected variation was found at amino acid residue position 225 after the starting methionine **(Figure 3.2**, position 225**)** that stabilizes the diphosphate leaving group in the first half-reaction (Pandit et al., 2000). An arginine residue is observed in the other 29 SQS sequences of plant, fungal, animal or algal origin (**Figure 3.4**, position 243**)**. Two other allelic differences in *sqs1*$_{Sc}$ are non-polar to non-polar substitutions in non-conserved regions of the protein **(Figure 3.3,** Val228Ala and Met338Ile**)**. Other than the allelic difference at position 225, the active sites and catalytic residues of SQS1$_{Sc}$ are identical to those of SQS isolated from other species (**Figure 3.2**; **Figure 3.4**).

```
SQShs  MEFVKCLGHPEEFYNLVRFRIGGKRKVMPKMDQDSLSSSLKTCYKYLNQTSRSFAAVIQALDGEMRNAVCIFYLVLRALDTLEDD 86
SQSrn  MEFVKCLGHPEEFYNLLRFRMGGRRNFIPKMDRNSLSNSLKTCYKYLDQTSRSFAAVIQALDGDIRHAVCVFYLILRAMDTVEDD
SQS1nt MGSLRAILKNPDDLYPLVKLKLAARHAEKQIPP---SPHWGFCYSMLHKVSRSFALVIQQLPVELRDAVCIFYLVLRALDTVEDD
SQS1sc MGTLRAILKNPDDLYPLIKLKLAARHAEKQIPP---EPHWGFCYLMLQKVSRSFALVIQQLPVELRDAVCIFYLVLRALDTVEDD
SQS2sc MGILRAILKHPEDIYPLLKLKVAARYAEKQIPS---QPHWAFCYIMLHKVSRSFSLVIKQLPVELRDAICIFYLVLRALDTVEDD
SQS4sc MELMQEILMHPDELYPLVKLMLTAKRVEKKTSVWLLQPYWAFCYATLRKVSRSFALVIQQLPSDLRNVVCVYYLVLRALDTVEDD
       XXRR motif                                      domain I                 domain II

SQShs  MTISVEKKVPLLHNFHSFLYQPDWRFMESKEKDRQVLEDFPTISLEFRNLAEKYQTVIADICRRMGIGMAEFLDHVTSEQEWDKK 170
SQSrn  MAISVEKKIPLLRNFHTFLYEPEWRFTESKEKHRVVLEDFPTISLEFRNLAEKYQTVIADICHRMGCGMAEFLNKDVTSKQDWDK
SQS1nt TSIPTDVKVPILISFHQHVYDREWHFSCGTKEYKVLMDQFHHVSTAFLELRKHYQQAIEDITMRMGAGMAKFICKEVETTDDYDE
SQS1sc TSIPTDVKVPILISFHQHVYDREWHFACGTKEYKVLMDQFHHVSTAFLELGKLYQQAIEDITMRMGAGMAKFICKEVETTDDYDE
SQS2sc TSVATEVKVPILMSFHRHVYDREWHFSCGTKDYKVLMDQFHHVSTAFLELGKHYKEAIEDITMRMGAGMAKFIYKEVETIDDYDE
SQS4sc TSLAIEVRVPILRNFYCNFYDPQWHFSCGTKAFKVLMDQFHHVSIAFLELDTNYQEVIKDITKGMGKGMAKFLCKEVETIDDYNE

SQShs  YCHYVAGLVGIGLSRLFSASEFEDPLVGEDTERANSMGLFLQKTNIIRDYLEDQQGG---REFWPQEVWSRYVKKLGDFAKP 252
SQSrn  YCHYVAGLVGIGLSRLFSASEFEDPIVGEDTECANSMGLFLQKTNIIRDYLEDQQEG---RQFWPQEVWGKYVKKLEDFVKP
SQS1nt YCHYVAGLVGLGLSKLFHASGKED---LASDSLSNSMGLFLQKTNIIRDYLEDINEVPKCRMFWPREIWSKYVNKLEELKYE
SQS1sc YCHYVAGLVGLGLSKLFHASGTED---LASDSLSNSMGLFLQKTNIIRDYLEDINEVPKCCMFWPREIWSKYVNKLEDLKYE
SQS2sc YCHHVAGQVGLGLSKLFHASGKED---VASDSLCNSMGLFLQKTNIIRDYLEDINEVPKCRMFWPRQIWSEYVDKLEDLKYE
SQS4sc YSFYASGLCGLGLSKFFYVSGRED---LAPESISISMGLFLQKISIIRDYLEDINEVPKCRMFWPRQIWSKYVNKLEDFKYE
         domain III                       domain IV

SQShs  ENIDLAVQCLNELITNALHHIPDVITYLSRLRNQSVFNFCAIPQVMAIATLAACYNNQQVFKGAVKIRKGQAVTLMMDATNMPA 336
SQSrn  ENVDVAVKCLNELITNALQHIPDVITYLSRLRNQSVFNFCAIPQVMAIATLAACYNNHQVFKGVVKIRKGQAVTLMMDATNMPA
SQS1nt DNSAKAVQCLNDMVTNALSHVEDCLTYMSALRDPSIFRFCAIPQVMAIGTLAMCYDNIEVFRGVVKMRRGLTAKVIDQTRTIAD
SQS1sc ENSVKAVQCLNEMVTNALSHVEDCLTYMFNLRDPSIFRFCAIPQVMAIGTLAMCYDNIEVFRGVAKMRRGLTAKVIDRTKTMAD
SQS2sc GNSVKAVQCLNEMVTNALSHAEDCLTFLSTLRDPTIFRFCAIPQAMAIGTLAKCYNNIEVFRGVVKMRRGLTAQVIDRTRNMAD
SQS4sc ENSVKAVQCLNEMVTNALLYVEDCLTSMSSLRDPAIFQFCAIPQIINMGNLSMYYNNVEIFKGVVEMRRGLCARIIDQTRTMAD

                                    domain V

SQShs  VKAIIYQYMEEIYHRIPDSDPSSSKTRQIISTIRTQNLPNCQLISRSHYSPIYLSFVMLLAALSWQYLTTLSQVTEDYVQTGEH 420
SQSrn  VKAIIYQYIEEIYHRVPNSDPSASKAKQLISNIRTQSLPNCQLISRSHYSPIYLSFIMLLAALSWQYLSTLSQVTEDYVQREH
SQS1nt VYGAFFDFSCMLKSKVNNNDPNATKTLKRLEAILKTCRDSGTLNKRKSYIIRSEPNYSPVLIVVIFIILAIILAQLSGNRS
SQS1sc VYGAFFDFSCILKSKVNNNDPNATKTLKRLDAILKTCRDSGTLNKRKSYIIRSEPNYSPVLIVVIFIILAIILAQLSGNRS
SQS2sc VYGAFFDFSCILKSKVEYKDPHVAKTLKRLEVILRTCKNSGTLNKRKSFVIKSGPNYNSTFVVVLVVLVAILLGYQSGNRT
SQS4sc VYGAFYDFCCVMESKVDRDDPNATSTLKRLEAILKTCRDSGTLNQRKSYTFSHQPNYNIPVLIIFFFIMMAILLSTKIP
                                                   domain VI (Transmembrane Domain)
```

Figure 3.2. Amino acid sequence alignment of SQS from *Homo sapiens*, *Rattus norvegicus*, *Nicotiana tabacum*, and *Solanum chacoense*. The deduced amino acid sequences of *sqs* isolated from *H. sapiens* (Pandit et al., 2000), *R. norvegicus* (McKenzie *et al.*, 1992) accession No. M95591, and *N. tabacum* (Yoshioka et al., 1999; Devarenne et al., 2002) GenBank accession No. U59683, are aligned with the sequences including SQS1$_{Sc}$ allele 2, SQS2$_{Sc}$ allele 2, and SQS4$_{Sc}$ allele 1. The conserved domains that contain residues of the active site are underlined, labeled, and have a gray background. The domains I-V were based on the structural comparison between *H. sapiens* and fungal SQS (Robinson et al., 1993). Residues predicted to be in the first half-reaction center are colored red. Residues predicted to be in the second half-reaction center are colored in green. The amino acids in the transmembrane domain are italicized. The position in the alignment for the *H. sapiens* sequence is indicated at the end of each row. Sequence differences in conserved domains are highlighted in purple.

SQS1$_{Sc}$: allele1/allele2

M**GTLR**AILKNPDDLYPLIKLKLAARHAEKQIPPEPHWGFCYLMLQK**VSRSF**ALVIQQLPVELRDAVCIF**YLVLRALDTVE 80**

DDTSIPTDVKVPILISFHQHVYDREWHFACGTKEYKVLMDQFHHVSTAFLELGKLYQQAIEDITMRMGAGMAKFICKEVE 160

TTDDYDE**YCHYVAGLVGLGL**SKLFHASGTEDLASDSLSNS**MGLFLQKTNIIRDYLED**INEVPKC **(R/C)** MFWPREIWSKYVNKLED 242

LKYEENSVKAVQCLNEMVTNALSHVEDCLTYMFNL**RDPSIFRFCAIPQVMAIGTL**AMCYDNIEVFRGV (V/A) KMRRGLTAKVIDRT 323

KTMADVYGAFFDFSC (M/I) LKSKVNNNDPNATKTLKRLDAILKTCRDSGTLNKRKSYIIRSEPNYS*PVLIVVIFIILAIILAQLSGNRS* 411

SQS2$_{Sc}$: allele1/allele2

MGILRAIL (K/R) HPEDIYPLLKLKVAARYAEKQIPSQPHWAFCYIMLHKVSRSFSLVIKQLPVELRDAICIF**YLVLRALDTVE** 80

DDTSVATEVKVPILMSFHRHVYDREWHFSCGTKDYKVLMDQFHHVSTAFLELGKHYKEAIEDITMRMGAGMAKFIYKEVE 160

TIDDYDE**YCHHVAG** (Q/L) **VGLGL**SKLFHASGKEDVASDSLCNS**MGLFLQKTNIIRDYLED**INEVPKC**R**MFWPRQIWS (K/E) YVDKL 240

EDLKYEGNSVKAVQCLNEMVTNALSHAEDCLTFLSTL**RDP (A/T) IFRFCAIPQAMAIGTL**AKCYNNIEVFRGVVKMRRGLTAQVIDR 324

TRNMAD (A/V) YGAFFDFSCILKSKVEYKDPHVAKTLKRLEVILRTCKNSGTLNKRKSFVIKSGPNYN*STFVVVLVVLVAILLGYQSGNRT 411*

SQS4$_{Sc}$: allele1/allele2

MELMQEILMHPDELYPLVKLMLTAKRVEKKTSVWLLQP (Y/H) WAFCYATLRK**VSRSF**ALVIQQLPSDLRNVVCVY**YLVLRALDTVEDD**T 86

SLAIEVRVPILRNFYCNFYDPQWHFSCGTKAFKVLMDQFHHVS (I/T) AFLELDTNYQEVIKDITK (G/R) MG (K/E) GMAKFLCKEV 162

ETIDD**YNEYSFYASGLCGLGL**SKFFYVSGREDLAPESISIS**MGLFLQK (I/M) SIIRDYLED**INEVPKC**R**MFWPRQIWSKYVNKLEDFKYE 249

ENSVKAVQCLNEMVTNALLYVEDCLTSMSSL**RDPAIF (Q/K) FCA (I/F) PQIINMGNL**SMYYNNVEIFKGVVEMRRGLCA (R/K) IIDQT 382

RTMADVYGAFY (D/Y) FCC (V/I) MESKVDRDDPNATSTLKRLEAILKTCRDSGTL (N/S) QRKSYTFSHQPNYN*IPVLIIFFFIMMAILLS* 404

Figure 3.3 Sequence comparisons of the alleles of SQS1$_{Sc}$, SQS2$_{Sc}$, and SQS4$_{Sc}$. The predicted amino acid sequences of both alleles of SQS1$_{Sc}$ are between position 161 and 411 because these are the only portions with allelic differences. The entire coding sequence of SQS2$_{Sc}$ and SQS4$_{Sc}$ are shown. The conserved amino acids are in bold. The non-polar domain is italicized. The sequence differences are indicated in red and defined by parentheses. The residue of allele type 1 is indicated first and that of allele 2 is second separated by a "/". SQS1$_{Sc}$ has 3 sequence differences at positions 225, 311, 341 relative to the starting methionine. SQS2$_{Sc}$ has 5 non-synonymous differences at positions 9, 175, 235, 281, and 341. There are 11 such differences between the two alleles of SQS4$_{Sc}$ at positions 39, 130, 149, 152, 211, 287, 291, 323, 340, 344, and 377.

SQS2$_{Sc}$ was compared to other SQS amino acid sequences. Histidine was observed instead of tyrosine at position 171. This different residue should not affect SQS function because it is observed in other enzymes that were observed as catalyzing the same head-to-head prenyl-diphosphate dimerization. Two examples are SQS from *Yarrowia* (Merkulov et al., 2000) and also in phytoene synthases from *S. lycopersicum* (Misawa et al., 1994). At position 179 of the isoform encoded by one allele of *sqs2*$_{Sc}$ glutamine was substituted for leucine found in the other allele and other SQS isoenzymes (**Figure 3.2**). Although the position has no known association with catalysis, it is a conserved residue that supports the second half-reaction. This polymorphism may affect the efficiency of the second half-reaction or the SQS reaction velocity of the allele. Two differences of Tyr171His and Leu179Gln were found in domain IV between the active sites (**Figure 3.3**). Excluding these differences, SQS2$_{Sc}$ does not vary in the catalytic residues from other SQS protein sequences. SQS2$_{Sc}$ has some differences in conserved domains, and also has and conspicuously higher isoelectric point than SQS1$_{\mathrm{Sc}}$, but there are no differences that should prevent catalytic activity.

SQS4$_{Sc}$ was also compared to other SQS amino acid sequences to predict catalytic activity. The residues that catalyze the first half-reaction in domain II and IV are identical to other SQS sequences (**Figure 3.2**). The residues in domain III and V, which are mostly involved with the second half reaction, had different amino acid sequences compared to other SQS sequences in the residues between the catalytic residues. For instance, a Val to Cys substitution at position 179 may be involved in the second half-reaction center (Pandit et al., 2000) (**Figure 3.2**). There were also sequence differences in the conserved domains flanking the catalytic residues (**Figure 3.2**). SQS4$_{Sc}$ has a reaction center that should be able to catalyze FPP → PSPP, but the residues that form the channel between the two reaction centers of SQS and the residues that make up the second half-reaction center do not match those observed with catalytically active SQS.

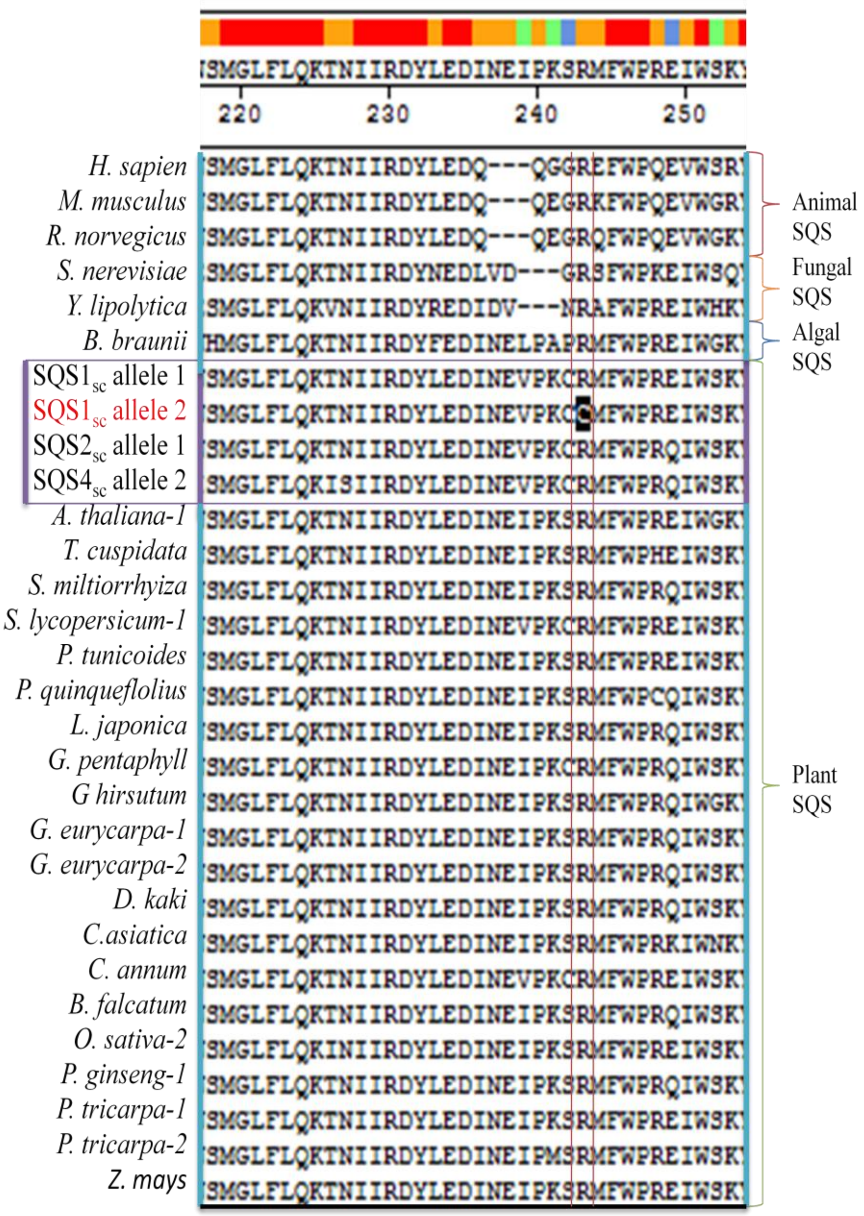

SMGLFLQKTNIIRDYLEDINEIPKSRMFWPREIWSK
220
230
240
250
H. sapien SMGLFLQKTNIIRDYLEDQ---QGGREFWPQEVWSR
M. musculus SMGLFLQKTNIIRDYLEDQ---QEGRKFWPQEVWGR
R. norvegicus SMGLFLQKTNIIRDYLEDQ---QEGRQFWPQEVWGK
S. nerevisiae SMGLFLQKTNIIRDYNEDLVD---GRSFWPKEIWSQ
Y. lipolytica SMGLFLQKVNIIRDYREDIDV---NRAFWPREIWHK
B. braunii HMGLFLQKTNIIRDYFEDINELPAPRMFWPREIWGK
SQS1sc allele 1 SMGLFLQKTNIIRDYLEDINEVPKCRMFWPREIWSK
SQS1sc allele 2 SMGLFLQKTNIIRDYLEDINEVPKCCMFWPREIWSK
SQS2sc allele 1 SMGLFLQKTNIIRDYLEDINEVPKCRMFWPRQIWSK
SQS4sc allele 2 SMGLFLQKISIIRDYLEDINEVPKCRMFWPRQIWSK
A. thaliana-1 SMGLFLQKTNIIRDYLEDINEIPKSRMFWPREIWGK
T. cuspidata SMGLFLQKTNIIRDYLEDINEIPKSRMFWPHEIWSK
S. miltiorrhyiza SMGLFLQKTNIIRDYLEDINEIPKSRMFWPRQIWSK
S. lycopersicum-1 SMGLFLQKTNIIRDYLEDINEVPKCRMFWPREIWSK
P. tunicoides SMGLFLQKTNIIRDYLEDINEIPKSRMFWPREIWSK
P. quinqueflolius SMGLFLQKTNIIRDYLEDINEIPKSRMFWPCQIWSK
L. japonica SMGLFLQKTNIIRDYLEDINEIPKSRMFWPRQIWSK
G. pentaphyll SMGLFLQKTNIIRDYLEDINEIPKCRMFWPRQIWSK
G hirsutum SMGLFLQKTNIIRDYLEDINEIPKSRMFWPRQIWGK
G. eurycarpa-1 SMGLFLQKTNIIRDYLEDINEIPKSRMFWPRQIWSK
G. eurycarpa-2 SMGLFLQKTNIIRDYLEDINEIPKSRMFWPRQIWSK
D. kaki SMGLFLQKTNIIRDYLEDINEIPKSRMFWPRQIWSK
C.asiatica SMGLFLQKTNIIRDYLEDINEIPKSRMFWPRKIWNK
C. annum SMGLFLQKTNIIRDYLEDINEVPKCRMFWPREIWSK
B. falcatum SMGLFLQKTNIIRDYLEDINEIPKSRMFWPRQIWSK
O. sativa-2 SMGLFLQKINIIRDYLEDINEIPKSRMFWPREIWSK
P. ginseng-1 SMGLFLQKTNIIRDYLEDINEIPKSRMFWPRQIWSK
P. tricarpa-1 SMGLFLQKTNIIRDYLEDINEIPKSRMFWPREIWSK
P. tricarpa-2 SMGLFLQKTNIIRDYLEDINEIPMSRMFWPREIWSK
Z. mays SMGLFLQKTNIIRDYLEDINEIPKSRMFWPREIWSK
Animal SQS
Fungal SQS
Algal SQS
Plant SQS

Figure 3.4 Alignment of domain IV and surrounding residues of 30 SQS isoenzymes showing a unique Cys residue in allele 2 of SQS1$_{Sc}$. SQS from three animal, two fungal, one bacterial, and 20 plant sources, and both alleles of SQS1$_{Sc}$, SQS2$_{Sc}$, SQS4$_{Sc}$, were aligned using Clustal W. The sequences of the SQS$_{Sc}$ proteins are outlined in a purple box. The Cys residue unique to SQS1$_{Sc}$ allele 2 has a black background between red bars to indicate that position in other isoforms. The sequences shown are an excerpt from an entire ORF alignment; the position in the alignment these are taken from is indicated at the top of the figure with the numbering relative to species. Sequences are arranged phylogenetically, indicated by brackets on the right. The species are listed to the left of the sequence; the accession number and complete names are indicated in **Appendix G**.

Localization

SQS proteins are known to localize to the endoplasmic reticulum (ER) (Busquets et al., 2008) and have not been described in any other sub-cellular location. A consensus of I used approximately 50 algorithms for subcellular sorting, to analyze the three SQS$_{Sc}$ and identify localization signals indicated that SQS1$_{\mathrm{Sc}}$ and SQS2$_{\mathrm{Sc}}$ are likely to be retained in the ER. The algorithm pSORT WoLF, identified N-terminal "double-arginine like motifs" which are sufficient for ER retention in plants (Schutze et al., 1994) were identified in SQS1$_{Nt}$ (GSLR), SQS1$_{Sc}$ (GTLR), and SQS2$_{Sc}$ (GILR). No plant localization motifs were identified in SQS4$_{Sc}$ suggesting that SQS4$_{Sc}$ is not retained in the ER by a similar mechanism as the other polypeptides. Indeed, the double-arginine motif is not an ER targeting sequence in animals (**Figure 3.2**) and no ER targeting sequence has been found in animal SQS using algorithms predicting protein localization in animal systems. In mammals SQS is synthesized directly on the ER membrane and may be retained by a protein interacting with the SQS C-terminal domain.

Heterologous gene expression constructs

To characterize the encoded SQS activity, I prepared inducible expression constructs containing the coding region of *sqs1*$_{Sc}$, *sqs2*$_{Sc}$ and *sqs4*$_{Sc}$. The *sqs*$_{Sc}$ expression constructs were based on other constructs used for *sqs* expression (Inoue et al., 1995; Hayashi et al., 1999; Akamine et al., 2003; Busquets et al., 2008; Lee and Chappell, 2008; Uchida et al., 2009). The *E. coli* expression vector pET32 was used by Akamine et al. (2003) and Uchida et al. (2009) for plant SQS enzyme assays. Gene expression constructs containing the entire coding region were poorly expressed in *E. coli* but removal of the membrane anchoring domain led to elevated expression levels and the accumulation of active enzyme. The N-terminus of SQS has no negative effects on the accumulation of SQS when expressed in *E. coli.*

The transmembrane domain (domain VI) in each of the three SQS proteins was identified in order to remove it at a natural junction from the remainder of the coding region. Hydrophobicity and surface probability plots of each protein (**Figure 3.5,** Protean, DNASTAR Lasergene program 8) identified a site after an "NY" motif, which introduces kinks in protein structure. A similar motif "HY" was observed in the animal SQS proteins (**Figure 3.2**). The *sqs* coding region for the expression vector was designed to begin at the starting methionine and end immediately before the transmembrane domain at position 388 of SQS1$_{Sc}$ and SQS2$_{Sc}$ and at position 391 for SQS4$_{Sc}$; the difference in position is due to an additional three amino acids in the N-terminal region.

The truncated coding regions were cloned into pET32a and the entire coding region of the vector was sequenced to verify that the inserts were in frame with the vector coding regions (**Appendix H**). SQS2$_{Sc}$ was sequenced in only a portion of the construct (**Appendix I**). SQS4$_{Sc}$ was not cloned into an expression vector. The molecular weight and isoelectric point of each recombinant protein was predicted using EMBOSS (**Table 3.1**)**.** The expression constructs were cloned into the BL21 (DE3) strain of *E. coli* and induced for expression of the SQS fusion protein. Crude protein extracts were prepared from the induced bacterial cultures, fractionated by denaturing protein gel electrophoresis and stained with Coomassie Brilliant Blue. I could not detect any novel bands from the IPTG-induced bacterial cultures that had the predicted mobility of the SQS fusion protein**.**

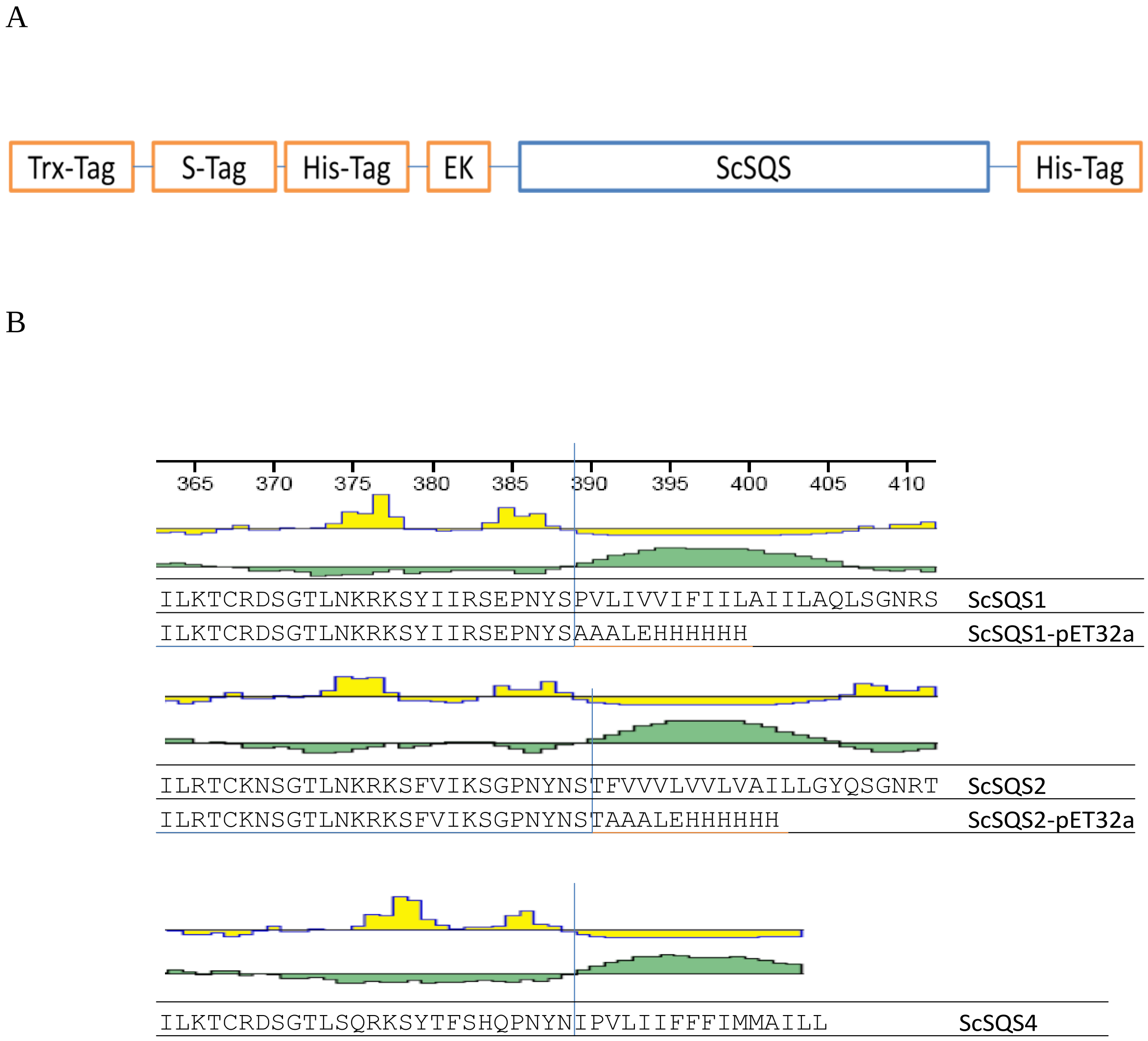

A
Trx-Tag
S-Tag
His-Tag
EK
ScSQS
His-Tag
B
365
370
375
380
385
390
395
400
405
410
ILKTCRDSGTLNKRKSYIIRSEPNYSPVLIVVIFIILAIILAQLSGNRS
ScSQS1
ILKTCRDSGTLNKRKSYIIRSEPNYSAAALEHHHHHH
ScSQS1-pET32a
ILRTCKNSGTLNKRKSFVIKSGPNYNSTFVVVLVVLVAILLGYQSGNRT
ScSQS2
ILRTCKNSGTLNKRKSFVIKSGPNYNSTAAALEHHHHHH
ScSQS2-pET32a
ILKTCRDSGTLSQRKSYTFSHQPNYNIPVLIIFFFIMMAILL
ScSQS4

Figure 3.5. Recombinant SQS$_{Sc}$ protein in *E. coli*. (A) The peptide domains of the recombinant SQS protein in pET32a with N-terminal and C-terminal peptide tags. Tags from the vector are bordered in orange. (B) The deduced amino acid sequences of the SQS$_{Sc}$ proteins were used to generate a Kyte-Doolittle hydrophobicity plot (yellow fill) and a surface probability plot (green fill). The point the transmembrane domain begins is indicated by a blue line. SQS1$_{Sc}$ and SQS2$_{Sc}$ are aligned to the deduced amino sequence of the gene expression construct.

Discussion

S. chacoense has three genes that have 99%, 81%, and 69% identity to the amino acid sequence of the potato SQS enzyme (Yoshioka et al., 1999). The nucleotide sequences described here can give insights into the catalytic capabilities of the three SQS isoforms based on our current understanding of the reaction mechanism and the predicted amino acid sequence of the three genes (Figure 3.1, Figure 3.2). One of the two alleles of *sqs1*$_{Sc}$ was identical to SQS1$_{Nt}$ in the deduced amino acid sequence of domains I – V, which contain the enzyme active sites. Because SQS1$_{Nt}$ has been shown to have *in vitro* SQS activity, the *sqs*1$_{Sc}$ allele should code for an active SQS enzyme. However, allele 2 of *sqs1*$_{Sc}$ has a SNP that may result in an inactive enzyme. The diphosphate leaving group formed with PSPP is stabilized by Arg225 and the Arg225Cys substitution in allele 2 of SQS1$_{Sc}$ is likely to affect this interaction (Pandit et al., 2000) (Figure 3.2).

Both alleles of SQS2$_{Sc}$ have all of the catalytic residues required for SQS activity (**Figure 3.2**). A targeting sequence identified in the N-terminal region is likely to confer retention of the enzyme to the ER. One of the differences between SQS2$_{Sc}$ and other SQS sequences is the Tyr171His substitution in domain III. This substitution has been found in a functional SQS isolated from *Yarrowia* (Merkulov et al., 2000) and is common in plant phytoene synthases (Misawa et al., 1994). Phytoene synthase catalyzes the same head-to-head dimerization reaction mechanism to condense GGPP instead of FPP. I conclude that because of the similarities SQS2$_{Sc}$ has to other SQS enzymes it should function as an SQS.

Unlike SQS1$_{Sc}$ and SQS2$_{Sc}$, SQS4$_{Sc}$ was shown to have active sites unlike other SQS enzymes, and possibly different subcellular localization. SQS4$_{Sc}$ has several features that indicate the enzyme functions differently than other SQS enzymes. WhileSQS1$_{Sc}$ and SQS2$_{Sc}$ both have double-arginine ER-retention tags (**Figure 3.2**) (Schutze et al., 1994), SQS4$_{Sc}$ was found to lack any ER targeting sequences based on bioinformatic predictions. SQSs are generally considered to be ER localizing proteins, and ER localization was demonstrated in *Arabidopsis thaliana* (Busquets et al., 2008) and *R. norvegicus* (Stamellos et al., 1993). If SQS4$_{Sc}$ does not localize to the ER, it may catalyze a novel reaction in a different organelle. To clarify localization, a GFP-SQS4 gene fusion construct could be expressed in hairy root cultures to identify the subcellular localization by fluorescent microscopy.

The trans-prenyl first-half reaction site of SQS4$_{Sc}$ is intact, indicating that this enzyme should be able to catalyze the condensation of two prenyl diphosphates such as FPP. However, the second half-reaction center does not resemble a SQS active site (**Figure 3.2**). It is unlikely that SQS4$_{Sc}$ can complete the formation of squalene and therefore PSPP is a possible product of SQS4$_{Sc}$. PSPP has not been shown to accumulate in cells, however, phosphatases that remove one phosphate group from PSPP have been identified in animal, plants, and yeast species (Theofilopoulos et al., 2008). Besides FPP, other potential substrates for this enzyme include IPP and DMAPP, though they are smaller than FPP. GGPP is produced in the ER in addition plastids and mitochondria (Okada et al., 2000). While SQS localizes to the ER, but because an ER targeting sequence has not been identified in SQS4$_{Sc}$, substrates such as GGPP or its precursor GPP should not be ruled out as candidate substrates.

In addition to the *in vitro* characterization of the SQS proteins, the phenotypes of transgenic plants altered in the expression of the *sqs* genes could also be studied to find biological functions. Systematic RNAi knockdown of each *sqs* gene in a doubled monoploid background of chc 8380-1, with analysis of selected metabolites such as SGAs, phytosterols or brassinosteroids will address the functions of each *sqs* allele. Alternatively, the SQS activity in individuals segregating and genotyped for the *sqs* alleles may reveal biochemical differences between the alleles.

In conclusion, the study has demonstrated that each *sqs* gene of *S. chacoense* has distinct properties that are likely to result in nonredundant functions in sterol and triterpene metabolism.

Materials and Methods

Alignment

Alignments were generated using Clustal W (Thompson J.D., 1994) in the Megalign program of DNASTAR Lasergene (Version 8).

Cloning truncated SQS1$_{Sc}$ and SQS2$_{Sc}$ for protein expression

Oligonucleotide primers were designed to introduce an *Eco*RI site immediately before the starting methionine and a *Not*I site at the beginning of the domain VI in order to exclude the transmembrane domain from the recombinant protein intended for prokaryotic expression. The primer pairs, templates, and PCR conditions for amplification of the coding region of *sqs* are described in **Table 3.2**. Accuprime kit (Invitrogen, Carlsbad, CA) was used following the manufacturer's protocol. The PCR products were separated by agarose gel electrophoresis and gel purified (Macheray-Nagel). Both pET32a and PCR products were separately digested overnight with *Eco*RI and *Not*I. The digested products of pET32a were separated on an agarose gel to isolate the 6 kbp fragment, which was gel purified and ligated with the digested PCR products overnight. Top10 *E. coli* cells were transformed with the ligation reaction. Plasmids isolated from the Top10 cell cultures were transformed into BL21 (DE3) *E. coli* cells (Invitrogen) and also sequenced to verify the region coding for the recombinant protein (**Appendix H**, **Appendix I**).

Table 3.1 Isoelectric point and molecular weight of SQS proteins

SQS	SQS$_{Hs}$	SQS$_{Rn}$	SQS1$_{Nt}$	SQS1$_{Sc}$	SQS2$_{Sc}$	SQS4$_{Sc}$
pI	6.1	6.61	7.91	6.79	8.58	6.06
MW (kD)	48	48	47	47	47	47
pI truncated				6.08	8.47	6.06
MW truncated (kD)				64	64	64

Table 3.2 PCR protocol for generating truncated SQS1$_{Sc}$ and SQS2$_{Sc}$

SQS isoform	Primer pair	Template used	PCR size	Vector size	PCR conditions
SQS1$_{Sc}$ allele 1	SQS1TruncF-SQS1TruncR	SQS1ORF-1-4	1.1 kb	6.0 kb	25 cycles 95°C : 15 sec, 55°C : 30 sec, 68°C: 60 sec
SQS2$_{Sc}$ allele 2	SQS2Trunc2F-SQS2Trunc2R	SQS2-ORF-5	1.1 kb	6.0 kb	

Appendices:

Appendix G: Accession number and species of each SQS isoform from Figure 3.4

Appendix H: Deduced amino acid sequence of protein produced by *sqs1*$_{Sc}$-pET32a

Appendix I: Deduced amino acid sequence of *sqs2*$_{Sc}$ in pET32a

References

A**kamine, S., Nakamori, K., Chechetka, S.A., Banba, M., Umehara, Y., Kouchi, H., Izui, K., and Hata, S.** (2003). cDNA cloning, mRNA expression, and mutational analysis of the squalene synthase gene of *Lotus japonicus*. Biochim. Biophys. Acta-Gene Struct. Expression **1626,** 97-101.

Benveniste, P. (2004). Biosynthesis and accumulation of sterols. Annu.Rev.Plant **55,** 429-457.

Bouvier, F., Rahier, A., and Camara, B. (2005). Biogenesis, molecular regulation and function of plant isoprenoids. Prog Lipid Res **44,** 357-429.

Bushway, A.A., Bushway, R.J., and Kim, C.H. (1990). Isolation, partial-purification and characterization of a potato peel α-solanine cleaving gycosidase. Am. Potato J. **67,** 233-238.

Busquets, A., Keim, V., Closa, M., del Arco, A., Boronat, A., Arro, M., and Ferrer, A. (2008). *Arabidopsis thaliana* contains a single gene encoding squalene synthase. Plant Mol. Biol. **67,** 25-36.

Choe, S. (2010). Brassinosteroid Biosynthesis and Metabolism. in: P. J. Davies (Ed.),. Plant Hormones**,** 156-178.

Choi, D., Ward, B.L., and Bostock, R.M. (1992). Differential induction and suppression of potato 3-hydroxy-3-methylglutaryl coenzyme-a reductase genes in response to *Phytophthora infestan*s and to its elicitor arachidonic-acid. Plant Cell **4,** 1333-1344.

Cooke, J., Nowak, M.A., Boerlijst, M., and MaynardSmith, J. (1997). Evolutionary origins and maintenance of redundant gene expression during metazoan development. Trends Genet. **13,** 360-364.

Coombs, J.J., Douches, D.S., Li, W.B., Grafius, E.J., and Pett, W.L. (2002). Combining engineered (Bt-cry3A) and natural resistance mechanisms in potato for control of Colorado potato beetle. J. Am. Soc. Hort. Sci. **127,** 62-68.

Coombs, J.J., Douches, D.S., Li, W.B., Grafius, E.J., and Pett, W.L. (2003). Field evaluation of natural, engineered, and combined resistance mechanisms in potato for Control of Colorado potato beetle. J. Am. Soc. Hortic. Sci. **128,** 219-224.

Devarenne, T.P., Ghosh, A., and Chappell, J. (2002). Regulation of squalene synthase, a key enzyme of sterol biosynthesis, in tobacco. Plant Physiol. **129,** 1095-1106.

Flagel, L.E., and Wendel, J.F. (2009). Gene duplication and evolutionary novelty in plants. New Phytol. **183,** 557-564.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151,** 1531-1545.

Gallova, J., Uhrikova, D., Kucerka, N., Doktorovova, S., Funari, S.S., Teixeira, J., and Balgavy, P. (2011). The effects of cholesterol and beta-sitosterol on the structure of saturated diacylphosphatidylcholine bilayers. Eur. Biophys. J. **40,** 153-163.

Gu, P., Ishii, Y., Spencer, T.A., and Shechter, I. (1998). Function-structure studies and identification of three enzyme domains involved in the catalytic activity in rat hepatic squalene synthase. (vol 273, pg 12515, 1998). J. Biol. Chem. **273,** 17296-17296.

Hahn, M.W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. J. Hered. **100,** 605-617.

Hata, S., Sanmiya, K., Kouchi, H., Matsuoka, M., Yamamoto, N., and Izui, K. (1997). cDNA cloning of squalene synthase genes from mono- and dicotyledonous plants, and expression of the gene in rice. Plant Cell Physiol. **38,** 1409-1413.

Hayashi, H., Hirota, A., Hiraoka, N., and Ikeshiro, Y. (1999). Molecular cloning and characterization of two cDNAs for *Glycyrrhiza glabra* squalene synthase. Biol. Pharm. Bull. **22,** 947-950.

Huang, Z.S., Jiang, K.J., Pi, Y., Hou, R., Liao, Z.H., Cao, Y., Han, X., Wang, Q., Sun, X.F., and Tang, K.X. (2007). Molecular cloning and characterization of the yew gene encoding squalene synthase from *Taxus cuspidata*. J. Biochem. Molec. Biol. **40,** 625-635.

Inoue, T., Osumi, T., and Hata, S. (1995). Molecular-cloning and functional expression of a cDNA for mouse squalene synthase. Biochim. Biophys. Acta-Gene Struct. Expression **1260,** 49-54.

Jensen, P.H., Pedersen, R.B., Svensmark, B., Strobel, B.W., Jacobsen, O.S., and Hansen, H.C.B. (2009). Degradation of the potato glycoalkaloid α-solanine in three agricultural soils. Chemosphere **76,** 1150-1155.

Kaneko, K., Tanaka, M.W., and Mitsuhashi, H. (1977). Dormantinol, a possible precursor in solanidine biosynthesis, from budding Veratrum grandiflorum. Phytochemistry **16,** 1247-1251.

Kertesz, S., Kerenyi, Z., Merai, Z., Bartos, I., Palfy, T., Barta, E., and Silhavy, D. (2006). Both introns and long 3 '-UTRs operate as cis-acting elements to trigger nonsense-mediated decay in plants. Nucleic Acids Res. **34,** 6147-6157.

Kim, T.D., Han, J.Y., Huh, G.H., and Choi, Y.E. (2011a). Expression and functional characterization of three squalene synthase genes associated with saponin biosynthesis in *Panax ginseng*. Plant Cell Physiol. **52,** 125-137.

Kim, Y.S., Cho, J.H., Park, S., Han, J.Y., Back, K., and Choi, Y.E. (2011b). Gene regulation patterns in triterpene biosynthetic pathway driven by overexpression of squalene synthase and methyl jasmonate elicitation in *Bupleurum falcatum*. Planta **233,** 343-355.

Korth, K.L., Stermer, B.A., Bhattacharyya, M.K., and Dixon, R.A. (1997). HMG-CoA reductase gene families that differentially accumulate transcripts in potato tubers are developmentally expressed in floral tissues. Plant Mol. Biol. **33,** 545-551.

Kribii, R., Arro, M., DelArco, A., Gonzalez, V., Balcells, L., Delourme, D., Ferrer, A., Karst, F., and Boronat, A. (1997). Cloning and characterization of the *Arabidopsis thaliana* SQS1 gene encoding squalene synthase - Involvement of the C-terminal region of the enzyme in the channeling of squalene through the sterol pathway. Eur. J. Biochem. **249,** 61-69.

Krits, P., Fogelman, E., and Ginzberg, I. (2007). Potato steroidal glycoalkaloid levels and the expression of key isoprenoid metabolic genes. Planta **227,** 143-150.

L. Calpe-Berdiel, J.C.E.-G., F. Blanco-Vaca,. (2009). New insights into the molecular actions of plant sterols and stanols in cholesterol metabolism. . Atherosclerosis **203,** 18-31.

Lee, J.H., Yoon, Y.H., Kim, H.Y., Shin, D.H., Kim, D.U., Lee, I.J., and Kim, K.U. (2002). Cloning and expression of squalene synthase cDNA from hot pepper (*Capsicum annuum L.*). Mol. Cells **13,** 436-443.

Lee, M.H., Jeong, J.H., Seo, J.W., Shin, C.G., Kim, Y.S., In, J.G., Yang, D.C., Yi, J.S., and Choi, Y.E. (2004). Enhanced triterpene and phytosterol biosynthesis in Panax ginseng overexpressing squalene synthase gene. Plant Cell Physiol. **45,** 976-984.

Lee, S., and Chappell, J. (2008). Biochemical and genomic characterization of terpene synthases in *Magnolia grandiflora*. Plant Physiol. **147,** 1017-1033.

Merkulov, S., van Assema, F., Springer, J., del Carmen, A.F., and Mooibroek, H. (2000). Cloning and characterization of the *Yarrowia lipolytica* squalene synthase (SQS1) gene and functional complementation of the *Saccharomyces cerevisiae* erg9 mutation. Yeast **16,** 197-206.

Mignone, F., Grillo, G., Licciulli, F., Iacono, M., Liuni, S., Kersey, P.J., Duarte, J., Saccone, C., and Pesole, G. (2005). UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res. **33,** D141-D146.

Misawa, N., Truesdale, M.R., Sandmann, G., Fraser, P.D., Bird, C., Schuch, W., and Bramley, P.M. (1994). Expression of a tomato cdna coding for phytoene synthase in Escherichia-coli, phytoene formation in-vivo and in-vitro, and functional-analysis of the various truncated gene-products. J Biochem **116,** 980-985.

Moore, R.C., and Purugganan, M.D. (2005). The evolutionary dynamics of plant duplicate genes. Curr. Opin. Plant Biol. **8,** 122-128.

Mweetwa, A.M. (2009). Biosynthesis and regulation of steroidal glycoalkaloids in wild potato, *Solanum chacoense* bitter (Blacksburg: Virginia Polytechnic Intitute and State University).

Normen, L., Johnsson, M., Andersson, H., van Gameren, Y., and Dutta, P. (1999). Plant sterols in vegetables and fruits commonly consumed in Sweden. Eur. J. Nutr. **38,** 84-89.

Nyiko, T., Sonkoly, B., Merai, Z., Benkovics, A.H., and Silhavy, D. (2009). Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. Plant Mol. Biol. **71,** 367-378.

Oda, Y., Saito, K., Ohara-Takada, A., and Mori, M. (2002). Hydrolysis of the potato glycoalkaloid alpha-chaconine by filamentous fungi. J. Biosci. Bioeng. **94,** 321-325.

Okada, K., Saito, T., Nakagawa, T., Kawamukai, M., and Kamiya, Y. (2000). Five geranylgeranyl diphosphate synthases expressed in different organs are localized into three subcellular compartments in Arabidopsis. Plant Physiol. **122,** 1045-1056.

Pandit, J., Danley, D.E., Schulte, G.K., Mazzalupo, S., Pauly, T.A., Hayward, C.M., Hamanaka, E.S., Thompson, J.F., and Harwood, H.J. (2000). Crystal structure of human squalene synthase - A key enzyme in cholesterol biosynthesis. J. Biol. Chem. **275,** 30610-30617.

Rangarajan, A., Miller, A.R., and Veilleux, R.E. (2000). Leptine glycoalkaloids reduce feeding by Colorado potato beetle in diploid *Solanum* sp hybrids. J. Am. Soc. Hort. Sci. **125,** 689-693.

Robinson, G.W., Tsay, Y.H., Kienzle, B.K., Smithmonroy, C.A., and Bishop, R.W. (1993). Conservation between human and fungal squalene synthetases - similarities in structure, function, and regulation. Molec. Cell. Biol. **13,** 2706-2717.

Ronning, C.M., Stommel, J.R., Kowalski, S.P., Sanford, L.L., Kobayashi, R.S., and Pineada, O. (1999). Identification of molecular markers associated with leptine production in a population of *Solanum chacoense* Bitter. Theor. Appl. Genet. **98,** 39-46.

Rose, A.B. (2004). The effect of intron location on intron-mediated enhancement of gene expression in Arabidopsis. Plant J **40,** 744-751.

Sandrock, R.W., DellaPenna, D., and VanEtten, H.D. (1995). Purification and characterization of beta(2)-tomatinase, an enzyme involved in the degradation of alpha-tomatine and isolation of the gene encoding beta(2)-tomatinase from *Septoria lycopersici*. Mol. Plant-Microbe Interact. **8,** 960-970.

Schutze, M.P., Peterson, P.A., and Jackson, M.R. (1994). An N-terminal double-arginine motif maintains type-II membrane-protein in the endoplasmic-reticulum. Embo. Journal **13,** 1696-1705.

Seo, J.W., Jeong, J.H., Shin, C.G., Lo, S.C., Han, S.S., Yu, K.W., Harada, E., Han, J.Y., and Choi, Y.E. (2005). Overexpression of squalene synthase in *Eleutherococcus senticosus* increases phytosterol and triterpene accumulation. Phytochemistry **66,** 869-877.

Shakya, R., and Navarre, D.A. (2008). LC-MS analysis of solanidane glycoalkaloid diversity among tubers of four wild potato species and three cultivars (*Solanum tuberosum*). J. Agric. Food Chem. **56,** 6949-6958.

Sinden, S.L., Deahl, K.L., and Aulenbach, B.B. (1976). Effect of glycoalkaloids and phenolics on potato flavor. J. Food Sci. **41,** 520-523.

Sinden, S.L., Sanford, L.L., and Osman, S.F. (1980). Glycoalkaloids and resistance to the Colorado potato beetle in *Solanum chacoense* Bitter. Am Potato J **57,** 331-343.

Sinden, S.L., Sanford, L.L., Cantelo, W.W., and Deahl, K.L. (1986). Leptine glycoalkaloids and resistance to the Colorado potato beetle (*Coleoptera, chrysomelidae*) in *Solanum chacoense*. Environ. Entomol. **15,** 1057-1062.

Stamellos, K.D., Shackelford, J.E., Shechter, I., Jiang, G.J., Conrad, D., Keller, G.A., and Krisans, S.K. (1993). Subcellular-localization of squalene synthase in rat hepatic cells - biochemical and immunochemical evidence. J. Biol. Chem. **268,** 12825-12836.

Theofilopoulos, S., Lykidis, A., Leondaritis, G., and Mangoura, D. (2008). Novel function of the human presqualene diphosphate phosphatase as a type II phosphatidate phosphatase in phosphatidylcholine and triacylglyceride biosynthesis pathways. Biochim. Biophys. Acta Mol. Cell Biol. Lipids **1781,** 731-742.

Thompson J.D., H.D.G., Gibson T.J. . (1994). CLUSTAL W improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acid. Res. **22,** 4673-4680.

Threlfall, D.R., and Whitehead, I.M. (1988). Coordinated inhibition of squalene synthetase and induction of enzymes of sesquiterpenoid phytoalexin biosynthesis in cultures of *Nicotiana tabacum*. Phytochemistry **27,** 2567-2580.

Tjamos, E., and Kuc, J.A. (1982). Inhibition of steroid glyco alkaloid accumulation by arachidonic-acid and eicosapentaenoic-acid in potato. Science (Washington D C) **217,** 542-544.

Uchida, H., Yamashita, H., Kajikawa, M., Ohyama, K., Nakayachi, O., Sugiyama, R., Yamato, K.T., Muranaka, T., Fukuzawa, H., and Takemura, M. (2009). Cloning and characterization of a squalene synthase gene from a petroleum plant, *Euphorbia tirucalli L*. Planta **229,** 1243-1252.

Valkonen, J.P.T., Keskitalo, M., Vasara, T., and Pietila, L. (1996). Potato glycoalkaloids: A burden or a blessing? Crit. Rev. Plant Sci. **15,** 1-20.

Wentzinger, L.F., Bach, T.J., and Hartmann, M.A. (2002). Inhibition of squalene synthase and squalene epoxidase in tobacco cells triggers an up-regulation of 3-hydroxy-3-methylglutaryl coenzyme A reductase. Plant Physiology **130,** 334-346.

Xu, R., Fazio, G.C., and Matsuda, S.P.T. (2004a). On the origins of triterpenoid skeletal diversity. Phytochemistry **65,** 261-291.

Xu, R., Fazio, G.C., and Matsuda, S.P.T. (2004b). On the origins of triterpenoid skeletal diversity. Phytochem. Rev. **65,** 261-291.

Xu, X., Pan, S.K., Cheng, S.F., Zhang, B., Mu, D.S., Ni, P.X., Zhang, G.Y., Yang, S., Li, R.Q., Wang, J., Orjeda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De la Cruz, G., Chakrabarti, S.K., Patil, V.U., Skryabin, K.G., Kuznetsov, B.B., Ravin, N.V., Kolganova, T.V., Beletsky, A.V., Mardanov, A.V., Di Genova, A., Bolser, D.M., Martin, D.M.A., Li, G.C., Yang, Y., Kuang, H.H., Hu, Q., Xiong, X.Y., Bishop, G.J., Sagredo, B., Mejia, N., Zagorski, W., Gromadka, R., Gawor, J., Szczesny, P., Huang, S.W., Zhang, Z.H., Liang, C.B., He, J., Li, Y., He, Y., Xu, J.F., Zhang, Y.J., Xie, B.Y., Du, Y.C., Qu, D.Y., Bonierbale, M., Ghislain, M., Herrera, M.D., Giuliano, G., Pietrella, M., Perrotta, G., Facella, P., O'Brien, K., Feingold, S.E., Barreiro, L.E., Massa, G.A., Diambra, L., Whitty, B.R., Vaillancourt, B., Lin, H.N., Massa, A., Geoffroy, M., Lundback, S., DellaPenna, D., Buell, C.R., Sharma, S.K., Marshall, D.F., Waugh, R., Bryan, G.J., Destefanis, M., Nagy, I., Milbourne, D., Thomson, S.J., Fiers, M., Jacobs, J.M.E., Nielsen, K.L., Sonderkaer, M., Iovene, M., Torres, G.A., Jiang, J.M., Veilleux, R.E., Bachem, C.W.B., de Boer, J., Borm, T., Kloosterman, B., van Eck, H., Datema, E., Hekkert, B.T.L., Goverse, A., van Ham, R., Visser, R.G.F., and Potato Genome Sequencing, C. (2011). Genome sequence and analysis of the tuber crop potato. Nature **475,** 189-U194.

Yang, Z.B., Park, H.S., Lacy, G.H., and Cramer, C.L. (1991). Differential activation of potato 3-hydroxy-3-methylglutaryl coenzyme-a reductase genes by wounding and pathogen challenge. Plant Cell **3,** 397-405.

Yoshioka, H., Yamada, N., and Doke, N. (1999). cDNA cloning of sesquiterpene cyclase and squalene synthase, and expression of the genes in potato tuber infected with *Phytophthora infestans*. Plant Cell Physiol. **40,** 993-998.

Zhang, D.L. (1993). Yeast squalene synthase: expression, purification, and characterization of soluble recombinant enzyme. Arch. Biochem. Biophys. **304,** 133.

Zhang, J.Z. (2003). Evolution by gene duplication: an update. Trends Ecol. Evol. **18,** 292-298.

Appendices

Appendix A: Genomic sequence of *sqs1*$_{Sc}$

*GAACAGTGTTAGAATTTGTTGAGAAGA*ATGGGAACATTGAGGGCGATTCTGAAGAATCCAGATGATTTGTATCCATTGAT 80

AAAGCTGAAACTAGCGGCTAGACATGCGGAAAAGCAGATCCCGCCTGAGCCACATTGGGGCTTCTGTTACTTAATGCTTC 160

AAAAGGTCTCTCGTAGTTTTGCTCTCGTCATTCAACAGCTTCCTGTCGAGCTTCGTGATGCTgtaagtttgtttttttc 240

ttcagaaaaatgctttctctggcatttatgctatagcgcttgcgattcgttaaatttctgattggttttcaatctgtttt 320

aaatttctgtgtgtgtgcgcgtatattcgttttgtagaagtttgtatttttgcttaattaggaagaatttatattctgct 400

tctgtgaattgatcggaattgatataatctattgttattccttttgatttgtgctaatttgctggatagttgtgttacta 480

ttactattgttttgaatttgtttttataacttccattgaacgttgcagGTATGCATATTCTATTTGGTCCTTCGAGCACT 560

GGACACTGTTGgtaagcttggttactatcatctaaatttgtttgtactttatgtattcttaggagatatgaaactcagaa 640

agcagatgactggtattagttcttcatttttgtgcaaactttggtgtgattaatttagtaatttagtcttctatctttga 720

aagagtaggttgaatcgttgattcagttcgtcatacttatcttctattgcttcttgcgaggaacgaaagatgccaagaga 800

gggaataatacctggttgaaaacttccaccctttttattttcaaaaagaggaaaagaagcaagttgttttcccttataaa 880

aaaaaagaagcaaattgttcttctcatactagttatattggtatttgatatttagtttatatgatacgagttgcctacca 960

gcttcctgttgaactcttcttggttataaacaagttttgactaacttttattgcttaaatactagtatgaattatcatat 1040

gctacttactatttcatttcattctagacattcataaggtagtttttaggttgtaatcttagttgattcaagggttgttg 1120

cactcatatttttgagaaacacataatttgcttattttcatattggagtcttattgtggacagAGGATGATACCAGCATA 1200

CCCACCGATGTTAAAGTACCTATTCTGATCTCTTTTCATCAGCATGTTTATGATCGCGAATGGCACTTCGCATgtaagtc 1280

tctgaatgcaacttgttgatctccctaaattctcaatattgcatgagtgtgcttttgcaattgaagaatcccaagttgga 1360

aagacttcaactacttttagaggttactaaggaaattattttgaataagtcagatggaaaccgatgcaaaatatttcac 1440

tgtctacccaaaaatgtttttcttttgctgcacaatgctgtaagttacyaatgatccaagattcacgtcatccactgttt 1520

tcttgcacatactgcaccagtcaccagcggtccataatgcatttcccacttctcaagttgtcgtctgtcaaaacacactt 1600

tgcaaaaggccatagattcctcttgaatttcctatttactaccatattagagacactgtcgacgtgagtcattcaacacc 1680

aaacaattcataatccttgaaatcctgttaaaaaaagcacactttctagttactgaattatatatgttgcagcacaccct 1760

ctcatgtgtgggtttgattctttttcatgggccaaacacatggaaatttattttgctttcttttttaggtggctgtgaga 1840

tttgaaatagggtctctttctctctaatatcatgttgaagtcttatttaaaaacttaaattgttagaaagagcacactat 1920

tgattacttaattatattatgtctcattctcaataaccttttgattttttttaaaatccatactttagggggggaggggg 2000

ataaaactaaagaccctatcattactagctcatcatctgtatgctgatagcaggtcgttgtgatttgtattgctagttat 2080

actgtatacatttgttggttcatgtataaacatggaggagctttcacatctgagggctggccagcacttacattgacctg 2160

aataagataatgtgtaaagttctaggacattctagtgggtcctctaatctaatttttcatttgctctgggtaaataactt 2240

```
atgttgattttatttctttcgctatacaaatagtaggatgccgatagggaaatgagcttcaaaatctatgttttaaccct 2320
ggaatggtttccgcattattagtggttgtaattggctcttttgtcaattttctcctgaaaatgtcttgacatgttggccc 2400
aatgggtgaacctgactcaaaggtaggatggtaaggggaagggagcaagctaggaatcttctgaaattcgtattacttgt 2480
ttatactgaagtttcttacactaaatagacgctctttttctctttcattatttatattggagtttctttttcttactcta 2560
atattttgaaatgcacattgccctccctacaaagacattctgtctagggtggaccatgatcatttttttcccgaataactt 2640
gaagcattatagactgatttcctaatcccaattttaacatgttagGTGGTACGAAGGAGTACAAGGTTCTCATGGACCAA 2720
TTCCATCATGTTTCGACTGCTTTTCTGGAACTTGGTAAACTgtgagttcttacccagcttttgtgtttttctgatactaa 2800
gattttgctcaattgaaaggttacaatcagcatattttgtaaagacatagtttttttccccaagaaatccgtctggggcca 2880
cccttagaaccaacaacaacattcaaatctcgggataatgggcactctcttctgaccttaaattttggtataaaataaat 2960
agttaaatcagaaatcgcgtgagtgttcttatgatgacaatttagaagcccctttatacttggaaacttgtaaaaagttt 3040
catgaaatagtgagcgtgaagctttctatcagggattttaggatatcattttgcatgaaatagatatctgttcatcactg 3120
gacagggttcttgctgttagaactagagactttcttgggtcttatttttttcattcaaataattgctttgtgcgaatttcc 3200
tgtttgtgatgattattccctggcaatccatttcaaataacgacaagtcttacatcaatacgagtcaatattcagacaca 3280
caataaatataakgtagcagtatgtcctccagaatttaaactcttttgcaaggcttggagaattccaaactcaactctca 3260
atccacgtgacctaatgtgaatcagaaatgttacttataagtggtgcagtttcctctgccctcctcaaccctaaaatcat 3340
attacttatgaaacaagaaatgtgaactcagattgtagtctgaatttcaaaacgtataaactctgcgggattacacgggg 3520
tctgatatctccagactccacttgtgggataacattgggtatgttgtttaaactctgggttccttttgtatgagtagtag 3600
gttataatgaagattatgttcttatatctgtgcagTTATCAGCAGGCAATTGAGGACATTACCATGAGGATGGGTGCAGG 3680
AATGGCAAAATTTATATGCAAGGAGgtatgcaagatataccaaaaagaacaatatcaaattttctgatactcacaaaatg 3760
caatttatacttgtttgctttttttcgtttgtatgaattggctagaatctcaagtttcccttttcatgcaattcctcgtct 3840
tcaagacattggattattccacctcctcaaaagtgtaaaccatggtctctgcttcagtacattgtgattggcgtggttac 3920
tagtacttatgctatttgagattaggtgtccttttcacccatatattccttggatattttactggaaaatttttggtata 4000
cgaagttgagcattgttagcctgtatgaatactatgatatttgttgaattattgtctcagtctttgtttgttaagcaaaa 4080
catggctattaacaatagaaaacaagtacatgttatttggttgactagcgctaatcttggtaggttttgaaaggtaaatt 4160
aaaactttaaccccgcctaatgggtaggttccttaccgaggtaataatttttttgcccaaggcttcattttttaaaatgct 4240
tttcaaccctgttttattcaggtttaaaaaatctttcctggcaGTGGAAACAACTGATGATTATGACGAATACTGTCACT 4320
ATGTAGCTGGGCTTGTTGGGCTAGGATTGTCAAAACTGTTCCATGCCTCGGGGACAGAAGATCTGGCTTCAGATTCTCTC 4400
TCCAACTCCATGGGTTTATTTCTTCAGgtttggaccttaacatgtgtggcaggcacaactgttgttcgttgagattcttc 4480
tttgttgagggcactaatactaacatatttgctatatattgcagAAAACAAACATTATCAGAGATTATTTGGAAGATATA 4560
AATGAAGTACCCAAGTGCCGTATGTTCTGGCCCCGTGAGATTTGGAGTAAATATGTTAACAAGCTTGAGgtttgagttcc 4640
```

```
tcatatccttggtcttgtatgaatttctcattttgttttcccctgacatcataggtatcgaattaacttgagtgcatggt 4720

agcaagtgctgttgtttggtccaagtccaaaaagggtttcaatttaaaaattaattgacatccccggttaattcccccca 4800

GACTTAAAGTACGAGGAGAACTCGGTTAAGGCAGTGCAATGTCTCAATGAAATGGTCACCAATGCTTTGTCACATGTAGA 4880

AGATTGTTTGACTTACATGTTCAATTTGCGTGATCCTTCCATCTTTCGATTCTGTGCCATTCCACAGgtatatcctcttg 4960

gttattttctaggctttttactttcaatttgtctatcaagctaaaagcttcctacatgaagacatttcttctaccacagt 5040

tgataattcctttttaactttttggacttccagttcttgttcttcccatctattcatttttgaagaagtgttcctaag 5120

atggaactttatgttctgtgtttggggaagtgctgatttgacagagtttgcttcagcaaatttggtacttccatattgtc 5200

ttgttgccatgattgatattgatgcttggttacttggccatggagttaagtgtggtgaatgaatagatagcgtacaattc 5280

cctttagcgatcaaaaaataaaatagattgtgcacatttctctaaatatgagcttgtaggctgggaggaagaatttttac 5360

agcgtccagaaacagtgtttttacactagtaatttgctaagcatccttctttaccacacttttgccttgttgtgttgtct 5440

aatggtatcagGTCATGGCAATTGGGACATTAGCTATGTGCTATGACAACATTGAAGTCTTCAGAGGAGTGGTAAAAATG 5520

AGGCGTGgtaagaactatttttgtcaacatgcatttcttgtgacttcatacccgattgcgatctgactccttgaagttcg 5600

gccattttgtttttgcgtcccatcctaatattgtcatatgcaaaatgactttcttcctctagaattcatcatgtaattga 5680

ggggttttaatcttaaacaccgtctactgtatattcattggttgtataggatctaacttcctcgccttatgttcattctt 5760

tcagatctttccacagctccttatatctatgtattattcatgcatctatatgttctactacatttgcagattgaaattat 5840

ttataatagattttgtgtttttttgaatattattgtggcttcagtttttaatatatctatgcttgttcggcggcttaatat 5920

atccatcatacccgtcataattaataacttttggtttcttctttctagGTCTTACTGCTAAGGTCATTGACCGGACCAAG 6000

ACTATGGCAGATGTATATGGTGCTTTTTTTGACTTTTCTTGTATGCTGAAATCCAAGgtgtggaacttatttatttctga 6080

tcatgcctaattgcccagttataatgtgataatgttgaatttgctgggtcatattgaggattactccttcgtttttgttc 6160

aatccaacaatccatgtcttgtcttggaagaagtatatgcgcagtgtgcctttgtgcctcttaggggtcgtttggtagag 6240

agtatttggagaaatagtccaagtattatgtgtggtattatttagtattatgtttggtaggaattttgggtctatgtata 6320

accaatccatggattagttatacaccctacatggtattatagggtgtataactaataccttccatttggtggtattagtt 6400

ctaataccattggactaatctaggtaaagacaaaaataccccctcaaatccttttaatcattttgtttattttcttgatt 6480

ttatgttttatatttataataataaatttatttaaataaattagcttacaaattttattatttagatataatttttttgtt 6560

tctaaaatatgagttatttaatctatttattattaatataaaatattataatccgactaatatgctaatgttgatggatg 6640

aatttaagaacaattcataagtcaaatattgcctcttaacggaagtccattaaacattacacaagtagtctaaaacaaga 7720

gcatatatatatatatataaatataaaaaaaacatctcgagtataaaaatagtttaatttattttgctatattttttttt 6800

tttttttttttatgataatgagtttttttatttatttattgatacaaaacaaagttcaataaattttcgctataaattat 6880

tatagttgtgtgaccttttgagctattaactccaatttattaagatgacaattatttactctcaaataatttaagtgaga 6960

aaatagtaagcatgacaataaaaattttattctcctagctagttaaaatgctatgaaaaaataatattgtccgtcaatta 7040
```

tctactttgacccaattacatgaaatagaaaatcccatataactcaagggtataattggaaagaatttttttgtagagt 7120

tttaaaccacacacaaaactaggtaagaaacaatgaaccaaacacctgataaaactaatcattgcattactaatccatgc 7200

attaccaatccctgcattactaatccatgcattattgatctttgtaccaaacgaccccttagtgtctatctggacccttg 7280

catcatgctacgagaaaccttgtgcatacaaatgcctgtcactctgtagatttagaagtcctcttccagaaatacttatt 7360

gttgtttgcaaattagttatcccagtaaactctctctctctctcttatgtacaacagtgtagttcaagaatagaacaa 7440

attagtacacttctaaatgtaacatgcgcctattataatactgatatctcttagcaattgaactgatacatctatcatct 7520

atttgaactgggggtacaaaattggggaagcttttgctcaatggccaatagtaatggggtttcattccgtttgctatctg 7600

tgtttttttttttatatcaaagttccccttcttcatgtaaagacttgccactaaagctgccttgttattgttttagGTT 7680

AATAATAACGATCCAAATGCAACAAAAACTTTGAAGAGGCTTGACGCCATCCTGAAAACTTGCAGAGACTCGGGAACCTT 7760

GAACAAAAGgtttgtacatatcctagttgctctcatcttcacaattctgtgaatatcaactaatgcatgtacctatcaac 7840

agGAAATCTTACATAATCAGGAGCGAGCCTAATTACAGTCCAGTTCTGgtaactgttcaatgctctgattgtttattaat 7920

gctttagatacaatatgtctctcgcattagatgttttcttattctctcaaatttactgcagATTGTTGTCATCTTCATCA 8000

TACTGGCTATTATTCTTGCACAACTTTCTGGCAACCGATCTTAG*ACCATTTgtaagtatctaatcatgagatacgtacat* 8080

gccaattatttagatgcatgcctcgtagttcagaaatataccctctatgcacctaagcttttgacttgatgtctaatgat 8160

aagcatgtgtatcattatatgaccttttttttaatctagcctttaaacaataagcacagtaattttccaaattatgactt 8240

gtattctctcttttcttttctttctattacctgctatttaagattgcattgttttttttaacgaacacaaaaacttttcc 8320

tcccaacttaacccaattccttttaaaaaaaaccaccactttcctcctaaccttgaacgtatttgacaactaatttcgg 8400

ttatgtcattcttctgccatcacagatttccaattctaagtgaaaaatgaacaaattatggaaaatgtgtatcaatttaa 8480

ggataactgtgttaagagtcagtcaacatagagacatggaaattgtatccctttcagttttatggtggagagttttaacc 8560

cggatttatttgtcctgatttgtagattgGTCTACAAAAATGAAGTATGGTCAAGGAAGACAGCACAAACTCTTGGCCAA 8640

TTATGTACTGCTAATTGTTATGTTTGTATTACTATGTTCATTAAGTTAATAGTTGCATCTTCAACCTGACTAGATAATTA 8720

CGAAAGCCTATTTATGGCAGTTAGTTTGGTATGTATTTGTTTGCAAGCTAGGAAAGCAAATTCCAAGTGTTGTAGAGTCG 8800

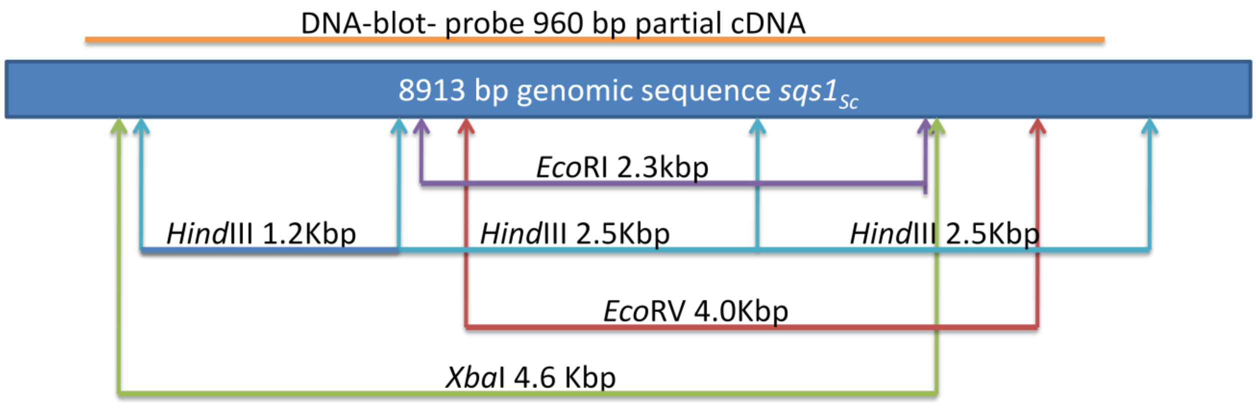
TTTTTCCGTAATGCACATTTCATTTTAATACTCTGTCGAATTTTGTGGTAAATTGACGTATTTACAGAGAGCCGTTGTAT 8880

TTGGACTAACACATTTTCAGAGGTTTTCCCAGAA 8914

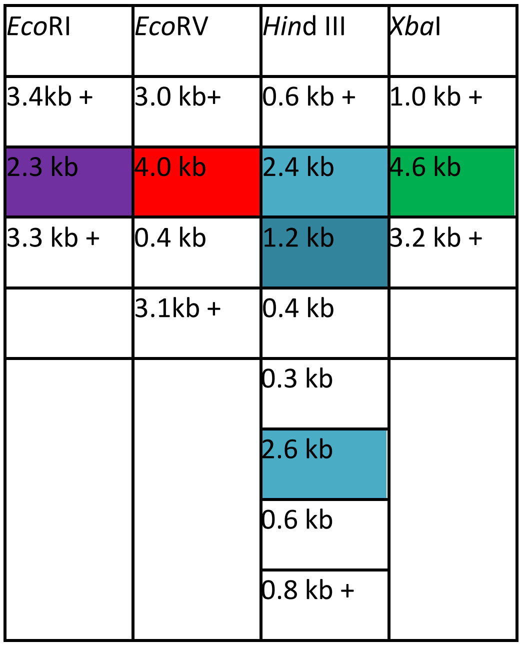
Appendix A. Genomic sequence of *sqs1*$_{Sc}$. The genomic sequence of *sqs1*$_{Sc}$ accession chc 80-1 is from the current (underlined) and previous study (Mweetwa, 2009). cDNA sequence is represented by capital letters. Intron regions are in lower case. The untranslated regions are in italics. The dinucleotide sequences at the ends of the intron are highlighted in red.

Appendix B: DNA blot verified

A

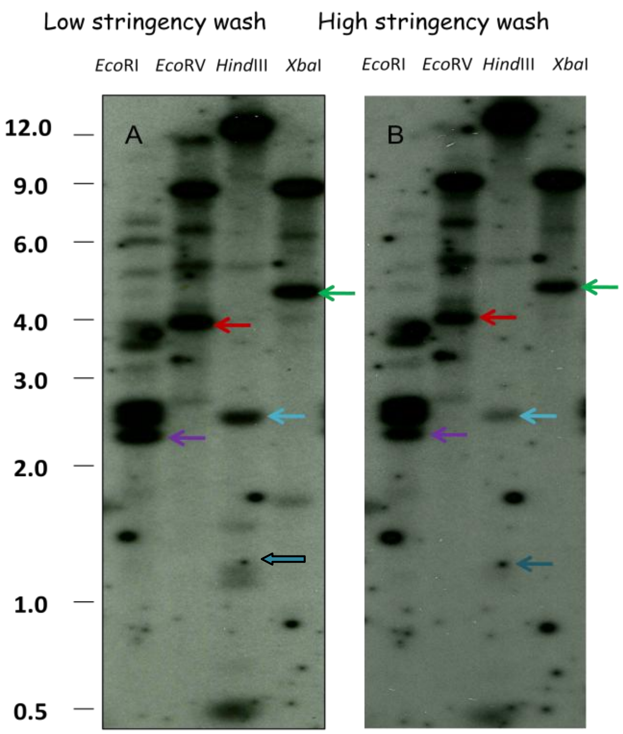


B



*Eco*RI	*Eco*RV	*Hind* III	*Xba*I
3.4kb +	3.0 kb+	0.6 kb +	1.0 kb +
2.3 kb	4.0 kb	2.4 kb	4.6 kb
3.3 kb +	0.4 kb	1.2 kb	3.2 kb +
	3.1kb +	0.4 kb	
		0.3 kb	
		2.6 kb	
		0.6 kb	
		0.8 kb +	

C



Appendix B DNA blot verified. Comparison of a restriction map of *$sqs1_{Sc}$* and the restriction digest of the DNA blot generated previously (Mweetwa, 2009). A) A restriction map of the genomic sequence for the indicated restriction enzymes. Arrows indicate restriction sites, and the colored coded bars connecting the arrows indicate which enzyme cuts at the site and the distance between the arrows. The 960 bp probe fragment is indicated. B) The expected sizes are displayed in a chart per enzyme. C) Arrows indicate band of the expected size in the DNA blot.

Appendix C: Alignment of the ORFs of the four predicted *sqs*$_{sp}$ genes and *sqs3*$_{sl}$

```
sqs1sp ATGGGAACATTGAGGGCGATTCTGAAGAATCCAGATGATTTGTATCCATTGATAAAGCTG        60
sqs2sp ATGGGGATTTTACGTGCAATTCTGAAGCATCCTGAAGATATTTATCCATTGTTGAAGCTG        60
sqs3sp ATGGGGG---------AGATTATGAAGCATCCAGATGAATTATATCCATTGATGAAGCTC        51
sqs3sl ATGGGGG---------AGACTATAAAGCATCCAGATGAATTTTATCCATTGATGAAGCTC        51
sqs4sp ATGGAGTTGATGCAGGAGATTTTGATGCATCCAGATGAATTATACCCATTGGTAAAGCTC        60

       ---------+---------+---------+---------+---------+---------+

sqs1sp AAACTAGCGGCTAGACATGCGGAGAAGCAGATCCC---------GCCGGAGCCACATTGG        111
sqs2sp AAGGTAGCGGCACGATATGCCGAAAAACAGATCCC---------TCCACAACCACATTGG        111
sqs3sp ATGTTATCGGCTAAACGCGTCGAGAAGAAGACTTCAGTGTGGCTGTTGCAGCCACACTGG        111
sqs3sl ATGTTATTGGCTAAACGCGTCGAGAAGAAGACGTCAGTGTGGCTATTGCAGCCACACTGG        111
sqs4sp ATGTTAACGGCAAAACGCGTCGAGAAGAAGACGTCAGTGTGGCTGTTGCAGCCACACTGG       120

       ---------+---------+---------+---------+---------+---------+

sqs1sp GGCTTCTGTTACTTAATGCTTCAAAAGGTCTCTCGTAGTTTTGCTCTCGTCATTCAACAG      171
sqs2sp GCCTTCTGTTACATCATGCTTCACAAGGTCTCTCGTAGCTTTTCTCTCGTCATTAAACAG      171
sqs3sp GCCTTCTGCTACGCTATTCTCCGTAAGGTGTCTCGTAGCTTTGCTCTTGTCATTCAACAA      171
sqs3sl GCCTTCTGTTACGCTACTCTCCGTAAGGTGTCTCGTAGCTTTGCTCTTGTCATTCAGCAA      171
sqs4sp GCCTTCTGCTACGCTACTCTCCGAAAGGTGTCTCGTAGCTTTGCTCTTGTAATTCAACAA      180

       ---------+---------+---------+---------+---------+---------+

sqs1sp CTTCCTGTCGAGCTTCGTGATGCTGTATGCATATTCTATTTGGTCCTTCGAGCACTTGAC      231
sqs2sp CTTCCTGTCGAGCTTCGCGACGCCATATGTATTTTCTATTTGGTTCTGCGTGCGCTTGAC      231
sqs3sp CTTCCTAGCGACCTT---------GTTTGTGTTTACTATTTGGTTCTTAGAGCACTTGAT      222
sqs3sl CTTCCTAGCGATCTT---------CGTTGTGTTTACTATTTGGTTCTTAGAGCACTTGAT      222
sqs4sp CTTCCTAGTGACCTT---------GTTTGTGTTTATTATTTGGTTCTTAGAGCACTTGAC      231

       ---------+---------+---------+---------+---------+---------+

sqs1sp ACTGTTGAGGATGATACCAGCATACCCACCGATGTTAAAGTACCTATTCTGATCTCTTTT      291
sqs2sp ACTGTCGAGGATGATACCAGTGTAGCGACAGAGGTGAAAGTACCAATTTTGATGTCCTTC      291
sqs3sp ACTGTTGAGGATGATACCAGCTTAACAACTGAAGTTAGAGTACCCATCTTAAAAGACTTT      282
sqs3sl ACTGTTGAGGATGATACCAGCTTAACAACTGAAGTTAGAGTACCCATTTTAAAAGACTTC      282
sqs4sp ACTGTTGAGGATGATACCAGCTTAGCCATTGAGGTTAGAGTACCTATTTTGAGAAATTTT      291

       ---------+---------+---------+---------+---------+---------+

sqs1sp CATCAGCATGTTTATGATCGTGAATGGCACTTTGCATGTGGTACGAAGGAGTACAAGGTT      351
sqs2sp CATCGCCATGTTTATGATCGTGAATGGCATTTTTCAGGCGGTACAAAGGACTACAAGGTT      351
sqs3sp TATAGCCATTTACATGATCCTGAATGGCATTTTTCAGGTGATACAATTGCCTTCAAAGTT      342
sqs3sl TATTGCCATTTACATGATCCTGAATGGCATTTTTCATGTGTTACAATGGCCTTCAAAGTT      342
sqs4sp TATTGCAACTTCTATGATCCTCAATGGCATTTTTCAGGTGGTACAAAGGCATTCAAGGTT      351

       ---------+---------+---------+---------+---------+---------+

sqs1sp CTCATGGACCAATTCCATCATGTTTCGACTGCTTTTCTGGAACTTGGTAAACTTTATCAG      411
sqs2sp CTTATGGATCAATTCCATCATGTTTCAACTGCTTTTCTGGAGCTAGGGAAACGTTACAAG      411
sqs3sp CTTATGGACCAATTCCATCATGTTTCCACTTCTCTCCTAGAGCTTGATACAGGTTATCAG      402
sqs3sl CTTATGGACCAATTCCATCATGTTTCCACTGCTTTTCTAGAGCTTGATACAGATTATCTG      402
sqs4sp CTTATGGACCAATTCCATCATGTTTCCACTGCTTTCCTAGAGCTTGATACAAGTTATCAA      411

       ---------+---------+---------+---------+---------+---------+

sqs1sp CAGGCAATTGAGGACATTACCATGAGGATGGGTGCAGGAATGGCAAAATTTATATGCAAG      471
sqs2sp GAAGCAATCGAGGACATTACCATGAGGATGGGTGCAGGAATGGCAAAGTTTATATACAAG      471
sqs3sp GAGGTGATTAAGGATATTACCAAGAGGATGGGTGAAGGAATGGCGAAATTTCTATGCAAG      462
```

```
sqs3sl  GAGGTGATTAAGGATATTACCAAGAGGATGGGTGAAGGAATGGCGAAATTTCTAAGCAAA      462
sqs4sp  GAGGTGATTAAGGATATTACCAAGAGAATGGGTGAAGGAATGGCGAAATTTCTATGCAAG      471

        ---------+---------+---------+---------+---------+---------+

sqs1sp  GAGGTGGAAACAACTGATGATTATGACGAATACTGTCACTATGTAGCTGGGCTTGTTGGG      531
sqs2sp  GAGGTTGAAACAATTGATGATTATGATGAATATTGTCACCATGTAGCTGGGCTCGTTGGA      531
sqs3sp  GAGGTAGAAACAATCAATGATTATAATGAATATTGTCACTATGCGGCTGGACTTTGTGGA      522
sqs3sl  GAGGTAGAAACAATCGATGATTATAATGAATATTGTCACTATGTAGCTGGACTTTGTGGA      522
sqs4sp  GAGGTAGAAACAATCGATGATTATAATGAATATAGTTTCTATGCATCTGGACTTTGTGGA      531

        ---------+---------+---------+---------+---------+---------+

sqs1sp  CTAGGATTGTCAAAACTGTTCCATGCCTCTGGGACAGAAGATCTGGCTTCAGATTCTCTC      591
sqs2sp  TTAGGCTTATCAAAACTTTTCCATGCCTCTGGGAAAGAAGATGTGGCTTCAGATTCTCTC      591
sqs3sp  TTAGGATTGTCAAAAAAAATTTATGCTTCTGGAAGAGAAGATTTAGCACCAGAATCCCTC      582
sqs3sl  TTAGGATTGTCAAAACTTTTCTATGCTTCTGGAAGAGAAGACTTAGCACCAGAATCCCTC      582
sqs4sp  TTAGGATTATCAAAGTTTTTTTATGTTTCTGGAAGAGAAGATTTAGCACCAGAATCCATT      591

        ---------+---------+---------+---------+---------+---------+

sqs1sp  TCCAACTCCATGGGTTTATTTCTTCAGAAAACAAACATTATCAGAGATTATTTGGAAGAT      651
sqs2sp  TGCAACTCCATGGGTTTATTTCTTCAGAAAACAAATATCATTAGAGATTATCTAGAAGAC      651
sqs3sp  TCGATTTCCATGGGTTTATTTCTTCAGAAAATAAGCATCATTAGAGATTATCTAGAGGAT      642
sqs3sl  TCGATTTCCATGGGTTTATTTCTTCAGAAAATAAGCATCGTTAGAGATTATCTAGAGGAC      642
sqs4sp  TCGATTTCCATGGGTTTATTTCTTCAGAAAATAAGCATCATTAGAGATTATCTAGAGGAC      651

        ---------+---------+---------+---------+---------+---------+

sqs1sp  ATAAATGAAGTACCCAAGTGCCGTATGTTCTGGCCCCGTGAGATTTGGAGTAAATATGTT      711
sqs2sp  ATAAATGAAGTACCCAAATGTCGCATGTTTTGGCCTCGTCAGATTTGGAGTAAATACGTT      711
sqs3sp  ATAAATGAAGTACCTAAATGTCGTATGTTTTGGCCTCGTCAAATTTGGAGTAAATATGTT      702
sqs3sl  ATAAATGAACTACCTAAATGTCGTATGTTTTGGCCCCGTCAAATTTGGAGTAAATATGTT      702
sqs4sp  ATAAATGAAGTACCTAAATGTCGTATGTTTTGGCCTCGTCAAATTTGGAGTAAATATGTT      711

        ---------+---------+---------+---------+---------+---------+

sqs1sp  AACAAGCTTGAGGACTTAAAGTACGAGGAGAACTCGGTTAAGGCAGTGCAATGTCTCAAT      771
sqs2sp  GACAAACTCGAGGACTTGAAGTATGAGGAGAACTCCGTCAAGGCAATTCAATGTCTGAAT      771
sqs3sp  AACAAACTCGAGGACTTTAAGTACGAGGAAAACTCGGTCAAGGCAGTACAGTGTCTGAAT      762
sqs3sl  AACAAACTCGAGGACTTTAAGTACGAGGATAACTCGGTCAAGGCAGTACAATGTCTGAAT      762
sqs4sp  AACAAACTCGAGGACTTTAAGTACGAGGAAAACTCGGTCAAGGCAGTACAATGTCTGAAT      771

        ---------+---------+---------+---------+---------+---------+

sqs1sp  GAAATGGTCACCAATGCTTTGTCACATGTAGAAGATTGTTTGACTTACTTGTTCAATTTG      831
sqs2sp  GAAATGGTCACTAATTCTTTGTCACATGTAGAAGATTGTTTAACTTTCTTGTCTACACTG      831
sqs3sp  GAA------------------------------GATTGTCTGGTTTACATGTCTAATTTA      792
sqs3sl  GAA------------------------------GATTGTCTGGTTTACATGTATAATTTA      792
sqs4sp  GAAATGGTCACTAATGCTTTATTATATGTAGAAGATTGTCTGACTAGCATGTCTAGTTTA      831
        ---------+---------+---------+---------+---------+---------+

sqs1sp  CGTGATCCTGCCATCTTTCGATTCTGTGCCATTCCACAGGTCATGGCAATTGGGACATTA      891
sqs2sp  CGGGATCCTGCTATCTTTCGATTCTGTGCTATTCCACAGGTCATGGCAATTGGGACCCTG      891
sqs3sp  CGAGATCCTTCTATCTTTCAGTTCTGTGCAATTCCGCTGGTCATAAACATGGGGAATTTG      852
sqs3sl  CGAGATCCTGCTATCTTTCAGTTCTGTGCAATTCCGCTGGTCATAAATATGGGGAATTTG      852
sqs4sp  CGCGATCCTGCTATCTTTCAGTTCTGTGCAATTCCACAGGTCATAAATATGGGGAATTTG      891
```

```
        ---------+---------+---------+---------+---------+---------+

sqs1sp  GCTATGTGCTACGACAACATTGAAGTCTTCAGAGGAGTGGTAAAAATGAGGCGTGGTCTT      951
sqs2sp  GCAAAGTGCTATAACAACATTGAAGTTTTCAGAGGAGTTGTGAAAATGAGACGTGGTCTC      951
sqs3sp  ATGATGTACTACAACAACGTTGAAATTTTCAAAGGTGTTGTTGAAATGAGGCGAGGTCTT      912
sqs3sl  ACGGTGTACTACAACAATGTTGTAATTTTCAAAGGTGTTGTAGAAATGAGACGAGGGCCT      912
sqs4sp  TCGATGTACTACAACAACGTTGAGATTTTCAAAGGCGTTGTTGAAATGAGACGAGGTCTC      951

        ---------+---------+---------+---------+---------+---------+

sqs1sp  ACTGCTAAGGTCATTGACCGGACCAAGACTATGGCAGATGTATATGGTGCTTTTTTTGAC      1011
sqs2sp  ACCGCTCAGGTTATTGACCGGACCAGGAACATGGCAGATGTATATGGTGCTTTCTTCGAC      1011
sqs3sp  TGTGCTAAGATTATTGATCAAACGAGGACAATGGCTGATGTCTACGGAGCTTTCTTTGAC      972
sqs3sl  TTGTCTAAGATTATTGATCAGACGAGGACGATGGCTGATGTCTACGGAGCTTTCTTTGAC      972
sqs4sp  TGTGCTAAGATTATTGATCAGACGAGGACGATGGCTGATGTCTACAGAGCTTTTTATGAC      1011

        ---------+---------+---------+---------+---------+---------+

sqs1sp  TTTTCTTGTATGCTGAAATCCAAGGTTAATAATAACGATCCAAATGCAACAAAAACTTTG        1071
sqs2sp  TTCTCGTGTATTCTGAAATCCAAGGTAGAGTATAAAGATCCTCAAGTGGCAAAAACTTTA       1071
sqs3sp  TTTTGTTGTATAATGGAATCTAAGGTTGATCGTGATGATCCAAATGCAACGAGTACTTTG       1032
sqs4sp  TTCTGTTGTATAATGGAATCTAAGGTTGATCGCGATGATCCAAATGCAACGAGTACATTG       1071
sqs3sl  TTTTGTTGTATAATGGAATCTAAGGTTGATCGTGATGATCCAAATGCAACGAGTACTTTG       1032

        ---------+---------+---------+---------+---------+---------+

sqs1sp  AAGAGGCTTGACGCCATCCTGAAAACTTGCAGAGACTCGGGAACCTTG---AACAAAAGA     1128
sqs2sp  AAGAGGCTTGAAGTGATCTTGAGAACTTGCAAAAACTCGGGAACCTTG---AACAAAAGG     1128
sqs3sp  AAGAGGCTAGAAGAAATTTCAAAAAATTGCAGAAACTCTGGAACCTTGATCAATCA---G     1089
sqs3sl  AAGAGGCTTGAAGAAATTTCAAAAACTTGCAGAGACTCTGGAACCTTACTCAATCAAAGG     1092
sqs4sp  AAGAGGCTTGAAGCAATTTTGAAAACTTGCAGAGACTCCGGAACCTTA---AATCAAAGG     1128

        ---------+---------+---------+---------+---------+---------+

sqs1sp  AAATCTTACATAATCAGGAGCGAGCCTAATTACAGT---CCAGTTCTG------ATTGTT     1179
sqs2sp  AAATCTTTCGTAATCAAGAGTGGACCTAATTACAAT---TCAACTTTG------GTTGTT     1179
sqs3sp  AAATCTTACACATTCAGCCATCAGCCTAATTATAATATTCCAGTTCTTTTGCAGATTATT     1149
sqs3sl  AAATCTTACACATTCAGTCATCAGCCAAATTATAATATTCCAGTTCTTTTGCAGATTATT     1152
sqs4sp  AAATCTTACACATTCAGCCATCAGCCTAATTATAATATTCCAGTTCTTTTGCAGATTATC     1188

      ---------+---------+---------+---------+---------+--------

sqs1sp  GTCATCTTCATCATACTGGCTATTATTCTTGCA-CAACTTTCTGGCAACCGATCTTAG       1236
sqs2sp  GTCCTTGTTGTCTTAGTGGCTATCCTTCTAGGA-TACCAATCTGGAAACCGGACTTAG       1236
sqs3sp  TTCTTTTTCATCATGTTGGCTATTCTTTTATCAACAAAAATACCTTAA                 1197
sqs3sl  TTCTTTTTCATCATGTTGGCTATTCTTTTATCAACAAAAATACCTCA                  1199
sqs4sp  TTCTTTTTCATCATGATGGCTATTCTTTTATCAACAAAAATACCTTAA                 1236
```

Appendix C. Alignment of the ORFs of the four predicted *sqs*$_{sp}$ genes and *sqs3*$_{sl}$. Clustal W (DNASTAR Megalign) was used to produce an alignment of the five ORFs. Gaps introduced in the alignment are indicated by dashes (-). The gene names are labeled on the left with a species abbreviation (sp: *S. phureja*; sl: *S. lycopersicum*). The position number of the nucleotide at the end of each row is numbered for each ORF. A graduating line above each stanza marks every tenth nucleotide. The nine nucleotides that translate to a WIL motif only found in *sqs3* and *sqs4* are indicated in red letters. A gap characteristic of *sqs3* in exon 8 is highlight in red (-).

Appendix D: Alignment of the alleles of each partial cDNA of *sqs*$_{sc}$

```
                ---------+---------+---------+---------+---------+---------+

                           10        20        30        40        50        60

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 ----------------------------------------------------GAACAGTG
cDNASQS1allele2 ----------------------------------------------------GAACAGTG          8

cDNASQS2allele1 CCTTTTACACAAATTAATTAAAGTGAATGTATATAGATGTGAATATCTGATTGATCAAAT
cDNASQS2allele2 CCTTTTACACAAATTAATTAAAGTGAATGTATATAGATGTGAATATCTGATTGATCAAAT         60

cDNASQS4allele1 -------------------------------------------TAAATTGACACTCCTTA
cDNASQS4allele2 -------------------------------------------TAAATTGACACTCCTTA         17

                ---------+---------+---------+---------+---------+---------+

                           70        80        90        100       110       120

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 TTTGAATTTGTTGAGAAGAATGGGAACATTGAGGGCGATTCTGAAGAATCCAGATGATTT
cDNASQS1allele2 TTTGAATTTGTTGAGAAGAATGGGAACATTGATGGCGATTCTGAAGAATCCAGATGATTT         68

cDNASQS2allele1 AACAATAATAATAATAATAATGGGGATTTTACGTGCAATTCTGAAGCATCCTGAAGATAT
cDNASQS2allele2 AGCAATAATAATAATAATAATGGGGATTTTACGTGCAATTCTGAGGCATCCTGAAGATAT        120

cDNASQS4allele1 ATTAAACAAATTTATAAAAATGGAGTTGATGCAGGAGATTTTGATGCATCCAGATGAATT
cDNASQS4allele2 ATTAAACAAATTTATAAAAATGGAGTTGATGCAGGAGATTTTGATGCATCCAGATGAATT         77

                ---------+---------+---------+---------+---------+---------+

                           130       140       150       160       170       180

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 GTATCCATTGATAAAGCTGAAACTAGCGGCTAGACATGCGGAAAAGCAGATCCC------
cDNASQS1allele2 GTATCCATTGATAAAGCTGAAACTAGCGGCTAGACATGCGGAAAAGCAGATCCC------        122

cDNASQS2allele1 TTATCCATTGTTGAAGCTGAAGGTAGCAGCACGATATGCCGAAAAACAGATCCC------
cDNASQS2allele2 TTATCCATTGTTGAAGCTGAAGGTAGCGGCACGATATGCCGAAAAACAGATCCC------        174

cDNASQS4allele1 ATACCCATTGGTAAAGCTCATGTTAACGGCAAAACGCGTTGAGAAGAAGACGTCAGTGTG
cDNASQS4allele2 ATACCCATTGGTAAAGCTCATGTTAACGGCAAAACGCGTCGAGAAGAAGACGTCAGTGTG        137

                ---------+---------+---------+---------+---------+---------+
                         190       200       210       220       230       240

                 ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 ---GCCGGAGCCACATTGGGGCTTCTGTTACTTAATGCTTCAAAAGGTCTCTCGTAGTTT
cDNASQS1allele2 ---GCCGGAGCCACATTGGGGCTTCTGTTACTTAATGCTTCAAAAGGTCTCTCGTAGTTT        170

cDNASQS2allele1 ---TTCACAACCACATTGGGCCTTCTGTTACATCATGCTTCACAAGGTCTCTCGTAGCTT
cDNASQS2allele2 ---TTCACAACCACATTGGGCCTTCTGTTACATCATGCTTCACAAGGTCTCTCGTAGCTT        231

cDNASQS4allele1 GCTGTTGCAGCCATACTGGGCCTTCTGCTACGCTACTCTCCGAAAGGTGTCTCGTAGCTT
cDNASQS4allele2 GCTGTTGCAGCCACACTGGGCCTTCTGCTACGCTACTCTCCGAAAGGTGTCTCGTAGCTT        197

                 ---------+---------+---------+---------+---------+---------+

                         250       260       270       280       290       300

                 ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 TGCTCTCGTCATTCAACAGCTTCCTGTCGAGCTTCGTGATGCTGTATGCATATTCTATTT
cDNASQS1allele2 TGCTCTCGTCATTCAACAGCTTCCTGTCGAGCTTCGTGATGCTGTATGCATATTCTATTT        239

cDNASQS2allele1 TTCTCTCGTCATTAAACAGCTTCCTGTCGAGCTTCGAGATGCCATATGTATTTTCTATTT
cDNASQS2allele2 TTCTCTCGTCATTAAACAGCTTCCTGTTGAGCTTCGCGATGCCATATGTATTTTCTATTT        291

cDNASQS4allele1 TGCTCTTGTAATTCAACAACTTCCTAGTGACCTTCGTAACGTGGTTTGTGTTTATTATTT
cDNASQS4allele2 TGCTCTTGTAATTCAACAACTTCCTAGTGACCTTCGTAACGTGGTTTGTGTTTATTATTT        257

                 ---------+---------+---------+---------+---------+---------+

                         310       320       330       340       350       360

                 ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 GGTCCTTCGAGCACTGGACACTGTTGAGGATGATACCAGCATACCCACCGATGTTAAAGT
cDNASQS1allele2 GGTCCTTCGAGCACTGGACACTGTTGAGGATGATACCAGCATACCCACCGATGTTAAAGT        299

cDNASQS2allele1 GGTTCTGCGTGCGCTTGACACTGTCGAGGATGATACCAGTGTAGCGACAGAGGTGAAAGT
cDNASQS2allele2 GGTTCTGCGTGCGCTTGACACTGTCGAGGATGATACCAGTGTAGCGACAGAGGTGAAAGT        351

cDNASQS4allele1 GGTTCTTAGAGCACTTGACACTGTTGAGGATGATACCAGCTTAGCCATTGAGGTTAGAGT
cDNASQS4allele2 GGTTCTTAGAGCACTTGACACTGTTGAGGATGATACCAGCTTAGCCATTGAGGTTAGAGT        317

                 ---------+---------+---------+---------+---------+---------+

                         370       380       390       400       410       420

                 ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 ACCTATTCTGATCTCTTTTCATCAGCATGTTTATGATCGCGAATGGCACTTCGCATGTGG
cDNASQS1allele2 ACCTATTCTGATCTCTTTTCATCAGCATGTTTATGATCGCGAATGGCACTTCGCATGTGG        359

cDNASQS2allele1 ACCAATTTTGATGTCCTTCCATCGCCATGTTTATGATCGTGAATGGCATTTTTCATGTGG
cDNASQS2allele2 ACCAATTTTGATGTCCTTCCATCGCCATGTTTATGATCGTGAATGGCATTTTTCATGTGG        411

cDNASQS4allele1 ACCTATTTTGAGAAATTTTTATTGCAACTTCTATGATCCTCAATGGCATTTTTCATGTGG
cDNASQS4allele2 ACCTATTTTGAGAAATTTTTATTGCAACTTCTATGATCCTCAATGGCATTTTTCATGTGG        377

                 ---------+---------+---------+---------+---------+---------+
                           430       440       450       460       470       480

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 TACGAAGGAGTACAAGGTTCTCATGGACCAATTCCATCATGTTTCGACTGCTTTTCTGGA
cDNASQS1allele2 TACGAAGGAGTACAAGGTTCTCATGGACCAATTCCATCATGTTTCGACTGCTTTTCTGGA        419

cDNASQS2allele1 TACAAAGGACTACAAGGTTCTTATGGATCAATTCCATCATGTTTCAACTGCTTTTCTGGA
cDNASQS2allele2 TACAAAGGACTACAAGGTTCTTATGGATCAATTCCATCATGTTTCAACTGCTTTTCTGGA        471

cDNASQS4allele1 TACAAAGGCATTCAAGGTTCTTATGGACCAATTCCATCATGTTTCTATTGCTTTCCTAGA
cDNASQS4allele2 TACAAAGGCGTTCAAGGTTCTTATGGACCAATTCCATCATGTTTCCACTGCTTTCCTAGA        437

                ---------+---------+---------+---------+---------+---------+

                           490       500       510       520       530       540

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 ACTTGGTAAACTTTATCAGCAGGCAATTGAGGACATTACCATGAGGATGGGTGCAGGAAT
cDNASQS1allele2 ACTTGGTAAACTTTATCAGCAGGCAATTGAGGACATTACCATGAGGATGGGTGCAGGAAT        479

cDNASQS2allele1 GCTAGGGAAACATTACAAGGAAGCAATCGAGGACATTACCATGAGGATGGGTGCAGGAAT
cDNASQS2allele2 GCTAGGGAAACATTACAAGGAAGCAATCGAGGACATTACCATGAGGATGGGTGCAGGAAT        531

cDNASQS4allele1 GCTTGATACAAATTACCAAGAGGTGATTAAGGATATTACCAAGGGAATGGGTAAAGGAAT
cDNASQS4allele2 GCTTGATACAAATTACCAAGAGGTGATTAAGGATATTACCAAGAGAATGGGTGAAGGAAT        497

                ---------+---------+---------+---------+---------+---------+

                           550       560       570       580       590       600

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 GGCAAAATTTATATGCAAGGAGGTGGAAACAACTGATGATTATGACGAATACTGTCACTA
cDNASQS1allele2 GGCAAAATTTATATGCAAGGAGGTGGAAACAACTGATGATTATGACGAATACTGTCACTA        539

cDNASQS2allele1 GGCAAAGTTTATATACAAGGAGGTTGAAACAATTGATGATTATGATGAATACTGTCACCA
cDNASQS2allele2 GGCAAAGTTTATATACAAGGAGGTTGAAACAATTGATGATTATGATGAATACTGTCACCA        591

cDNASQS4allele1 GGCGAAATTTCTATGCAAGGAGGTAGAAACAATCGATGATTATAATGAATATAGTTTCTA
cDNASQS4allele2 GGCGAAATTTCTATGCAAGGAGGTAGAAACAATCGATGATTATAATGAATATAGTTTCTA        557

                ---------+---------+---------+---------+---------+---------+

                           610       620       630       640       650       660

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 TGTAGCTGGGCTTGTTGGGCTAGGATTGTCAAAACTGTTCCATGCCTCGGGGACAGAAGA
cDNASQS1allele2 TGTAGCTGGGCTTGTTGGGCTAGGATTGTCAAAACTGTTCCATGCCTCGGGGACAGAAGA        599

cDNASQS2allele1 TGTAGCTGGGCAAGTTGGATTAGGCTTATCAAAACTTTTCCATGCCTCTGGGAAAGAAGA
cDNASQS2allele2 TGTAGCTGGGCTAGTTGGATTAGGCTTATCAAAACTTTTCCATGCCTCTGGGAAAGAAGA        651

cDNASQS4allele1 TGCATCTGGACTTTGTGGATTAGGATTATCAAAGTTTTTTTATGTTTCTGGAAGAGAAGA
cDNASQS4allele2 TGCATCTGGACTTTGTGGATTAGGATTATCAAAGTTTTTTTATGTTTCTGGAAGAGAAGA        617

               ---------+---------+---------+---------+---------+---------+
                           670       680       690       700       710       720

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 TCTGGCTTCAGATTCTCTCTCCAACTCCATGGGTTTATTTCTTCAGAAAACAAACATTAT
cDNASQS1allele2 TCTGGCTTCAGATTCTCTCTCCAACTCCATGGGTTTATTTCTTCAGAAAACAAACATTAT        659

cDNASQS2allele1 TGTAGCTTCAGATTCTCTCTGCAACTCCATGGGATTATTTCTACAGAAAACAAATATCAT
cDNASQS2allele2 TGTAGCTTCAGATTCTCTCTGCAACTCCATGGGATTATTTTTACAGAAAACAAATATCAT        711

cDNASQS4allele1 TTTAGCACCAGAATCCATCTCGATTTCCATGGGTTTATTTCTTCAGAAAATAAGCATCAT
cDNASQS4allele2 TTTAGCACCAGAATCCATCTCGATTTCCATGGGTTTATTTCTTCAGAAAATGAGCATCAT        677

                ---------+---------+---------+---------+---------+---------+

                           730       740       750       760       770       780

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 CAGAGATTATTTGGAAGATATAAATGAAGTACCCAAGTGCCGTATGTTCTGGCCCCGTGA
cDNASQS1allele2 CAGAGATTATTTGGAAGATATAAATGAAGTACCCAAGTGCTGTATGTTCTGGCCCCGTGA        719

cDNASQS2allele1 TAGAGATTATCTAGAAGACATAAATGAAGTACCCAAATGTCGTATGTTTTGGCCTCGTCA
cDNASQS2allele2 TAGAGATTATCTAGAAGACATAAATGAAGTACCCAAATGTCGTATGTTTTGGCCTCGTCA        771

cDNASQS4allele1 TAGAGATTATCTAGAGGACATAAATGAAGTACCTAAATGTCGTATGTTTTGGCCTCGTCA
cDNASQS4allele2 TAGAGATTATCTAGAGGACATAAATGAAGTACCTAAATGTCGTATGTTTTGGCCTCGTCA        737

                ---------+---------+---------+---------+---------+---------+

                           790       800       810       820       830       840

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 GATTTGGAGTAAATATGTTAACAAGCTTGAGGACTTAAAGTACGAGGAGAACTCGGTTAA
cDNASQS1allele2 GATTTGGAGTAAATATGTTAACAAGCTTGAGGACTTAAAGTACGAGGAGAACTCGGTTAA        779

cDNASQS2allele1 GATTTGGAGTAAATACGTTGACAAGCTCGAGGACTTGAAGTATGAGGGGAACTCTGTCAA
cDNASQS2allele2 GATTTGGAGTGAATACGTTGACAAGCTCGAGGACTTGAAGTATGAGGGGAACTCTGTCAA        831

cDNASQS4allele1 AATTTGGAGCAAATATGTTAACAAACTCGAGGACTTTAAGTACGAGGAAAACTCGGTCAA
cDNASQS4allele2 AATTTGGAGCAAATATGTTAACAAACTCGAGGACTTTAAGTACGAGGAAAACTCGGTCAA        79

                ---------+---------+---------+---------+---------+---------+

                         850       860       870       880       890       900

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 GGCAGTGCAATGTCTCAATGAAATGGTCACCAATGCTTTGTCACATGTAGAAGATTGTTT
cDNASQS1allele2 GGCAGTGCAATGTCTCAATGAAATGGTCACCAATGCTTTGTCACATGTAGAAGATTGTTT        839

cDNASQS2allele1 GGCAGTTCAATGTCTGAATGAAATGGTCACTAATGCTTTGTCACATGCAGAAGATTGTTT
cDNASQS2allele2 GGCAGTTCAGTGTCTGAATGAAATGGTCACTAATGCTTTGTCACATGCAGAAGATTGTTT        891

cDNASQS4allele1 GGCAGTACAATGTCTGAATGAAATGGTCACTAATGCTTTATTATATGTAGAAGATTGTCT
cDNASQS4allele2 GGCAGTACAATGTCTGAATGAAATGGTCACTAATGCTTTATTATATGTAGAAGATTGTCT        857

                ---------+---------+---------+---------+---------+---------+
                         910       920       930       940       950       960
                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 GACTTACATGTTCAATTTGCGTGATCCTTCCATCTTTCGATTCTGTGCCATTCCACAGGT
cDNASQS1allele2 GACTTACATGTTCAATTTGCGTNATCCTTCCATCTTTCGATTCTGTGCCATTCCACAGGT        899

cDNASQS2allele1 GACTTTCTTGTCTACACTGCGGGATCCTGCTATCTTTCGATTCTGTGCTATTCCACAGGC
cDNASQS2allele2 GACTTTCTTGTCTACACTGAGGGATCCTACTATCTTTCGATTCTGTGCTATTCCACAGGC        951

cDNASQS4allele1 GACTAGCATGTCTAGTTTACGCGATCCTGCTATCTTTCAGTTCTGTGCAATTCCACAGAT
cDNASQS4allele2 GACTAGCATGTCTAGTTTACGGGATCCTGCTATCTTTAAGTTCTGTGCATTTCCACAGAT        917

                ---------+---------+---------+---------+---------+---------+
                         970       980       990       1000      1010      1020
                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 CATGGCAATTGGGACATTAGCTATGTGCTATGACAACATTGAAGTCTTCAGAGGAGTGGT
cDNASQS1allele2 CATGGCAATTGGGACATTAGCTATGTGCTATGACAANATTGAAGTCTTCAGAGGAGTNGN        959

cDNASQS2allele1 CATGGCAATTGGAACTCTGGCAAAGTGCTATAACAACATTGAAGTCTTCCGAGGAGTTGT
cDNASQS2allele2 CATGGCAATTGGAACTCTGGCAAAGTGCTATAACAACATTGAAGTCTTCCGAGGAGTTGT       1011

cDNASQS4allele1 CATAAATATGGGGAATTTGTCGATGTACTACAATAACGTTGAGATTTTCAAAGGCGTTGT
cDNASQS4allele2 CATAAATATGGGGAATTTGTCGATGTACTACAACAACGTTGAAATTTTCAAAGGCGTTGT        977

                ---------+---------+---------+---------+---------+---------+
                         1030      1040      1050      1060      1070      1080
                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 AAAAATGAGGCGTGGTCTTACTGCTAAGGTCATTGACCGGACCAAGACTATGGCAGATGT
cDNASQS1allele2 AAAAATGAGGCGTGGTCTTACTGCTAAGGTCATTGACCGGACCAAGACTATGGCAGATGT       1019

cDNASQS2allele1 GAAAATGAGACGTGGTCTCACCGCTCAGGTTATTGACCGGACCAGGAACATGGCAGATGC
cDNASQS2allele2 GAAAATGAGACGTGGTCTCACCGCTCAGGTTATTGACCGGACCAGGAACATGGCAGATGT       1071

cDNASQS4allele1 TGAAATGAGACGAGGCCTCTGTGCTAGGATTATTGATCAGACGAGGACGATGGCTGATGT
cDNASQS4allele2 TGAAATGAGACGAGGCCTCTGTGCTAAGATTATTGATCAGACGAGGACGATGGCTGATGT       1037

                ---------+---------+---------+---------+---------+---------+
                         1090      1100      1110      1120      1130      1140
                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 ATATGGTGCTTTTTTTGACTTTTCTTGTATGCTGAAATCCAAGGTTAATAATAACGATCC
cDNASQS1allele2 ATATGGTGCTTTTTTTGACTTTTCTTGTATTNTGAAATCCAAGGTTAATAATAACGATCC       1079

cDNASQS2allele1 ATATGGTGCTTTCTTCGACTTCTCGTGTATTCTGAAATCCAAGGTAGAGTATAAAGATCC
cDNASQS2allele2 ATATGGTGCTTTCTTTGACTTCTCGTGTATTCTGAAATCCAAGGTAGAGTATAAAGATCC       1131

cDNASQS4allele1 CTACGGAGCTTTTTATGACTTCTGTTGTGTAATGGAATCTAAGGTTGATCGCGATGATCC
cDNASQS4allele2 CTACGGAGCTTTTTATTACTTCTGTTGTATAATGGAATCTAAGGTTGATCGCGATGATCC       1097

                ---------+---------+---------+---------+---------+---------+
                         1150      1160      1170      1180      1190      1200

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 AAATGCAACAAAAACTTTGAAGAGGCTTGACGCCATCCTGAAAACTTGCAGAGACTCGGG
cDNASQS1allele2 AAATGCAACAAAAACTTTGAAGAGGCTTGACGCCATCCTGAAAACTTGCAGAGACTCGGG       1139

cDNASQS2allele1 ACATGTGGCAAAAACTTTAAAGAGGCTTGAAGTGATCTTGAGAACTTGCAAAAACTCGGG
cDNASQS2allele2 ACATGTGGCAAAAACTTTAAAGAGGCTTGAAGTGATCTTGAGAACTTGCAAAAACTCGGG       1191

cDNASQS4allele1 AAATGCAACGAGTACTTTGAAGAGGCTTGAAGCAATCTTGAAAACTTGCAGAGATTCCGG
cDNASQS4allele2 AAATGCAACGAGTACATTGAAGAGGCTTGAAGCAATTTTGAAAACTTGCAGAGACTCCGG       1157

               ---------+---------+---------+---------+---------+---------+

                         1210      1220      1230      1240      1250      1260

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 AACCTTGAACAAAAGGAAATCTTACATAATCAGGAGCGAGCCTAATTACAGT---CCAGT
cDNASQS1allele2 AACCTTGAACAAAAGGAAATCTTACATAATCAGGAGCGAGCCTAATTACAGT---CCAGT       1196

cDNASQS2allele1 AACCTTGAACAAAAGGAAATCTTTTGTAATCAAGAGTGGACCTAATTACAAT---TCAAC
cDNASQS2allele2 AACCTTGAACAAAAGGAAATCTTTTGTAATCAAGAGTGGACCTAATTACAAT---TCAAC       1248

cDNASQS4allele1 AACCTTGAATCAAAGGAAATCTTACACATTCAGCCATCAGCCTAATTATAATATTCCAGT
cDNASQS4allele2 AACCTTAAGTCAAAGGAAATCTTACACATTCAGCCATCAGCCTAATTATAATATTCCAGT       1217

                ---------+---------+---------+---------+---------+---------+

                         1270      1280      1290      1300      1310      1320

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 TCTGATTGTTGTCATCTTCATCATACTGGCTATTATTCTTGCACAACTTTCTGGCAACCG
cDNASQS1allele2 TCTGATTGTTGTCATCTTCATCATACTGGCTATTATTCTTGCACAACTTTCTGGCAACCG       1256

cDNASQS2allele1 TTTCGTTGTGGTCCTTGTTGTCTTAGTGGCTATCCTTCTAGGATACCAATCTGGAAACCG
cDNASQS2allele2 TTTCGTTGTGGTCCTTGTTGTCTTAGTGGCTATCCTTCTAGGATACCAATCTGGAAACCG       1308

cDNASQS4allele1 TTTGATTATCTTCTTTTTCATCATGATGGCTATTCTTTTATCAACAAAAATACCTTAA
cDNASQS4allele2 TTTGATTATCTTCTTTTTCATCATGATGGCTATTCTTTTATC                         1259

                ---------+---------+---------+---------+---------+---------+

                         1330      1340      1350      1360      1370      1380

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 ATCTTAGACCATTTATTGGTCTACA---AAAATGAACTATGGTCAACGAA---GACAGCA
cDNASQS1allele2 ATCTTAGACCATTTATTGGTCTACA---AAAATGAACTATGGTCAACGAA---GACAGCA       1310

cDNASQS2allele1 GACTTAGACCATGTGTAAGACTGTTGTTAGACGAAAGTCCGGTAAAAAAAAGAGATGGCA
cDNASQS2allele2 GACTTAGACCATGTGTAAGACTGTTGTTAGACGAAAGTCCGGTAAAAAAA-GAGATGGCA       1367

                ---------+---------+---------+---------+---------+---------+
                         1390      1400      1410      1420      1430      1440

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 CAAACTCTTGGCCAATTATGTACTGCTAATTGTTATGTTTGTATTACTATGTTCATTAAG
cDNASQS1allele2 CAAACTCTTGGCCAATTATGTACTGCTAATTGTTATGTTTGTATTACTATGTTCATTAAG       1370

cDNASQS2allele1 CAAGCTCTTGGACGAGTGTGTGATAGCTGCAGATTTTGTCATC
cDNASQS2allele2 CAAGCTCTTGGACGAGTGTGTGATAGCTGCAGATTTTGTCATC                        1410

                ---------+---------+---------+---------+---------+---------+

                         1450      1460      1470      1480      1490      1500

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 TTAATAGTTGCATCTTCAACCTGACTAGATAATTACGAAAGCCTATCTATGGCAGTTAGT
cDNASQS1allele2 TNAATAGTTGCATCTTCAACCTGACTAGATAATTACGAAAGCCTATCTATGGCAGTTAGT       1430

                ---------+---------+---------+---------+---------+---------+

                         1510      1520      1530      1540      1550      1560

                ---------+---------+---------+---------+---------+---------+

cDNASQS1allele1 TTGGTATGTATTTGTTTGCAAGCTAGGAAAGCAAATTCCAAGTGTTGTAGAGTCGTTTTT
cDNASQS1allele2 TTGGTATGTATTTGTTTGCAAG                                             1452
```

Appendix D. Alignment of the alleles for each partial cDNA of *sqs$_{Sc}$*. Sequences of the partial cDNA of each allele of *sqs$_{Sc}$*1, *sqs$_{Sc}$*2, and *sqs$_{Sc}$*4 were aligned (ClustalW). Allelic differences are highlighted. Positions are numbered on the horizontal axis by place in the alignment. Positions in the alleles are numbered on the right side of the sequence. The beginning and end of the coding regions is indicated by red highlight of the start and stop codons.

Appendix E: Alignment of the 3’UTR of *sqs1$_{Sc}$* and *sqs1$_{Ca}$*

```
                             10        20        30        40        50        60

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     ATGGGAACATTGAGGGCGATTCTGAAGAATCCAGATGATTTGTATCCATTGATAAAGCTG
sqs1sc cDNA         ATGGGAACATTGAGGGCGATTCTGAAGAATCCAGATGATTTGTATCCATTGATAAAGCTG
sqs1cacDNA          ATGGGGACTTTGAGAGCGATTTTGAAGAATCCAGATGATTTGTATCCATTGATAAAGCTA

                    ---------+---------+---------+---------+---------+---------+

                             70        80        90        100       110       120

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     AAACTAGCGGCTAGACATGCGGAAAAGCAGATCCCGCCTGAGCCACATTGGGGCTTCTGT
sqs1sc cDNA         AAACTAGCGGCTAGACATGCGGAAAAGCAGATCCCGCCTGAGCCACATTGGGGCTTCTGT
sqs1cacDNA          AAACTAGCGGCTCGACATGCCGAAAAGCAGATCCCGCCGGAGCCACATTGGGGATTCTGT

                    ---------+---------+---------+---------+---------+---------+

                             130       140       150       160       170       180

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     TACTTAATGCTTCAAAAGGTCTCTCGTAGTTTTGCTCTCGTCATTCAACAGCTTCCTGTC
sqs1sc cDNA         TACTTAATGCTTCAAAAGGTCTCTCGTAGTTTTGCTCTCGTCATTCAACAGCTTCCTGTC
sqs1cacDNA          TACTTAATGCTTCAAAAAGTCTCTCGTAGTTTTGCTCTCGTCATTCAACAGCTTCCTGTC

                    ---------+---------+---------+---------+---------+---------+

                             190       200       210       220       230       240

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GAGCTTCGTGATGCTGTATGCATATTCTATTTGGTCCTTCGAGCACTGGACACTGTTGAG
sqs1sc cDNA         GAGCTTCGTGATGCTGTATGCATATTCTATTTGGTCCTTCGAGCACTGGACACTGTTGAG
sqs1cacDNA          GAGCTTCGTGATGCTGTATGCATATTCTATTTGGTCCTTAGAGCACTTGACACTGTCGAG

                    ---------+---------+---------+---------+---------+---------+

                             250       260       270       280       290       300

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GATGATACCAGCATACCCACCGATGTTAAAGTACCTATTCTGATCTCTTTTCATCAGCAT
sqs1sc cDNA         GATGATACCAGCATACCCACCGATGTTAAAGTACCTATTCTGATCTCTTTTCATCAGCAT
sqs1cacDNA          GATGATACCAGCATTCCCACGGATGTTAAAGTACCTATTCTGATCTCTTTTCATCAGCAT

                    ---------+---------+---------+---------+---------+---------+

                             310       320       330       340       350       360

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GTTTATGATCGCGAATGGCACTTCGCATGTGGTACGAAGGAGTACAAGGTTCTCATGGAC
sqs1sc cDNA         GTTTATGATCGCGAATGGCACTTCGCATGTGGTACGAAGGAGTACAAGGTTCTCATGGAC
sqs1cacDNA          ATCTATGATCGTGAATGGCACTTTTCATGTGGTACAAAGGAGTACAAGGTTCTCATGGAC

                    ---------+---------+---------+---------+---------+---------+
                             370       380       390       400       410       420

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     CAATTCCATCATGTTTCGACTGCTTTTCTGGAACTTGGTAAACTTTATCAGCAGGCAATT
sqs1sc cDNA         CAATTCCATCATGTTTCGACTGCTTTTCTGGAACTTGGTAAACTTTATCAGCAGGCAATT
sqs1cacDNA          CAGTTCCATCATGTCTCAACTGCTTTTCTGGAACTTGGAAAAAATTATCAGCAAGCAATT

                    ---------+---------+---------+---------+---------+---------+

                             430       440       450       460       470       480

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GAGGACATTACCATGAGGATGGGTGCAGGAATGGCAAAATTTATATGCAAGGAGGTGGAA
sqs1sc cDNA         GAGGACATTACCATGAGGATGGGTGCAGGAATGGCAAAATTTATATGCAAGGAGGTGGAA
sqs1cacDNA          GAGGATATTACCATGAGGATGGGTGCAGGAATGGCAAAATTTATATGCAAGGAGGTGGAA

                    ---------+---------+---------+---------+---------+---------+

                             490       500       510       520       530       540

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     ACAACTGATGATTATGACGAATACTGTCACTATGTAGCTAGGCTTGTTGGGCTAGGATTG
sqs1sc cDNA         ACAACTGATGATTATGACGAATACTGTCACTATGTAGCTAGGCTTGTTGGGCTAGGATTG
sqs1cacDNA          ACAACCGATGATTATGACGAATATTGTCACTACGTAGCTGGGCTTGTTGGGCTAGGATTG

                    ---------+---------+---------+---------+---------+---------+

                             550       560       570       580       590       600

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     TCAAAACTGTTCCATGCCTCGGGGACAGAAGATCTGGCTTCAGATTCTCTCTCCAACTCC
sqs1sc cDNA         TCAAAACTGTTCCATGCCTCGGGGACAGAAGATCTGGCTTCAGATTCTCTCTCCAACTCC
sqs1cacDNA          TCAAAACTGTTCCATGCATCTGGGAAAGAAGATCTGGCTTCAGATTCTCTCTCCAACTCC        6

                    ---------+---------+---------+---------+---------+---------+

                             610       620       630       640       650       660

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     ATGGGTTTATTTCTTCAGAAAACAAACATTATCAGAGATTATTTGGAAGATATAAATGAA
sqs1sc cDNA         ATGGGTTTATTTCTTCAGAAAACAAACATTATCAGAGATTATTTGGAAGATATAAATGAA
sqs1cacDNA          ATGGGTTTATTTCTTCAGAAAACAAACATCATTAGAGATTATCTGGAAGACATAAATGAA

                    ---------+---------+---------+---------+---------+---------+

                             670       680       690       700       710       720

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GTACCCAAGTGCCGTATGTTCTGGCCCCGTGAGATTTGGAGTAAATATGTTAACAAGCTT
sqs1sc cDNA         GTACCCAAGTGCCGTATGTTCTGGCCCCGTGAGATTTGGAGTAAATATGTTAACAAGCTT
sqs1cacDNA          GTACCCAAGTGCCGTATGTTTTGGCCCCGTGAGATTTGGAGTAAATATGTTAACAAGCTT

                    ---------+---------+---------+---------+---------+---------+
                             730       740       750       760       770       780

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GAGGACTTAAAGTACGAGGAGAACTCGGTTAAGGCAGTGCAATGTCTCAATGAAATGGTC
sqs1sc cDNA         GAGGACTTAAAGTACGAGGAGAACTCGGTTAAGGCAGTGCAATGTCTCAATGAAATGGTC
sqs1cacDNA          GAGGAGTTAAAGTATGAGGAGAACTCGGTCAAGGCAGTGCAATGTCTTAATGACATGGTC

                    ---------+---------+---------+---------+---------+---------+

                             790       800       810       820       830       840

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     ACCAATGCTTTGTCACATGTAGAAGATTGTTTGACTTACATGTTCAATTTGCGTGATCCT
sqs1sc cDNA         ACCAATGCTTTGTCACATGTAGAAGATTGTTTGACTTACATGTTCAATTTGCGTGATCCT
sqs1cacDNA          ACCAATGCTTTGTCACATGTAGAAGATTGTTTGATTTACATGTCCAATTTGCGTGATCCT

                    ---------+---------+---------+---------+---------+---------+

                             850       860       870       880       890       900

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     TCCATCTTTCGATTCTGTGCCATTCCACAGGTCATGGCAATTGGGACATTAGCTATGTGC
sqs1sc cDNA         TCCATCTTTCGATTCTGTGCCATTCCACAGGTCATGGCAATTGGGACATTAGCTATGTGC
sqs1cacDNA          GCCATCTTTCGATTCTGTGCTATTCCACAGGTCATGGCAATTGGGACTTTAGCTATGTGC

                    ---------+---------+---------+---------+---------+---------+

                             910       920       930       940       950       960

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     TATGACAACATTGAAGTCTTCAGAGGAGTGGTAAAAATGAGGCGTGGTCTTACTGCTAAG
sqs1sc cDNA         TATGACAACATTGAAGTCTTCAGAGGAGTGGTAAAAATGAGGCGTGGTCTTACTGCTAAG
sqs1cacDNA          TATGACAACATTGAAGTCTTCAGAGGAGTGGTTAAAATGAGACGTGGTCTGACAGCTAAG

                    ---------+---------+---------+---------+---------+---------+

                             970       980       990       1000      1010      1020

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GTCATTGACCGGACCAAGACTATGGCAGATGTATATGGTGCTTTTTTTGACTTTTCTTGT
sqs1sc cDNA         GTCATTGACCGGACCAAGACTATGGCAGATGTATATGGTGCTTTTTTTGACTTTTCTTGT
sqs1cacDNA          GCCATTGACCGGACTAGAACTATGGCTGATGTATATGGTGCTTTTTTTGACTTCTCTTGT

                    ---------+---------+---------+---------+---------+---------+

                             1030      1040      1050      1060      1070      1080

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     ATGCTGAAATCCAAGGTTAATAATAACGATCCAAATGCAACAAAAACTTTGAAGAGGCTT
sqs1sc cDNA         ATGCTGAAATCCAAGGTTAATAATAACGATCCAAATGCAACAAAAACTTTGAAGAGGCTT
sqs1cacDNA          ATGCTGAAATCCAAGGTTAATAATAATGATCCAAATGCAACAAAAACTTTGAAGAGGCTT

                    ---------+---------+---------+---------+---------+---------+
                             1090      1100      1110      1120      1130      1140

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GACGCCATCCTGAAAACTTGCAGAGACTCGGGAACCTTGAACAAAAGGAAATCTTACATA
sqs1sc cDNA         GACGCCATCCTGAAAACTTGCAGAGACTCGGGAACCTTGAACAAAAGGAAATCTTACATA
sqs1cacDNA          GAAGCAATCCTGAAAACTTGCAGAGACTCGGGAACCTTGAATAAAAGGAAATCTTACGTA        1

                    ---------+---------+---------+---------+---------+---------+

                             1150      1160      1170      1180      1190      1200

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     ATCAGGAGCGAGCCTAATTACAGTCCAGTTCTGATTGTTGTCATCTTCATCATACTGGCT
sqs1sc cDNA         ATCAGGAGCGAGCCTAATTACAGTCCAGTTCTGATTGTTGTCATCTTCATCATACTGGCT
sqs1cacDNA          ATCAAGAGCGAGCCTACTTACAGTCCAGTTCTGATCTTTGTCATCTTCATCATACTGGCT

                    ---------+---------+---------+---------+---------+---------+

                             1210      1220      1230      1240      1250      1260

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     ATTATTCTTGCACAACTTTCTGGCAACCGATCTTAGACCATTTTAAGTATCTAATCATGA
sqs1sc cDNA         ATTATTCTTGCACAACTTTCTGGCAACCGATCTTAGACCATTT-----------------
sqs1cacDNA          ATTATTCTTGCACACCTATCTGGAAACCGCTCTTAGATGATCT---------------



                    ---------+---------+---------+---------+---------+---------+

                             1270      1280      1290      1300      1310      1320

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GATACTTGAGTACATGCCAATTATTTAGATGCATGCCTCGTAGTTCAGAAATATACCCTC
sqs1sc cDNA         ------------------------------------------------------------
sqs1cacDNA          ------------------------------------------------------------

                    ---------+---------+---------+---------+---------+---------+

                             1330      1340      1350      1360      1370      1380

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     TATGCACCTAAGCTTTTGACTTGATGTCTAATGATAAGCATGTGTATCATTATATGACCT
sqs1sc cDNA         ------------------------------------------------------------
sqs1cacDNA          -----------------------------------------------------GTGACCA

                    ---------+---------+---------+---------+---------+---------+

                             1390      1400      1410      1420      1430      1440

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     TTTTTTTTAATCTAGCCTTTAAACAATAAGCACAGTAATTTTCCAAATTATGATTTGTAT
sqs1sc cDNA         ------------------------------------------------------------
sqs1cacDNA          TCTTTT------------------------------------------------------

                    ---------+---------+---------+---------+---------+---------+
                             1450      1460      1470      1480      1490      1500

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     TCTCTCTTTTCTTTTCTTTCTATTACCTGCTATTTAAGATTGCATTGTTTTTTTTAATAA
sqs1sc cDNA         ------------------------------------------------------------
sqs1cacDNA          ------------------------------------------------------------

                    ---------+---------+---------+---------+---------+---------+

                             1510      1520      1530      1540      1550      1560

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     ACACAAAAACTTTTCCTCCCAACTTAACCCAATTCTTTTTTTAAAAAAACCACCACTTTC
sqs1sc cDNA         ------------------------------------------------------------
sqs1cacDNA         ------------------------------------------------------------

                    ---------+---------+---------+---------+---------+---------+

                             1570      1580      1590      1600      1610      1620

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     CTCCTAACCTTGAACGTATTTGACAACTAATTTCGGTTATGTCATTCTTCTGCCATCACA
sqs1sc cDNA         ------------------------------------------------------------
sqs1cacDNA          ------------------------------------------------------------

                    ---------+---------+---------+---------+---------+---------+


                             1630      1640      1650      1660      1670      1680

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GATTTCCAATTCTAAGTGAAAAAAGAACAAATTATGGAAAATGTGTATCAATTTAAGGAT
sqs1sc cDNA         ------------------------------------------------------------
sqs1cacDNA          ------------------------------------------------------------

                    ---------+---------+---------+---------+---------+---------+

                             1690      1700      1710      1720      1730      1740

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     AACTGTGCTAAGAGTCAGTCAACATAGAGACATGGAAATTGTATCCCTTTCAGTTTTATG
sqs1sc cDNA         ------------------------------------------------------------
sqs1cacDNA          ----GGGTTAGAAGTT--------------------------------------------

                    ---------+---------+---------+---------+---------+---------+

                             1750      1760      1770      1780      1790      1800

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     GTGGAGAGTGTTAACCCGGATTTATTTGTCCTGATTTGTAGATTGATTGGTCTACAAAAA
sqs1sc cDNA         ---------------------------------------------ATTGGTCTACAAAAA
sqs1cacDNA          ------------------------------------------------------------

                    ---------+---------+---------+---------+---------+---------+
                             1810      1820      1830      1840      1850      1860

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     TGAACTATGGTCAACGAAGACAGCACAAACTCTTGGCCAATTATGTACTGCTAATTGTTA
sqs1sc cDNA         TGAACTATGGTCAACGAAGACAGCACAAACTCTTGGCCAATTATGTACTGCTAATTGTTA
sqs1cacDNA          -------TGGTCAAGGAGGTCAAT-------------TATGTGATTAATACAAATTGTCA

                    ---------+---------+---------+---------+---------+---------+

                             1870      1880      1890      1900      1910      1920

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     TGTTTGTATTACTATGTTCATTAAGTTAATAGTTGCATCTTCAACCTGACTAGATAATTA
sqs1sc cDNA         TGTTTGTATTACTATGTTCATTAAGTTAATAGTTGCATCTTCAACCTGACTAGATAATTA
sqs1cacDNA          TGTTTGTATTAGTATGT--ATTAAGT-GATAGTTGCACCTTCAACCTGACAG----ATAA

                    ---------+---------+---------+---------+---------+---------+

                             1930      1940      1950      1960      1970      1980

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     CGAAAGCCTATC-TATGGCAGTTAGTTTGGTATGTATTTGTTTGCAAGCT---AGGAAAG
sqs1sc cDNA         CGAAAGCCTATC-TATGGCAGTTAGTTTGGTATGTATTTGTTTGCAAGCT---AGGAAAG
sqs1cacDNA          CGAAAGCCTATTATCTGGTAGTTTGTTGAGTATGTACTTGTTTGCAAGCTGCTACGAAAG

                    ---------+---------+---------+---------+---------+---------+


                             1990      2000      2010      2020      2030      2040

                    ---------+---------+---------+---------+---------+---------+

sqs1sccDNA+in13     CAAATTCCAAGTGTTGTAGA-GTCGTTTTTCCGTAATGCACATTTCATTTTAA
sqs1sc cDNA         CAAATTCCAAGTGTTGTAGA-GTCGTTTTTCCGTAATGCACATTTCATTTTAA
sqs1cacDNA          CAAATTCCAATTGTTGTAGAAGTCGGTTTACCGTAATATACATTTCATTGTAACAGCTTG
```

Appendix E. Alignment of the 3'UTR of *sqs1$_{Sc}$* and *sqs1$_{Ca}$*. The top row is the alignment of the *sqs1$_{Sc}$* coding region plus the genomic sequence of intron 13. The coding region and 3'UTR of two *sqs1* homologs are compared to show that *sqs1$_{Ca}$* does not have an intron in the 3'UTR. Intron 13 in the genomic sequence is blue. *sqs1$_{Ca}$* fragments aligned in the middle of the intron are red.

Appendix F: RNA-seq data for *hmg*, *sqs*, and *sqe* genes in potato

Gene	roots	stolons	tubers	stems	leaves	Flower	sepals	petals	stamen	carpels	Scaffold
hmg	21	70	50	51	8	196	203	703	88	88	DMG400009924
hmg	39	266	132	25	5	1477	207	91	5380	372	DMG400003461
sqs1											not found
sqs2	0	0	0	0	0	7	0	0	6	0	DMG400003408
sqs3	0	0	0	0	0	0	0	0	0	0	DMG400008184
sqs4	0	0	0	0	0	0	0	0	0	0	DMG400039005
spe	49	19	41	29	19	70	80	39	112	36	DMG400003324
spe	325	63	111	10	142	340	213	62	235	185	DMG400004923
spe	0	0	0	0	0	1	2	0	0	0	DMG400005105

Appendix F RNA-seq data of *hmg*, *sqs* and *spe* genes in potato. Quantified RNA-seq values of annotated *3-hydroxy-3-methyl-glutaryl-CoA reductase* (*hmg*), squalene synthase (*sqs)*, and *squalene epoxidase* (*spe*) are presented. The scaffolds (version 3) are listed on the right. The values for each tissue are listed in the columns. The members of the *sqs* gene family were given numbers. Only three *sqs* genes were annotated. The tissue of highest RNA-seq value is colored in pink. RNA-seq was done by the Potato Genome Sequencing Consortium (http://www.potatogenome.net/)

Appendix G: Accession number and species of origin for each SQS from Figure 3.4

Species-Isoform	Accession	Species-Isoform	Accession
Homo sapiens	2004281A	*Gossypium hirsutum-2*	EF688567.1
Rattus norvegicus	NM_019238.2	*Glycyrrhiza eurycarpa-1*	AM182331.1
Mus musculus	2105185A	*Glycyrrhiza eurycarpa-2*	AM182332.1
Saccharomyces cerevisiae	ACD03847	*Diospyros kaki*	FJ687954.1
Yarrowia lipolytica	AAD22408	*Capsicum annuum*	AF124842.1
Botryococcus braunii	AAF20201	*Centella asiatica*	AY787628.1
Arabidopsis thaliana	NM_119630	*Bupleurum falcatum*	AY964186.1
Taxus cuspidata	DQ836053	*Oryza sativa*-1	AB007501
Salvia miltiorrhiza	FJ768961.1	*Panax ginseng-1*	AB010148.1
Psammosilene tunicoides	EF585250.1	*Populus trichocarpa*-1	XM_002305419
Panax quinquefolius	AM182457.1	*Populus trichocarpa*-2	XM_002313729
Gynostemma pentaphyllum	FJ906799.1	*Zea mays*	BAA22558

Appendix G: Accession number and species of origin for each SQS isoform from Figure 3.4. The species name of each organism the SQS sequences used in **Figure 3.4** were isolated from species with multiple *sqs* genes, the isoform used is indicated after a dash after the name. The accession numbers for each gene are presented.

Appendix H: Deduced amino acid sequence of protein produced by *sqs1*$_{Sc}$-pET32a

```
MSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLL 80

LFKNGEVAATKVGALSKGQLKEFLDANLAGSGSGHMHHHHHHSSGLVPRGSGMKETAAAKFERQHMDSPDLGTDDDDKAM 160

ADIGSEFMGTLRAILKNPDDLYPLIKLKLAARHAEKQIPPEPHWGFCYLMLQKVSRSFALVIQQLPVELRDAVCIFYLVL 240

RALDTVEDDTSIPTDVKVPILISFHQHVYDREWHFACGTKEYKVLMDQFHHVSTAFLELGKLYQQAIEDITMRMGAGMAK 320

FICKEVETTDDYDEYCHYVAGLVGLGLSKLFHASGTEDLASDSLSNSMGLFLQKTNIIRDYLEDINEVPKCRMFWPREIW 400

SKYVNKLEDLKYEENSVKAVQCLNEMVTNALSHVEDCLTYMFNLRDPSIFRFCAIPQVMAIGTLAMCYDNIEVFRGVVKM 480

RRGLTAKVIDRTKTMADVYGAFFDFSCMLKSKVNNNDPNATKTLKRLDAILKTCRDSGTLNKRKSYIIRSEPNYSAAALE 560

HHHHHH.                                                                          572
```

Appendix H Deduced amino acid sequence of *sqs1*$_{Sc}$ in pET32a. Sequence was obtained from cloned expression vectors. The recombinant enzyme is comprised of allele one of truncated *sqs1*$_{Sc}$ in the center, which is underlined, and part of the expression vector that introduces tags to the protein. Amino acid position is numbered at the right.

Appendix I: Deduced amino acid sequence of *sqs2*$_{Sc}$ in pET32a

```
GILRAILRHPEDIYPLLKLKVAARYAEKQIPSQPHWAFCYIMLHKVSRSFSLVIKQLPVELRDAICIFYLVLRALDTVED 80

DTSVATEVKVPILMSFHRHVYDREWHFSCGTKDYKVLMDQFHHVSTAFLELGKHYKEAIEDITMRMGAGMAKFIYKEVET 160

IDDYDEYCHHVAGLVGLGLSKLFHASGKEDVASDSLCNSMGLFLQKTNIIRDYLEDINEVPKCRMFWPRQIWSEYVDKLE 240

DLKYEGNSVKAVQCLNEMVTNALSHAEDCLTFLSTLRDPTIFRFCAIPQAMAIGTLAKCYNNIEVFRGVVKMRRGLTAQV 360

IDRTRNMADVYGAFFDFSCILKSKVEYKDPHVAKTLKRLEVILRTCKNSGTLNKRKS                        377
```

Appendix I Deduced amino acid sequence of *sqs2*$_{Sc}$ in pET32a. Sequence was obtained from cloned expression vectors. The *sqs2*$_{Sc}$ portion of the *sqs2*$_{Sc}$-pET32 construct was sequenced. All the sequence represents SQS2$_{Sc}$ that is in pET32a. The sequence of the SQS2$_{Sc}$ coding region in the vector is not complete.