Plant Biotechnology Journal

Plant Biotechnology Journal (2015) **13**, pp. 177–187 doi: 10.1111/pbi.12238

The *QQS* orphan gene of Arabidopsis modulates carbon and nitrogen allocation in soybean

Ling Li* and Eve Syrkin Wurtele*

Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA, USA

Received 20 March 2014;
revised 30 June 2014;
accepted 3 July 2014.
**Correspondence* (Tel +1 515 294 6236, +1 515 294 8989; fax +1 515 294 1337; emails liling@iastate.edu and mash@iastate.edu)

Accession numbers: Sequence data from this article can be found in The Arabidopsis Genome Information Resource under the following accession numbers: *QQS* (At3g30720).

Keywords: *QQS*, orphan, carbon and nitrogen allocation, protein, starch, *Glycine max*.

Summary

The genome of each species contains as high as 8% of genes that are uniquely present in that species. Little is known about the functional significance of these so-called species specific or orphan genes. The *Arabidopsis thaliana* gene Qua-Quine Starch (*QQS*) is species specific. Here, we show that altering *QQS* expression in Arabidopsis affects carbon partitioning to both starch and protein. We hypothesized *QQS* may be conserved in a feature other than primary sequence, and as such could function to impact composition in another species. To test the potential of *QQS* in affecting composition in an ectopic species, we introduced *QQS* into soybean. Soybean T1 lines expressing *QQS* have up to 80% decreased leaf starch and up to 60% increased leaf protein; T4 generation seeds from field-grown plants contain up to 13% less oil, while protein is increased by up to 18%. These data broaden the concept of *QQS* as a modulator of carbon and nitrogen allocation, and demonstrate that this species-specific gene can affect the seed composition of an agronomic species thought to have diverged from Arabidopsis 100 million years ago.

Introduction

The ability to optimize protein productivity of plant-based foods could have far-ranging impacts to both world health and to sustainability (Godfray *et al.*, 2010; Heitschmidt *et al.*, 1996; Pimentel and Pimentel, 2003). Dietary protein is essential for animals, whereas photosynthetic organisms can biosynthesize all amino acids required for protein synthesis. Over four billion of the seven billion people on our planet obtain the majority of their dietary protein from plants (Pimentel and Pimentel, 2003; Young and Pellett, 1994). However, for many people, protein intake is insufficient, and its deficiency results in mental retardation, stunting of growth, and greatly increased susceptibility to disease, predominantly affecting children (Gomes *et al.*, 2009; Muller and Krawinkel, 2005; Victora *et al.*, 2010).

Because plants are solar-powered heterotrophs in the food web, consumption of plant-derived proteins has far less impact on the environment than consumption of animal protein sources, especially considering the earth's dwindling water resources (Pimentel and Pimentel, 2003). Thus, increasing the use of plants as a protein source, rather than animals, would have a major ecological significance.

Plant composition is determined by a metabolic network that mediates the conversion of imported, photosynthetically-derived carbon and nitrogen into protein, oil and carbohydrate. Many of the pathways by which plants synthesize or degrade protein, oil and starch have been delineated; however, much less is understood about the mechanisms that integrate these pathways and that regulate carbon and nitrogen allocation to and within this network (Eastmond, 2006; Eastmond *et al.*, 1997; Higashi *et al.*, 2006; Li *et al.*, 2009, 2012; Schiltz *et al.*, 2004; Sulpice *et al.*, 2013). Understanding this process holistically is a major biological challenge.

To identify genes that impact plant composition, our strategy leveraged the model species Arabidopsis. Based on the postulate that the basic molecular genetic mechanisms for homeostasis are conserved across species (Li *et al.*, 2009), we selected Arabidopsis single-gene mutants that appeared morphologically like the wild type (WT) control, but differed in composition, and then determined the transcripts whose expression was impacted in the mutants. We anticipated identifying a combination of metabolic and regulatory genes by this approach. The *ss3* knockout mutant of Arabidopsis is high in starch, but has a normal morphological phenotype (Zhang *et al.*, 2005, 2008); among the genes whose expression is altered in *Atss3* mutants relative to WT plants is Qua-Quine Starch (*QQS*, locus At3g30720) (Li *et al.*, 2009). Reduction of *QQS* expression in transgenic *QQS* RNAi lines of Arabidopsis resulted in plants that were morphologically indistinguishable from control lines, but that expressed a 15%–30% increase in leaf starch content (Li *et al.*, 2009).

This study further defines how *QQS* functions in Arabidopsis, showing that altering expression of *QQS* in either over-expression or *QQS* RNAi lines results in shifts in leaf protein and starch content.

The *QQS* gene encodes a protein of only 59 amino acids whose homolog is not identifiable by primary sequence comparisons to any other sequenced species, not even the closely related *Brassica napus* (Li *et al.*, 2009) or *Arabidopsis lyrata*; as such, *QQS* is considered an orphan gene. Orphan genes (also referred to as species-specific genes, or, in prokaryotes, ORFans) can be defined as genes that encode proteins that are unique to a given species, having no identifiable sequence homologs in other species (Gollery *et al.*, 2006, 2007; Li *et al.*, 2009). The concept of orphan genes was first described by Fischer and Eisenberg in 1999 from studies of microbial genomes (Fischer and Eisenberg, 1999).

Although many have predicted that genes considered species specific would later turn out to be an artifact of sparse genome sequence, this has proved not to be the case (Arendsee *et al.*, 2014; Gollery *et al.*, 2006, 2007; Marsden *et al.*, 2006; Neme and Tautz, 2013; Silveira *et al.*, 2013; Tautz and Domazet-Loso, 2011). Orphan genes appear to be present in all species, and represent a significant fraction (approximately 0.5% to >8%) of analysed eukaryotic and prokaryotic genomes. The function of most orphan genes is obscure; however, they have been considered to be a determinant of species character (Gollery *et al.*, 2006; Tautz and Domazet-Loso, 2011).

Orphan genes are thought to arise by multiple mechanisms (Carvunis *et al.*, 2012; Donoghue *et al.*, 2011; Tautz and Domazet-Loso, 2011; Wissler *et al.*, 2013). Nongenic sequence can be defined as the sequence that is not a part of an organism's genes (Adler, 1992). In *Saccharomyces cervesiae*, nongenic sequences have been shown to be transcribed widely (Nagalakshmi *et al.*, 2008) and some is also translated (Carvunis *et al.*, 2012). Genes could arise *de novo* from such nongenic sequences via a noncoding or coding proto-gene that becomes more stabilized during evolution into an orphan gene (Carvunis *et al.*, 2012). Orphan genes also arise via pre-existing genes, whose sequence can be highly modified by combinations of gene duplication, domain shuffling, shifting of location of translation frames and subsequent diversification (Carvunis *et al.*, 2012; Ohno, 1987; Tautz and Domazet-Loso, 2011). The lack of even remote footprints of any *A. thaliana* genic sequence within *QQS* indicates it is likely an orphan gene that arose *de novo* (Silveira *et al.*, 2013).

Another feature of orphans may be a general lack of co-expression with other genes. One way to consider this feature is to evaluate their representation in regulons. Regulons of eukaryotes can be defined as clusters of genes that have a prevailing pattern of co-expression *across multiple/thousands of diverse conditions* (Biehl *et al.*, 2005; Feng *et al.*, 2012; Mentzen *et al.*, 2008); these have been detailed for the Arabidopsis transcriptome using a Markov chain cluster (MCL) approach (Mentzen and Wurtele, 2008). Only some of the genes of an organism can be classified into regulons; others do not have a prevailing pattern of co-expression with other genes (Feng *et al.*, 2012; Mentzen and Wurtele, 2008). The Arabidopsis transcriptome can be partitioned into 872 regulons; about 49% of all Arabidopsis genes are members of these regulons (Mentzen and Wurtele, 2008). However, only 4.3% of orphans are members of regulons. Thus orphans are highly underrepresented among the co-expressed gene clusters.

We conjecture that some orphan genes might have arisen and stabilized because they confer a selective advantage by interacting with a previously existing protein. This previously existing protein could be relatively conserved, that is, present in many lineages. Because *QQS* influences composition, a process critical to plants in general, our working hypothesis is that *QQS* contains structural features that would interact with a conserved protein, such that introduction of the *QQS* gene into another species would influence its compositional traits.

We tested the hypothesis that *QQS* could function to regulate composition in another species by introducing the *QQS* gene into the major food crop, soybean. Here, we demonstrate the expression of the *QQS* gene in soybean increases leaf and seed protein and decreases leaf and seed carbohydrate. These results reveal that expression of the *QQS* gene increases carbon and nitrogen allocation to protein and that it can exert this function even in an ectopic species.

Results

Arabidopsis plants over-expressing and under-expressing *QQS* have altered carbon and nitrogen allocation to protein and starch

Our findings that down-regulation of *QQS* increases starch content in Arabidopsis (Li *et al.*, 2009) might indicate that altered *QQS* impacts starch content without changing other aspects of leaf composition. Alternately, *QQS* might have a more general effect on carbon and nitrogen allocation, affecting protein and/or lipid content. To distinguish between these alternatives, we evaluated the protein and starch content of transgenic Arabidopsis plants that either over-expressed [under control of the constitutive cauliflower mosaic virus (CaMV) 35S promoter, see Figure S1] or suppressed (using *QQS* RNAi, Li *et al.*, 2009) the accumulation of the *QQS* coding sequence. A total of ten independent *QQS* over-expression mutant lines and four independent *QQS* RNAi mutant lines were grown and evaluated using a randomized complete block design. The visual phenotype of these transgenic Arabidopsis lines appeared identical to the control lines throughout development, from seedlings to senescence (Figure 1a,b). However, leaf starch content in lines that over-expressed *QQS* (*QQS-OE*) was decreased by up to 23% (Figure S2). Conversely, *QQS* RNAi lines showed an increase in leaf starch content (Figures 1b and S2), consistent with our previous publication (Li *et al.*, 2009). Leaf protein content was altered in each of the *QQS-OE* and *QQS* RNAi mutant lines that we tested; specifically, protein content was increased by about 3% in *QQS-OE* mutants and decreased by 3%–7% in *QQS* RNAi mutants (Figure 1c). These data indicate that *QQS* acts either directly or indirectly as a regulator of carbon and nitrogen metabolism and affects not only the accumulation of starch, but also protein.

The BLink algorithm (NCBI, http://www.ncbi.nlm.nih.gov/sutils/blink.cgi?mode=query) identifies 1155 orphan genes in Arabidopsis, based on the gene models described in Arabidopsis using the TAIR10 genome release (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR10_genome_release/), including 30 mitochondrial genes (as dated on September 23, 2013). In addition to these 1115 *A. thaliana*-specific genes, 839 genes can be identified using BLink that are unique to *A. thaliana* and *A. lyrata* but not identifiable outside of the Arabidopsis genus; these can be referred to as clade-specific genes or genus-specific genes, rather than the more restrictive term, orphans.

Orphans are typically shorter than typical genes (Amiri *et al.*, 2003; Knowles and McLysaght, 2009; Wu *et al.*, 2011). In *A. thaliana*, the median length of the predicted protein models from the 1155 orphan genes is 57 amino acid (range, 16–445 amino acid) while the median length of predicted protein models from all genes is 349 amino acid (range, 16–5393 amino acid) (Figure 2). When an alternate, more inclusive, algorithm for orphan genes is used, the predicted proteins have a similarly short length (Gollery *et al.*, 2006).

Eukaryotic orphan genes generally tend to have greater tendency of being associated with transposable elements, relatively low expression that is organ specific, and less helical and sheet structure (as predicted computationally) (Arendsee *et al.*, 2014; Donoghue *et al.*, 2011; Wilson *et al.*, 2007). *QQS* has only some of these characteristics. Like a typical orphan gene, it is a short peptide (Figure 2) with a GC content of 40%, slightly higher than the average for Arabidopsis genes (this average GC

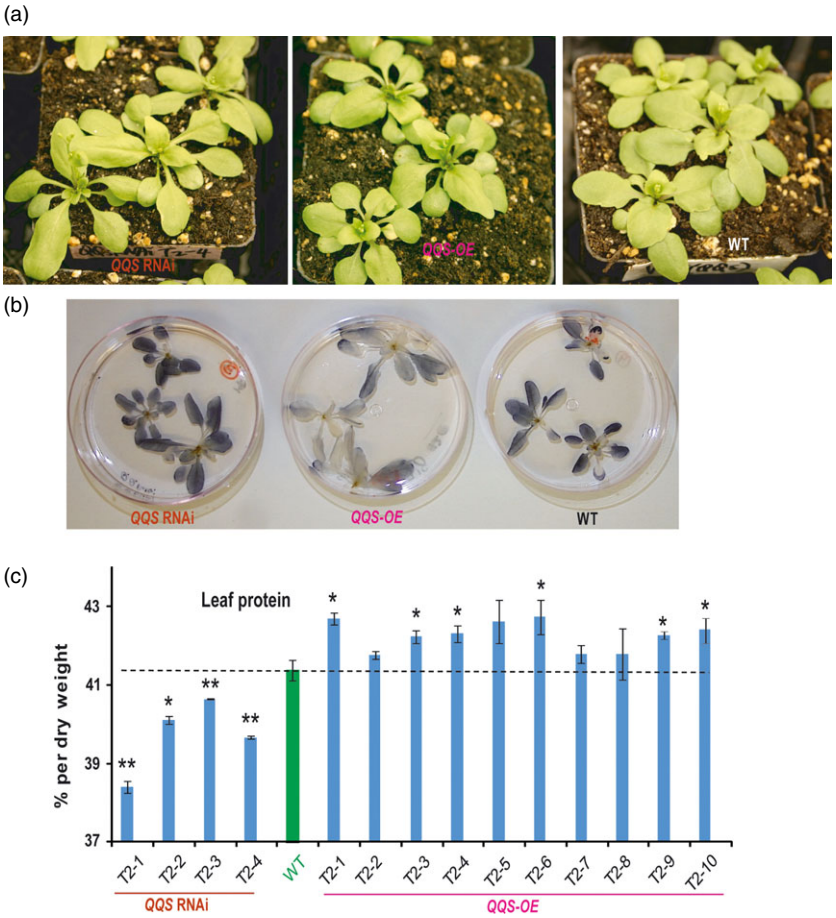

Figure 1 Phenotype and leaf composition of transgenic lines of Arabidopsis with down-regulation or over-expression of Qua-Quine Starch (*QQS*). Seedling shoots *of QQS* RNAi plants (Li *et al.*, 2009) derived from four independent transformation events, and *QQS-OE* plants derived from ten independent transformation events, together with wild type (WT) control plants, were sampled at the end of the light period. *QQS* RNAi plants are known to have increased starch (Li *et al.*, 2009). (a) All of these lines of transgenic *QQS* RNAi and *QQS-OE* plants are similar in morphology to WT plants throughout development. This phenotype (i.e. the *QQS-OE* plants were indistinguishable from WT in appearance) contrasts with the 35S:*QQS* transgenic phenotype of slower growth and rounded leaves that was described by Seo *et al.* (2011); possibly this is due to a difference in experimental conditions in which the plant were grown. (b) Starch staining shows increased starch in *QQS* RNAi mutants and decreased starch in *QQS-OE* mutants compared with WT. (c) Leaf protein is significantly decreased in *QQS* RNAi mutants and increased in *QQS-OE* mutants compared with WT. *QQS* mutants are each in the T2 generation, and each independent transformation events are designated by T2-independent transformation event #. All data in bar charts show mean ± SE (standard error), $n = 3$. Student's *t*-test was used to compare *QQS* RNAi and *QQS-OE* with WT, $*P < 0.05$; $**P < 0.01$.

content is 36% for Arabidopsis, Mishra *et al.*, 2009). *QQS* is embedded in a neighbourhood of chromosome 3 that is highly enriched in transposons (Figure S3) (Li *et al.*, 2009) and generally unusual. In the 3′ direction, *QQS* is 2.5 kb from At3g30725, a 'glutamine dumper-like' protein of unclear function; however, in the 5′ direction, *QQS* is a surprising 77.6 kb from the nearest gene, At3g30705, which is also an orphan gene and has no known function. The sequence immediately 5′ of the *QQS* start site contains multiple 5′ flanking siRNA repeated elements (Lister *et al.*, 2008; Silveira *et al.*, 2013).

Unlike what has been reported as general orphan characteristics, *QQS* expression in Arabidopsis ecotype Columbia (Col-0) is relatively not low and *QQS* is expressed in most plant organs; furthermore, its expression responds strongly to genetic and environmental perturbations (Li *et al.*, 2009). We evaluated qualitatively whether *QQS* expression is affected under two environmental conditions that are known to increase starch accumulation in Arabidopsis: high sucrose (Aloni *et al.*, 1997) and low temperature (Espinoza *et al.*, 2010) using transgenic plants containing the *QQS* promoter driving the GUS coding sequence (Figure S4). *QQS* expression is decreased in plants grown in medium contain 5% sucrose compared with medium without added sucrose. *QQS* expression is also decreased in plants growing at 4 °C compared with 22 °C. These results further demonstrate the strong sensitivity of *QQS* expression to environmental perturbations.

Disorder prediction tools (http://www.disprot.org/predictors.php) indicate that QQS protein has a disordered N-terminal tail (approximately 20 first residues). The remainder of QQS protein is predicted to contain two α helixes.

Soybean plants expressing *QQS* have increased leaf protein

To test the hypothesis that *QQS* could affect carbon and nitrogen allocation in an ectopic species, we determined whether expression of the *QQS* transgene would impact

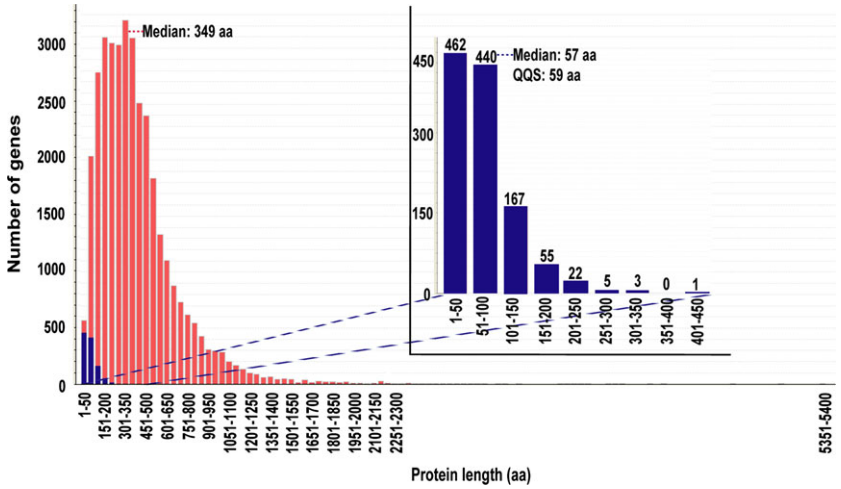

Figure 2 Distribution of orphan protein lengths in Arabidopsis. The median length of the predicted protein models from *Arabidopsis thaliana* orphan genes (blue bar) is smaller (57 amino acid) than the median length of all *A. thaliana* predicted protein models (orange bar) (349 amino acid) (range of orphan protein models: 16–445 amino acid; range of all protein models: 16–5393 amino acid). The *X*-axis represents the protein length (in amino acids); the *Y*-axis represents the number of genes. Orphan genes are predicted by NCBI BLink (http://www.ncbi.nlm.nih.gov/sutils/blink.cgi?mode=query). The insert provides a magnified view of the distribution of the sizes of orphan genes.

composition in the major crop plant, soybean (*Glycine max*). We chose soybean based on its evolutionary divergence from Arabidopsis and because of its importance as a direct or indirect source of dietary protein. Indeed, soy is the predominant source of protein for humans (Pathan and Sleper, 2008). In addition to its potential agronomic value, increasing the protein content in an already protein-rich crop could provide a stringent test of how *QQS* may affect compositional traits. Soybean (cultivar Williams 82) was transformed with the *QQS* coding sequence under the control of the constitutive CaMV 35S promoter (Figure 3). Transgenic soybean lines that expressed the *QQS* transgene were identified by selection for herbicide resistance conferred by the introduced vector, followed by real-time PCR analysis (Table S1).

T1-generation lines of soybean expressing the *QQS* gene survived from herbicide selection, from each of five independent transformation events (a total of 33 lines), together with Williams 82 controls, were evaluated after they were grown in growth chambers in a completely randomized design within each chamber. Visual examinations of the plants throughout development from seedling to senescence indicated that the morphology of the transgenic soybean plants expressing the *QQS* transcript was similar to the WT control plants (Figure 3a). However, in leaves of the transgenic *QQS*-expressing lines (*QQS-E*), starch content was reduced to levels as low as 20% of the WT control plants (Figure 3b). In contrast, leaf protein content was increased by up to 60% in the transgenic soybean plants (Figure 3c). The effect of *QQS* on the starch and protein accumulation in lines from the same transformation event may vary as the expression level of the transgene in different lines of the same transformation event may be different (Shou *et al.*, 2004). The plots of leaf starch and protein versus *QQS* transcript accumulation in Arabidopsis and in soybean indicate that for both species, when *QQS* RNA accumulation is elevated, protein accumulation is increased and starch accumulation is decreased. The relationship between *QQS* RNA level and starch and protein concentrations was not linear over the range of *QQS* RNA accumulation tested (Figure S5).

Soybean plants expressing *QQS* have increased seed protein

The increase in leaf protein associated with *QQS* gene expression in soybean led us to assess whether *QQS* had a broader impact, also affecting seed composition. As a preliminary indication of whether *QQS* expression alters seed protein content, T2 seeds from soybean plants that had been grown in growth chambers and survived from herbicide selection were screened by Nuclear Magnetic Resonance Spectrometry (NMR). These data (Figure S6) indicated a significant increase in seed protein in the *QQS*-expressing lines. To evaluate this finding in more detail, plants were propagated via single-seed descent. Progeny were grown in a completely randomized design in a greenhouse. The transgenic soybean plants expressing the *QQS* transcript survived from herbicide selection were indistinguishable in morphology and development from the WT control plants (Figure S7a). T3 seeds from the greenhouse-grown lines (seeds from the offspring of the same T1 plant were pooled and harvested together) were evaluated for composition by destructive chemical analysis. These seeds also showed a significant increase in seed protein in the *QQS*-expressing lines, compared with WT controls; the seed oil content of these lines was similar or slightly decreased (Figure S7b).

Based on the results of these analyses, we tested the effect of *QQS* expression in field-grown plants. Segregating seeds of independent transgenic lines were planted in a field in a randomized complete block design. Plants were monitored weekly during development. Plants were sprayed by herbicide and data on survival was monitored (Table S2). One line is possibly a homozygous line (16-6); in that line, all plants were herbicide resistant. WT nontransgenic siblings were identified by PCR analysis of DNA of leaf pieces from individual progeny of self-propagated T3 generation plants derived from three independent transformation events. Morphology and development of field-grown plants was indistinguishable among the populations of these WT-sibling lines and *QQS*-expressing lines (Figures 4a and S8a).

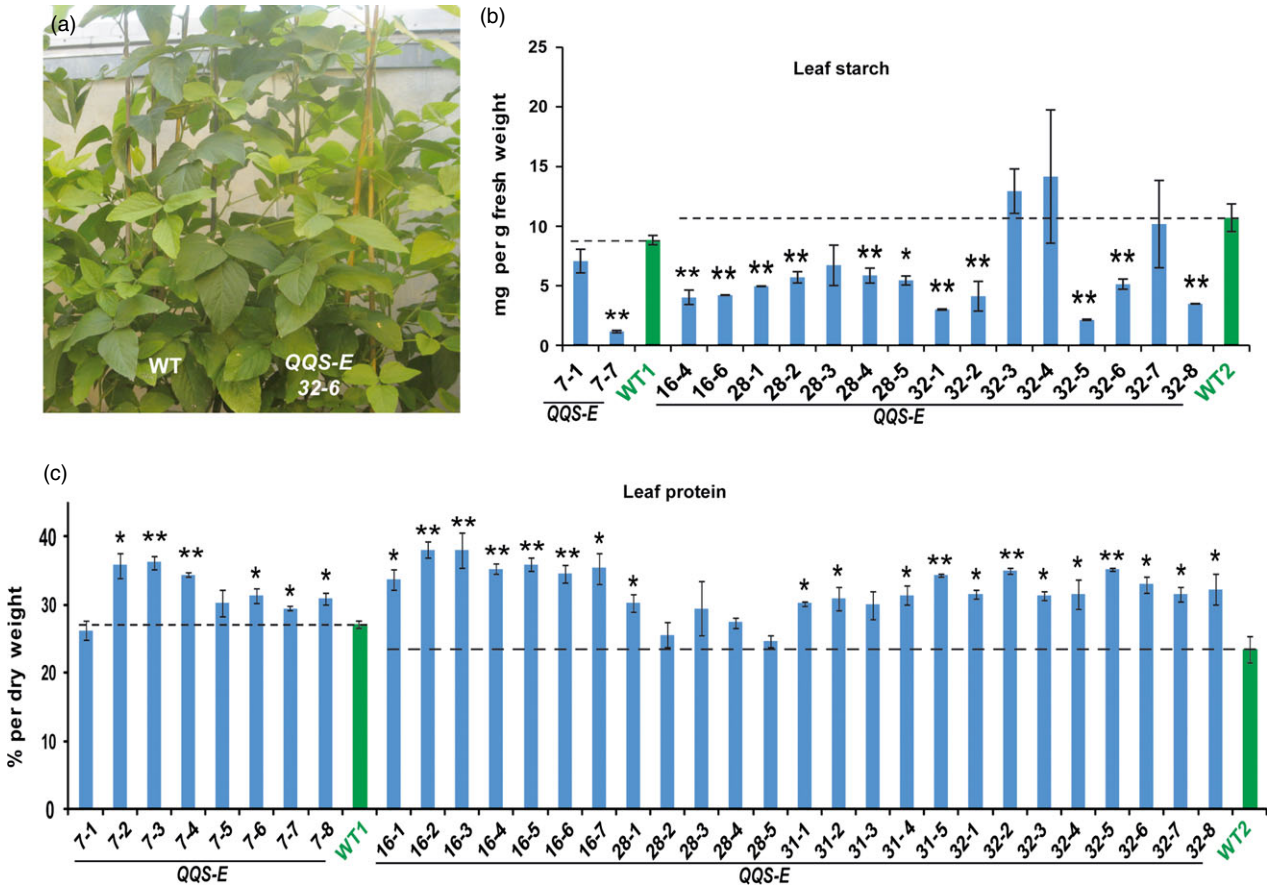

Figure 3 Characterization of leaves of growth chamber-grown transgenic soybean plants expressing Qua-Quine Starch (*QQS*). (a) Transgenic *QQS-E* plants (engineered in the Williams 82 background) are not visually distinguishable from Williams 82 control plants. (b) Leaf starch is decreased in *QQS-E* plants compared with wild type (WT) controls. (c) Leaf protein is increased in *QQS-E* plants compared with WT controls. WT1 and WT2, Williams 82 controls from two growth chambers; *QQS-E*, transgenic Williams 82 lines expressing the *QQS* coding sequence, selected by herbicide resistance. *QQS-E* mutants of T1 generation derived from 17 lines of four independent transformation events (for starch) and from 33 lines of five independent transformation events (for protein) are designated by: independent transformation event #–line #. All data in bar charts show mean ± SE, $n = 3$. Student's *t*-test was used to compare *QQS-E* and WT, $*P < 0.05$; $**P < 0.01$.

Seeds were harvested at maturity. Seed morphology, seed size and shape, seed weight per seed and per plant, and moisture content were similar among the populations of field-grown plants (Figures 4b and S8b,c). However, seed composition was significantly affected, as determined by independent methods: near infrared spectroscopy (NIRS, for eight lines, Table S3) and destructive chemical analyses (for four lines) to mature pooled seeds from herbicide-resistant mutant plants transformed with *QQS* compared to mature pooled seeds from WT plants. Protein content was from 10% to 18% higher in seeds of *QQS*-expressing lines as compared to those of WT-sibling controls (Figures 4c and S8c). The levels of several free amino acids were also increased (Table S4). The increase in the seed protein content did not affect the relative distribution of amino acids in hydrolysed proteins while the amino acid contents are significantly increased (Figure S8d and Table S5). Seed oil content in different lines ranged from a value similar to that of WT-sibling seeds to a 13% decrease (Figures 4c and S8c). The level of C16:0 is slightly decreased in the mutants (Table S6). There was no detectable change in levels of free fatty acids. Carbohydrate and fibre contents were decreased (Figure S8c). The total protein and oil content in seeds of *QQS*-expressing soybean was increased by up to 6.5% as compared with that of the WT-siblings (Figures 4c and S8c). Thus, individual lines of soybean expressing the *QQS* transgene appeared morphologically indistinguishable but showed increases of 18.0%, 10.3%, 13.1%, 12.9% seed protein and 6.5%, 6.3%, 5.5% and 4.3 % seed protein + oil, compared with their sibling WTs. Segregation study of individual transgenic plants and their segregated siblings identified by PCR analysis of DNA of leaf pieces, and these individual plants' seed composition analysis by NIRS indicated that the high-protein trait was associated with *QQS* expression (Table S7).

Discussion

Orphan gene function in plant biology

Little is understood about the functional significance of the vast majority of orphan genes in any species. Indeed, to our knowledge, *QQS* is the only plant orphan gene that has been studied in any detail. In general, orphan genes have been considered to confer a species-specific function, for example, immunity to particular pathogens, self-recognition, or resistance/adaptation to an environmental stress (Gollery *et al.*, 2007; Khalturin *et al.*, 2009). Of the plant orphan genes that anything is known about

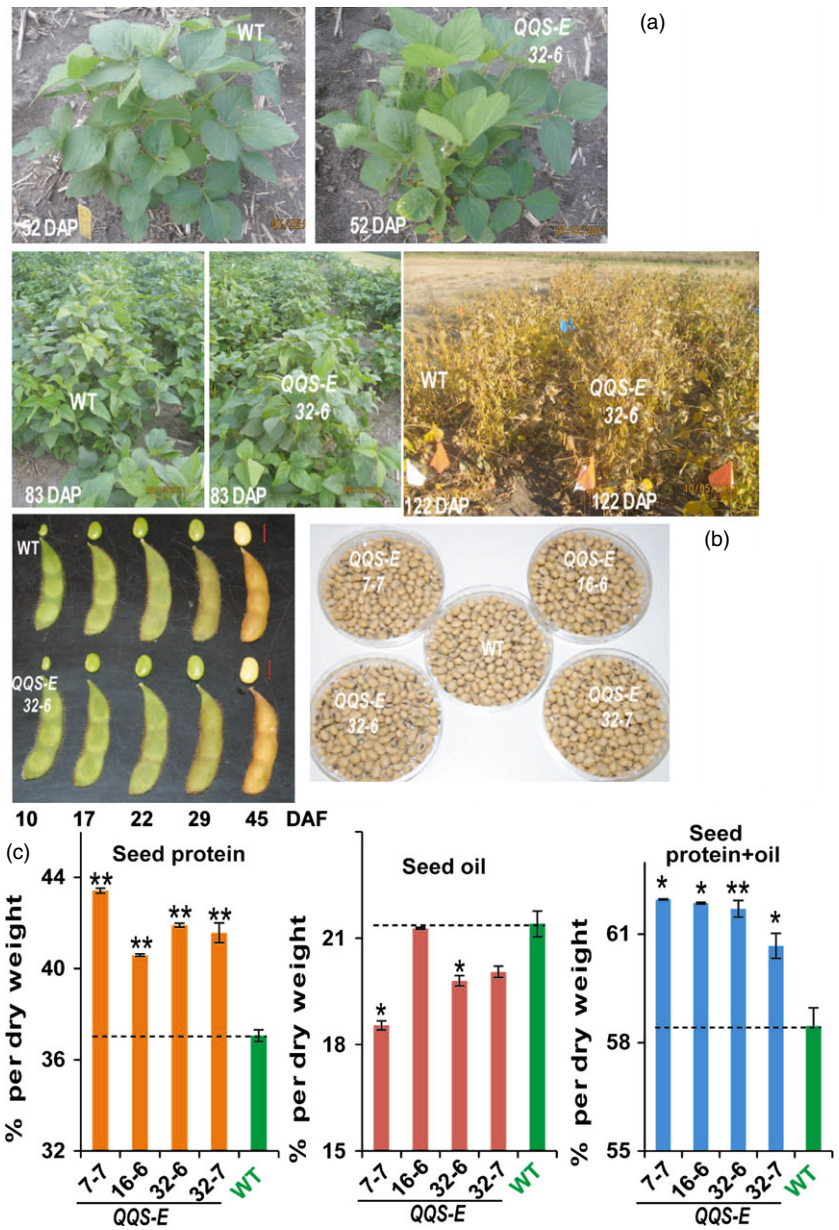

Figure 4 Characterization of seeds of field-grown transgenic soybean plants expressing Qua-Quine Starch (*QQS*). *QQS-E* plants derived from three independent transformation events (a total of four lines) and segregating wild type (WT)-sibling control plants were grown in a randomized block design in the field. Composition was determined in mature T4 seeds by chemical methods. Figure S8c shows near infrared spectroscopy analysis of different seed batches. (a) Transgenic *QQS-E* soybean plants are similar in morphology from WT-sibling plants throughout development and in visual phenotypes (Figure S8a). (b) Seed development and seed size are similar in *QQS-E* compared to WT-siblings. (c) Seed composition in *QQS-E* compared with that of the WT-sibling controls: seed protein content is increased, oil content is similar or decreased, and protein + oil content is increased. DAP, days after planting; DAF, days after flowering. WT, Williams 82 siblings identified by PCR from segregating populations of T3 plants (from transformation events 7 and 32); *QQS-E*, transgenic Williams 82 expressing the *QQS* coding sequence, selected by herbicide resistance. *QQS-E* mutants are designated by independent transformation event #–line #. All data in bar charts show mean ± SE, $n = 3$. Student's *t*-test was used to compare *QQS-E* and WT, $*P < 0.05$; $**P < 0.01$. Scale bar, 1 cm.

(other than *QQS*), most have been identified in mutant screens for genes that alter resistance to abiotic or biotic stresses (Gollery *et al.*, 2006, 2007; Luhua *et al.*, 2013). In the most comprehensive analysis of plant orphan genes to date, Mittler and colleagues (Luhua *et al.*, 2013) evaluated the responses to abiotic stresses of knockout mutants of 1007 Arabidopsis genes randomly selected from those annotated as of 'unknown function' (TAIR, 2005 gene model release). Of these genes annotated as of 'unknown function', 12 were also orphan genes. Knockout mutants of nine of these 12 orphan genes conferred an altered response to one or more abiotic stresses (Luhua *et al.*, 2013).

Among the 839 clade-specific genes that are common to *A. thaliana* and *A. lyrata* are a small cluster of genes that, although divergent across the two species, are still recognizable by sequence; these genes have been shown to play a role in self-recognition (Takeuchi and Higashiyama, 2012). The genes encode

cysteine-rich peptides (CRP810_1; AtLUREs) that are among the 300 defensin-like (*DEFL*) genes in Arabidopsis and function in self-recognition associated with pollen attraction (Takeuchi and Higashiyama, 2012). AtLURE1 (AT5G43285) from *A. thaliana,* when introduced into *Torenia fournieri*, enables *A. thaliana* pollen to be attracted to and penetrate *T. fournieri* ovules (Takeuchi and Higashiyama, 2012). Thus, with the possible exception of *QQS*, the few plant orphan or near-orphan (clade-specific) genes with functional information appear to play a role in recognition or defence-related processes (Gollery *et al.*, 2007; Kim *et al.*, 2009; Li *et al.*, 2009).

QQS and carbon and nitrogen allocation

QQS plays a role, direct or indirect, in regulating carbon and nitrogen allocation to starch (Li *et al.*, 2009) and protein, a process that would be expected to have considerable commonality among plant species. Although carbon and nitrogen allocation might be considered as very distinct from species-specific recognition, immune response, or defence, it actually is closely intertwined. Plants are 'planted' in one place, and therefore must respond with tremendous sensitivity to environmental cues. Global approaches to understand the processes of photosynthesis, nutrient supply and carbon and nitrogen allocation are revealing the intricate relationship among what some might consider distinct processes (Stitt *et al.*, 2010; Sulpice *et al.*, 2013; Thum *et al.*, 2008).

Among the most pronounced of the extremely varied pattern of *QQS* expression across changes in genotypes, environments and developments (Li *et al.*, 2009), *QQS* expression level is strongly changed in a variety of knockout mutants. For example, *QQS* expression is significantly higher (compared with WT controls) in mutants of genes as diverse as *PEN3,* a putative ATP binding cassette transporter that contributes to pathogen resistance (Stein *et al.*, 2006); starch synthase 3 (*SS3*) (Li *et al.*, 2009); *WIN1*, involved in regulating cuticular wax deposition (Kannangara *et al.*, 2007); and the brassinosteroid-induced *FER*, which functions in pollen tube-ovule interaction (Guo *et al.*, 2009). Plants expressing the *NahG* transgene, a bacterial gene that hydroxylates salicylic acid (SA) and reduces SA-mediated signalling (Takahashi *et al.*, 2004), have increased *QQS* expression (ArrayExpress experiment ID 'E-GEOD-5727', data submitted by Buchanan-Wollaston). SA plays a role in plant defense against pathogens (Lin *et al.*, 2013). Consistent with our direct demonstration that *QQS* impacts carbon and nitrogen allocation, *QQS* has been implicated in the ability of Arabidopsis to adjust to reduced carbon and energy environments, based on its altered expression in knockout lines of the EXORDIUM-LIKE1 (*EXL1*) gene (At1g35140) (Schroder *et al.*, 2011). In addition, the inverse relationship between *QQS* expression and starch accumulation across multiple environments fits our model that plants adjust *QQS* expression in response to stresses. Overall, the extremely variable expression pattern of *QQS*, combined with its lack of co-expression with other genes and association with compositional changes under changing environments (Li *et al.*, 2009), is consistent with the concept that the *QQS*-induced compositional changes in *A. thaliana* may aid in the metabolic adaptation of that species to its environment.

Ectopic function of *QQS*

Our demonstration that the introduction of the *QQS* gene into soybean results in a significant increase in protein accumulation and decrease in lipid accumulation in seeds indicate the potential of *QQS* as a molecular tool to increase the protein content of agronomic species. Comparison of composition in plants grown in growth chamber, greenhouse and field materials indicates that the general trend of high-protein content in *QQS-E* soybean holds for plants grown under these very different environments. Protein (and oil) content typically is extremely responsive to both genotype and environment (Arslanoglu *et al.*, 2011; Jing *et al.*, 2003; Singh *et al.*, 1993). Fertilizers containing nitrogen were applied to growth chamber- and greenhouse-grown soybeans, while no fertilizer was applied to field-grown soybeans. Therefore, the differences in overall composition (e.g. seed protein and oil) between the growth chamber-, greenhouse- and field-grown material, as well as the difference in composition between the *QQS-E* and control lines could be due to either genotype or environment, or a genotype and environment effect.

QQS expression in soybean causes greater increases in protein in soybean than in Arabidopsis. One possible explanation for the larger effect of *QQS* on soybean is that over time Arabidopsis has evolved mechanisms for homeostatic balances for the *QQS* gene; however, these mechanisms are not present in soybean. An alternate explanation is that the signalling mechanisms in soybean respond with different sensitivity than those in Arabidopsis.

Interestingly, leaves of *QQS-E* soybean lines with a ratio as low as 1.9 QQS RNA/18S rRNA by real-time PCR display the high-protein, low-starch trait. However, *QQS-E* 7-1 did not have high leaf protein, despite having significant *QQS* expression. Possible explanations are that Event 7 had multiple transgene insertion sites, and these multiple transgenes may have been retained in line 7-1, but not in 7-7. It may be that the insertion site of one of these transgenes interfered with some metabolic process, or that the expression of *QQS* was not stable in *QQS-E* 7-1. Indeed, there does not appear statistically significant linear regression of leaf starch and protein contents with leaf *QQS* expression (as determined by transcript level) in either Arabidopsis (R^2 = 0.83 for starch and 0.64 for protein) or soybean (R^2 = 0.78 for starch and 0.49 for protein); these calculations are complicated by the fact that the relationship between the composition traits and *QQS* transcript level is not necessarily linear, but complex, and have not yet been fully defined. This is perhaps not surprising given the very low level of QQS protein that accumulates in Arabidopsis even in mutant lines that highly express *QQS* (Li *et al.*, 2009). It is not unusual for regulatory proteins to have very low levels of expression (Nagaraj *et al.*, 2011). Thus, it may be that only a very small concentration of QQS saturates the QQS receptor, and any increases over this concentration do not affect protein and starch contents. Other possible explanations for a lack of strong linear correlation between levels of *QQS* transcript and composition are that QQS translational efficiency or stability or the effectiveness of the QQS protein to biochemically express its function or post-translational modification is limiting; also, a variety of post-translational regulatory mechanisms can come into play, as have been described for other transgenes (Lillo *et al.*, 2004; Vaucheret *et al.*, 2001). These considerations present interesting questions about the mechanism by which *QQS* acts in soybean; our current working hypothesis is that soybean and Arabidopsis have a common protein with which QQS interacts, and that QQS-interactor becomes saturated at low levels of *QQS* expression.

The experiments described do not distinguish whether the high-protein trait in the seeds is a maternal effect or a seed effect. The 35S promoter, which we are using in these studies, drives GUS expression in both leaf and in seed (Anderson and Botella, 2007; Wu *et al.*, 2010), which is consistent with either a

seed or a maternal effect. Thus, it is possible that the *QQS* expression in the seed causes the high-protein trait (a seed effect). Alternately, a signalling molecule or a larger flux of organic carbon and nitrogen from the leaves might drive the high-protein trait in the soybean seeds- this would represent a maternal effect of *QQS*.

An increase in soybean protein of the magnitude reported in this study has societal relevance, as soybean provides a major source of global dietary vegetable protein (Pathan and Sleper, 2008; Wilson, 2008). Over 70 years of soybean breeding efforts have not been able to break the inverse relationship between seed protein content and oil content, or the inverse relationship between seed protein content and yield. However, the transgenic expression of *QQS* increases seed protein content in soybean grown under three diverse conditions (growth chamber, greenhouse and field) without detectably affecting plant or seed morphology or seed weight (these factors were determined under all three conditions). In seeds of field-grown plants, for which we made more detailed determinations, there were no significant differences in seed yield per plant, the relative composition of amino acids in the hydrolysed protein from seeds, the moisture content or the yield. We analysed free amino acids to determine whether the increase in protein we observed was associated with a very substantial increase in any free amino acid; this does not seem to be the case. Some free amino acid levels were altered in *QQS-E* soybean seeds. Glutamic acid and arginine are reported to help to avoid protein aggregating and precipitating (Golovanov *et al.*, 2004); it is possible that these shifts in free glutamic acid and arginine might help to adapt to the increased protein content. Lysine and arginine are among the essential amino acid group according to their importance to nutrition and physiology values (Belitz *et al.*, 2009). Thus increased free lysine and arginine in *QQS-E* seeds could potentially provide an increased value; however, the absolute level of free lysine and arginine is very low.

Taken together, our findings suggest that *QQS* could be introduced by breeding or transformation into an elite soybean variety with specific desirable agronomic traits to increase protein content. For example, *QQS* could be introduced into an Iowa soybean variety that is resistant to the soybean cyst nematode but relatively low in protein (http://www.cad.iastate.edu/gensoyrel.html), or in African soybean varieties highly resistant to rust, bacterial blight and leaf spot (http://www.iita.org/soybean-asset/-/asset_publisher/t3fl/content/better-soybean-varieties-offer-african-farmers-new-opportunities?redirect=%2Fsoybean#.U8ib1fldWSp).

Conclusions

Our data demonstrate that *QQS* expression alters plant composition. We show that expressing this gene in a plant species that has no *QQS* sequence homolog increases protein content of leaves and seeds, yet the morphology and development of the soybean expressing *QQS* cannot be distinguished from the WT sibling controls. Thus, the *QQS*-expressing mutant appears to preserve overall homeostasis while selectively effecting composition. *QQS* might be acting upstream of the process that controls carbon partitioning, or might be central to this process. These results also illustrate that orphan genes, although often poorly annotated and even ignored, may provide a valuable resource for new traits.

The evolutionary changes that resulted in the *de novo* origin (Silveira *et al.*, 2013) of the *QQS* gene of *A. thaliana* must have been rapid and extensive as there is no gene homolog in even the closely related species, *A. lyrata*. Yet, soybean, a species that diverged from Arabidopsis approximately 100 million years ago (Hedges and Kumar, 2009), appears to contain a conserved receptor or mechanism that recognizes *QQS* and responds to its occurrence by conferring a compositional phenotype. This research reveals the fundamental capacity of a species-specific gene to act across species to impact the major metabolic function of carbon and nitrogen allocation.

Experimental procedures

Construction of *QQS* over-expression vector and Arabidopsis transformation and selection are provided in Appendix S1.

Soybean transformation, selection and nomenclature

Glycine max cultivar Williams 82 was transformed by Agrobacterium-mediated soybean transformation using half-seed explants (Paz *et al.*, 2004). The transformation and selection of T1 plants were performed at the Plant Transformation Facility at Iowa State University (ISU) (http://www.agron.iastate.edu/ptf/index.aspx). Soybean plants from independent transformation events were selected based on herbicide resistance (segregated WTs were killed) and confirmed by real-time PCR analysis to identify transgenic lines that expressed the *QQS* transgene.

The progeny of each independent soybean transformation is referred to as an 'event' (a transformation event is considered independent if it is taken from an individual plate). Each plant germinated from one T1 seed is called a 'line' and the line designation continues throughout generations. So multiple lines stem from one independent transformation event; because each line is the result of sexual reproduction; these lines may not be genetically identical. A total of 33 lines from five independent transformation events were confirmed on the basis of BAR selection followed by PCR analysis for presence of the *QQS* gene.

Plant growth

This study used WT *A. thaliana* ecotype Columbia (Col-0), and transgenic lines derived from Col-0. Detailed information is provided in Appendix S1.

Detailed information about transgenic *QQS*-expressing (*QQS-E*) soybean grown in growth chambers (T1 generation) and in a greenhouse (T2 generation, selected on herbicide resistance and segregated WTs were killed), is provided in Appendix S1.

Transgenic *QQS-E* soybean (T3 generation) and WT (Williams 82) plants were planted at a randomized block design in the field at ISU Curtiss Farm in Ames, IA. One line (60 seeds) was planted in one row, with a total of three replicates in three rows. Each row was ten feet long and the rows were 2.5 feet apart. The criteria used for selecting the events-lines to study further in the field were: having sufficient seeds for NIRS analysis and planting in the field; having an increased leaf protein/decreased leaf starch trait; and the presence of the *QQS* transgene (as determined by PCR). The field conditions were harsh prior to germination, and there was considerable flooding; some seeds in flooded area were eaten by ground squirrels or other animals, and a number of lines were lost (including *QQS-E* 7-1). Eight mutant lines that survived germination were sprayed by herbicide, and numbers of herbicide-resistant and herbicide-sensitive progeny were counted. For yield trials, only seeds harvested from plants from the middle seven feet were used for yield estimate (grams of seed weight per plant). The seeds from different plants of the same line were pooled and used as different replicates for seed

composition analysis by NIRS (all eight lines) and by chemical methods (four lines). Some plants were randomly marked for genotype screening by PCR to identify transgenic plants and their WT-siblings. Seeds from these plants were harvested per individual plant and were not pooled with seeds from other plants.

QQS determination by PCR is provided in Appendix S1.

Leaf composition analyses

For screening starch using I_2/KI staining in Arabidopsis, shoots were harvested at the end of the light period, and processed as described before (Li *et al.*, 2007). Detailed information is provided in Appendix S1.

Leaf starch and protein were determined at the end of the light period in Arabidopsis seedling shoots of 20 days after planting (DAP) grown in a growth chamber, and in soybean leaves that were newly, fully expanded, harvested from branches two to four from shoot apex from 58-DAP T1 plants (starch: four independent transformation events, protein: five). Three (for starch) or five (for protein) plants per replicate and three replicates from each independent T2 lines (Arabidopsis), and leaves from three positions that were used as three replicates in the T1 plants (soybean), were analysed. Detailed information about leaf starch and protein determination is provided in Appendix S1.

Soybean seed composition analyses

Information about NMR and NIRS screening is provided in Appendix S1.

Destructive chemical analyses were mostly conducted at Eurofins (Des Moines, IA). Methods were: protein content, AOCS Ba 4e-93 (American Oil Chemists' Society, 1997); oil content, AOCS Ac 3-44 (American Oil Chemists' Society, 1997); moisture content, AOCS Ac 2-41 (American Oil Chemists' Society, 1997); hydrolysed fatty acid profiling, AOCS Ce 2-66 and AOCS Ce 1-62 (American Oil Chemists' Society, 1991); free amino acid profiling, AOAC 999.13 modified (Fontaine *et al.*, 2000). About 30 g of seeds (approximately 170 seeds from nine plants, for protein content), 10 g (for oil content), 10 g (for moisture content), 10 g (for hydrolysed fatty acid profiling) and 20 g (for free amino acid profiling) per replicate were tested, respectively, with three biological replicates for each sample.

Hydrolysed amino acid composition analysis was conducted at the Experiment Station Chemical Laboratories, University of Missouri (http://www.aescl.missouri.edu/), using method AOAC 982.30 E (a,b,c) Ch. 45.3.05 (Association of Official Analytical Chemists (AOAC), 2006). About 30 g of seeds per replicate were tested, with three biological replicates for each sample.

Statistical analyses

For each experiment, plants were collected and analysed in a randomized complete block design or completely randomized design. Plant composition tests were conducted with a minimum of three biological tests. For all composition analyses, plant samples were assigned randomized numbers and provided to the analysis facilities for determination in a randomized order with no designator of genotype.

Data are presented as mean ± SE. Two sets of independent samples were compared using Student's *t*-test (two-tailed) with assumption of equal variances ($n = 3$). $P < 0.05$ was considered significant (*); $P < 0.01$ was considered very significant (**).

Bioinformatics analyses is provided in Appendix S1.

Acknowledgements

We are grateful to Walter Fehr for advice on soybean breeding, Zebulun Arendsee and Ruoran Li for orphan gene analyses, and Basil Nikolau and Jianming Yu for helpful suggestions on the manuscript. We thank Kan Wang and Diane Luth from the ISU Plant Transformation Facility for generating transgenic lines of soybean; Jianling Peng for the *QQS*-OE construct; Taner Sen and Vladimir Uversky for information on prediction of QQS protein structure and disorder; Kent Berns for field management; Wenguang Zheng, Sheng Huang, Marah Hoel, Xiaoran Shang, Alan Kading, Le Song, Hiwot Abebe, Ana Boehm and Sean Wefel for assistance with soybean growth and harvest; Charles Hurburgh and Glen Rippke from ISU Grain Quality Laboratory, for NIRS analysis; Dan Duvick for seed free fatty acid determinations; Jian Li for statistical consultation; Waitent Sow, Kevin Wenceslao, Dallas Jones, and ISU W. M. Keck Plant Metabolomics Laboratory for contributions to composition analysis; Eurofins (Des Moines, IA) for analysis of seed composition; ISU Soil and Plant Analysis Laboratory for leaf combustion analysis; Experiment Station Chemical Laboratories at University of Missouri for amino acid profiling; Pennington Caroline from UEA Consulting Limited for PCR analyses of *QQS* transcript level; and Sabry Elias, Oregon State University Seed Lab for NMR Spectrometry screening. This material is based in part upon work supported by National Science Foundation EEC-0813570 (to E.S.W.) and MCB-0951170 (to E.S.W. and L.L.); United Soybean Board award 2287 (to L.L.); ISU Research Foundation (to L.L.); and Center for Metabolic Biology (to E.S.W.). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Author contributions

L.L. and E.S.W. conceived the project. L.L. performed the experiments. L.L and E.S.W. co-supervised the project and contributed equally to data analysis and preparation of the paper.

References

Adler, R.G. (1992) Genome research: fulfilling the public's expectations for knowledge and commercialization. *Science*, **257**, 908–914.

Aloni, B., Karni, L., Zaidman, Z. and Schaffer, A.A. (1997) The relationship between sucrose supply, sucrose-cleaving enzymes and flower abortion in pepper. *Ann. Bot.* **79**, 601–605.

American Oil Chemists' Society (AOCS). (1991) *Official and Tentative Method of the American Oil Chemists' Society*, 4th edn. Champaign, IL: AOCS Press.

American Oil Chemists' Society (AOCS). (1997) *Official Methods and Recommended Practices of the American Oil Chemists' Society*, 5th edn. Champaign, IL: AOCS Press.

Amiri, H., Davids, W. and Andersson, S.G. (2003) Birth and death of orphan genes in Rickettsia. *Mol. Biol. Evol.* **20**, 1575–1587.

Anderson, D.J. and Botella, J.R. (2007) Expression analysis and subcellular localization of the *Arabidopsis thaliana* G-protein β-subunit AGB1. *Plant Cell Rep.* **26**, 1469–1480.

Arendsee, Z., Li, L. and Wurtele, E.S. (2014) Coming of age: the species-specific (orphan) genes of plants. *Trends Plant Sci.* In press.

Arslanoglu, F., Aytac, S. and Oner, E.K. (2011) Effect of genotype and environment interaction on oil and protein content of soybean (*Glycine max* (L.) Merrill) seed. *Afr. J. Biotechnol.* **10**, 18409–18417.

Association of Official Analytical Chemists (AOAC) (2006) *Official Methods of Analysis of the Association of Official Analytical Chemists*, Gaithersburg, MD: AOAC International.

Belitz, H.D., Grosch, W. and Schieberle, P. (2009) Amino acids, peptides, protein. In *Food Chemistry*, 4th revised and extended edn (Belitz, H.D., Grosch, W. and Schieberle, P., eds), pp. 8–34. Berlin: Springer.

Biehl, A., Richly, E., Noutsos, C., Salamini, F. and Leister, D. (2005) Analysis of 101 nuclear transcriptomes reveals 23 distinct regulons and their relationship to metabolism, chromosomal gene distribution and co-ordination of nuclear and plastid gene expression. *Gene*, **344**, 33–41.

Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B., Brar, G.A., Weissman, J.S., Regev, A., Thierry-Mieg, N., Cusick, M.E. and Vidal, M. (2012) Proto-genes and de novo gene birth. *Nature*, **487**, 370–374.

Donoghue, M.T., Keshavaiah, C., Swamidatta, S.H. and Spillane, C. (2011) Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol. Biol.* **11**, 47.

Eastmond, P.J. (2006) SUGAR-DEPENDENT1 encodes a patatin domain triacylglycerol lipase that initiates storage oil breakdown in germinating Arabidopsis seeds. *Plant Cell*, **18**, 665–675.

Eastmond, P.J., Dennis, D.T. and Rawsthorne, S. (1997) Evidence that a malate/inorganic phosphate exchange translocator imports carbon across the leucoplast envelope for fatty acid synthesis in developing castor seed endosperm. *Plant Physiol.* **114**, 851–856.

Espinoza, C., Degenkolbe, T., Caldana, C., Zuther, E., Leisse, A., Willmitzer, L., Hincha, D.K. and Hannah, M.A. (2010) Interaction with diurnal and circadian regulation results in dynamic metabolic and transcriptional changes during cold acclimation in Arabidopsis. *PLoS ONE*, **5**, e14101.

Feng, Y., Hurst, J., Almeida-De-Macedo, M., Chen, X., Li, L., Ransom, N. and Wurtele, E.S. (2012) Massive human co-expression network and its medical applications. *Chem. Biodivers.* **9**, 868–887.

Fischer, D. and Eisenberg, D. (1999) Finding families for genomic ORFans. *Bioinformatics*, **15**, 759–762.

Fontaine, J., Eudaimon, M., Fontaine, J. and Eudaimon, M. (2000) Liquid chromatographic determination of lysine, methionine, and threonine in pure amino acids (feed grade) and premixes: collaborative study. *J. AOAC Int.* **83**, 771–783.

Godfray, H.C., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M. and Toulmin, C. (2010) Food security: the challenge of feeding 9 billion people. *Science*, **327**, 812–818.

Gollery, M., Harper, J., Cushman, J., Mittler, T., Girke, T., Zhu, J.K., Bailey-Serres, J. and Mittler, R. (2006) What makes species unique? The contribution of proteins with obscure features. *Genome Biol.* **7**, R57.

Gollery, M., Harper, J., Cushman, J., Mittler, T. and Mittler, R. (2007) POFs: what we don't know can hurt us. *Trends Plant Sci.* **12**, 492–496.

Golovanov, A.P., Hautbergue, G.M., Wilson, S.A. and Lian, L.Y. (2004) A simple method for improving protein solubility and long-term stability. *J. Am. Chem. Soc.* **126**, 8933–8939.

Gomes, S.P., Nyengaard, J.R., Misawa, R., Girotti, P.A., Castelucci, P., Blazquez, F.H., de Melo, M.P. and Ribeiro, A.A. (2009) Atrophy and neuron loss: effects of a protein-deficient diet on sympathetic neurons. *J. Neurosci. Res.* **87**, 3568–3575.

Guo, H., Li, L., Ye, H., Yu, X., Algreen, A. and Yin, Y. (2009) Three related receptor-like kinases are required for optimal cell elongation in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **106**, 7648–7653.

Hedges, S.B. and Kumar, S. (2009) *The Timetree of Life*. New York, NY: Oxford University Press.

Heitschmidt, R.K., Short, R.E. and Grings, E.E. (1996) Ecosystems, sustainability, and animal agriculture. *J. Anim. Sci.* **74**, 1395–1405.

Higashi, Y., Hirai, M.Y., Fujiwara, T., Naito, S., Noji, M. and Saito, K. (2006) Proteomic and transcriptomic analysis of Arabidopsis seeds: molecular evidence for successive processing of seed proteins and its implication in the stress response to sulfur nutrition. *Plant J.* **48**, 557–571.

Jing, Q., Jiang, D., Dai, T. and Cao, W. (2003) Effects of genotype and environment on wheat grain quality and protein components. *Chin. J. Appl. Ecol.* **14**, 1649–1653.

Kannangara, R., Branigan, C., Liu, Y., Penfield, T., Rao, V., Mouille, G., Hofte, H., Pauly, M., Riechmann, J.L. and Broun, P. (2007) The transcription factor WIN1/SHN1 regulates Cutin biosynthesis in *Arabidopsis thaliana*. *Plant Cell*, **19**, 1278–1294.

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. and Bosch, T.C. (2009) More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* **25**, 404–413.

Kim, M.J., Shin, R. and Schachtman, D.P. (2009) A nuclear factor regulates abscisic acid responses in Arabidopsis. *Plant Physiol.* **151**, 1433–1445.

Knowles, D.G. and McLysaght, A. (2009) Recent de novo origin of human protein-coding genes. *Genome Res.* **19**, 1752–1759.

Li, L., Ilarslan, H., James, M.G., Myers, A.M. and Wurtele, E.S. (2007) Genome wide co-expression among the starch debranching enzyme genes AtISA1, AtISA2, and AtISA3 in *Arabidopsis thaliana*. *J. Exp. Bot.* **58**, 3323–3342.

Li, L., Foster, C.M., Gan, Q., Nettleton, D., James, M.G., Myers, A.M. and Wurtele, E.S. (2009) Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *Plant J.* **58**, 485–498.

Li, Z., Gao, J., Benning, C. and Sharkey, T.D. (2012) Characterization of photosynthesis in Arabidopsis ER-to-plastid lipid trafficking mutants. *Photosynth. Res.* **112**, 49–61.

Lillo, C., Meyer, C., Lea, U.S., Provan, F. and Oltedal, S. (2004) Mechanism and importance of post-translational regulation of nitrate reductase. *J. Exp. Bot.* **55**, 1275–1282.

Lin, J., Mazarei, M., Zhao, N., Zhu, J.J., Zhuang, X., Liu, W., Pantalone, V.R., Arelli, P.R., Stewart Jr, C.N. and Chen, F. (2013) Overexpression of a soybean salicyclic acid methlyltransferase confers resistance to soybean cyst nematode. *Plant Biotechnol. J.* **11**, 1135–1145.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.

Luhua, S., Hegie, A., Suzuki, N., Shulaev, E., Luo, X., Cenariu, D., Ma, V., Kao, S., Lim, J., Gunay, M.B., Oosumi, T., Lee, S.C., Harper, J., Cushman, J., Gollery, M., Girke, T., Bailey-Serres, J., Stevenson, R.A., Zhu, J.K. and Mittler, R. (2013) Linking genes of unknown function with abiotic stress responses by high-throughput phenotype screening. *Physiol. Plant.* **148**, 322–333.

Marsden, R.L., Lee, D., Maibaum, M., Yeats, C. and Orengo, C.A. (2006) Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. *Nucleic Acids Res.* **34**, 1066–1080.

Mentzen, W.I. and Wurtele, E.S. (2008) Regulon organization of Arabidopsis. *BMC Plant Biol.* **8**, 99.

Mentzen, W.I., Peng, J., Ransom, N., Nikolau, B.J. and Wurtele, E.S. (2008) Articulation of three core metabolic processes in Arabidopsis: fatty acid biosynthesis, leucine catabolism and starch metabolism. *BMC Plant Biol.* **8**, 76.

Mishra, A.K., Agarwal, S., Jain, C.K. and Rani, V. (2009) High GC content: critical parameter for predicting stress regulated miRNAs in *Arabidopsis thaliana*. *Bioinformation*, **4**, 151–154.

Muller, O. and Krawinkel, M. (2005) Malnutrition and health in developing countries. *CMAJ*, **173**, 279–286.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S. and Mann, M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7**, 548.

Neme, R. and Tautz, D. (2013) Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, **14**, 117.

Ohno, S. (1987) Early genes that were oligomeric repeats generated a number of divergent domains on their own. *Proc. Natl Acad. Sci. USA*, **84**, 6486–6490.

Pathan, M.S. and Sleper, D.A. (2008) Advances in soybean breeding. In *Genetics and Genomics of Soybean* (Stacey, G., ed.), pp. 113–134. New York: Springer.

Paz, M.M., Shou, H., Guo, Z., Zhang, Z., Banerjee, A.K. and Wang, K. (2004) Assessment of conditions affecting Agrobacterium-mediated soybean transformation using the cotyledonary node explant. *Euphytica*, **136**, 167–179.

Pimentel, D. and Pimentel, M. (2003) Sustainability of meat-based and plant-based diets and the environment. *Am. J. Clin. Nutr.* **78**, 660S–663S.

Schiltz, S., Gallardo, K., Huart, M., Negroni, L., Sommerer, N. and Burstin, J. (2004) Proteome reference maps of vegetative tissues in pea. An investigation of nitrogen mobilization from leaves during seed filling. *Plant Physiol.* **135**, 2241–2260.

Schroder, F., Lisso, J. and Mussig, C. (2011) EXORDIUM-LIKE1 promotes growth during low carbon availability in Arabidopsis. *Plant Physiol.* **156**, 1620–1630.
Seo, P.J., Kim, M.J., Ryu, J.Y., Jeong, E.Y. and Park, C.M. (2011) Two splice variants of the IDD14 transcription factor competitively form nonfunctional heterodimers which may regulate starch metabolism. *Nat. Commun.* **2**, 303.
Shou, H., Frame, B.R., Whitham, S.A. and Wang, K. (2004) Assessment of transgenic maize events produced by particle bombardment or Agrobacterium-mediated transformation. *Mol. Breed.* **13**, 201–208.
Silveira, A.B., Trontin, C., Cortijo, S., Barau, J., Del Bem, L.E., Loudet, O., Colot, V. and Vincentz, M. (2013) Extensive natural epigenetic variation at a de novo originated gene. *PLoS Genet.* **9**, e1003437.
Singh, K.B., Bejiga, G. and Malhotra, R.S. (1993) Genotype–environment interactions for protein content in chickpea. *J. Sci. Food Agric.* **63**, 87–90.
Stein, M., Dittgen, J., Sanchez-Rodriguez, C., Hou, B.H., Molina, A., Schulze-Lefert, P., Lipka, V. and Somerville, S. (2006) Arabidopsis PEN3/PDR8, an ATP binding cassette transporter, contributes to nonhost resistance to inappropriate pathogens that enter by direct penetration. *Plant Cell*, **18**, 731–746.
Stitt, M., Lunn, J. and Usadel, B. (2010) Arabidopsis and primary photosynthetic metabolism—more than the icing on the cake. *Plant J.* **61**, 1067–1091.
Sulpice, R., Flis, A., Ivakov, A.A., Apelt, F., Krohn, N., Encke, B., Abel, C., Feil, R., Lunn, J.E. and Stitt, M. (2013) Arabidopsis coordinates the diurnal regulation of carbon allocation and growth across a wide range of photoperiods. *Mol. Plant*, **7**, 137–155.
Takahashi, H., Kanayama, Y., Zheng, M.S., Kusano, T., Hase, S., Ikegami, M. and Shah, J. (2004) Antagonistic interactions between the sa and ja signaling pathways in Arabidopsis modulate expression of defense genes and gene-for-gene resistance to cucumber mosaic virus. *Plant Cell Physiol.* **45**, 803–809.
Takeuchi, H. and Higashiyama, T. (2012) A species-specific cluster of defensin-like genes encodes diffusible pollen tube attractants in Arabidopsis. *PLoS Biol.* **10**, e1001449.
Tautz, D. and Domazet-Loso, T. (2011) The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702.
Thum, K.E., Shin, M.J., Gutierrez, R.A., Mukherjee, I., Katari, M.S., Nero, D., Shasha, D. and Coruzzi, G.M. (2008) An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis. *BMC Syst. Biol.* **2**, 31.
Vaucheret, H., Béclin, C. and Fagard, M. (2001) Post-transcriptional gene silencing in plants. *J. Cell Sci.* **114**, 3083–3091.
Victora, C.G., de Onis, M., Hallal, P.C., Blossner, M. and Shrimpton, R. (2010) Worldwide timing of growth faltering: revisiting implications for interventions. *Pediatrics*, **125**, e473–e480.
Wilson, R.F. (2008) Soybean: market driven research needs. In *Genetics and Genomics of Soybean* (Stacey, G., ed.), pp. 3–16. New York: Springer.
Wilson, G.A., Feil, E.J., Lilley, A.K. and Field, D. (2007) Large-scale comparative genomic ranking of taxonomically restricted genes (TRGs) in bacterial and archaeal genomes. *PLoS ONE*, **2**, e324.
Wissler, L., Gadau, J., Simola, D.F., Helmkampf, M. and Bornberg-Bauer, E. (2013) Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome Biol. Evol.* **5**, 439–455.
Wu, L., EL-Mezawy, A., Duong, M. and Shah, S. (2010) Two seed coat-specific promoters are functionally conserved between *Arabidopsis thaliana* and *Brassica napus*. *In Vitro Cell Dev. Biol. Plant*, **46**, 338–347.
Wu, D.D., Irwin, D.M. and Zhang, Y.P. (2011) De novo origin of human protein-coding genes. *PLoS Genet.* **7**, e1002379.
Young, V.R. and Pellett, P.L. (1994) Plant proteins in relation to human protein and amino acid nutrition. *Am. J. Clin. Nutr.* **59**, 1203S–1212S.
Zhang, X., Myers, A.M. and James, M.G. (2005) Mutations affecting starch synthase III in Arabidopsis alter leaf starch structure and increase the rate of starch synthesis. *Plant Physiol.* **138**, 663–674.
Zhang, X., Szydlowski, N., Delvalle, D., D'Hulst, C., James, M.G. and Myers, A.M. (2008) Overlapping functions of the starch synthases SSII and SSIII in amylopectin biosynthesis in Arabidopsis. *BMC Plant Biol.* **8**, 96.

Supporting information

Additional Supporting information may be found in the online version of this article:

Figure S1 *QQS-OE* construct used to transform *Arabidopsis thaliana* and soybean (*Glycine max*).
Figure S2 Starch quantification shows increased starch in *QQS* RNAi mutants and decreased starch in *QQS-OE* mutants compared to WT.
Figure S3 Mapping of *Arabidopsis thaliana* orphan genes and transposons on chromosomes.
Figure S4 Spatial expression of *QQS* and starch accumulation with 5% sucrose and in the cold.
Figure S5 The plots of leaf starch and protein versus *QQS* transcript in Arabidopsis and in soybean.
Figure S6 Composition screening of seeds from transgenic lines of growth chamber-grown transgenic soybean plants expressing *QQS*.
Figure S7 Phenotype and seed composition of transgenic lines of greenhouse-grown soybean plants expressing *QQS*.
Figure S8 Phenotype and seed composition of transgenic lines of field-grown soybean plants expressing *QQS*.
Table S1 *QQS* transcript accumulation (by real-time PCR) in leaves of *QQS*-expressing transgenic lines and WT lines.
Table S2 Chi-square analysis assuming a 3:1 (Resistant:Sensitive) ratio for T3 generation of progeny of *QQS-E* containing 35S:QQS and pNos:Bar.
Table S3 Seed composition of transgenic lines of field-grown soybean plants expressing *QQS*.
Table S4 Profiles of most abundant free amino acids from seeds of *QQS*-expressing transgenic lines, compared with WTs.
Table S5 Profiles of amino acids from hydrolysed proteins: almost all amino acids analysed are increased in seeds of *QQS*-expressing transgenic lines, compared with WTs.
Table S6 Hydrolysed fatty acid profiles in the seeds of *QQS*-expressing transgenic lines, compared with WTs.
Table S7 Seed composition of individual plants of transgenic lines and their segregated siblings of field-grown soybean plants expressing *QQS*.
Appendix S1 Supplementary experimental procedures.

