ENDOGENOUS SHORT INTERFERING RNAS THAT SILENCE CHALCONE SYNTHASE
RESULT IN PIGMENT PATTERN FORMATION ON SOYBEAN SEED COATS

BY

YOUNG B. CHO

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Crop Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Master’s Committee:
Professor Lila Vodkin, Chair
Professor Brian Diers
Associate Professor Stephen Moose
ABSTRACT

In soybean, dominant alleles of the I locus inhibit pigmentation of the seed coat while the homozygous recessive i allele results in fully pigmented seed. Previous work showed that dominant alleles correspond to duplications of chalcone synthase (CHS) genes leading to production of CHS siRNAs which in turn degrade CHS mRNAs resulting in yellow seed coats (Tuteja et al., 2009 and Vodkin et al., 2012). However, it was not known if the same phenomenon applied to the yellow and pigmented regions within the same seed coats that have homozygous i-i or i-k alleles that restrict pigment to the hilum and saddle regions of the seed coat, respectively. Here, I describe the results of Illumina high throughput sequencing of small RNA populations from pigmented and yellow regions within seed coats with the same genotype. The level of CHS siRNAs is much greater in the yellow versus the pigmented region. CHS siRNAs from another genotype (i-i, kI), which also produces a pigmented saddle on the seed coat, are also much more abundant in the yellow versus the pigmented region. Small RNA blots show CHS siRNAs accumulated only in the yellow region and confirm the sequencing data. Thus, these data demonstrate that CHS siRNAs result in pigment pattern formations on soybean seed coats. In addition, small RNA sequencing data from twenty-one samples in six different tissues including cotyledons, roots, stems, and leaves showed that CHS siRNA expression is limited to the yellow seed coats confirming the tissue specificity of the generation of the CHS siRNAs. The differential expression of CHS siRNAs during seed coat development was examined by eight small RNA libraries from eight developmental stages.
# TABLE OF CONTENTS

CHAPTER 1 LITERATURE REVIEW

- DISCOVERIES OF SHORT INTERFERING RNA (siRNA) .................................................. 1
- DISCOVERIES OF SOYBEAN SEED COAT PIGMENTATION ......................................... 2

CHAPTER 2 MATERIALS AND METHODS .......................................................... 5

- TISSUE COLLECTION ........................................................................................................... 5
- SMALL RNA ISOLATION AND SEQUENCING DATA ANALYSIS ........................................ 5
- SMALL RNA BLOTTING ....................................................................................................... 7
- MEssenger RNA ISOLATION AND SEQUENCING DATA ANALYSIS .................................... 7

CHAPTER 3 RESULTS .......................................................................................... 9

- TISSUE SPECIFICITY OF CHS siRNA EXPRESSION ....................................................... 9
- DIFFERENTIAL EXPRESSION OF CHS siRNA DURING SEED COAT DEVELOPMENT ............. 9
- THE ROLE OF CHS siRNA IN PIGMENT PATTERN FORMATION ON SEED COATS WITH THE SAME GENOTYPE ................................................................. 10
- CHS siRNAs RESULT IN THE BLACK HILUM PIGMENT PATTERN FORMATION OF WILLIAMS 43 ........................................... 11
- CHS siRNAs RESULT IN THE SADDLE PATTERN FORMATION OF CLARK 8 .................. 13
- CHS siRNAs RESULT IN THE SADDLE PATTERN FORMATION OF CLARK 18a ................ 14
- SMALL RNA BLOTS CONFIRM THAT CHS siRNAs ACCUMULATE ONLY IN THE NON-PIGMENTED REGIONS ................................................................. 16
- TRANSCRIPTOME DATA SHOWS CHS siRNA AND CHS mRNA EXPRESSION LEVELS ARE INVERSELY CORRELATED ......................................................... 16

CHAPTER 4 DISCUSSION .................................................................................. 18

- TISSUE-SPECIFIC BIOGENESIS OF CHS siRNAs ......................................................... 18
- BIOGENESIS OF CHS siRNAs IS REGULATED BY DEVELOPMENTAL PROGRAM OF SOYBEAN .......................................................... 19
- SPECIFIC CHS siRNAs CHANGE DURING DEVELOPMENT BUT BIOLOGICAL REPEATS SHOW SIMILAR PATTERNS FOR MANY OF THE ABUNDANT CHS siRNAs ................................................................. 20
- CHS siRNAs ARE STRONGLY ASSOCIATED WITH THE YELLOW SEED COAT AND NOT THE PIGMENTED HILUM REGION IN THE DOMINANT WILLIAMS i-i GENOTYPE ......................................................... 20
- CHS siRNAs ARE EXPRESSED AND DEGRADE TARGET CHS mRNAs IN THE NON-PIGMENTED REGION OF THE SADDLE PATTERN PHENOTYPES ......................................................... 21
- UNIDENTIFIED K LOCUS REGULATES THE PATTERN-SPECIFIC CHS siRNA EXPRESSION .......................................................................................................................... 23

FIGURES AND TABLES .................................................................................. 25

LITERATURE CITED .................................................................................. 58
CHAPTER 1 LITERATURE REVIEW

Commercially grown soybean (Glycine max) varieties have non-pigmented yellow seeds to reduce the undesirable effects of the black anthocyanin pigments during processing of soybean products (Palmer et al., 2004). However, some soybean varieties are pigmented and accumulate pigment within the epidermal layer of the seed coat. This naturally occurring mutation of the soybean seed coat is known very well since many discoveries were reported during the last twenty years. Recently, it was discovered that CHS siRNA down-regulates genes at the I locus, leading to pigmentation changes in a tissue-specific manner (Tuteja et al., 2009). Here, I review discoveries about small RNAs and soybean seed coat pigmentation.

DISCOVERIES OF SHORT INTERFERING RNA (siRNA)

In early 1990’s, researchers found a very interesting phenomenon which they called “co-suppression” in flower pigmentation (Van der Krol et al., and Napoli et al., 1990). They introduced extra CHS gene copies into transgenic lines in order to intensify flower pigmentation. Unexpectedly, pigment expression was suppressed, leading to completely white flowers. This was termed co-suppression and later RNA interference. David Baulcombe and his colleagues in the late 1990’s used northern blots to detect the type of RNA which mediates RNA interference (Hamilton and Baulcombe, 1999). Researchers also determined the size of the RNA, which is 21-25 nucleotides (nt). This is much smaller than regular mRNAs, so the type of RNA which is associated with RNA interference has been called small RNA (smRNA). Andrew Fire and Craig Mello discovered that double-stranded RNAs are triggers of RNA interference (Fire et al., 1998). They won Nobel prizes for their discovery in 2006. Since smRNA was detected first in the late 1990’s, much research was conducted in plant and animal systems. Several types of smRNAs were reported. One type is a short interfering RNA (siRNA). Double-stranded RNA is processed to siRNA by several different proteins such as Argonaute (AGO) and Dicer like proteins. Double-stranded RNA can be generated from several different pathways such as endogenous
inverted repeats. Recently, it was discovered that soybean seed coat pigmentation is regulated by siRNAs which are generated from inverted repeats (Tuteja et al., 2009).

DISCOVERIES OF SOYBEAN SEED COAT PIGMENTATION

Soybean seed coat pigmentation is regulated by the I locus (Palmer et al., 2004). The I (Inhibitor) locus has four alleles that determine the pattern of pigment along with the spatial distribution of the pigments. Unlike other plants in which the presence of pigment is usually the dominant phenotype, the absence of pigments in soybean seed coats is dominant because of the I allele, which inhibits pigment accumulation, resulting in a non-pigmented yellow seed coat. Varieties that are homozygous recessive (i) are pigmented and have pigment distributed across the entire seed coat. Varieties with the i-i and i-k alleles have pigment partially restricted to the hilum and saddle region, respectively. Dominance relations between the alleles are in the order: I, i-i, i-k, i.

Both anthocyanins and proanthocyanidins are present in black seed coats with the homozygous recessive i genotype (Todd and Vodkin, 1993). Proanthocyanidins are not precursors of anthocyanins but are synthesized as polymeric condensation products of leucoanthocyanins, the precursors of both anthocyanins and proanthocyanidins (Wang and Vodkin, 1994). The I locus controls the absence or presence of proanthocyanidins. The expression of several key mRNAs were examined which are involved in flavonoid synthesis in near-isogenic lines containing different I alleles (Wang and Vodkin, 1994). It was discovered that lack of CHS is the primary cause for the inhibition of anthocyanins and proanthocyanidins synthesis by the dominant I allele. Polymorphisms are found in the CHS gene family in near-isogenic lines that carry the I allele by Southern blot experiments. CHS is a key enzyme in the flavonoid pathway which produces anthocyanins and proanthocyanidins (Figure 1.1). Thus, the I locus represents an example of a dominant allele that affects expression of a specific gene, CHS, which regulates the flavonoid pathway.

CHS constitutes a nine member gene family in soybean (Clough et al., 2004; Tuteja et al., 2008). Every CHS gene except CHS7/8 has >95% sequence similarity. The relative expression
profiles of CHS gene family members were examined by quantitative RT-PCR in the seed coats of two near-isogenic pairs that result from independently occurring mutations of the dominant I allele to the recessive i allele or of the dominant i-i allele to the recessive i allele. Decreased expression of the CHS7/8 genes results in the lack of pigmentation in the yellow seed coats (I and i-i). The relative expression of CHS gene family members was measured in the seven different tissues of one near-isogenic pair that resulted from independently occurring mutations of the dominant I allele to the recessive i allele by not only quantitative RT-PCR but also mRNA blots. The conclusions from these data were that the CHS genes were expressed in the pigmented seed coat in a tissue-specific manner.

The I locus was reported first as a region of duplicated and inverted CHS genes (CHS1, CHS3 and CHS4) (Todd and Vodkin, 1996) by fine PCR mapping analyzing near-isogenic pairs. The deletions of CHS genes that are related to mutations of the I locus were found in the CHS1-3-4 genes which are located close by each other in the inverted repeat clusters. Naturally occurring mutations of the dominant I or i-i alleles to the recessive i alleles involve the deletion of CHS promoter sequences from CHS4-3-1 which enhanced CHS7/CHS8 transcripts in pigmented soybean seed coats (Todd and Vodkin, 1996; Tuteja et al., 2004). Sequencing BACs (Bacterial Artificial Chromosomes) from the Williams variety containing the i-i allele revealed that five (CHS1, CHS3, CHS4, CHS5, and CHS9) of the nine CHS gene family members are located in a 230 kb region (Clough et al., 2004; Tuteja and Vodkin, 2008). CHS1, CHS3, and CHS4 are located in two 10.91 kb perfect and base-by-base inverted repeat clusters separated by 5.87 kb of intevening sequence. The I locus is defined by this region in that deletions of this region occur in recessive i mutations. Based on BLAST searches to the Glyma models of the Williams 82 reference soybean genome at the Department of Energy Joint Genome Institute, the clustered CHS region of the I locus maps to chromosome Gm8. Other CHS family members, CHS2, CHS6, CHS7, and CHS8, are located in different chromosomes, Gm5, Gm9, Gm1, and Gm11, respectively.

The dominant nature of the I locus with the presence of inverted repeats raised the possibility of the presence of the siRNAs (short interfering RNAs) since inverted repeats can generate dsRNAs which process to siRNAs. Also, nuclear run-on assays provided an evidence which suggests a post-transcriptional mechanism of CHS silencing in another isogenic pair of I
locus alleles (Senda et al., 2004). Both small RNA blots and high throughput small RNA sequencing from three genotypes (I, i-i and i alleles) of soybeans revealed that CHS siRNAs accumulated only in the yellow seed coats containing either dominant I or i-i alleles and not in the pigmented seed coats with homozygous recessive i genotypes (Tuteja et al., 2009). Interestingly, the CHS siRNAs are generated in a tissue-specific manner. CHS siRNAs are not accumulated in the cotyledons of the genotype with dominant I or i-i alleles. Both sense and antisense strand CHS siRNAs were abundant and mostly matched to the exon 2 domain of the CHS genes, with fewer to exon 1 and none to the intron, indicating that siRNAs are generated after RNA splicing. The predominant size of CHS siRNAs is 21 nucleotides. CHS7/8 siRNAs, which are more distantly related to CHS1, 3, and 4 siRNAs are the most abundant. Figure 1.2 is a proposed generation pathway of CHS siRNAs involved in silencing of the I locus (Tuteja et al., 2009). A dsRNA generated from within a 27 kb inverted CHS region comprised of two clusters of the CHS1-3-4 genes is cleaved into primary siRNAs representing both strands that are amplified by RNA-dependent RNA polymerase (RdRP) to generate secondary CHS siRNAs capable of down-regulating all members of the CHS gene family. CHS genes are highly expressed in the pigmented seed coats in which CHS siRNA production has been abolished by deletion in the mutant i allele. However, for genotypes in which pigment is restricted to the hilum and saddle regions (i-i and i-k), the role of CHS siRNAs has not yet been identified.

Here, I present the results of next generation high-throughput sequencing of CHS siRNA populations from pigmented and non-pigmented regions within seed coats with the same genotype. These results demonstrate that the CHS siRNAs accumulated only in the non-pigmented region of both genotypes (i-i and i-k). Also, the population of CHS siRNAs from another genotype (i-i, k1), which produces the same saddle region phenotype as i-k, was examined by high-throughput sequencing technology. I also extend previous studies to examine tissue specificity of CHS siRNA expression. Small RNA sequencing data from twenty-one libraries in six different tissues were analyzed in this study to show the CHS siRNA tissue specificity. Differential expression of CHS siRNAs during seed coat development was examined by smRNA sequencing data of eight different developmental stages.
CHAPTER 2 MATERIALS AND METHODS

TISSUE COLLECTION

Three lines (Williams 43, Clark 8 and Clark 18a) of soybean (*Glycine max*) were used in this study. Williams 43 is an internal lab number of Williams variety which is maturity group III. Clark 8 and Clark 18a are near isogenic lines in the Clark variety which is maturity group IV. All lines were homozygous for the loci indicated in Table 2.1. Soybean plants were grown in the greenhouse and immature seeds were harvested over the course of several weeks. More than 40 beans were harvested to represent the expression level of siRNA of each line. The seeds were shelled, pooled together, and then sorted by weight range. The seeds were first dissected to separate the seed coat from the cotyledon then were dissected again to separate the pigmented and the non-pigmented regions of the seed coat with each part placed in different 15-ml tubes. Figure 2.1 shows an illustration of the dissection method of this study. The seed coats were dissected without any visible cutting line since the pigmentation cannot be observed before the 300mg fresh seed weight stage. The saddle pigmented regions were dissected with a triangle shaped region to enclose the hilum to avoid mixing with the non-pigmented region. The seed coats were frozen in liquid nitrogen for 10 minutes and stored in the freezer (-80°C) until they were lyophilized.

SMALL RNA ISOLATION AND SEQUENCING DATA ANALYSIS

Proanthocyanidins bind to the RNA in the pigmented seed coat regions of these three soybean varieties. The modified RNA isolation method was designed specifically to overcome the problem of proanthocyanidins binding to the RNA (Wang and Vodkin, 1994). The seed coats were ground to a powder using a mortar, pestle, and autoclaved sand. 0.75g of hydrated polyvinylpolypyrrolidone (PVPP) (Sigma, St. Louis, MO) and 1ml of proanthocyanidin binding protein buffer (Complete RNA Extraction Buffer [100 mM Tris-HCl, pH 9.0; 200 mM NaCl; 20mM EDTA; 10mM dithiothreitol (DTT); 16 mM mercaptobenzothiazol; and 2% sarkozyl]; 10mg/ml heparin; 2mg/ml polyproline; and 5% bovine serum albumin (BSA)) was ground with
the sample for 1 minute, then 4 ml of the Complete RNA Extraction Buffer was added and the sample was ground for 1 minute again. The sample was transferred to a 50-ml tube and was mixed with 100 µl of 10mg/ml proteinase K (Invitrogen, Carlsbad, CA). The tube was incubated at 37°C with gentle shaking (80rpm) for 20 minutes. After the incubation, the tube was centrifuged for 10 minutes at 5,000 rpm at 4°C. The supernatant was withdrawn using a glass pipette and transferred to 15-ml tubes containing 4 ml of saturated phenol. The RNA was extracted with phenol and chloroform:isoamyl alcohol and then centrifuged in a Beckman J2-21M/E centrifuge at 5,000 rpm. A one-tenth volume of 3M sodium acetate and a double volume of 100% ethanol were added and the RNA was precipitated out at -80°C overnight. The sample was then centrifuged at 8,000 rpm for 20 minutes to pellet the RNA, which was subsequently dissolved in nuclease-free water.

Small RNA libraries and high-throughput sequencing were performed with the Genome Analyzer-II and HiSeq-2000 (Illumina, San Diego, CA) by the Keck Center (University of Illinois, Urbana, IL) using standard Illumina protocol. A total of three to eighty million reads were obtained from these deep sequencing libraries. Adapter trimming was performed using Illumina’s Flicker pipeline which finds the presence of the adapter in each read by finding the best alignment of the adapter to the read, and removing it from the read. The adapter sequence has changed over time in each library (5’-TCGTATGCCGTTCTTGTTG-3’, 5’-ATCTCGATGCGCTTCTGCTTG-3’, 5’-TGGAAATTCTCGGGTCCAAGGAACTCCA GTCAC-3’). The sizes of the small RNAs after adapter trimming ranged from 14 to 33 nucleotides with the majority in the range of 18 to 25 nucleotides. The range was modified to include 16 to 25 nucleotides since the primary size of CHS siRNAs was known and also the amount of raw data was increased significantly. Adapter trimmed sequences were compared to obtain the number of unique sequences and occurrences of each unique sequence. Alignments of siRNA sequences to Glyma models (Phytozome, Joint Genome Institutes) from the Williams 82 reference genome of Glycine max were performed using the Bowtie program (Langmead et al., 2009). Alignments were made to individual CHS sequences with no mismatches allowed. The results from CHS siRNA data were analyzed and shown with Microsoft Excel program.
SMALL RNA BLOTTING

For small RNA blot analysis, 43 µg of total RNA concentrated in 25% formamide was denatured at 70°C for 15 minutes. Denatured RNAs were fractionated with the XCell SureLock Mini Cell device (Invitrogen, Carlsbad, CA) through 15% TBE-Urea polyacrylamide 7M urea denaturing gel cassettes (Invitrogen, Carlsbad, CA) then transferred to Hybond-N membrane (Amersham-GE Healthcare, Buckingham, UK) via capillary action with the Bio-Rad Trans-Blot apparatus (Bio-Rad, Richmond, CA) at 90V for 1 hour. The membrane was equilibrated on 20X SSC saturated filters and cross-linked with UV radiation by a UV-crosslinker (Stratagene Technologies, Santa Clara, CA). The membrane was stored to dry before blotting. Pre-hybridization was performed for longer than two hours in a standard Northern pre-hybridization solution containing 6X SSC (Saline sodium citrate), 5X Denhart (1% Ficol Type 400, 1% PVP-360, 1% BSA, 100 ml water), 0.5% SDS (Sodium dodecyl sulfate), and 100 µg/ml denatured calf thymus DNA at 40°C. The [γ-32P] dATP labeled oligoprobe (mixture of s3_803CHS_RC: 5’-TTTGTATGAGCTTTGTTTGGAC-3’ and s3_913CHS_RC: 5’-CAACTTGTGGAATTGGGTCAG-3’) was added to the pre-hybridization mixture and later hybridized at 40°C overnight. The blot was washed for 15 minutes at 40°C with Wash buffer solution (2X SSC, 0.2% SDS) and exposed to Hyper-film (Amersham-GE Healthcare, Buckingham, UK) at -80°C. For more details on small RNA blotting and protocols refer to Campos (2011).

MESSENGER RNA ISOLATION AND SEQUENCING DATA ANALYSIS

The small RNA extraction sample was processed further for mRNA isolation. After re-suspension of the pellet, the RNA was precipitated overnight again using 2M lithium chloride. The sample was then centrifuged at 8000 rpm to pellet the RNA, which was subsequently dissolved in de-ionized water. A one-tenth volume of 3M sodium acetate and a double volume of 100% ethanol were added and the RNA was precipitated at -80°C for 45 minutes. The pellet was dried in the Speed Vac (Savant Instruments, Holbrook, NY) for 5 minutes. The sample was re-suspended in de-ionized water.

Transcriptome libraries and high-throughput sequencing were performed with the HiSeq-2000 (Illumina, San Diego, CA) by the Keck Center (University of Illinois, Urbana, IL) using
standard Illumina protocol. A total of fifty to eighty million reads of 75 bp each were obtained from these deep sequencing libraries. Alignments of mRNA sequences to Glyma models (Phytozome, Joint Genome Institutes) from the Williams 82 reference genome of *Glycine max* were performed using the Bowtie program (Langmead et al., 2009). Transcriptome data was normalized in reads per kilobase of gene model per million mapped reads (RPKM) as the transcriptome depends on the transcript length (Mortazavi et al., 2008). The results from transcriptome data analysis were shown with Microsoft Excel program.
CHAPTER 3 RESULTS

TISSUE SPECIFICITY OF CHS siRNA EXPRESSION

Data from multiple small RNA libraries were provided by the Vodkin lab and sequenced deeply by the Illumina high-throughput sequencing technology. In order to ascertain the tissue specific expression of CHS siRNAs, the small RNAs were aligned with Bowtie to the CHS genes. Twenty one small RNA libraries were obtained from six different tissues of the Williams 43 variety of soybean including yellow seed coat, cotyledon, unifoliate, trifoliate, stem and root. The total numbers of small RNA reads were from three million to thirty million, providing a wealth of data to show the tissue specificity of CHS siRNA (Table 3.1). The total counts of CHS7 siRNA were normalized per million reads to compare the expression level of CHS7 siRNA in different libraries. CHS7 siRNA is considered very critical to soybean seed coat pigmentation since its target, CHS7 mRNA, is a main player in soybean seed coat pigmentation among the nine CHS family members.

Tissue specificity of CHS7 siRNA can be clearly seen in Figure 3.1. The normalized total counts of CHS7 siRNAs are 11566, 9078 and 4031 in three yellow seed coat samples: 50-75mg fresh seed weight seed coat of Williams 43; 50-75mg fresh seed weight seed coat of Richland; and 75-100mg fresh seed weight seed coat of Williams 43, respectively. However, the expression levels of CHS7 siRNAs were less than 20 counts in five other tissues: cotyledon, unifoliate, trifoliate, stem and root. CHS siRNAs are significantly abundant only in yellow seed coats among six different soybean tissues.

DIFFERENTIAL EXPRESSION OF CHS siRNA DURING SEED COAT DEVELOPMENT

Data from multiple small RNA libraries were provided by Sarah Jones in the Vodkin lab and were annotated to characterize the expression patterns of CHS siRNAs during seed coat development. Eight small RNA libraries were obtained from Williams 43 seed coat in eight developmental stages: 4 DAF (days after flowering) whole seed, 12-14 DAF whole seed, 22-24 DAF whole seed, 5-6mg whole seed, 50-75mg fresh seed weight seed coat, 75-100mg fresh seed
weight seed coat, 200-300mg fresh seed weight seed coat and 300-400mg yellow (desiccating) seed coat. In early developmental stages through 5-6mg, whole seeds were used since seed coats are the biggest part of the whole seed in early stages. Also, soybean seeds obtained at younger stages are too tiny to be dissected. First 3 stages are determined by day after flowering since we can not measure the weight of seed. After first 3 stages, the weight of seed is more accurate to show developmental stages. The total numbers of small RNA reads are from three million to eighty million, providing a wealth of data to show the expression level of CHS siRNAs in each stage. The total counts of CHS7 siRNA were normalized per million reads to compare the expression level of CHS7 siRNA.

CHS7 siRNAs were found to be most abundant in the 50-75mg stage among eight developmental stages (Figure 3.2). The normalized total counts of CHS7 siRNAs are relatively static at first and gradually increase up to 2054 in the 5-6mg whole seed stage. Then, the expression level of CHS7 siRNAs dramatically increases up to 11,566 in the 50-75mg stage before decreasing again. However, we don’t know precisely when CHS siRNAs are expressed most highly since there is no sequencing data of Williams 43 seed coat between 5-6mg and 50-75mg stages.

To ascertain the expression level of CHS siRNAs more precisely, additional sequencing data was obtained from the 25-50mg fresh seed weight stage and the 50-100mg fresh seed weight stage. These small RNA libraries were generated from the non-pigmented region of the Williams 43 seed coat. Figure 3.3 shows that CHS siRNA is more abundant in the 25-50mg developmental stage than in the 50-100mg developmental stage. Thus, CHS siRNAs are highly expressed in seed coats at an earlier developmental stage, at 25-50mg seed weight.

THE ROLE OF CHS siRNA IN PIGMENT PATTERN FORMATION ON SEED COATS WITH THE SAME GENOTYPE

It was previously mentioned that soybean seed coat pigmentation is affected by the I locus (Palmer et al., 2004). There is another locus, K, that determines the distribution of the pigments (Figure 3.4). Varieties that have homozygous dominant I alleles with K1 alleles have yellow seed coats. However, dominant I alleles with recessive k1 alleles are self-colored and
have pigment distributed across the entire seed coat. Also, another soybean variety with the \(i-i\) and \(KI\) alleles has pigment restricted to the hilum, but varieties with \(k1\) alleles have pigment distributed to the saddle region.

Previous studies (Tuteja et al., 2009) compared Williams \((i-i; \text{pigmented hilum})\) to Williams 54 \((i; \text{pigmented black mutation})\), and Richland \((I; \text{non-pigmented yellow})\) to T157 \((i; \text{self-colored mutation of Richland})\). This study compared two different pigmented and non-pigmented regions in the same genotype within the same seed coat tissue. In this study, three soybean lines, Williams 43, Clark 8 and Clark18a, were used to study the role of \(CHS\) siRNA in the formation of pigment patterns on seed coats. The phenotype of Williams 43 has pigment present in the hilum where the seed coat attaches to the pod with an otherwise yellow, non-pigmented seed coat (Figure 3.5). Small RNAs were extracted from the pigmented hilum and the non-pigmented seed coat separately to compare the quantity of \(CHS\) siRNA.

Clark 8 and Clark 18a are both isolines in a Clark background. They have the same phenotype in which pigment is restricted to the saddle regions but they have different genotypes. Clark 8 has homozygous \(i-k\) and dominant \(KI\) alleles (Figure 3.6). Clark 18a has homozygous \(i-i\) and recessive \(k1\) alleles (Figure 3.7). They are completely different varieties with USDA germplasm numbers L70-4204 and L67-3469, respectively. Small RNAs were extracted from the pigmented saddle region and the non-pigmented seed coat separately to compare the quantity of \(CHS\) siRNA in the same seed coat tissue.

\(CHS\) siRNAs RESULT IN THE BLACK HILUM PIGMENT PATTERN FORMATION OF WILLIAMS 43

Four small RNA libraries were generated and sequenced deeply by the Illumina high-throughput technology to determine the role of \(CHS\) siRNAs in the formation of pigment patterns restricted to the hilum. Multiple \(CHS\) siRNAs that map to the coding region of the nine-member \(CHS\) gene family were obtained from the pigmented hilum and the non-pigmented seed coat of Williams 43. Alignments were made individually for each library to each \(CHS\) gene. Williams 43 soybean seeds were harvested at two different developmental stages, 25-50mg and 50-100mg. The total number of siRNA reads was obtained about eighty million, providing a wealth of data
to determine the role of CHS siRNAs in the formation of the hilum pigmented seed coat (Table 3.2).

All CHS siRNAs are more abundant in the non-pigmented seed coat than in the pigmented hilum of 25-50mg Williams 43 soybean seeds (Figure 3.8). The total numbers of CHS1 siRNAs from the non-pigmented seed coat and the pigmented hilum are 30098 and 8821, respectively, which is a three-fold difference. The expression levels of CHS2, 3, 4, 5, 6 and 9 siRNAs show a similar pattern in that CHS siRNAs are more abundant in non-pigmented seed coats than the pigmented hilum. CHS7 and CHS8 siRNAs, which are the most critical CHS siRNAs since their target mRNAs have a critical role in soybean seed coat pigmentation among the nine CHS family members, also show similar expression patterns. The total numbers of CHS7 siRNAs in the non-pigmented seed coat and the pigmented hilum are 67495 and 29416, respectively, which is about a two-fold difference. The total numbers of CHS8 siRNAs in the non-pigmented seed coat and the pigmented hilum are 59174 and 27627, respectively, which is also about a two-fold difference. The expression level of all CHS siRNAs is much higher in the non-pigmented seed coat than in the pigmented hilum.

In the 50-100mg Williams 43 soybean seed coats, the populations of all CHS siRNAs are also much higher in the non-pigmented seed coat than in the pigmented hilum (Figure 3.9). The total reads of CHS1 siRNA from the non-pigmented seed coat and the pigmented hilum are 9912 and 584, respectively, which is about a seventeen-fold difference. The expression levels of CHS2, 3, 4, 5, 6 and 9 siRNAs show a similar pattern, that CHS siRNAs are more abundant in the non-pigmented seed coat than the pigmented hilum. CHS7 and CHS8 siRNAs, which are more highly expressed than any other CHS siRNAs, also show similar expression patterns. The total counts of CHS7 siRNAs in the non-pigmented seed coat and the pigmented hilum are 42408 and 2160. The total counts of CHS8 siRNAs in the non-pigmented seed coat and the pigmented hilum are 37799 and 2031. CHS7 and CHS8 siRNAs have about twenty-fold and nineteen-fold differences between the pigmented and the non-pigmented regions, respectively.

The predominant size of CHS siRNA in the non-pigmented region of Williams 43 is 21-nucleotides (Figure 3.10). This figure shows the size distribution of each CHS siRNA in the 50-100mg Williams 43 soybean seed coats. CHS siRNAs were extracted from the non-pigmented region. Small RNA libraries were filtered to identify those with 100% identity to individual CHS
genes. Previous studies reported that 18 to 25nt is the primary size of CHS siRNAs and 21nt is the predominant size in Williams 43 whole seed coat (Tuteja et al., 2009). Figure 3.10 shows 21-nucleotides CHS siRNAs are predominantly abundant in the Williams 43 non-pigmented region. Together, these data (Figure 3.8~10) strongly support that CHS siRNAs result in this pigment pattern formation, leading to the black hilum pigment pattern.

CHS siRNAs RESULT IN THE SADDLE PATTERN FORMATION OF CLARK 8

Four small RNA libraries from two biological repeats were obtained from the pigmented saddle region and the non-pigmented seed coat of the Clark 8 near isogenic line. Small RNA sequencing data in which the total number of reads is about twenty million is sufficient to characterize the role of CHS siRNAs in the formation of the saddle pigment pattern on soybean seed coats (Table 3.3).

All of CHS siRNAs are more abundant in the non-pigmented seed coat than in the pigmented saddle region of Clark 8 soybean seeds (Figure 3.11). The total numbers of CHS1 siRNAs read by the Illumina sequencer from the non-pigmented seed coat and the pigmented saddle region are 1408 and 249, respectively, which is about a six-fold difference. The expression levels of CHS2, 3, 4, 5, 6 and 9 siRNAs show a similar pattern, that CHS siRNAs are more abundant in the non-pigmented seed coat than the pigmented saddle region. Also, CHS7 and CHS8 siRNAs show similar expression patterns, that CHS siRNAs are more abundant in the non-pigmented seed coat than the pigmented saddle region. The total numbers of CHS7 siRNAs in the non-pigmented seed coat and the pigmented saddle region are 6533 and 423, respectively, which is about a fifteen-fold difference. The total numbers of CHS8 siRNAs in the non-pigmented seed coat and pigmented saddle region are 5888 and 380, respectively, which is also about a fifteen-fold difference. The expression levels of all CHS siRNAs are much higher in the non-pigmented seed coat than in the pigmented saddle.

The biological repeat of Clark 8 shows a similar expression pattern, that the population of every CHS siRNA is higher in the non-pigmented seed coat than in the pigmented saddle region (Figure 3.12). The total reads of CHS1 siRNAs from the non-pigmented seed coat and the pigmented saddle region are 925 and 304, respectively, which is about a three-fold difference.
The expression levels of CHS2, 3, 4, 5, 6 and 9 siRNAs show a similar pattern that CHS siRNAs are more abundant in the non-pigmented seed coat than the pigmented saddle region. CHS7 and CHS8 siRNAs, which are more highly expressed than any other CHS siRNAs, also show similar expression patterns, that CHS siRNAs are more abundant in the non-pigmented seed coat than the pigmented saddle region. The total counts of CHS7 siRNAs in the non-pigmented seed coat and the pigmented saddle region are 2577 and 591. The total counts of CHS8 siRNAs in the non-pigmented seed coat and the pigmented saddle region are 2132 and 464. Both CHS 7 and 8 siRNAs have about a four-fold difference between the pigmented and the non-pigmented region.

The predominant size of CHS siRNA in the non-pigmented region of Clark 8 is 21-nucleotides. Figure 3.13 shows the size distribution of each CHS siRNA in the non-pigmented region of Clark 8 soybean seed coats. Small RNA libraries were filtered to identify those with 100% identity to individual CHS genes. The Clark 8 non-pigmented region also has 21nt CHS siRNAs predominantly. The conclusion from the data (Figure 3.11~13) is that CHS siRNAs result in the saddle pattern formation of Clark 8 near isogenic lines.

CHS siRNAs RESULT IN THE SADDLE PATTERN FORMATION OF CLARK 18a

Four small RNA libraries from two biological repeats were generated and sequenced deeply to characterize the role of CHS siRNAs in the formation of the saddle pattern of Clark 18a. CHS siRNA libraries that map to the coding region of the nine-member CHS gene family were obtained from the pigmented saddle region and non-pigmented seed coat of the Clark 18a soybean. Alignments were made individually for each library to each CHS gene. The total number of siRNA reads was about eighty million, providing strong data to determine the role of CHS siRNAs in saddle pattern formation on soybean seed coats (Table 3.4).

All CHS siRNAs are more abundant in the non-pigmented seed coat than in the pigmented saddle region of the Clark 18a soybean seed coat (Figure 3.14). The total numbers of CHS1 siRNAs from the non-pigmented seed coat and the pigmented saddle region are 10327 and 2428, respectively, which is about a four-fold difference. The expression levels of CHS2, 3, 4, 5, 6 and 9 siRNAs show a similar pattern, that CHS siRNAs are more abundant in the non-pigmented seed coat than the pigmented saddle region. Also, the total counts of CHS7 and CHS8
siRNAs show similar expression patterns, that \textit{CHS} siRNAs are more abundant in the non-pigmented seed coat than the pigmented saddle region. The total numbers of \textit{CHS7} siRNAs in the non-pigmented seed coat and the pigmented saddle region are 56932 and 2067, respectively, which is about a twenty-eight-fold difference. The total numbers of \textit{CHS8} siRNAs in the non-pigmented seed coat and the pigmented saddle region are 50558 and 1877, respectively, which is about a twenty-seven-fold difference. The expression level of every \textit{CHS} siRNA is much higher in the non-pigmented seed coat than in the pigmented saddle region.

Also, in the biological repeat of Clark 18a, the population of every \textit{CHS} siRNA is much higher in the non-pigmented seed coat than in the pigmented saddle region (Figure 3.15). The total reads of \textit{CHS1} siRNAs from the non-pigmented seed coat and the pigmented saddle region are 14023 and 1257, respectively, which is about an eleven-fold difference. The expression levels of \textit{CHS2}, 3, 4, 5, 6 and 9 siRNAs show a similar pattern, that \textit{CHS} siRNAs are more abundant in the non-pigmented seed coat than the pigmented saddle region. Also, \textit{CHS7} and \textit{CHS8} siRNAs, which are more highly expressed than any other \textit{CHS} siRNAs, show a similar expression pattern, that \textit{CHS} siRNAs are more abundant in the non-pigmented seed coat than the pigmented saddle region. The total counts of \textit{CHS7} siRNAs in the non-pigmented seed coat and the pigmented saddle region are 58862 and 1144. The total counts of \textit{CHS8} siRNAs in the non-pigmented seed coat and the pigmented saddle region are 51834 and 1042. \textit{CHS} 7 and 8 siRNAs have about fifty-one-fold and fifty-fold differences between the pigmented and the non-pigmented region, respectively.

The predominant size of \textit{CHS} siRNA in the non-pigmented region of Clark 18a is 21 nucleotides (Figure 3.16). This figure shows the size distribution of each \textit{CHS} siRNA in Clark 18a soybean seed coats. All of these data demonstrate (Figure 3.14~16) that \textit{CHS} siRNAs result in the saddle pigment pattern formation of Clark 18a.

So, the conclusion is that \textit{CHS} siRNAs have a critical role in the formation of pigment pattern on soybean seed coats.
SMALL RNA BLOTS CONFIRM THAT CHS siRNAs ACCUMULATE ONLY IN THE NON-PIGMENTED REGIONS

Small RNA blots confirm that CHS siRNAs result in the formation of pigment patterns on the soybean seed coats. Figure 3.17 shows the differential expression of CHS siRNAs in the pigmented and the non-pigmented regions of Clark 8 and Clark 18a soybean seed coats. Figure 3.17 clearly shows that CHS siRNAs accumulated only in the non-pigmented region of Clark 18a. CHS siRNAs also accumulated only in the non-pigmented region of biological repeats of Clark 8. These results were in accordance with previous sequencing data. Together, the sequencing data and small RNA blots demonstrate that CHS siRNAs have a critical role in the formation of pigment patterns on the soybean seed coats.

TRANSCRIPTOME DATA SHOWS CHS siRNA AND CHS mRNA EXPRESSION LEVELS ARE INVERSELY CORRELATED

Four transcriptome data sets from two biological repeats were generated and sequenced to characterize the relationship of CHS mRNA and CHS siRNA in the formation of the saddle pattern of Clark 18a. CHS mRNA libraries that map to the coding region of the nine member of CHS gene family were obtained from the pigmented saddle region and the non-pigmented seed coat of Clark 18a. Alignments were made individually for each library to each CHS gene. RNA sequencing data in which the total numbers of reads are about fifty to seventy-five million provide sufficient data to determine the relationship of CHS mRNA and CHS siRNA in the formation of the saddle pattern of Clark 18a (Table 3.5). The total counts of CHS mRNAs were normalized by reads per kilobase per million mapped reads (RPKM) to compare the expression levels of CHS mRNAs.

CHS7/8 mRNAs, the main players of soybean seed coat pigmentation, are more highly expressed in the pigmented saddle region (Table 3.6). The normalized counts of CHS7 mRNAs in the pigmented saddle region and the non-pigmented region are 170 and 52, respectively. The normalized counts of CHS8 mRNAs in the pigmented saddle region and the non-pigmented region are 128 and 38. These data represent the expression levels of CHS7 and CHS8 mRNAs have about three to four-fold differences between the pigmented and the non-pigmented region,
respectively. The expression levels of CHSI, 2, 3, 4, 5, 6 and 9 mRNAs are much lower than CHS7/8 mRNAs in both the pigmented and the non-pigmented regions. Biological repeats show similar expression patterns of CHS mRNAs.

Figure 3.18 clearly shows that the expression levels of CHS7 siRNAs and their target mRNA are inversely correlated. The normalized total number of reads of CHS7 siRNAs and mRNAs are compared between the pigmented and the non-pigmented regions of two biological repeats of Clark 18a. CHS7 siRNAs are more highly expressed in the non-pigmented region and the expression level of CHS7 mRNA is low. However, in the pigmented regions, CHS7 mRNAs are highly expressed and siRNAs are not. The biological repeat shows a similar expression pattern for both CHS7 siRNA and mRNA. All of these data demonstrate that CHS siRNAs have a critical role in the formation of the pigment pattern on soybean seed coats.
CHAPTER 4 DISCUSSION

TISSUE-SPECIFIC BIOGENESIS OF CHS siRNAs

Previous studies demonstrated the tissue specificity of CHS siRNA expression (Tuteja et al., 2009 and Campos, 2011). In this expanded study with high throughput sequencing, 21 small RNA libraries from six different tissues clearly showed and confirmed the tissue specificity of CHS siRNA. However, it is not understood how CHS siRNAs are expressed in a tissue-specific manner. Several hypotheses have been put forward to explain the tissue-specific biogenesis of CHS siRNAs (Tuteja et al., 2009). In summary, one postulated mechanism (a) could involve the action of a cell or tissue-specific transcription factor or DNA binding protein that initiates production of the dsRNA progenitor molecules only in the seed coats and not in other tissues of varieties with the dominant I and i-i alleles. Alternatively, in mechanism (b) the dsRNA could be formed in other tissues but not cleaved properly by Dicer to yield primary siRNA, or (c) the primary CHS siRNAs might not be amplified to high levels of secondary siRNAs due to lack of an RdRP (RNA dependent RNA Polymerase) enzyme in other tissues.

It is not likely that other tissues of the soybean are lacking in Dicer or RdRP as there are numerous miRNAs and other siRNAs found in the cotyledon, leaf, and root for example, that would require the function of these basic steps in processing small RNAs. In addition, most of the CHS genes have been shown to be induced to high levels in pathogen-challenged leaf tissues (Zabala et al., 2006) indicating that sufficient CHS mRNA substrates for secondary amplification of a low level of primary CHS siRNAs in the leaves would be possible. Both of these observations suggest that mechanism (a) biogenesis of the CHS siRNA in a highly controlled developmental manner was more likely.

Using RNAse protection experiments, a recent study has detected CHS dsRNA not only in the seed coat but also in the cotyledon and leaf tissues (Kurauchi et al., 2011) of lines with a dominant I allele. Their study suggested that the biogenesis of CHS siRNA could be regulated in a tissue-specific manner after dsRNA is generated. This speculation does not account for why CHS dsRNAs formed in non seed-coat tissues would not be processed and amplified when many
other miRNAs and siRNAs are fully processed and amplified in other tissues such as cotyledons, leaves, and roots. Perhaps there are additional steps that require transport of the dsRNA from the nucleus to the cytoplasm. Whether any of these earlier steps could be specific to the CHS dsRNA generated at the I locus as opposed to all other types of dsRNA generated by other endogenous loci in the cell remains to be determined. Whatever the mechanism, further elucidation of the tissue-specific biogenesis of CHS siRNAs could demonstrate novel processes of small RNA generation, useful to target tissue-specific gene silencing.

**BIOGENESIS OF CHS siRNAs IS REGULATED BY THE DEVELOPMENTAL PROGRAM OF SOYBEAN**

Previously, it was shown with RNA blots that CHS mRNAs are most prevalent at 10-25mg, 25-50mg and 50-75mg weight ranges of the seed coats of mid-maturation embryos with pigmented genotypes. Using TaqMan real time RT-PCR, it was shown that CHS7 and CHS8 mRNAs are highest at the 50-75 mg stage of Williams 55 pigmented seed coats with the homozygous recessive i genotype (Tuteja et al., 2004). It was also known by high throughput sequencing that CHS siRNAs are prevalent in the seed coats at the 50-75 mg seed weight range (Tuteja et al., 2009). In this study, the expression of CHS siRNAs was examined by high throughput sequencing of nine developmental stages from a few days post flowering until the mature dry seed stage (courtesy of small RNAs isolated by Sarah Jones in the Vodkin laboratory). As shown in Figures 3.2 and 3.3, the normalized counts of CHS7 siRNAs rise to their highest levels at the 25-50mg and 50-75 mg weight ranges and decline during later stages. They are most abundant at the 25-50mg weight range. In this regard, the profile of the CHS siRNAs resembles many other genes expressed during seed coat development including many in the isoflavone and anthocyanin pathways (Wang et al., 1994). Thus, the most abundant levels of CHS siRNAs in the i-i genotype with yellow seed coats generally coincide with the time of appearance of the CHS7 and CHS8 target mRNAs in the recessive i genotype that has pigmented seed coats. Thus, in contrast with the down-regulation of the pathway by CHS siRNA-targeted destruction of CHS mRNAs in the yellow seed coats, the CHS transcripts continue to produce CHS isoforms after the
50 mg stage with the resulting accumulation of large amounts of anthocyanins leading to the pigmented seed coat.

**SPECIFIC CHS siRNAs CHANGE DURING DEVELOPMENT BUT BIOLOGICAL REPEATS SHOW SIMILAR PATTERNS FOR MANY OF THE ABUNDANT CHS siRNAs**

Figure 4.1 shows the abundance of each siRNA in the Williams 43 in five of the developmental samples that were shown previously in Figure 3.2 as normalized total counts with respect to developmental time. In Figure 4.1, the position of each siRNA is shown where it matches to the target CHS7 mRNA, one of the main players in soybean seed coat pigmentation. Blue dots represent antisense strand siRNAs and red dots represent sense strand siRNAs. Clearly the patterns differ in the developmental samples. On the other hand, Figure 4.2 shows the distribution of CHS7 siRNA of biological repeats in the non-pigmented region of Clark 18a. Compared to the developmental series, the expression pattern for the two biological repeats is very similar not only in the abundance but also in the position. Most of the abundant sense strand siRNAs match to 520bp from start codon and antisense strand siRNAs match mostly to position 350bp in both biological repeats. These data could suggest that CHS siRNAs are generated in the same pattern at the same developmental stage.

Future work will include a detailed analysis of the pattern for all CHS genes from early development through maturity and identification of the CHS siRNAs that are unique to each CHS family member versus those that are shared in common between the family members as conducted in Tuteja et al., 2009 for the 50-75 mg weight range seed coats.

**CHS siRNAs ARE STRONGLY ASSOCIATED WITH THE YELLOW SEED COAT AND NOT THE PIGMENTED HILUM REGION IN THE DOMINANT WILLIAMS i-i GENOTYPE**

The CHS siRNAs from the pigmented hilum versus the yellow seed coat of Williams 43 were described in the results for two developmental stages, 25-50 mg and 50-100 mg seed fresh weight. In both instances, the levels of CHS siRNAs are much more abundant in the yellow seed
coat than in the pigmented hilum region with a twenty-fold difference in the 50-100 mg range (Figure 3.9) versus two-fold for the earlier stage of 25-50 mg (Figure 3.8). The experimental error during dissection could explain the increased presence of CHS siRNAs in the pigmented region at the 25-50mg stage. Small amounts of the non-pigmented region could be mixed with the pigmented region, since seed coats were dissected by hand at a time before the pigment is present. The dissection is more difficult when the seeds are small. This experiment will likely be repeated in future research with increased attention to localizing only the center region of the hilum area. Although difficult, laser capture microdissection would be one way to definitively separate out only specific cells from the pigmented hilum versus the seed coat proper.

**CHS siRNAs ARE EXPRESSED AND DEGRADE TARGET CHS mRNAs IN THE NON-PIGMENTED REGION OF THE SADDLE PATTERN PHENOTYPES**

From high-throughput small RNA sequencing data, 20 to 80 million reads from both the pigmented and non-pigmented regions of three genotypes (i-i, K1 pigmented hilum; i-k, K1 pigmented saddle pattern; i-i, k1 pigmented saddle pattern) were obtained. This presented sufficient data to determine the role of CHS siRNAs in pigment pattern formation. Based on alignments of these small RNA sequences to the multiple CHS genomic sequences in Figures 3.11 to 3.16 and on data from RNA gel blots (Figure 3.18), I conclude that presence and absence of CHS siRNAs are responsible for pigment pattern formations in both of the saddle pattern genotypes. Also, transcriptome data confirm an inverse relationship between CHS mRNA and CHS siRNA (Figure 3.18). CHS siRNAs are expressed and degrade target CHS mRNAs in the non-pigmented region. In the pigmented region, CHS mRNAs are expressed highly enough to produce pigmentation since CHS siRNAs are not highly generated. Figure 4.3 depicts a model for the generation of CHS siRNAs in two saddle pattern genotypes (i-k, K1 and i-i, k1).

The generation mechanism of CHS siRNAs was proposed in previous work (Tuteja et al., 2009). The long inverted repeat contains I cluster A (CHS1-3-4) and I cluster B (CHS4-3-1) on chromosome Gm8 and likely forms the CHS dsRNA, although the exact size is unknown. The cleavage of CHS dsRNA by Dicer Like proteins (DCL) generates the primary CHS siRNAs, resulting in the degradation of CHS transcripts. These cleaved CHS mRNAs might be substrates
for further RdRP and DCL activity. After cleavage at the mRNA site targeted by the primary CHS siRNAs, an RdRP could synthesize dsRNAs from the cleaved CHS mRNA. The secondary CHS siRNAs generated from this dsRNA could target additional CHS mRNAs, amplifying the silencing response as well as spreading it over a larger region.

As shown in this work, the production of small RNAs is now shown to be pattern-specific within very similar regions of the seed coat within the same genotype in addition to being tissue specific. Possibly, CHS siRNAs are not being amplified to effective levels for lack of an RdRP enzyme in the pigmented region. A recent study reported that CHS dsRNA was detected not only in the seed coat but also in the cotyledon and leaf tissues (Kurauchi et al., 2011). However, examination of the transcriptome data from the saddle pattern genotypes in this report showed that there is no significant difference in RdRP mRNA level between the pigmented and non-pigmented regions. Another hypothesis is that a transcription factor or DNA binding protein may interact with the I locus cluster and potentially other genes to create the pattern. The pattern-specific DNA transcription factor or binding protein generates the dsRNA or the primary CHS siRNAs in the non-pigmented region.

The structure of heterochromatin is a well known epigenetic regulation system which affects the promoter region where DNA binding proteins bind. The heterochromatin structure system is defined by a greater degree of compactibility than other genomic regions, leading to lower accessibility than other regions for DNA binding protein. The heterochromatin structure is influenced by epigenetic information, including modification of histones and methylation of cytosine bases in DNA (Bernstein et al., 2006). Genes with methylated promoters had both lower expression levels and tissue-specific expression patterns in Arabidopsis (Zhang et al., 2006 and Zilberman et al., 2007). The promoter region in the long inverted repeat which contains the I cluster could be methylated and inhibit generation of dsRNAs in the pigmented region. Additionally, siRNAs can provide sequence specificity to guide epigenetic modifications in plants (Henderson and Jacobsen, 2007). Small RNAs including CHS siRNAs possibly affect the density of cytosine methylation of DNA in the non-pigmented regions. However, studies in the Vodkin laboratory using methylation sensitive enzymes previously showed that there is no difference in the methylation status between the pigmented and non-pigmented genotypes (Chan, 1998). The promoters in the I cluster were unmethylated and the Exon 2 of each CHS gene was
methylated. Thus, the CHS genes in the I cluster appeared to have an open chromatin structure. However, these studies included only approximately 12 enzyme sites. With advances in soybean genomics, a more thorough study can be done with whole genome resequencing to survey the entire methylation status of the I locus alleles.

THE UNIDENTIFIED K LOCUS REGULATES THE PATTERN-SPECIFIC CHS siRNA EXPRESSION

Genetic studies have shown that the K locus also determines the distribution of the pigments on the seed coats and interacts with the I locus (Palmer et al., 2004). This study showed that the K locus is involved in the pattern-specific expression of the CHS siRNAs. However, there is no mechanism as to how the K locus regulates this pattern-specific expression of the CHS siRNAs. Figure 4.4 depicts how the unidentified K locus might regulate various steps in the generation of CHS siRNAs. One hypothesis is that the K locus encodes some inhibitors of CHS siRNA generation that prevent the production of the CHS siRNAs to degrade the target CHS genes.

Pattern formation as reflected by the saddle pigment pattern is likely to involve more genes than just the production of the soybean CHS siRNAs. For example, the soybean seed coat includes two vascular bundles (the phloem and xylem elements) at the hilum, the point of attachment to the pod (Thorne, 1981) as well as in the seed coat. There is likely differentiation of these structures within the seed coat and also the potential that nutrients and other materials from other tissues transported from hilum to seed coats may influence gene expression in the seed coats. Also, the seed coat is an end point of phloem transport and is not likely able to transport CHS siRNAs or other siRNAs backward out of the seed coat into other tissues.

To summarize, I have described that CHS siRNAs are expressed in the non-pigmented regions, thereby inhibiting pigmentation of the seed coats. I present clear evidence that a large number of siRNAs with sequences identical to multiple members of the CHS gene family accumulated in the non-pigmented region. Further research to aid in elucidating the mechanism will compare all small RNAs and the mRNA transcriptome in the pigmented and non-pigmented regions of all three pigment pattern genotypes. Also, the seed coat small RNAs and mRNA
transcriptome can be compared with other tissues. Further study of this system should provide more insight into the epigenetic mechanism of pattern formation regulated by small RNAs and more broadly the blueprint of genes involved in seed coat expression and development.
Figure 1.1. The chalcone synthase (CHS) is a key enzyme in the flavonoid pathway which produces anthocyanins and proanthocyanidins (Zabala and Vodkin, 2005).
Figure 1.2. A schematic illustration of generation pathway of CHS siRNAs involved in silencing of the I locus (Tuteja et al., 2009).
<table>
<thead>
<tr>
<th>Lab number</th>
<th>USDA#</th>
<th>Genotype</th>
<th>Phenotype</th>
<th>Origin/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clark 8</td>
<td>L70-4204 (PI 547450)</td>
<td>i-k R T K1</td>
<td>black saddle</td>
<td>L66-14,i(6) x Black Eyebrow</td>
</tr>
<tr>
<td>Clark 18a</td>
<td>L67-3469 (PI 547438)</td>
<td>i-i R T k1</td>
<td>black saddle</td>
<td>mutation in Clark</td>
</tr>
<tr>
<td>Williams 43</td>
<td>Williams (PI 548631)</td>
<td>i-i R T K1</td>
<td>black hilum</td>
<td>Wayne x L57-0034 (Clark x Adams), 1971</td>
</tr>
</tbody>
</table>

Table 2.1. The soybean lines used in this study with USDA germplasm numbers, genotype, phenotype and source/origin.
Figure 2.1. An Illustration of the dissection method of soybean seed coats.
Table 3.1. Summary of 21 small RNA sequencing libraries from 6 different soybean tissues. The total numbers of small RNA reads were from three million to thirty million, providing sufficient data to show the tissue specificity of CHS siRNAs. (DAP=Day After Planting, M=Millions)

<table>
<thead>
<tr>
<th>Library</th>
<th>Variety</th>
<th>Tissue (Developmental Stage)</th>
<th># of Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>D01_S3</td>
<td>Williams</td>
<td>Seed Coat (50-75mg)</td>
<td>2.9 M</td>
</tr>
<tr>
<td>D02_S4</td>
<td>Richland</td>
<td>Seed Coat (50-75mg)</td>
<td>2.9 M</td>
</tr>
<tr>
<td>D04_JT1</td>
<td>Williams</td>
<td>Cotyledon (50-75mg)</td>
<td>3.0 M</td>
</tr>
<tr>
<td>D05_ConStem</td>
<td>PI194639</td>
<td>Stem (2 weeks old plant)</td>
<td>3.5 M</td>
</tr>
<tr>
<td>D06_StemInn</td>
<td>PI194639</td>
<td>Stem (2 weeks old plant)</td>
<td>3.5 M</td>
</tr>
<tr>
<td>D07_FLAM-P</td>
<td>Flambeau</td>
<td>Unifoliate (10 days seedling)</td>
<td>3.8 M</td>
</tr>
<tr>
<td>D08_W82-P</td>
<td>Williams 82</td>
<td>Unifoliate (10 days seedling)</td>
<td>4.3 M</td>
</tr>
<tr>
<td>D09_TO</td>
<td>PI462312</td>
<td>Trifoliate (1 Mo. after planting)</td>
<td>6.1 M</td>
</tr>
<tr>
<td>D10_INF</td>
<td>PI462312</td>
<td>Trifoliate (1 Mo. after planting)</td>
<td>5.4 M</td>
</tr>
<tr>
<td>D12_GCOT</td>
<td>Williams</td>
<td>Germinating Cotyledon</td>
<td>12.4 M</td>
</tr>
<tr>
<td>D14_SCN123</td>
<td>JX4</td>
<td>Roots (5 days old etiolated seedling)</td>
<td>12.2 M</td>
</tr>
<tr>
<td>D15_CS</td>
<td>Clark Standard</td>
<td>Trifoliate (1cm)</td>
<td>15.2 M</td>
</tr>
<tr>
<td>D16_CG</td>
<td>Clark Glaborous</td>
<td>Trifoliate (1cm)</td>
<td>16.2 M</td>
</tr>
<tr>
<td>D18_102GR</td>
<td>Williams</td>
<td>Germinating Cotyledon (10 DAP)</td>
<td>16.4 M</td>
</tr>
<tr>
<td>D19_104ROOT</td>
<td>Williams</td>
<td>Roots (10 DAP)</td>
<td>17.3 M</td>
</tr>
<tr>
<td>D22_OHSCN8</td>
<td>JX4</td>
<td>Seedling roots</td>
<td>33.4 M</td>
</tr>
<tr>
<td>D23_OHBSH8</td>
<td>JX4</td>
<td>Tissue culture of hairy roots</td>
<td>34.4 M</td>
</tr>
<tr>
<td>D24_OHFLH8</td>
<td>JX4</td>
<td>Tissue culture of hairy roots</td>
<td>34.2 M</td>
</tr>
<tr>
<td>D25_OHTAH8</td>
<td>JX4</td>
<td>Tissue culture of hairy roots</td>
<td>34.0 M</td>
</tr>
<tr>
<td>D33_WM43SC</td>
<td>Williams</td>
<td>Seed Coat (75-100mg)</td>
<td>30.1 M</td>
</tr>
<tr>
<td>D34_WM43COT</td>
<td>Williams</td>
<td>Cotyledon (100-300mg)</td>
<td>31.9 M</td>
</tr>
</tbody>
</table>
Figure 3.1. Comparison of *CHS* 7 siRNA counts from libraries of 6 different soybean tissues that map to the coding region of *CHS* 7. Counts are normalized by reads per million. 6 different tissues include yellow seed coat, cotyledon, unifoliate, trifoliate, stem and root as shown in Table 3.1. Numbers above yellow bar is the normalized total counts of *CHS* 7 siRNAs.
Figure 3.2. Comparison of CHS 7 siRNA counts from 8 developmental stages of Williams 43 soybean seed libraries that map to the coding region of the CHS 7. The total counts are normalized by reads per million. The 8 developmental stages are from 4 DAF (Days After Flowering) whole seed, 12-14 DAF whole seed, 22-24 DAF whole seed, 5-6mg fresh seed weight seed coat (approximately 30-35 DAF), 50-75mg fresh seed weight seed coat, 75-100 mg fresh seed weight seed coat, 200-300mg fresh seed weight seed coat, 300-400mg yellow (start to dessicating) seed coat. First 3 stages are determined by day after flowering since we can not measure the weight of seed. After first 3 stages, the weight of seed is more accurate to show developmental stages. (small RNA sequencing data provided by Sarah Jones in the Vodkin lab)
Figure 3.3. Comparison of $CHS$ siRNA counts from 25-50mg and 50-100mg stages of the non-pigmented (yellow) seed coat region of Williams 43. Small RNA sequencing libraries were mapped to the coding region of the nine-member $CHS$ gene family. Alignments were made individually for each library to each $CHS$ gene.
Figure 3.4. The $K$ locus affects pigment distribution in soybean seeds along with the $I$ locus. This figure shows various phenotypes with the two loci.
Figure 3.5. Williams 43 seed and its genotype. Williams 43 is an internal lab number of Williams. $i-i$ and $K1$ alleles specify pigment present in the hilum where the seed coat attaches to the pod with an otherwise yellow, non-pigmented seed coat.
Figure 3.6. Clark 8 seed and its genotype. Clark 8 is an internal lab number of L70-4204. $i$-$k$ and $K1$ alleles produce this saddle pattern.
Figure 3.7. Clark 18a seed and its genotype. Clark 18a is an internal lab number of L67-3469. $i-i$ and $k1$ alleles produce this saddle pattern.
Table 3.2. Summary of smRNA sequencing libraries of Williams 43 indicates USDA germplasm numbers, source/origin, genotype, tissues, and the total number of reads of small RNA. Developmental stage is determined by seed weight. (M=Millions)

<table>
<thead>
<tr>
<th>Variety</th>
<th>USDA#</th>
<th>Source/Origin</th>
<th>Genotype</th>
<th>Tissue (Developmental Stage)</th>
<th># of Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Williams43</td>
<td>Williams</td>
<td>Wayne x L57-0034 (Clark x Adams), 1971</td>
<td>i-i, K1</td>
<td>Seed Coat Hilum (25-50mg)</td>
<td>85.0 M</td>
</tr>
<tr>
<td>Williams43</td>
<td>Williams</td>
<td>Wayne x L57-0034 (Clark x Adams), 1971</td>
<td>i-i, K1</td>
<td>Seed Coat w/o Hilum (25-50mg)</td>
<td>86.6 M</td>
</tr>
<tr>
<td>Williams43</td>
<td>Williams</td>
<td>Wayne x L57-0034 (Clark x Adams), 1971</td>
<td>i-i, K1</td>
<td>Seed Coat Hilum (50-100mg)</td>
<td>85.7 M</td>
</tr>
<tr>
<td>Williams43</td>
<td>Williams</td>
<td>Wayne x L57-0034 (Clark x Adams), 1971</td>
<td>i-i, K1</td>
<td>Seed Coat w/o Hilum (50-100mg)</td>
<td>83.9 M</td>
</tr>
</tbody>
</table>
Figure 3.8. Comparison of CHS siRNA counts from the pigmented (Black) hilum and the non-pigmented (Yellow) seed coat regions of 25-50mg Williams 43. Small RNA libraries were mapped to the coding region of the nine-member CHS gene family. Alignments were made individually for each library to each CHS gene. The small RNA library was filtered to identify those with 100% identity to individual CHS genes. All CHS siRNAs are much more abundant in the non-pigmented region. CHS 7/8 siRNAs are about two-fold more abundant in the non-pigmented region.
Figure 3.9. Comparison of \( CHS \) siRNA counts from the pigmented (Black) hilum and the non-pigmented (Yellow) seed coat regions of 50-100mg Williams 43 soybean. Small RNA libraries were mapped to the coding region of the nine-member \( CHS \) gene family. Alignments were made individually for each library to each \( CHS \) gene. The small RNA library was filtered to identify those with 100% identity to individual \( CHS \) genes. All \( CHS \) siRNAs are much more abundant in the non-pigmented region. \( CHS7/8 \) siRNAs are about twenty-fold more abundant in the non-pigmented region.
Figure 3.10. Size distribution of CHS siRNAs for each CHS gene in Williams 43 CHS-silenced non-pigmented (Yellow) seed coats. The predominant size of CHS siRNAs is 21 nucleotides in the Williams 43 non-pigmented (Yellow) seed coat.
Table 3.3. Summary of small RNA sequencing libraries of Clark 8 indicates USDA germplasm numbers, source/origin, genotype, tissues, and the total number of reads of small RNA. Clark 8 is a near isogenic line. BR1 and BR2 are biological repeats which were grown at different times. (M= Millions)
Figure 3.11. Comparison of CHS siRNA counts from the pigmented (Black) saddle and the non-pigmented (Yellow) seed coat regions of Clark 8^{BR1}. All CHS siRNAs are much more abundant in the non-pigmented region. CHS7/8 siRNAs are about fifteen-fold more abundant in the non-pigmented region. (BR1 : First biological repeat.)
Figure 3.12. Comparison of CHS siRNA counts from the pigmented (Black) saddle and the non-pigmented (Yellow) seed coat regions of Clark 8 BR2. All CHS siRNAs are much more abundant in the non-pigmented region. (BR2 : Second biological repeat.)
Figure 3.13. Size distribution of CHS siRNAs for each CHS gene in Clark 8 CHS-silenced non-pigmented (Yellow) seed coats. The predominant size of CHS siRNAs is 21 nucleotides in the Clark 8 CHS-silenced non-pigmented (Yellow) region. Another biological repeat shows a similar pattern.
Table 3.4. Summary of smRNA sequencing libraries of Clark 18a indicates USDA germplasm numbers, source/origin, genotype, tissues, and the total number of reads of small RNA. Clark 18a is a near isogenic line. BR1 and BR2 are biological repeats which were grown at different times. (M=Millions)

<table>
<thead>
<tr>
<th>Variety</th>
<th>USDA#</th>
<th>Source/Origin</th>
<th>Genotype</th>
<th>Tissue (Developmental Stage)</th>
<th># of Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clark18a BR1</td>
<td>L67-3469</td>
<td>mutation in Clark</td>
<td>i-i,kl</td>
<td>Seed Coat Black Saddle (100-200mg)</td>
<td>83.8 M</td>
</tr>
<tr>
<td>Clark18a BR1</td>
<td>L67-3469</td>
<td>mutation in Clark</td>
<td>i-i,kl</td>
<td>Seed Coat Yellow Segment (100-200mg)</td>
<td>87.3 M</td>
</tr>
<tr>
<td>Clark18a BR2</td>
<td>L67-3469</td>
<td>mutation in Clark</td>
<td>i-i,kl</td>
<td>Seed Coat Black Saddle (100-200mg)</td>
<td>88.0 M</td>
</tr>
<tr>
<td>Clark18a BR2</td>
<td>L67-3469</td>
<td>mutation in Clark</td>
<td>i-i,kl</td>
<td>Seed Coat Yellow Segment (100-200mg)</td>
<td>88.0 M</td>
</tr>
</tbody>
</table>
Figure 3.14. Comparison of $CHS$ siRNA counts from pigmented (Black) saddle and nonpigmented (Yellow) seed coat regions of Clark 18a $^{BR1}$. All $CHS$ siRNAs are much more abundant in non-pigmented region. $CHS7/8$ siRNAs are about twenty-eight-fold more abundant in the non-pigmented region. ($BR1$ : First biological repeat.)
Figure 3.15. Comparison of CHS siRNAs counts from pigmented (Black) saddle and nonpigmented (Yellow) seed coat region of Clark 18a BR2. All CHS siRNAs are much more abundant in non-pigmented region. CHS7/8 siRNAs are about fifty-fold more abundant in the non-pigmented region. (BR2 : Second biological repeat.)
Figure 3.16. Size distribution of *CHS* siRNAs for each *CHS* gene in Clark 18a *CHS*-silenced non-pigmented (Yellow) seed coats. The predominant size of *CHS* siRNAs is 21 nucleotides in Clark 18a *CHS*-silenced non-pigmented (Yellow) seed coats. Another biological repeat shows a similar pattern.
Figure 3.17. CHS siRNAs accumulate in the non-pigmented (Yellow) region but not in the pigmented (Black) regions. (UC8=Clark8, UC18a=Clark 18a)
Table 3.5. Summary of mRNA sequencing libraries of Clark 18a indicates USDA germplasm numbers, source/origin, genotype, tissues, and the total number of reads of mRNA. Clark 18a is a near isogenic line. BR1 and BR2 are biological repeats which were grown at different times. (M=Millions)

<table>
<thead>
<tr>
<th>Variety</th>
<th>USDA#</th>
<th>Source/Origin</th>
<th>Genotype</th>
<th>Tissue (Developmental Stage)</th>
<th># of Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clark18a&lt;sup&gt;BR1&lt;/sup&gt;</td>
<td>L67-3469 mutation in Clark</td>
<td>i-i,ki</td>
<td>Seed Coat Black Saddle (100-200mg)</td>
<td>74.8 M</td>
<td></td>
</tr>
<tr>
<td>Clark18a&lt;sup&gt;BR1&lt;/sup&gt;</td>
<td>L67-3469 mutation in Clark</td>
<td>i-i,ki</td>
<td>Seed Coat Yellow Segment (100-200mg)</td>
<td>70.6 M</td>
<td></td>
</tr>
<tr>
<td>Clark18a&lt;sup&gt;BR2&lt;/sup&gt;</td>
<td>L67-3469 mutation in Clark</td>
<td>i-i,ki</td>
<td>Seed Coat Black Saddle (100-200mg)</td>
<td>75.7 M</td>
<td></td>
</tr>
<tr>
<td>Clark18a&lt;sup&gt;BR2&lt;/sup&gt;</td>
<td>L67-3469 mutation in Clark</td>
<td>i-i,ki</td>
<td>Seed Coat Yellow Segment (100-200mg)</td>
<td>52.8 M</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.6. Comparison of CHS mRNAs counts from the pigmented (Black) saddle and non-pigmented (Yellow) seed coat regions of Clark 18a. CHS7/8 mRNAs are more abundant in the non-pigmented region. BR1 and BR2 are biological repeats which were grown at different times.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene Model</th>
<th>cds Length</th>
<th>cDNA Length</th>
<th>Black Saddle Region</th>
<th>Yellow Segment Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHS1</td>
<td>Glyma08g11620.1</td>
<td>1167</td>
<td>1512</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>CHS2</td>
<td>Glyma05g28610.1</td>
<td>1167</td>
<td>1486</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>CHS3</td>
<td>Glyma08g11630.1</td>
<td>1167</td>
<td>1337</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>CHS5</td>
<td>Glyma08g11520.1</td>
<td>1167</td>
<td>1680</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>CHS6</td>
<td>Glyma01g22880.1</td>
<td>1167</td>
<td>1555</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>CHS7</td>
<td>Glyma01g43880.1</td>
<td>1170</td>
<td>1496</td>
<td>170</td>
<td>357</td>
</tr>
<tr>
<td>CHS8</td>
<td>Glyma11g01350.1</td>
<td>1170</td>
<td>1725</td>
<td>128</td>
<td>274</td>
</tr>
<tr>
<td>CHS9</td>
<td>Glyma08g11610.1</td>
<td>1167</td>
<td>1536</td>
<td>5</td>
<td>8</td>
</tr>
</tbody>
</table>
Figure 3.18. The expression levels of CHS siRNA and CHS mRNA are inversely correlated.
Figure 4.1. Diagrams representing abundance and alignment of CHS siRNAs of 5 different developmental stages of Williams 43 seed coat library with sequence signatures identical to CHS7 gene. (Developmental stages from the top, D52_SJ22DC: 22-24 DAF; D56_SJ5SCB: 5-6mg fresh seed weight; D001_S3: 50-75mg fresh seed weight; D33_WM43SC: 75-100mg fresh seed weight and D54_SJ200SCC: 200-300mg fresh seed weight)
Figure 4.1. (Continued)
Figure 4.2. Diagrams representing abundance and alignment of CHS siRNAs of Clark 18a CHS-silenced non-pigmented (Yellow) seed coat library with sequence signatures identical to CHS7 gene. (D39_C18a2B and D50_C18a3B are small RNA libraries of biological repeats of Clark 18a which were harvested at 100-200mg fresh seed weight stage)
Figure 4.3. An illustration of the role of CHS siRNAs in saddle pattern formation. CHS siRNAs are expressed and degrade target CHS mRNAs in the non-pigmented region only.
Figure 4.4. Unidentified $K$ locus regulates the expression of $CHS$ siRNA along with $I$ locus
LITERATURE CITED


