COMPARATIVE GENOMICS OF *GOSSYPIUM SPP.* THROUGH GBS AND CANDIDATE GENES – DELVING INTO THE CONTROLLING FACTORS BEHIND PHOTOPERIODIC FLOWERING

A Dissertation

by

CARLA JO LOGAN YOUNG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Alan E. Pepper
Committee Members,	Konstantin V. Krutovsky
	John Z. Yu
	Keerti S. Rathore
	Wayne K. Versaw
Head of Department,	U. J. McMahan

August 2013

Major Subject: Biology

ABSTRACT

Cotton has been a world-wide economic staple in textiles and oil production. There has been a concerted effort for cotton improvement to increase yield and quality to compete with non-natural man-made fibers. Unfortunately, cultivated cotton has limited genetic diversity; therefore finding new marketable traits within cultivated cotton has reached a plateau. To alleviate this problem, traditional breeding programs have been attempting to incorporate practical traits from wild relatives into cultivated lines. This incorporation has presented a new problem: uncultivated cotton hampered by photoperiodism.

Traditionally, due to differing floral times, wild and cultivated cotton species were unable to be bred together in many commercial production areas world-wide. This worldwide breeding problem has inhibited new trait incorporation. Before favorable traits from undomesticated cotton could be integrated into cultivated elite lines using marker-assisted selection breeding, the markers associated with photoperiod independence needed to be discovered. In order to increase information about this debilitating trait, we set out to identify informative markers associated with photoperiodism.

This study was segmented into four areas. First, we reviewed the history of cotton to highlight current problems in production. Next, we explored cotton's floral development through a study of floral transition candidate genes. The third area was an in-depth analysis of *Phytochrome C* (previously linked to photoperiod independence in

other crops). In the final area of study, we used Genotype-By-Sequencing (GBS), in a segregating population, was used to determine photoperiod independence associated with single nucleotide polymorphisms (SNPs).

In short, this research reported SNP differences in thirty-eight candidate gene homologs within the flowering time network, including photoreceptors, light dependent transcripts, circadian clock regulators, and floral integrators. Also, our research linked other discrete SNP differences, in addition to those contained within candidate genes, to photoperiodicity within cotton. In conclusion, the SNP markers that our study found may be used in future marker assisted selection (MAS) breeding schemas to incorporate desirable traits into elite lines without the introgression of photoperiod sensitivity.

DEDICATION

To Ryan

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Alan Pepper, and my committee members, Dr. Wayne Versaw, Dr. Keerti Rathore, Dr. Konstantin Krutovsky, and Dr. John Yu, for their guidance and support throughout the course of my research. I would like to impart my deepest thanks to Richard Percy for providing the cotton material for this research.

Thanks go to my friends, colleagues, lab members, USDA-ARS-SPARC employees, and the department faculty and staff for making my time at Texas A&M University a great experience. I also want to extend my gratitude to the Cotton Inc., which provided funding to complete this study. My appreciation goes out to the USDA-ARS for their collaboration in this project.

I give my thanks to all the members in the Thesis/Dissertation group at the Texas A&M University Writing Center. I would like to recognize to Dr. Patricia Goodson and Dr. Dominique Chlup for their writing studios. Also, I would like to thank the Texas A&M University Writing Center for their help and motivation in both my DATA sessions with Mary Beth Schaefer and in the Thesis/Dissertation Boot Camp staff.

Finally, thanks to my mother and father for their encouragement, my family for their support, and to my husband for his patience and love.

I would like to thank the UMID Presidential Fund of the Government of Uzbekistan for providing support for the I.Y.A. to conduct research at Texas A&M University. Also, I thank the USDA-ARS International Research Programs for providing

research grant support for this study under the project P-121. I would like to thank Texas A&M University and Cotton Incorporated for their support by providing a Texas A&M University Regent's Fellowship and a Cotton Incorporated Fellowship (Project 08-380). I would like to give thanks to A. Millie Burrell for critically reading our manuscript and supporting my through the years. I would also like to thank Natalie Ware for the years of help that she provided in our continued research.

Cotton Inc. provided support to conduct our research here at Texas A&M University. Dr. Richard Percy from USDA-ARS-SPARC provided the parental lines PS-6 and K-46 that he has been using in his wild Germplasm conversion since the 1990's. Dr. Richard Percy from USDA-ARS-SPARC provided the conversion lines of BCF4 (PS-5xwild) and BCF2 (PS-6xwild) populations that he has been working on since the 1990's. We wish to thank Kostantin V. Krutovsky for helpful advice on strategies to discover homeologs of candidate genes. We would like to thank Texas A&M University undergraduate Chris Lyle for helping collect all the plant populations DNA during 2008 and Hurricane Ike. We would like to thank, Texas A&M University undergraduate and USDA ARS student employee, Kara Allen for all her hard work and dedication in helping analyze and organize the dN/dS and TGBS data.

NOMENCLATURE

amiRNA	Artificial micro-RNA
B	Blue Light
BC4F2	Back-Crossed 4 Generations and Filial Generation 2
BC4F5	Back-Crossed 4 Generations and Filial Generation 5
bp	Base Pair(s)
Btn	Biotin
dN	Non-synonymous Nucleotide Substitution Rate
dS	Synonymous Nucleotide Substitution Rate
FR	Far-Red Light
GBS	Genotype-By-Sequencing
InDel	Insertion/Deletion Polymorphism
IPGB	Institute for Plant Genomics and Biotechnology
Ka	Non-synonymous Nucleotide Substitution Rate
kb	Kilobase(s)
kDa	KiloDalton
Ks	Synonymous Nucleotide Substitution Rate
LD	Long Day
LD	Linkage Disequilibrium
MAS	Marker Assisted Selection
miRNA	Micro-RNA

MYA	Million Years Ago
NJ	Neighbor Joining
nt	Nucleotide
nM	Nano-Molar
PCR	Polymerase Chain Reaction
PD	Photoperiod Dependent
PI	Photoperiod Independent
PR	Progeny
R	Red Light
RNAi	RNA Interference
SD	Short Day
SID	Single Insertion/Deletion Polymorphism
SNP	Single Nucleotide Polymorphism
TGBS	Targeted Genotype-By-Sequencing
US	United States of America
USDA-ARS-SPARC	United States Department of Agriculture – Agricultural Research Service – Southern Plains Agricultural Research Center

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF EQUATIONS

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Introduction

Cotton has been grown world-wide for many centuries, and it has been the largest natural fiber production crop in the world and of enormous economic importance, as a cash crop, to many developing nations [1]. Cotton not only was utilized for fiber, but has played a huge role in seed oil production, seed stocks for animal feed, and even bio-diesel [2-4]. In the United States, cotton has helped anchor the US economy by being one of the largest contributors to the US gross national product. Each year, the cotton industry has produced over $100 billion in agriculture and textiles [5].

Modern cultivated cotton has limited genetic diversity, so there has been a strong need to develop practical traits from wild relatives. These untapped wild genetic resources had valuable assets which should be incorporated into traditional breeding programs [6]. A few desirable traits of introgression that piqued the interest of breeding programs are higher disease resistance and pest tolerance. Traditionally, desirable traits were bred into elite cultivars using marker-assisted selection (MAS) [7-10]. Consequently, achieving markers representing desirable traits to utilize in MAS breeding from uncultivated cotton species was not straightforward.

Most commercial cotton producing areas in the world have not provided light conditions that allowed the wild cotton species to flower in the span of a growing season under today's current cultivation practices. Uncultivated cotton has been hampered by

photoperiod sensitivity [11]. Since undomesticated species typically flowered under short day conditions, these ‘wild cotton taxa’ required the shortening of red light to nearly eleven hours in order to flower. Cultivated cotton did not have that limitation, it was able to establish flowering under early maturation. Wild and cultivated species have been bred together, but they have resulted in offspring that exhibit photoperiod sensitivity, which rendered them useless for commercial production.

What this Study Asked

Before novel traits from undomesticated species could be integrated into cultivated elite lines, the floral transition network must be analyzed within cotton. Currently, a paucity of information existed within the cotton network. This study reported single nucleotide polymorphism (SNP) differences in thirty-eight homologs of genes within the flowering time network, including photoreceptors, light dependent transcripts, circadian clock regulators, and floral integrators. This research asked if these genes are sound candidates for the photoperiod independence caused during the domestication process. In addition, it queried how different genes show a difference in selection pressures from uncultivated plants to modern domesticated plants. Moreover, this research used redundancy measures to link other discrete SNP differences (unassociated with candidate genes) to photoperiodicity within cotton through Genotype-By-Sequencing (GBS).

Domestication

The conversion of native plants into domesticated species has been propelled through modern breeding techniques resulting in crop improvement. These phenotypic changes have been shown to be exemplary models for studying the genetics of rapid evolutionary responses to natural selection [12, 13]. Studies have examined the genetic basis behind traits for domestication during the first agricultural eras of maize, rice, and wheat [14, 15].

Known Genes behind Other Plant Domestications

Modern domesticated maize was shown to emerge from a single monophyletic lineage of the proximal ancestor *Zea maize spp. parviglumis* in the central region of the Balsas River, México [16]. One locus, *BE518938*, involved in the domestication of corn, has homology to lysine decarboxylase. This locus lowered alkaloid (metallic-like) content in the seeds, improving the taste, which increased its value for human consumption [17].

Wheat's (*Triticum spp.*) point of origin was located west of Diyarbakir in Turkey. This location, where cereals were first domesticated, was determined by the high genetic similarity between wheat predecessors (*einkorn* and *emmer*). Previous research has determined that the tetraploid and hexaploid free-threshing wheat Q allele mutations (*AP2*-like transcription factor) are identical to each other. In summation, the free-threshing 15 loci and a dominant mutation at the Q locus occurred only once [18-21]. This free-threshing gene allowed for humans to easily separate the seeds from the wheat stalks during harvesting time.

Finally, rice (*Orzya sativa*) had several sites of domestication from which two subspecies *Orzya sativa* (ssp. *indica* and *japonica*) were formed. The rise of *O. sativa indica* appeared closely related to *O. nivara*, while *O. sativa japonica* was more closely related to *O. rufipogon* [22-26]. Prostrate growth 1 (*PROG 1*) was identified as a key gene for domestication in cultivated rice for erect stalk growth. A secondary key gene for domestication in rice was shattering locus on Chromosome 4 (*SH4*) [22]. *PROG 1*, on Chromosome 7, allowed erect rice stalks to increase grain yield because of a more stable plant structure. In turn, increased grain yield led to greater human consumption. When *SH4* reduced seed shattering, grain stayed on the stalks longer allowing humans to harvest the rice seeds more efficiently. Without these genetic domestication changes in all of these crops, farming and permanent civilizations might never have arisen.

Cotton's Domestication

In many articles and books, the domestication of different cotton taxa has been introduced by different researchers [5, 27-36]. Hutchison et al. was able to show the movement and evolution of diploid and allotetraploid cotton species throughout the New and Old World [33]. He explained how all wild *Gossypium spp.* were distributed in the arid regions of the tropical and sub-tropical zones. Hutchison et al. gave different theories of domestication for both the Old and New World varieties. At first, it was thought that the Indus civilization around 2000 B.C. was responsible for the domestication of Old World species, but this was disproven by the evidence presented in the cytogenetic work by Beasley et al. in 1942 [37]. This cytogenetic work showed more primitive characteristics of the Old World diploid species were located around the

Arabian Sea [32]. As for the New World species, the oldest cotton textiles (*Gossypium barbadense* L.) were discovered by Bird and Mahler (1951-1952). These textiles were found in Northern Peru with the Huaca Prieta civilization circa 2400 B.C. Wendel et al. was able to narrow down geographical regions for the domestication of *Gossypium hirsutum* L. by allozyme divergence to México or Guatemala [35]. In 1994, Brubaker et al. was able to establish that *Gossypium hirsutum* L. was first domesticated near the Yucatán peninsula [27]. These papers showed how cotton species diverged, and genetic bottlenecks occurred, but researchers floundered when determining the genetic basis for early flowering.

Humans began selecting for early flowering cotton as a result of harvest schedules approximately 5000 years ago in both the old and new worlds. Meso-Americans began domesticating cotton (*Gossypium hirsutum* L. and *Gossypium barbadense* L.) during their proto-agriculture phase. While indigenous people gathered *Teosinte* (primitive corn), *Canvalia* (beans), and other crops for food, they unintentionally selected cotton that produced bolls at an earlier time, rather than photoperiod-sensitive cotton [5, 11, 32]. Concurrently, aboriginal tribes in Africa and Asia began domesticating *Gossypium herbaceum* L. and *Gossypium arboreum* L. [38].

To study the genetic changes that occurred during the domestication process, it was necessary to evaluate the evolutionary foundation for how modern cultivated cotton was developed. Modern cotton diverged from a common ancestor by dividing into two uniquely different lineages (A and D) around 7 to 8 million years ago (MYA) [39]. During a natural hybridization event in Central and South America, the two lineages

[*Gossypium herbaceum* L. (A genome) and *Gossypium raimondii* U. (D genome)] intermixed [5, 32, 33, 40]. Thus, the AD genome was created about 1 to 2 MYA [5]. Following this event, a subsequent whole genome duplication event occurred yielding an allotetraploid cotton species from which modern cottons are derived.

Genetics Helps to Determine Domestication

While looking for the genetic support underlying the domestication process, two main designs became evident: 1) forward genetics, and 2) reverse genetics. Forward genetics used phenotypic traits and genetic loci variants to narrow down the region that controls this phenotype. Conversely, reverse genetics presumed that a molecular base change in a genome would eventually lead back to a phenotype [12-15, 41]. In some species (i.e. *Zea maize*), reverse genetics has successfully identified genes in the domestication process [17, 41, 42], but reverse genetics has been less successful in locating similar genes of domestication in other species (i.e. *Sorghum bicolor*) [43, 44]. In 2012, persistent research efforts to identify the genes behind domestication of a previously indeterminable species (i.e. *Oryza sativa)* were finally successful [22]. Therefore, it became important to strive to determine the underlying genetic support behind phenotypic traits in different species, which applied to cotton.

Flowering

Flowering, which is initiated usually via plant photoperiod perception, has determined a plant's ability to calculate the amount of daylight hours per day over a growing season. This perception occurs through a quantitative process via plant photoreceptors which show altered gene expression in central circadian clock genes

dependent on the amount of light absorbed. The regulatory network, which consisted of several biochemical pathways and controls flowering time, has been documented clearly in the model plant *Arabidopsis thaliana* [45]. *Arabidopsis thaliana* has been identified in the same phylogenetic clade as cotton, Eurosid II [46]. Since these two species were in the same clade, a high gene homology and correspondence for cotton traits were expected to be seen. Therefore, those genes within this documented regulatory network represented ideal candidates for involvement in photoperiod independence of cotton.

Long Day versus Short Day Plants

Although a high gene homology between Arabidopsis and cotton was expected, Arabidopsis and Cotton vary in their photoperiod perception to flower. *Arabidopsis thaliana* has been a long-day (LD) plant, while primitive accessions of *G. barbadense* and *G. hirsutum* flowered under short-day (SD) control [47-49]. Flowering in LD plants occurred when a maximum threshold for day-length is attained. In the northern hemisphere, this happened during the late spring or early summer as it approaches summer solstice (June 21st). In the southern hemisphere, seasons were opposite to the northern hemisphere. Therefore, LD plants flowered closer to Dec. 21st [50].

For flowering to occur in SD plants, a critical light reduction from the maximum day-length has to happen. In cotton, day-length must be reduced to 10 hours of day-light with an uninterrupted dark cycle before floral initiation starts. Disruption of SD floral initiation has occurred if an artificial light sources transpired during the night [50]. In the northern hemisphere, SD plants flowered after June 21st in the late-summer or fall.

Depending on the latitude, the time delay for floral initiation increased with the plant's proximity to the north and south poles.

When a plant has not flowered according to photoperiodism (indifferent to the amount of day and/or night hours), this is known as day-neutrality (photoperiod insensitivity). These day-neutral plants initiated flowering after reaching a certain developmental age or stage [50]. Interestingly, modern domesticated varieties of *G. barbadense* and *G. hirsutum* changed from SD plants to display day-neutrality [48, 49]. By understanding the molecular-genetic determents behind day-neutrality in modern cotton and SD primitive cotton, new strategies to introgress valuable genetic traits from wild Germplasm for crop improvement have be applied [48, 49].

Current Knowledge about the Flowering Process

In *Arabidopsis*, floral transition has been identified as a network of individual pathways comprised of 173 genes [51]. The transition from vegetative to flowering was induced under the lengthening of daylight, hence altering the expression levels in the circadian clock [52-54]. Ergo, the research on Arabidopsis laid a foundation for all photoperiod studies in plants.

To understand the regulatory network controlling the initiation of flowering, a general biochemical process must be understood. The regulatory network consists of photoreceptors which absorbed the light, circadian clock regulators that calculated the amount of light received, and the activation of floral transition factors.

Photoperiodic Light Genes

Light has been shown to affect many living organisms. During the course of a growing season, plants were able to recognize both dawn and dusk. As seen in many species, plants sensed the lengthening of daylight hours in the spring and the reduction of daylight hours in the fall [55-58]. How were plants able to do this? They achieved this by taking in different light wavelengths through photoreceptors in the leaves that regulate plant development. The main levels of light that plants perceived are: Blue (B), Red (R), Far-Red (FR), and ultraviolet-A/B (UV-A/B). These light levels were present at different times of the day. At the lowest light levels before dawn and after dusk, UV-B (282-320nm) was recognized by the newly identified photoreceptor *UV Resistance Locus 8* (*UVR8*) [59]. Blue light/UV-A (320-500nm) was highest during sunrise and dusk each day. While far-red light (700-750 nm) peaked during the early morning and late afternoon hours (730nm). During mid-day, red light (600-700 nm) was at topmost intensity (660 nm) [60, 61]. Each light level activated a particular photoreceptor gene in the leaf. The specialized photoreceptors were key components for a plant's ability to calculate their life span, activation of developmental processes (ie, flowering), membrane signaling, and other processes (Figure 1).

There are thirteen known photoreceptors in the model plant species Arabidopsis that perceived light. Photoreceptors ensnaring blue light were: cryptochromes [*Cryptochrome 1* (*CRY1*), *Cryptochrome 2* (*CRY2*), and *Cryptochrome 3* (*CRY3*)], phototrophins [*Phototrophin 1* (*PHOT1*) and *Phototrophin 2* (*PHOT2*)], and *LOV/F-box/Kelch* domains [*Zeitlupe* (*ZTL*), *Flavin-Binding Kelch Repeat F-Box 1* (*FKF1*), and *Lov Kelch Repeat Protein 2* (*LKP2*)] [61, 62]. Phytochrome photoreceptors which captured red light (R and FR) were: *Phytochrome A* (*PHYA*), *Phytochrome B* (*PHYB*), *Phytochrome C* (*PHYC*), *Phytochrome D* (*PHYD*), and *Phytochrome E* (*PHYE*) [63, 64].

These phytochromes played an integral role in the day-night cycling of the circadian clock [65, 66]. *PHYA*, *CRY1*, and *CRY2* prevented the degradation of *Constans* (*CO*) protein [67]. *PHYB*, *PHYD*, and *PHYE* acted as a redundant network to repress flowering by determining the R/FR light ratio [68-71]. The photoreceptors, shown to be key controlling factors in flowering, were *CRY2*, *PHYB*, and *PHYC* [72-74].

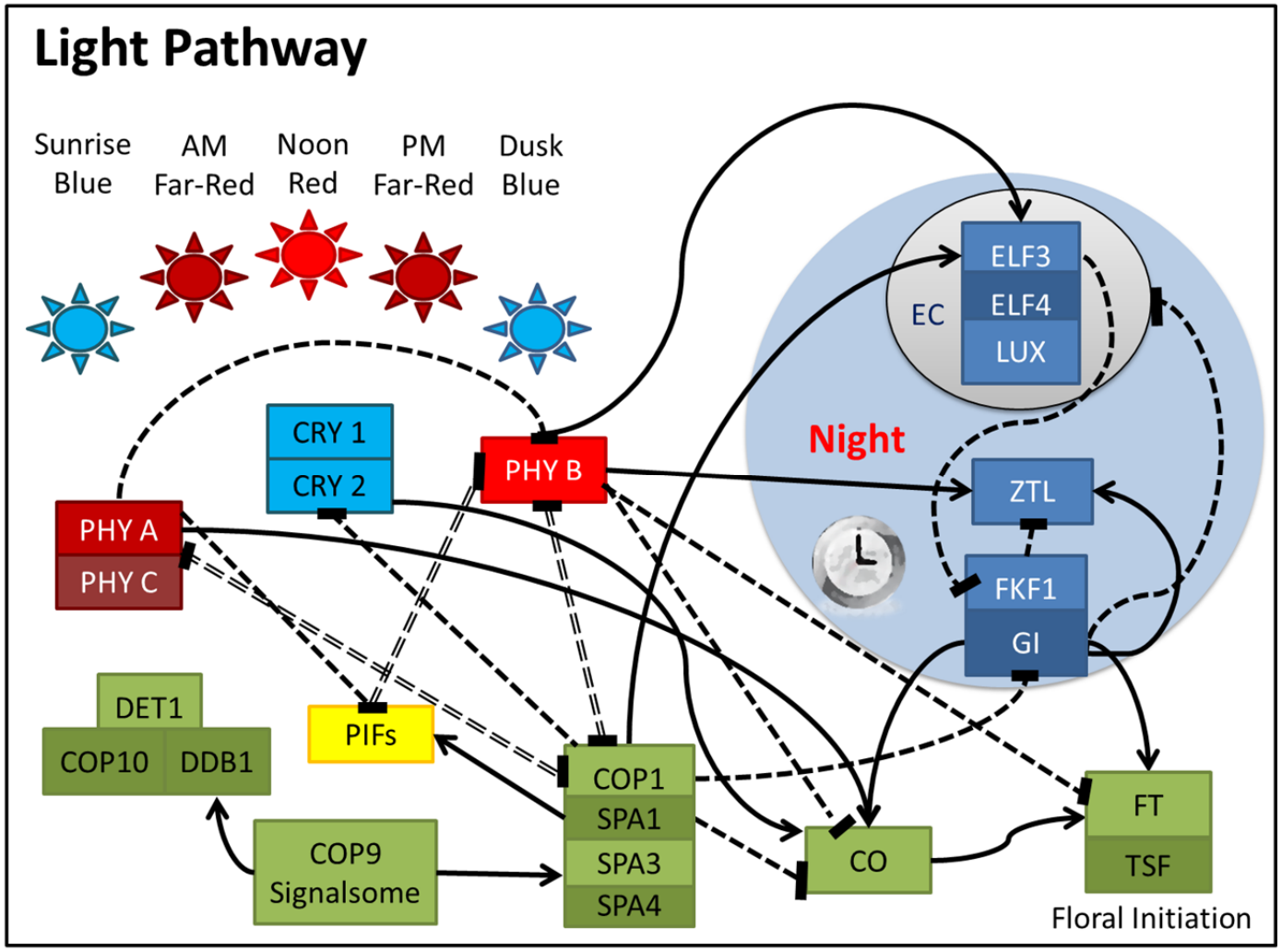


Figure 1 Photoreceptor Light Signaling Pathway

CO has been identified as one bottleneck in the flowering network, where information from both the circadian clock and the photoperiodic pathways converge. This was a key step in the induction of floral initiation (Figure 2). Some elements have by-passed *CO*, but floral initiation was usually the onset of *Flowering Locus T* (*FT*), a consequence of *CO* transcription.

One example of *CO* interacting with *FT* to initiate flowering outside the model species Arabidopsis was in rice [75]. The ortholog of *CO* in rice was *Heading date 1* (*Hd1*), while the ortholog of *FT* was *Heading date 3a* (*Hd3a*) [75, 76]. *Hd1* activated *Hd3a* under SD conditions to induce flowering, but suppressed *Hd3a* in LD conditions [76-78]. This was uniquely different from Arabidopsis, in which *CO* activated *FT* only under LD conditions. A similarity between rice and Arabidopsis though was that phytochromes and circadian clock members manage floral initiation through *CO* and *FT*. *PHYB* helped regulate *Hd1* and *Hd3a* to initiate flowering [54, 76, 79-84]. Tamaki et al. showed that once Hd3a activates, it moved from leaf to shoot to initiate flowering under SD conditions [85].

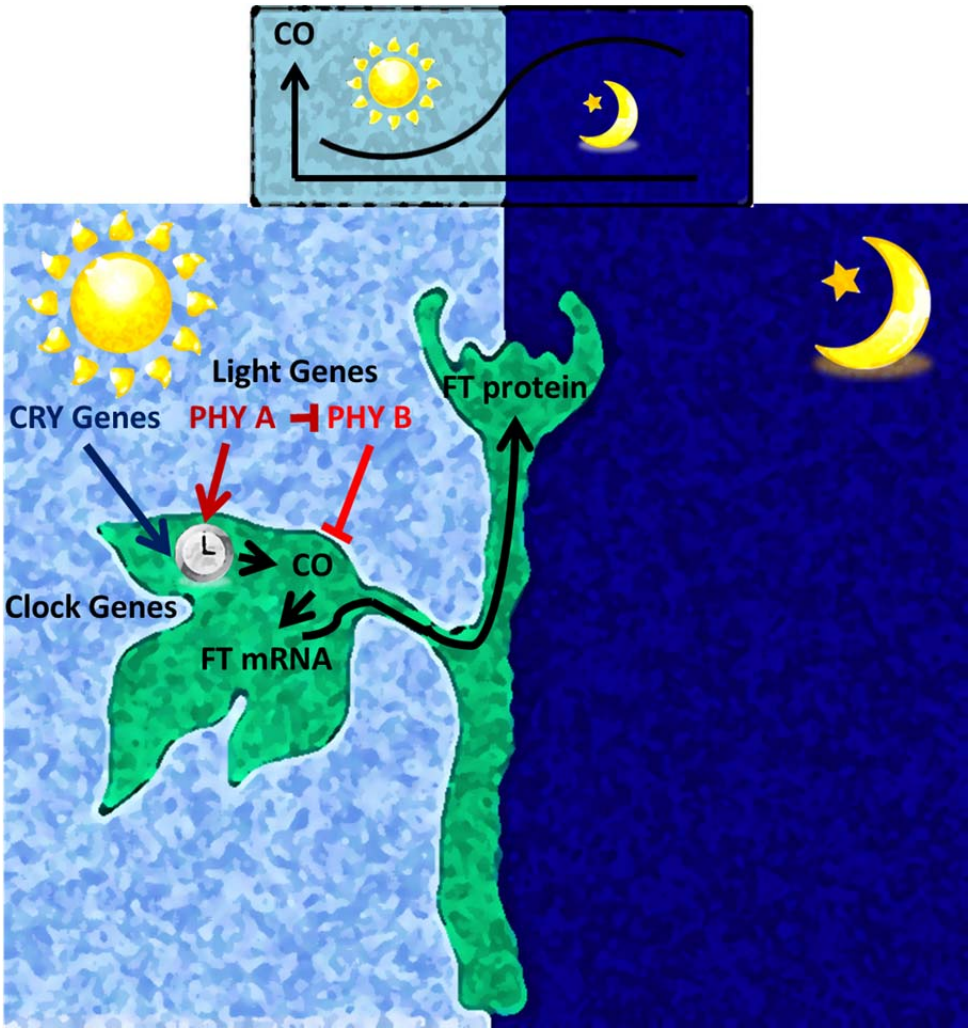


Figure 2 Constans Level Fluctuations during a 24 Hour Period

To attain transcription of *CO*, there were several genes and pathways that hindered and helped the regulation of *CO*. Cryptochromes stabilized *Constans* (*CO*), so that floral initiation begins [67]. In Arabidopsis, *CRY2* protein levels cycled through a

blue light dependent phase to determine photoperiodic clock regulation [72, 86]. This clock regulation told *Gigantea* (*GI*) when to activate *CO* [87]. *CRY2's* role in floral initiation was first seen in a day-neutral early flowering Arabidopsis plant from Cape Verde Islands in 1998 [88]. This photoperiod-insensitive phenotype was caused by a substitution in *CRY2* [72]. *PHYB* acted to suppress the *CO* protein, so that flowering was not always being initiated [67]. Another negative regulator of *CO* was the *COP1/SPA1* complex [89, 90].

To regulate the effects of *CRY2*, *PHYB*, and the *COP1/SPA1* complexes upon *CO*, feedback inhibition loops were created. In the phytochrome pathway, a feedback loop was created to regulate the influence of *PHYB* on the circadian clock. *Constitutive Photomorphogenic 1 and Suppressor of PHYA-105 complex* (*COP1/SPA1*) positively influenced *Phytochrome-Interacting Factors* (*PIFs*), while *PHYA* and *PHYB* halt *PIFs* phosphorylation activity [61, 91]. *PHYA* and *PHYB* negatively regulated *COP1* ability. *PIFs* help *COP1* to downgrade the *PHYB* protein [47, 61]. In turn, *COP1* cooperated with *ELF3* to communicate day-length cues created by *CRY2* [92]. Once this day-length cue was voiced, *COP1* helped degrade *CRY2* in the feedback loop [62].

Other cryptochromes may sway floral initiation, but were not key factors. *CRY1* produced Arabidopsis plants with late flowering in some experiments, but not all [86, 93-97]. Although in one double mutant Arabidopsis plant for *CRY1* and *CRY2*, the floral initiation process exhibited later than the wild type. This mutant study implicated *CRY1* as a backup for *CRY2* in photoperiod flowering [98, 99]. *CRY3* biochemically acted as DNA photolysis for single strands and may act as a back-up photoreceptor for blue light

in the mitochondria and chloroplast. This cryptochrome function was still enigmatic, but does not appear to be influential in the regulation of flowering [62, 100].

Even though *CO* played a central role in floral initiation, other cues stimulated *FT*. Independent from *CO* activation, *CRY1* and *CRY2* control, *FT* was stimulated by *microRNA 172* (*miRNA 172*) [101]. Also, *GI* has been found to stimulate *FT* without the presence of *CO* [101]. In short, alternate methods to stimulate flowering has been imperative to maintain a reproductive life in a plant. Since some positive stimuli have been shown to by-pass *CO* to activate flowering, *PHYB* has been shown to act downstream of *CO* to stifle the transcription of *FT* [102].

One of the other main influences of floral initiation was through the role of *PHYC*. Mutant studies in Arabidopsis and rice, portended a loss of *PHYC* results in early flowering. Research by Franklin et al. showed that Arabidopsis *PHYC*-1 mutants had larger primary and mature leaves, while Monte et al. Arabidopsis *PHYC* mutants showed late flowering under long day (LD) conditions and early flowering under short day (SD) conditions. Takano et al. displayed earlier flowering of the PHYC-antisense transgenic rice lines, than the Nipponbare lines exhibited under LD conditions. These studies suggested that *PHYC* plays a part in LD sensing and were required for SD photoperiod perception [103-105].

In Arabidopsis and pearl millet, *PHYC* steered phenotypic flowering variation naturally. Balasubramanian et al. showed that 29 Arabidopsis varieties have varying levels of linkage disequilibrium (*LD)* between *PHYC*'s presence and early SD flowering was dependent on elevation and latitude. The research by Samis et al. explored

longitudinal effects on Arabidopsis at similar latitudes and showed that Arabidopsis's photoperiod adaptability was caused by genetic differences related to the *PHYC* genotype under SD conditions. The research on pearl millet by Saidou et al. demonstrated a strong *LD* between *PHYC*'s presence and the three agro-ecological zones (differing in rainfall). The relation of *PHYC*'s *LD* showed that pearl millet flowers later in wetter regions, while flowering earlier in drier regions [73, 106, 107].

Few articles have described the effects of photoreceptors in the *Malvaceae* family, despite evidence from physiology investigations. These experiments suggested photoreceptors play cardinal pieces in cotton development: drought resistance, seed dormancy**,** plant architecture, photoperiodic flowering, and fiber elongation [108-111]. The loss of photoperiodism, in some major crops (i.e. sorghum, barley, rice, and soy), has been attributed to mutational changes in these photoreceptors [112-115]. Childs et al. reported that a mutation creating *Ma3* (a homolog of *PHYB*) in sorghum causes a frame shift mutation disrupting the photoreceptor. This resulted in photoperiod independent sorghum plants. In barley, Hanumappa et al. exposed *BMDR-1* as a mutant in barley that lacks a functioning *PHYB*. This mutant produced a photoperiod independent phenotype in barley. According to Izawa et al., the *SE5* mutant demonstrated complete loss of photoperiodism in 24 hours of constant white light. All phytochromes in *SE5* mutant rice showed reductions in expression levels and complete loss of *PHYA*. The study conducted by Izawa et al. pointed to phytochromes for the sacrifice of photoperiod sensitivity. Research by Liu et al. on photoperiod independence in soybean demonstrated that a secondary *PHYA* homeolog has a recessive locus *E4* for photoperiod sensitivity. The

study exposed a retro-transposon located at the *E4* locus causing a frame shift mutation truncating the protein. Thus, soybean with homozygous loci for *E4* conferred photoperiod insensitivity in the plant.

Circadian Clock

Light, the pinnacle environmental signal, has allowed organisms to coordinate their daily cycles regulated by the circadian clock and their physiological activities with environmental changes (Figure 3) [116-118]. During the perceived shortening of daylight, some genes in the circadian clock demonstrated altered expression levels. Most living things had some sort of circadian clock to regulate different processes in their daily biological operations [119, 120]. Thus, those genes affected by this clock regulation were *Early Flowering 4* (*ELF4*), *Early Flowering 3* (*ELF3*), *Circadian Clock Associated 1* (*CCA1*), *Late Elongated Hypocotyl* (*LHY*), *Timing of CAB Expression 1* (*TOC1*), and *Constans* (*CO*) [121-124].

To stimulate flowering in a plant *CO* first must be induced. The circadian clock played a key role on influencing the expression of *CO*. The circadian clock was controlled by an oscillating system between day and night in a 24 hour period. The core components in the circadian clock oscillator were *CCA1*, *LHY*, *TOC1*, and *CCA1 Hiking Expedition* (*CHE*). To achieve night/day oscillations the activation of *CCA1* and *LHY* was required. After activation of *CCA1* and *LHY*, the *COP10-DET1-DDB1* (*CDD*) complex was conscripted to help repress *TOC1* and *GI* transcription [125]. During the dwindling hours of twilight, the daytime repression of *TOC1*, *Lux Arrhythmo* (*LUX*), *ELF3*, *ELF4*, *GI*, *"Night" Brother of Lux Arrhythmo* (*NOX*), and *CHE* was nullified

[126, 127]. The night genes were then actively being transcribed and reciprocally, *TOC1* represses *CCA1* and *LHY* transcription [125].

TOC1, a main control factor in the clock oscillator, acted as a repressor of other daytime clock factors, such as *PRR9*. In the morning loop of the circadian clock, *TOC1* binds to *Pseudo-response Regulators PRR9* and *PRR7*. The initiation of *PRR9* was still unknown, but *Light-Regulated WD1* (*LDW1* – a protein involved in period length regulation and photoperiod flowering) was thought to be a good candidate in triggering *PRR9 [128]*. *PRR9*, *PRR7*, and *PRR5* must then move to control and regulate the expression of *CCA1* and *LHY* [129].

In the evening, *LUX* combined with *ELF3* and *ELF4* to form the *evening complex* (*EC*) [121, 123, 126, 127, 130-132]. The EC bound to *PIF4* and *PIF5* to regulate hypocotyl growth, while *ELF3* stifled *PRR9* expression [126, 129, 133-137]. After the threshold for consecutive dark hours was reached, *GI* and *FKF1* were activated to promote *CO* expression and *FKF1* inhibits *Zeitlupe's* (*ZTL*) expression. If the dark hour threshold for floral initiation was not reached, then *GI* stimulated *ZTL* to repress *TOC1* for the end of the cycle [138].

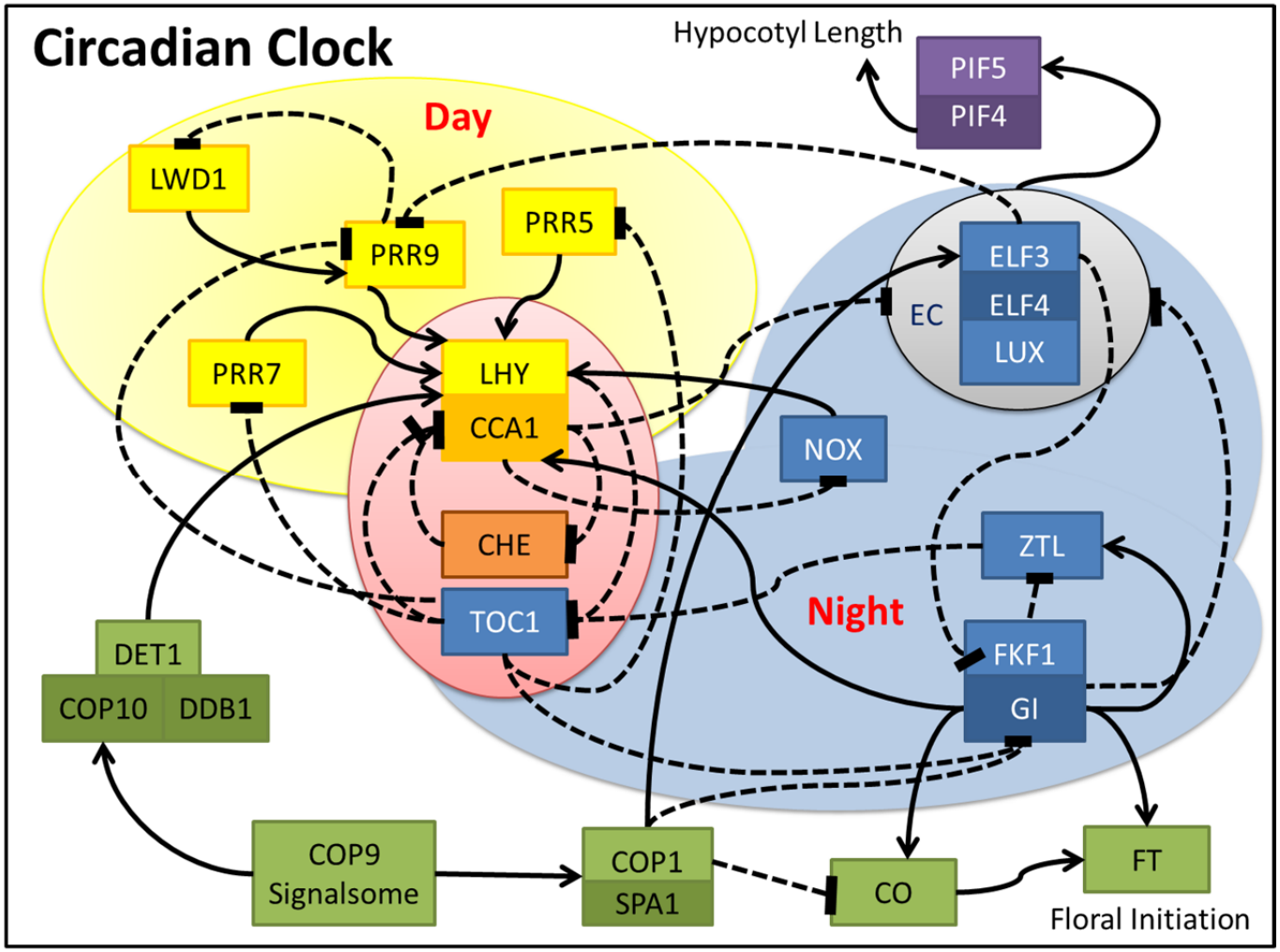


Figure 3 Circadian Clock Pathway

Floral Initiation/Integrators

The first key genes involved during floral transition from vegetative to flowering were *CO*, *Flowering D transcript* (*FD*), and *Flowering Locus T* (*FT)*. The accumulation of the *CO* transcription factor activated the signaling molecule of *FT*. The *FD* transcript was also a positive regulator of *FT*. The florigen *FT* protein molecule then travelled through the phloem to the shoot apical meristem (SAM) [12, 85, 139-141].

When *FT* reached the SAM, it activated the *Suppressor of Overexpression of Constans 1* (*SOC1*, *MADs box Transcription Factor*) and *Apetala1* (*AP1*) [142-144]. *SOC1* and *AP1* induced the expression of the meristem protein *Leafy* (*LFY*) which in turn signaled the genes regulating the metamorphosis of a vegetative meristem into an inflorescence meristem [145].

Induction of *FT* was the primary method for controlling floral initiation. Premature induction of *FT* activated a myriad of transcripts which prevented premature flowering differentiation. This suppression complex to prevent premature flowering consisted of: 1) *Short Vegetative Phase* (*SVP*), 2) *Agamous-like 24* (*AGL24*), and 3) *SOC1*. Another suppression group of *FT* contained *Terminal Flower 1* (*TFL1*) and *Embryonic Flower 1* (*EMF1*). *TFL1* had a similar amino acid sequence to *FT*, but it acted to suppress flower initiation. Up-regulation of *EMF1* determined that the plant's shoot cells were in the vegetative state. With the induction of the suppression complex or *TFL1* & *EMF1*, flowering initiation was controlled [146-153].

Flower Developmental Genes

After floral initiation, floral organs arose in a definitive order to produce a flower. This was known as the ABC model of floral organ development [154-157]. After *TFL1* and *Embryonic Flower 1* (*EMF1*) were stopped, 'A' genes [*AP1*, *Agamous-like 8 Fruitfull* (*AGL8_FUL*), and *Apetela 2* (*AP2*)] inducted sepal formation [145, 154, 158-160]. Next, 'B' genes [*Apetela 3* (*AP3*) and *Pistilatta* (*PI*)] were conscripted to create petals [154, 157]. Finally, *Agamous* (*AG*) was enacted to actuate stamen and carpel development [157, 158]. This ABC model has now added an 'E' gene group in angiosperms [154, 161]. The 'E' gene group was made up of *Agamous-like 9 Sepallata 3* (*AGL9_SEP3*), *Agamous-like 4 Sepallata 2* (*AGL4_SEP2*), *Agamous-like 3 Sepallata 4* (*AGL3_SEP4*), *Agamous-like 2 Sepallata 1* (*AGL2_SEP1*). This group functioned by helping properly co-ordinate ABC function and development by influencing *AGL8_FUL* to convert leaves into floral organs [154, 155]. The development of floral organs by ABC has varied between different species, so some 'A', 'B', or 'C' regions in development may overlap. This was best illustrated in Litt et al. in 2010 [154].

CHAPTER II

DUPLICATION, DIVERGENCE, AND PERSISTENCE IN THE PHYTOCHROME PHOTORECEPTOR GENE FAMILY OF COTTONS *(GOSSYPIUM SPP.)* [*]

Synopsis of Phytochrome C

Phytochromes, specialized photoreceptors, interpreted light frequencies to regulate plant development, such as floral initiation and circadian rhythms [162-164]. Current research showed a dearth of information known about phytochromes. Previous physiological experiments implicated phytochromes in managing aspects of cotton development: 1) drought resistance, 2) seed dormancy**,** 3) plant architecture, 4) photoperiodic flowering, and 5) fiber elongation [108-111].

Phytochrome genes were classified into two evolutionary clades of sub-families in angiosperms: Clade 1 (*PHYB*, *PHYD*, and *PHYE)* and Clade 2 *(PHYA* and *PHYC)* [63, 64, 165]. These clades influenced the circadian clock and help govern the inception of floral development. Clade 1 regulated floral initiation by preventing degradation of the *CONSTANS* (*CO*) protein, while Clade 2 induced degradation of *CO*, a floral inducer [166]. Floral initiation was usually the onset of *Flowering Locus T* (*FT*), a consequence of *CO* transcription.

---

[*] Part of the data reported in this chapter is reprinted with the permission from "Duplication, Divergence and Persistence in the Phytochrome Photoreceptor Gene Family of Cottons (*Gossypium spp.*) " by Abdurakhmonov I, Buriev Z, Logan-Young C; Abdukarimov A, and Pepper A, 2010, *BMC Plant Biology*, 10(1):119

In Arabidopsis, *CO* activated *FT* only under long day (LD) conditions. Although in rice, *Heading date 1 (Hd1)*, an ortholog of *CO*, activated *Heading date 3a (Hd3a)*, an ortholog of *FT*, under SD conditions to induce flowering, but suppressed *Hd3a* in LD conditions [76-78]. Tamaki et al. demonstrated parallelism between (1) *FT's* movement from the leaf to the shoot apical meristem (SAM) in Arabidopsis under LD conditions and (2) *Hd3a's* progression from leaf to shoot which initiates flowering under SD conditions [85]. Even in rice, floral initiation is under the influence of phytochromes. *PHYB* helped regulate *Hd1* and *Hd3a* to initiate flowering [54, 76, 79-84, 167].

Understanding the key photoperiodic flowering differences at the molecular-genetic level in cotton helped create strategies for crop improvement of cultivated varieties by integrating valuable traits from undomesticated 'wild' germplasm [48, 49]. To bring in those valuable traits from 'wild' cotton germplasm, it was imperative to establish modern cotton's evolution. During the domestication process of allotetraploid cotton, both *Gossypium hirsutum* L. and *Gossypium barbadense* L. photoperiod independent flowering arose. Modern cultivated *Gossypium hirsutum* L. and *Gossypium barbadense* L. displayed photoperiod independence, but undomesticated accessions of *Gossypium hirsutum* L. and *Gossypium barbadense* L. retained the original pre-domestication condition of short day (SD) photoperiodic control [48, 49]. Since phytochromes regulated floral initiation and the circadian clock activities in other plants, a mutational change in a phytochrome may lie behind the rise of photoperiod independence during cotton's domestication.

Mutational changes altering phytochrome function, in several major crops (sorghum, barley, rice, and soy) result in loss of photoperiod sensitivity [112-115]. Childs et al. illustrated how a frame shift mutation in the Ma3 (*PHYB* ortholog) locus disrupts the photoreceptor resulting in photoperiod independent sorghum plants. Hanumappa et al. suggested that a mutation at the BMDR-1 locus causes *PHYB* to be non-functioning and results in photoperiod independent barley. Izawa et al. explained that a mutation at the *SE5* locus of rice generates lower phytochrome expression levels and loss of expression in *PHYA*. The loss of photoperiod sensing in the research by Izawa et al. was due to this *SE5* mutational change. Liu clarifies how a single retro-transposon in the *E4* locus located on a secondary *PHYA* causes a frame shift mutation giving rise to photoperiod independence in homozygous recessive soybeans. These previous studies showed the large impact phytochromes have on flowering [112-115].

Another phytochrome that played a role in phenotypic flowering variation is *Phytochrome C (PHYC)*. Previous studies show that genetic variation at the *PHYC* locus naturally guided phenotypic variation of flowering time in Arabidopsis and pearl millet [73, 106, 107, 168]. The study by Balasubramanian et al. highlighted that varying levels linkage disequilibrium (*LD)* between the presence of *PHYC* and early SD floral initiation depends on the elevation and latitude of the Arabidopsis plant. The research by Samis et al. depicted Arabidopsis's SD photoperiod adaptability is due to genetic differences of *PHYC* at different longitudes at corresponding latitudes. In Africa, the research of Saidou et al. presented a strong *LD* correlation between the presence of PHYC and three agro-ecological zones; thereby conveying that the presence of *PHYC* in regions of heavy

rainfall caused pearl millet to flower later, while in more arid regions pearl millet flowered earlier [73, 106, 107, 168].

Recent mutant analysis studies implicate *PHYC's* role in photoperiodic control of floral initiation [103-105]. The Arabidopsis *PHYC* loss of function mutants by Monte et al. conveyed early flowering under short day (SD) conditions. The Arabidopsis phyC-1 mutants by Franklin et al. showed earlier fully developed primary and mature leaves inferring quicker plant maturity. Takano et al. illustrated their *PHYC*-antisense transgenic rice exhibit earlier maturity and flowering, than traditional Nipponbare rice lines in long day (LD) conditions. Research in both Arabidopsis and rice, has elicited that early flowering is the result loss of *PHYC* [103-105].

My research laid the foundation for determining the biological function of *PHYC* in cotton and illuminating *PHYC*'s role as a candidate gene behind photoperiod independence during the domestication of allotetraploid cotton. This study was a PCR-based approach using one low-degeneracy primer and one highly-correlated primer to cotton to obtain amplified gene fragments of *PHYC*'s molecular composition in New World allotetraploid cottons (*Gossypium hirsutum* L. and *Gossypium barbadense* L.) and in Old-World diploids (*Gossypium herbaceum* L. and *Gossypium raimondii* Ulbr.). Finally, markers based on *PHYC* 'candidate gene' amplified fragments may prove useful transferring valuable traits from photoperiodic 'wild' cottons into cultivated elite cotton varieties for crop improvement.

Previously, several sets of degenerate primer pairs for *PHYC* were designed on the conserved HYPATDIP and PFPLRYAC regions, but had failed to produce

amplification fragments during PCR in *Gossypium* [169]. In failing to amplify this region, *PHYC* was proposed to be elusive and possibly not there.

However, the failure to obtain *PHYC* hinge amplification with several sets of both universal (e.g. PHYdeg-F/PHYdeg-R) and rosid specific primers was due to substantial nucleotide differences in *PHYC* of *Gossypium spp.* versus Arabidopsis. This divergence from Arabidopsis was best illustrated in Figure 6 and Table 2 by Abdurakhmonov et. al [169]. The degenerate primers, like PHYdeg-R primer, had many mismatched nucleotides, including transitions and transversions, with the cotton *PHYC* genes. Although mismatches occurred, these changes were located at invariant (e.g. non-degenerate) nucleotide positions and did not alter the amino acid sequence (PFPLRYAC) [169].

During my research, I identified a small EST clone (GenBank CO121409) with similarity to Arabidopsis *PHYC* (E value = 7e-119) in a library from *G. raimondii* floral tissue [170]. Using this EST clone allowed for the development of primer, PHYC_1R_DFCI, within the C-terminal domain to be designed (Table 1) [169].

Table 1 Primers Correlated to *Phytochrome C*

Primer name	Sequence 5' to 3'	Fold-degeneracy
PHYdeg-F	CAYTAYYCIGCIACIGAYATHCC	768
PHYC-1R-DFCI	GGTCCGCCTGATTGAGACTGC	0

I corresponds to inosine. R, Y, M, K, S, W correspond to the IUPAC-IUB ambiguity set.

With different *Gossypium spp.*, we used PHYC_1R_DFCI in combination with PHYdeg-F to amplify a ~1 kb fragment exonic coding sequence from the first exon of *PHYC*, including the hinge region. (Figure 4) [169]. These cloned sequences had a scoring similarity to Arabidopsis *PHYC* at E value ~ 1e-172. The sequences in Sequencher 4.8 [171] created a single consensus contig from each of the following: 1) diploid species 'A' *G. herbaceum*, 2) diploid species 'D' *G. raimondii*, 3) allotetraploid species 'A' *G. hirsutum*, 4) allotetraploid species 'D' *G. hirsutum*, 5) allotetraploid species 'A' *G. barbadense*, and 6) allotetraploid species 'D' *G. barbadense*. An alignment of these consensus sequences for each putative *PHYC* contigs yielded a 1,022 bp alignment with an average pairwise sequence similarity of 99.1%, 1,002 sites (98.0%) identical across all taxa, with no insertions/deletions (InDels) or stop codons in any taxa [169].

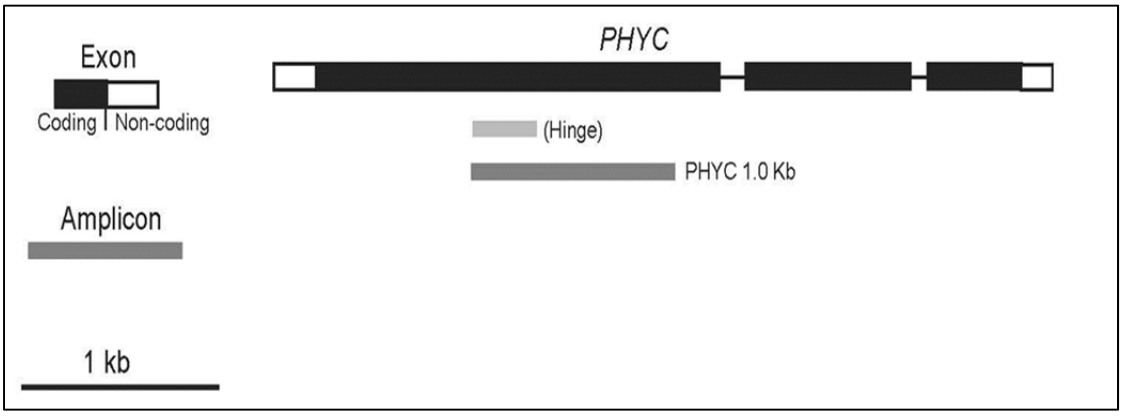


Figure 4 Coding and Non-coding Region for *Phytochrome C*

From a phylogenetic analysis (Figure 5), two major clades ('A' and 'D') emerged from the *PHYC* consensus sequences with 100% bootstrap support [169]. Clade 'A' contained: 1) the diploid species 'A' *G. herbaceum* contig, 2) the allotetraploid species 'A' *G. hirsutum* contig, and 3) the allotetraploid species 'A' *G. barbadense* contig. This clade was designated *PHYC.A*. Clade 'D', designated *PHYC.D*, included: 1) the diploid species 'D' *G. raimondii* contig, 2) the allotetraploid species 'D' *G. hirsutum* contig, and 3) the allotetraploid species 'D' *G. barbadense* contig. This data indicated that the A- and D-genome ancestors had a single copy of the *PHYC* gene. During the ancestral hybridization and polyploidization event, each diploid ancestor contributed a single copy of PHYC to the *Gossypium* allotetraploid ancestor [169].

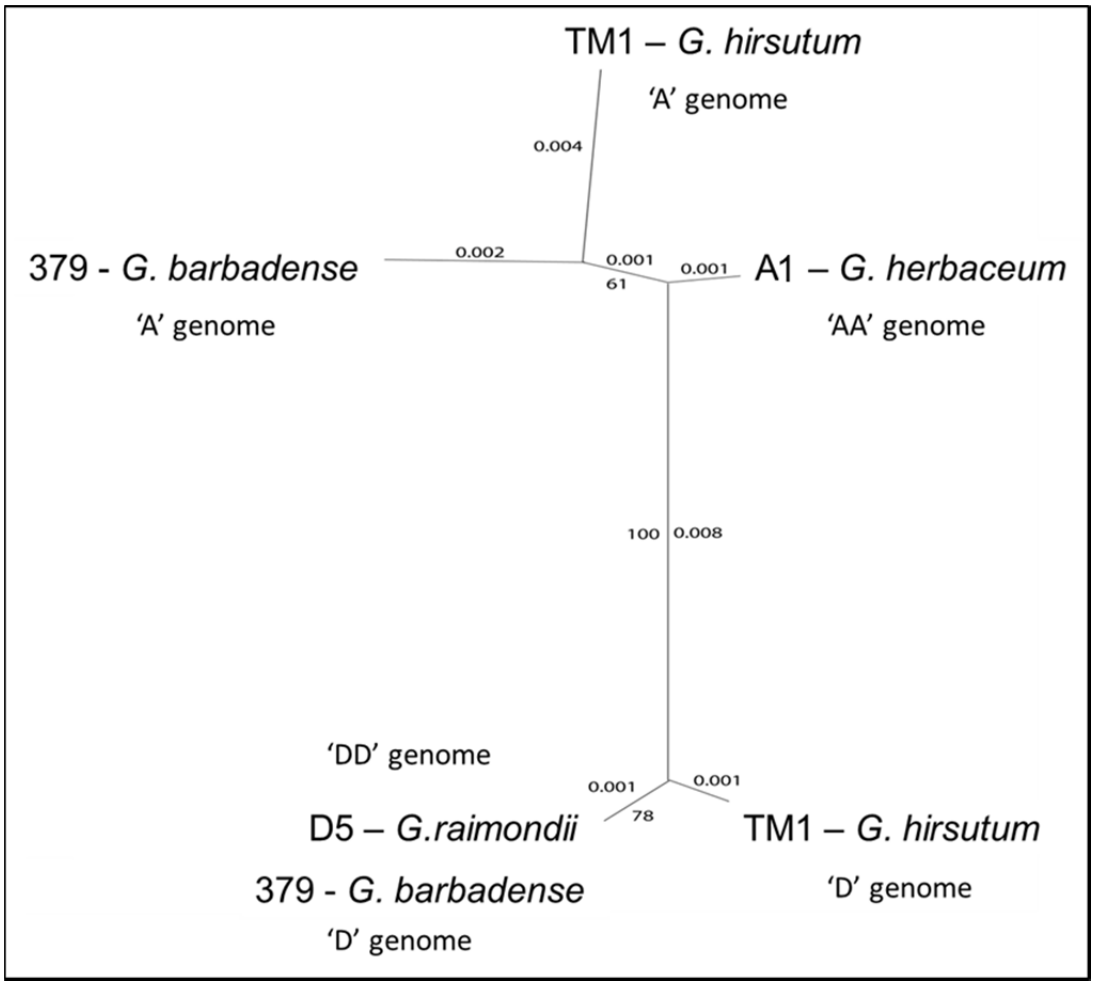


Figure 5 Phylogenetic Clade for *Phytochrome C*

In conclusion, *PHYC* illustrated a resolution between the evolutionary relationships of diploid and allotetraploid cottons. The nucleotide diversity found within the *PHYC* hinge region clarified the evolutionary pattern of inheritance during the

ancestral allotetraploidization event. The two clades, designated *PHYC.A* and *PHYC.D*, define an evolutionary pattern corresponding to purifying selection ($K_A/K_S = 0.184$ over 340 codons) (Table 2) [169]. The rate of evolution in both clades does differ. These findings suggesting the *PHYC.D* clade is evolving more quickly with (8 parsimonious substitutions: 6 non-synonymous, 2 synonymous), while the *PHYC.A* clade has only two parsimonious substitutions (2 synonymous) [169]. In *PHYC.D*, functional divergence with a relaxation on purifying selection may be occurring. Other phytochrome studies have shown faster amino acid evolution occurred in *PHYC* of cultivated *Sorghum bicolor* than those in wild accessions [172]. *PHYC* had a higher K_A/K_S ratio in the C-terminal signaling domain, which may reveal a downstream change in conformation of a protein signaling partner [162-164, 173-175].

Table 2 Synonymous and Non-Synonymous Values for *Phytochrome C* Hinge

Sequence	Comparison	S dif	S pos	K_S	NS dif	NS pos	K_A	K_A/K_S	P
***PHYC* Hinge**	D-D	3	60.83	0.051	1	230.17	0.004	0.086	0.032
	D-T	3.25	60.67	0.055	1.25	230.09	0.005	0.103	0.035
	T-T	3.5	60.5	0.06	1.5	230	0.007	0.119	0.038

D-D indicates means of comparisons within extant diploids, D-T indicates means of comparisons of extant diploids with tetraploids, T-T indicates means of comparisons within tetraploids. S dif, synonymous differences; S pos, synonymous positions; NS dif, non-synonymous differences; NS pos, non-synonymous positions. P indicates significance as determined by Fisher's exact test.

Duplication, Divergence and Persistence in the Phytochrome Photoreceptor Gene Family of Cottons (*Gossypium spp.*)

Overview

Overview Background

Phytochromes are a family of red/far-red photoreceptors that regulate a number of important developmental traits in cotton (*Gossypium spp.*), including plant architecture, fiber development, and photoperiodic flowering. Little is known about the composition and evolution of the phytochrome gene family in diploid (*G. herbaceum*, *G. raimondii*) or allotetraploid (*G. hirsutum*, *G. barbadense*) cotton species. The objective of this study was to obtain a preliminary inventory and molecular-evolutionary characterization of the phytochrome gene family in cotton.

Overview Results

We used comparative sequence resources to design low-degeneracy PCR primers that amplify genomic sequence tags (GSTs) for members of the *PHYA*, *PHYB/D*, *PHYC* and *PHYE* gene sub-families from A- and D-genome diploid and AD-genome allotetraploid *Gossypium* species. We identified two paralogous *PHYA* genes (designated *PHYA1* and *PHYA2*) in diploid cottons, the result of a *Malvaceae*-specific *PHYA* gene duplication that occurred approximately 14 million years ago (MYA), before the divergence of the A- and D-genome ancestors. We identified a single gene copy of *PHYB*, *PHYC*, and *PHYE* in diploid cottons. The allotetraploid genomes have largely retained the complete gene complements inherited from both of the diploid genome ancestors, with at least four *PHYA* genes and two genes encoding *PHYB*, *PHYC* and

PHYE in the AD-genomes. We did not identify a *PHYD* gene in any cotton genomes examined.

Overview Conclusions

Detailed sequence analysis suggests that phytochrome genes retained after duplication by segmental duplication and allopolyploidy appear to be evolving independently under a birth-and-death-process with strong purifying selection. Our study provides a preliminary phytochrome gene inventory that is necessary and sufficient for further characterization of the biological functions of each of the cotton phytochrome genes, and for the development of 'candidate gene' markers that are potentially useful for cotton improvement via modern marker-assisted selection strategies.

Background

Phytochromes are specialized photoreceptors that perceive and interpret light signals from the environment to regulate virtually all aspects of plant development, including seed germination, chloroplast development, tropisms, shade avoidance responses, floral initiation, circadian rhythms, pigmentation, and senescence [162-164]. The phytochromes have a primary role in sensing red (R) and far-red (FR) light, and also play a role in the perception of blue (B) and ultraviolet (UV) light [173]. The active phytochrome molecule consists of a large (~110 kDa) apoprotein bound to a phycobilin chromophore [174, 175]. The phytochrome apoproteins are encoded by a small gene family in all plant taxonomic divisions, including parasitic plants, mosses, cryptogams, and green algae [165, 176-181]. In angiosperms, the phytochrome apoprotein genes have been classified into four or five gene sub-families based on sequence similarity to the

five phytochrome genes of Arabidopsis: *PHYA*, *PHYB*, *PHYC*, *PHYD*, and *PHYE* [63, 64]. All five Arabidopsis phytochromes share an amino acid sequence similarity of 46-56%, with the exception of except *PHYB* and *PHYD*—which are the result of recent gene duplication and share ~80% amino acid identity [63, 182]. Thus, the five Arabidopsis genes are often assigned to four subfamilies: *PHYA*, *PHYB*/*D*, *PHYC*, and *PHYE* [183]. The Arabidopsis *PHYB/D* subfamily is more closely related to *PHYE* gene (~55% nt identity) than to the *PHYA* and *PHYC* genes (~47% nt identity), which together form a separate ancient evolutionary clade [63, 165].

Having presumably arisen by gene duplication and subsequent sub-functionalization and/or neo-functionalization, the phytochrome gene family *in toto* performs a complex network of redundant, partially redundant, non-overlapping, and in some cases antagonistic regulatory functions throughout plant development [65, 68-70, 103, 184-196]. For example, all Arabidopsis phytochromes play diverse and interacting roles in photoperiodic regulation of floral initiation. *PHYA*, *PHYB*, *PHYD* and *PHYE* act partially redundantly in the light-dependent entrainment of the circadian clock [65, 66], which in turn regulates transcription of the floral inducer *CONSTANS* (*CO*) in a circadian manner [166]. In Arabidopsis, *PHYA*, in conjunction with blue-light dependent cryptochrome photoreceptors *CRY1* and *CRY2,* promotes flowering by inhibiting the degradation of *CO* protein, while *PHYB* acts antagonistically to stimulate *CO* degradation [67]. In addition, *PHYB*, *PHYD* and *PHYE* act partially redundantly as repressors of flowering that are dependent on R/FR ratio [68-71]. In this role, *PHYB* also acts downstream of *CO* as a negative regulator of transcription of the ‘florigen’ molecule

FT (the target of *CO*) in a tissue specific manner [102]. Mutant analyses indicate that *PHYC* also plays a role in photoperiodic flowering [103, 104]. Further, genetic variation at the *PHYC* locus underlies some of the natural phenotypic variation in flowering time in Arabidopsis [73, 106].

In angiosperms, the composition of phytochrome gene family varies significantly among taxonomic lineages. Although a single *PHYA* gene is present in most flowering plants, some plant families, such as carnation (Carryophyllaceae) and legumes (Fabaceae), have two distinct *PHYA* genes [179]. Similarly, several plant lineages have gained multiple *PHYB*-like genes through independent gene duplications of *PHYB* [63, 179, 183, 197-200]. For example, tomato has two *PHYB* genes (designated *PHYB*1 and *PHYB*2) that are not directly orthologous to Arabidopsis *PHYB* and *PHYD,* respectively [197]. While most angiosperms have a single *PHYC* gene, species in some families such as *Fabaceae* and *Salicaceae* appear to have lost *PHYC* during evolution [179, 200]. Although a single *PHYE*-like gene is present in most flowering plants, *PHYE* is completely absent in poplar (*Salicaceae*), in the *Piperales*, and some monocots such as maize [179, 200]. Finally, the novel *PHYF* subfamily, which groups with *PHYA*/*C* clade, has been identified in tomato [197].

Little is known about the composition of the phytochrome gene family in cultivated cottons or their wild relatives (*Gossypium spp.*) in the *Malvaceae* family. This is despite the fact that physiological experiments suggest that phytochromes regulate economically important aspects of cotton development, including drought resistance, seed dormancy**,** plant architecture, photoperiodic flowering, and fiber elongation [108-

111]. For example, R/FR photon ratio influences the length and diameter of developing seed fiber; fibers exposed to a high R/FR photon ratio during development were longer than those that received lower R/FR ratio, implicating the involvement of a phytochrome [110, 111].

While modern domesticated varieties of the major cultivated cottons *G. hirsutum* L. and *G. barbadense* L. exhibit photoperiod independent flowering, wild and 'primitive' accessions of *G. hirsutum* and *G. barbadense* flower under short-day photoperiodic control [48, 49]. An understanding of the molecular-genetic basis of differences in photoperiodic flowering in cottons will accelerate strategies for improvement of cultivated varieties through the introgression of valuable genetic traits from wild germplasm [48, 49]. In this regard, it is important to note that mutational changes in phytochrome function have been implicated in the loss of photoperiod sensitivity in several major crops including sorghum, barley, rice, and soy [112-115].

A thorough characterization of the phytochrome gene family in cotton species is necessary for understanding the molecular basis of photoperiodic flowering, the influences of light quality on cotton fiber elongation, and other aspects of cotton development. Any inventory of phytochrome genes of cottons is complicated by the fact that the major cultivated species, *G. hirsutum* and *G. barbadense* are allotetraploids. Diploid species in the genus *Gossypium* are categorized into eight genome groups (designated A through G, and K) based on cytogenetic and phylogenetic criteria [35, 40, 201-203]. The old-world A genome group and the new world D genome group diverged from each other on the order of 1-7 MYA [35], then underwent hybridization and

polyploidization creating an AD allopolyploid lineage ancestral to *G. hirsutum* (designated AD_1) and *G. barbadense* (designated AD_2) on the order of 1 MYA [38, 203].

In this study, we utilized a PCR-based approach with low-degeneracy primers to obtain gene fragments, or 'genome sequence tags' (GSTs) that yield an initial description of the composition and evolution of the phytochrome gene family in the New World allotetraploid cottons *Gossypium hirsutum* and *G. barbadense*, and in the Old-World diploids *G herbaceum* L. and *G. raimondii* Ulbr., which are considered to be extant relatives of the A- and D-genome diploid ancestors (respectively) of the allotetraploid lineage. This study provides a necessary foundation for studies of the specific biological functions of each of the phytochrome genes in cotton species, and helps to illuminate the evolutionary patterns of duplicated genes in complex genomes, as well as the evolutionary history of the world's most important fiber crop species.

Results

Because our results were derived from PCR, our inventory of the phytochrome gene family in *Gossypium spp.* is provisional. All sequences have been submitted to GenBank (accession numbers HM143735-HM143763).

Phytochrome Hinge Amplification using 'Universal' Primers

Between N-terminal 'photoperception domain' and C-terminal 'signaling domain' of the phytochrome apoprotein is a short 'hinge region' (Figure 6) that shows relatively high sequence variation, and has proven useful for characterization of the phytochrome gene complement in a variety of plant species, and for robust phylogenetic analyses [179]. To amplify the hinge region of all cotton phytochromes, we used an

alignment of eudicot phytochrome sequences to design a 768-fold degenerate PCR primer (designated PHYdeg-F) based on the conserved HYPATDIP peptide in the N-terminal domain, and a 16,384-fold degenerate PCR primer (designated PHYdeg-R), based on the conserved PFPLRYAC peptide in the C-terminal domain (Table 3).

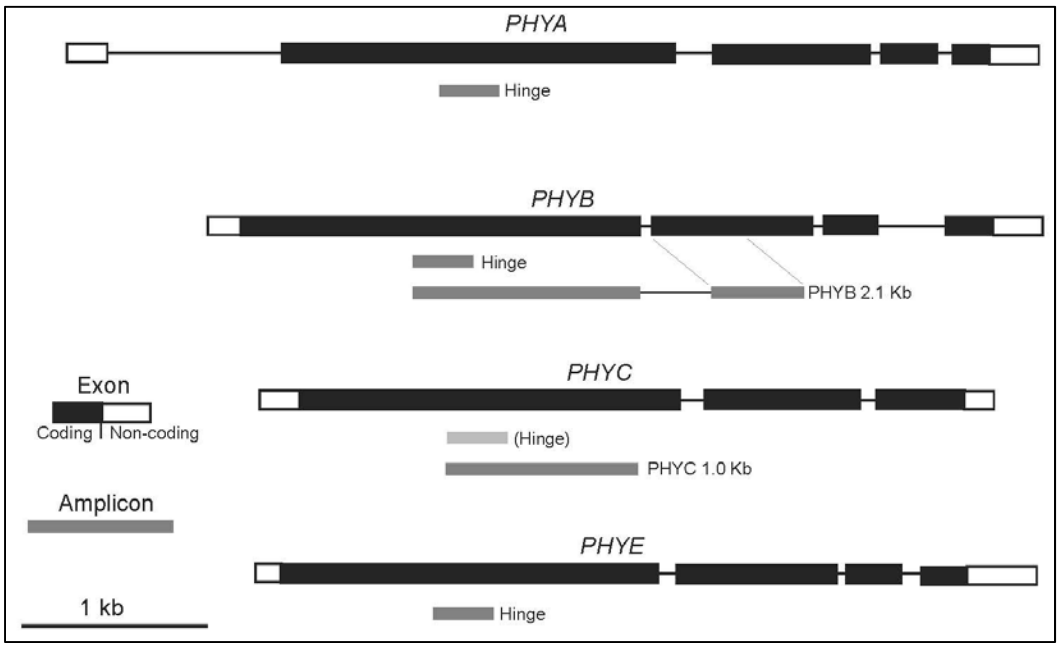


Figure 6 Phytochrome Coding and Non-coding Regions Compared with Sequenced Amplicon Region.

Table 3 Primers used to Amplify Cotton Phytochrome Gene Family.

Primer Name	Sequence 5' to 3'	Fold-Degeneracy
PHYdeg-F	CAYTAYYCIGCIACIGAYATHCC	768
PHYdeg-R	CRCAIGCRTAICKARIGGRWAIGG	16,384
PHYABnondeg-F	GCATTATCCTGCTACTACTGATATT	0
PHYAdeg-R	CAWGCATACCTWAGMGGRAAI	64
PHYBdeg-R	AACAACIAIICCCCAIAGCCTCAT	64
1010-F	GTTYTTGTTTAAGCARAACCG	4
1910-R	GAGTCWCKCAGAATAAGC	4
1910-F	AGCTTATTCTGMGWGACTC	4
2848-R	TAACCCKCTTRTTTGCAGTCA	2
PHYC-1R-DFCI	GGTCCGCCTGATTGAGACTGC	0

I corresponds to inosine. R, Y, M, K, S, W correspond to the IUPAC-IUB ambiguity set.

Amplification across the hinge region using Taq DNA polymerase yielded PCR products from all taxa. We cloned the amplification products from each taxon into an *E. coli vector*, then sequenced ~40 clones for each taxon. For all taxa, a majority (>60%) of clones showed the highest similarity in BLAST searches to Arabidopsis *PHYE* (*E* value ~ $1e^{-40}$). For each taxon, only a minority of clones showed high-scoring similarity to

Arabidopsis *PHYA* or *PHYB*. This apparently skewed distribution of amplification products — observed across all taxa — suggested an amplification bias in favor of *PHYE* amplicons. No clones were obtained from any taxon that had high-scoring similarity to Arabidopsis *PHYC* or *PHYD.* No new phytochrome sub-families were observed.

Amplification of the *PHYA* Gene Sub-Family

Because of possible biased amplification, we designed new less-degenerate hinge-region primer sets for the *PHYA, PHYB/D*, and *PHYC* sub-families (Table 3) using available phytochrome sequences from species in the rosid clade, which includes both cotton and Arabidopsis [204, 205].

The hinge regions of *PHYA* genes were amplified using PHYABnondeg-F and PHYAdeg-R (Table 1), yielding a ~360 bp amplification product from all accessions. In BLAST database searches, all clones had a high-scoring pair relationship with Arabidopsis *PHYA* (*E* value ~ $2e^{-63}$). Sequences from a total of more than 200 clones across all taxa yielded two distinct consensus contigs from each of the diploids *G. herbaceum* and *G. raimondii*, and four distinct contigs from the allotetraploids *G. barbadense* and *G. hirsutum*. When aligned across all taxa, these contigs yielded a 315 bp consensus alignment that had an average pairwise sequence similarity of 94.6%, with 282 sites (89.5%) identical across all taxa, and no stop codons or InDels in any taxa. Distance analysis (Figure 7) showed two well-separated gene sub-clades (100% bootstrap support). These sub-clades were designated *PHYA1* and *PHYA2*. The level of hinge-region differentiation between these two sub-clades was far greater than that seen

in other cotton phytochrome gene sub-families (discussed below), with an uncorrected "p" distance of 0.086, corresponding to 28 nt changes (9%) based on parsimony.

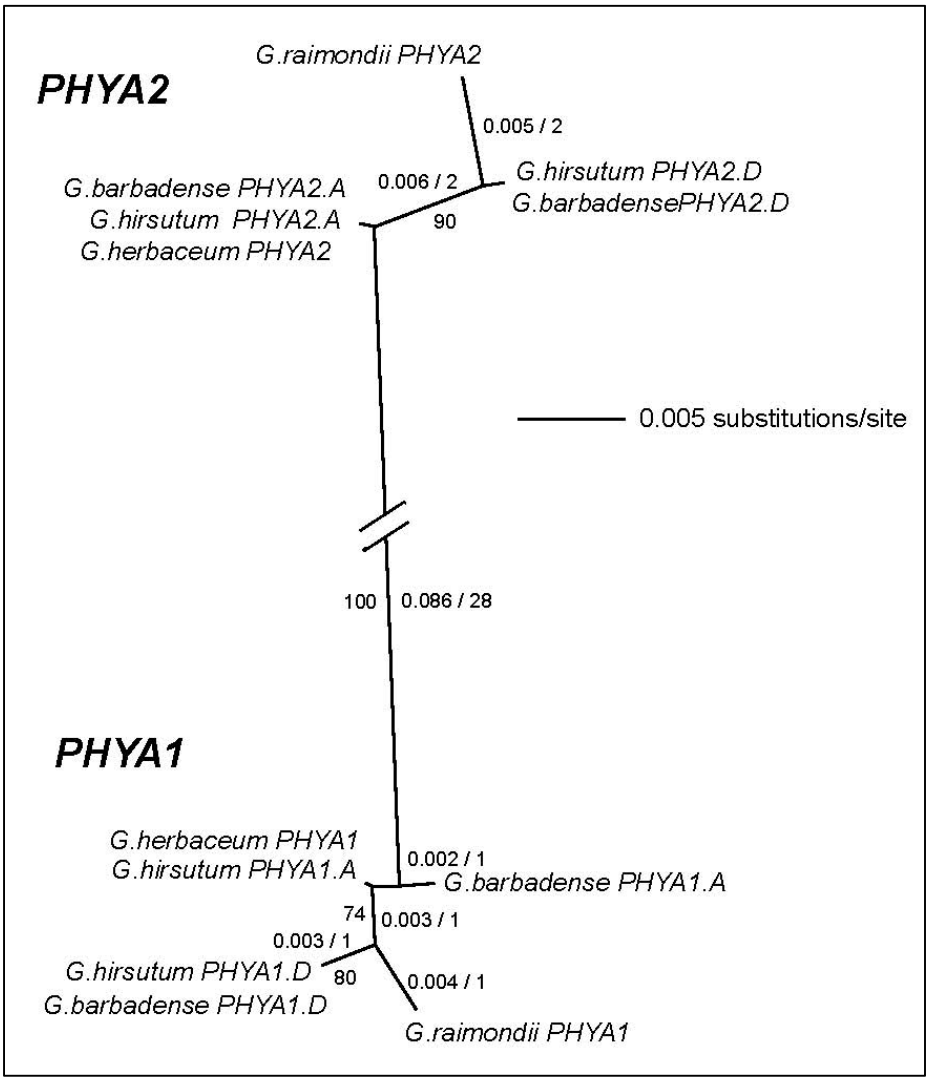


Figure 7 Phylogenetic Divergence of *Phytochrome A* in Cotton

These data indicated that a single *PHYA* gene underwent duplication after the divergence of the cotton and Arabidopsis lineages, but prior to the divergence of A-genome and D-genome lineages, leaving each of the modern diploids in our study (and presumably the ancestors to the AD allotetraploids) with a complement of two *PHYA* paralogs (*PHYA-1* and *PHYA-2*). Indeed, four distinct contigs were observed in both the inbred *G. hirsutum* cultivar TM-1 and in the doubled-haploid line *G. barbadense* 3-79. For each allotetraploid taxon, two contigs fell into each of the *PHYA-1* and *PHYA-2* clades (Figure 7). A conservative inventory of available EST sequences indicated that at least two distinct *PHYA* loci are expressed in *G. hirsutum* (Supplemental Table 1).

Within each of the *PHYA1* and *PHYA2* clades, the level of nucleotide diversity was very low, with at most four parsimonious nucleotide changes separating each contig. However, within the *PHYA1* clade, the contigs resolved into two subclades (74% bootstrap support) that each contained a single contig from one of the diploid taxa and one contig from each of the allotetraploids. For example, *G. raimondii* (D-genome) *PHYA1* grouped in a single contig from each of *G. hirsutum* and *G. barbadense.* Based on this grouping, the latter contigs were assigned the provisional designation of *PHYA1.D*. Similarly, *G. herbaceum* (A-genome) grouped with *G. hirsutum PHYA1.A* and *G. barbadense PHYA1.A*. Based on similar criteria, the *PHYA2* clade was also divided into *PHYA2.A* and *PHYA2.D* subclades (90% bootstrap support). The phylogenetic resolution of A- and D-genome subclades supported the hypothesis that each of the A- and D-genome diploids contributed both *PHYA1* and *PHYA2* to the allotetraploid lineage. Thus, although hinge-region nucleotide diversity within each of

the *PHYA1* and *PHYA2* clades was low, it was sufficient to resolve a tentative *PHYA* gene complement for each taxon, as well as the pattern of gene inheritance through the allopolyploidization event.

Amplification of the *PHYB/D* Gene Sub-Family

A ~320 bp fragment from the *PHYB/D* hinge region was obtained by amplification using primers PHYABnondeg-F and PHYBdeg-R (Table 3). Sequences from a total of 80 clones yielded a single consensus contig from each of the diploid cottons *G. herbaceum* and *G. raimondii*, and from the allotetraploid *G. hirsutum*. Two distinct contigs were assembled from clones derived from the allotetraploid *G. barbadense*. These clone sequences shared ~85% nucleotide identity with the Arabidopsis *PHYB* gene and ~78% nt identity with Arabidopsis *PHYD*. All clones had a high-scoring pair relationship with the Arabidopsis *PHYB* gene (*E* value ~ $1e^{-71}$) as well as significant similarity to the Arabidopsis *PHYD* gene (*E* value ~ $3e^{-55}$). Consensus sequences were aligned across all taxa, yielding a 319 bp alignment with an average pairwise sequence similarity of 99.8%, with 317 sites (99.4%) identical across all taxa, no stop codons and no InDels. Although these data indicated the presence of at least one *PHYB* gene in each of the A- and D-genome diploid plants and in *G. hirsutum*, and at least two genes *PHYB* genes in the *G. barbadense*, the low level of nucleotide differentiation observed within the hinge region yielded insufficient phylogenetic information to characterize the *PHYB* gene complement in any of the study taxa.

To obtain better resolution of the *PHYB* gene complement, additional low degeneracy primers 1010-F, 1910-F, 1910-R, and 2840-R (Table 4) were used along with primer PHYABnondeg-F to create a 2.1 kb long series of overlapping amplicons corresponding to approximately 1.8 kb of the Arabidopsis *PHYB* gene and extending from the hinge, through the first intron and into the second exon (Figure 6). After amplification, cloning and sequencing, the amplicons were assembled for each taxon. In all *Gossypium* taxa examined, the first intron was ~300 bp longer than the first intron of *PHYB* from Arabidopsis.

Unlike the other phytochrome amplicons, we detected a high frequency of PCR-mediated recombination events within the *PHYB2.1* kb fragment resulting from amplifications using *G. barbadense* as template. The recombination detection algorithm RDP3 [206] identified a number of clones resulting from apparent recombination between the A-genome and D-genome derived homeologous sequences, with predicted breakpoints ($P = 0$) between nucleotides 1000 and 1700 of the alignment. After omission of these recombinant clones, composite amplicon sequences from each taxon were aligned, creating a consensus alignment of 2,061 bp with 98.8% average pairwise similarity and 2,007 identical sites (97.4%). Overall, the cotton *PHYB* genes shared 65% nucleotide identity with the Arabidopsis *PHYB* ortholog. No stop codons or InDels were detected in exon sequences. A 2 bp putative deletion was observed in one contig (designated *PHYB.D*) from *G. hirsutum*. In addition, a 1 bp indel was polymorphic between the PHYB.A and PHYB.D clades. Finally, *PHYB* of *G. raimondii* had an additional 1 bp insertion. All InDel polymorphisms were located within first introns.

Table 4 Nucleotide Divergence in Phytochrome Genes in Comparisons of A- and D-Genome Derived Homeologs in Diploid and Allotetraploid Cottons

Sequence	Comparison	S dif	S pos	K_s	NS dif	NS pos	K_A	K_A/K_s	*P*
PHYA1 Hinge	D-D	2	67.33	0.303	0	241.67	0	**0**	**0.049**
	D-T	2.25	67.17	0.102	0	241.84	0	**0**	**0.039**
	T-T	2.5	67	0.038	0	242	0	**0**	**0.03**
PHYA2 Hinge	D-D	1	66.42	0.015	3	242.58	0.125	**8.224**	**0.622**
	D-T	1	66.38	0.015	2	242.63	0.065	**4.247**	**0.504**
	T-T	1	66.33	0.015	1	242.67	0.004	**0.27**	**0.386**
PHYB 2.1 kb	D-D	8	377.33	0.022	7	1293.67	0.005	**0.251**	**0.01**
	D-T	9	377.34	0.024	8	1293.63	0.006	**0.256**	**0.006**
	T-T	10	377.42	0.027	9	1293.58	0.007	**0.3**	**0.004**
PHYC Hinge	D-D	3	60.83	0.051	1	230.17	0.004	**0.086**	**0.032**
	D-T	3.25	60.67	0.055	1.25	230.09	0.005	**0.103**	**0.035**
	T-T	3.5	60.5	0.06	1.5	230	0.007	**0.119**	**0.038**
PHYC 1.0 kb	D-D	7	224.67	0.032	3	795.33	0.004	**0.12**	**0.002**
	D-T	8	224.84	0.037	4.5	795.17	0.006	**0.156**	**0.003**
	T-T	9	225	0.041	6	795	0.008	**0.184**	**0.002**
PHYE Hinge	D-D	4	60.42	0.069	1	206.58	0.005	**0.071**	**0.012**
	D-T	3.75	60.46	0.065	0.5	206.54	0.003	**0.042**	**0.015**
	T-T	3.5	60.5	0.06	0	206.5	0	**0**	**0.008**

D-D indicates means of comparisons within extant diploids, D-T indicates means of comparisons of extant diploids with tetraploids, T-T indicates means of comparisons within tetraploids. S dif, synonymous differences; S pos, synonymous positions; NS dif, non-synonymous differences; NS pos, non-synonymous positions. P indicates significance as determined by Fisher's exact test.

Detailed phylogenetic analyses of the 2,061 bp contigs from A-, D-, and AD–genome cottons (Figure 8) indicated the presence of least one *PHYB* locus in the two diploid cottons, *G. herbaceum* and *G. raimondii*, and at least two *PHYB* loci in both allotetraploid cottons. The *G. hirsutum* and *G. barbadense* sequence contigs each grouped into two sub-clades (tentatively designated *PHYB.A* and *PHYB.D*). The single *PHYB* contig from *G. herbaceum* was used to define the *PHYB.A* cluster (99% bootstrap support), while the single *PHYB* contig from *G. raimondii* anchored the *PHYB.D* cluster. From these results, we concluded that *PHYB.A* and *PHYB.D,* which shared ~98% nucleotide sequence identity, arose as orthologs at the time of divergence of the A- and D-genome diploid lineages. We surmised that *PHYB.A* was contributed to the allotetraploids via the A-genome ancestor and *PHYB.D* was contributed via the D-genome ancestor. Available EST sequences indicated that at least one *PHYB* locus is expressed in *G. hirsutum* (Supplemental Table 1).

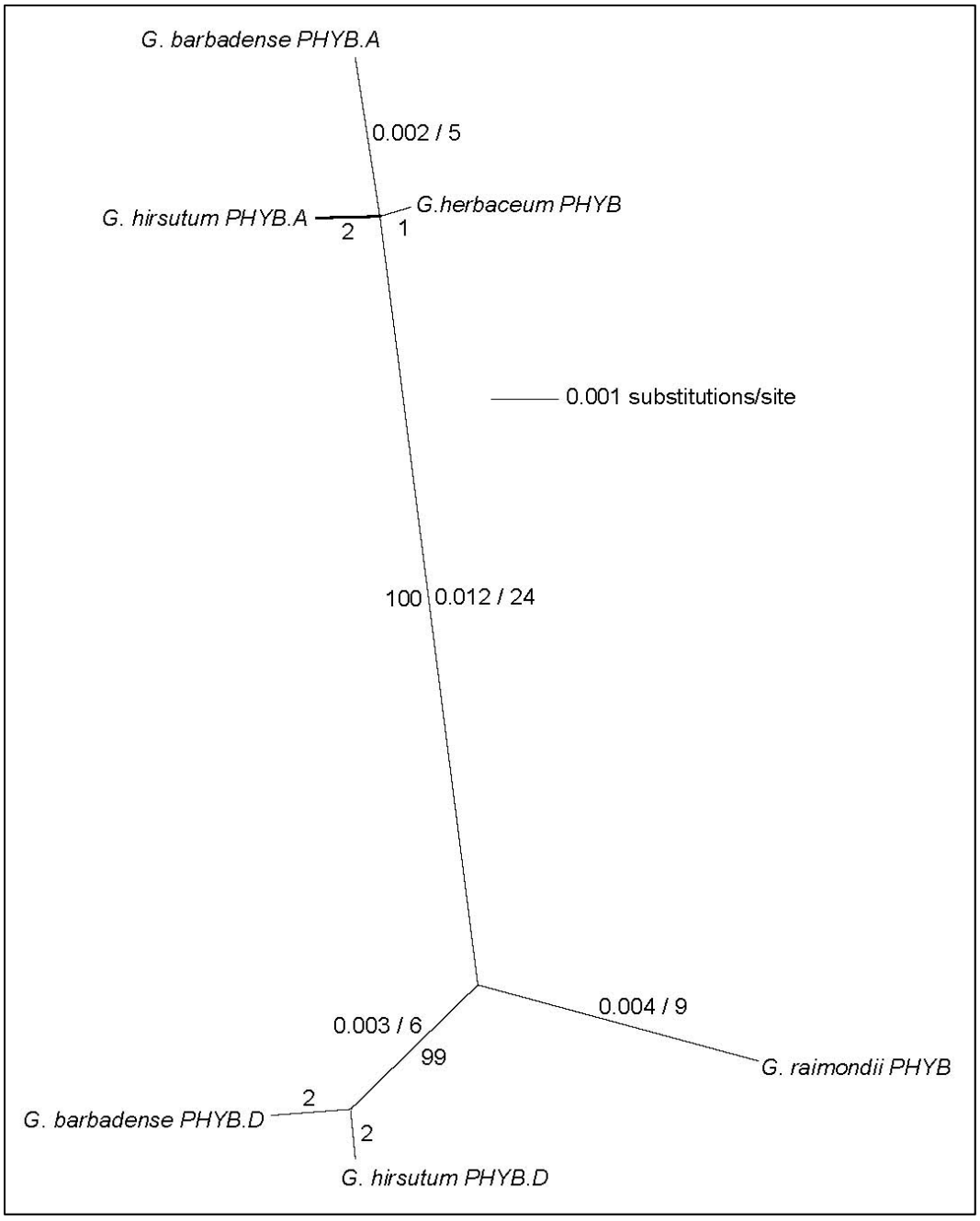


Figure 8 Phylogenetic Divergence of *Phytochrome B* in Cotton

Amplification from the *PHYC* Gene Sub-Family

Several sets of degenerate primer pairs that were designed on the basis of the conserved HYPATDIP and PFPLRYAC regions — including several designed from rosid *PHYC* nucleotide sequences — failed to produce detectable PCR amplification products from the *Gossypium* species tested (data not shown). However, the identification of a small EST clone (GenBank CO121409) with similarity to Arabidopsis *PHYC* (*E* value = $7e^{-119}$) in a library from *G. raimondii* floral tissue [170], allowed us to design the primer PHYC_1R_DFCI within the C-terminal domain (Table 3). When used in combination with PHYdeg-F, this primer amplified a ~1 kb fragment composed entirely of coding sequence from the first exon of *PHYC*, including the hinge (Figure 6). All clones obtained using this primer pair had a high-scoring similarity to Arabidopsis *PHYC* (*E value* ~ $1e^{-172}$). From these clones, we assembled a single consensus contig from each of the diploid species *G. herbaceum* and *G. raimondii*, and two distinct consensus contigs from each of the allotetraploids *G. hirsutum* and *G. barbadense*. Consensus sequences for each of the putative *PHYC* contigs were aligned across all taxa, yielding a 1,022 bp alignment with an average pairwise sequence similarity of 99.1%, 1,002 sites (98.0%) identical across all taxa, with no indels or stop codons in any taxa.

In phylogenetic analyses (Figure 9), the *PHYC* consensus sequences grouped into two major clades (100% bootstrap support). One of these clades contained the *G. herbaceum* contig and one contig from each of *G. hirsutum* and *G. barbadense*. This clade was designated *PHYC.A*. The other clade, designated *PHYC.D*, included the *G. raimondii* contig along with the other of the two contigs from each of *G. hirsutum* and

G. barbadense. These data indicated that both the A- and D-genome ancestors had one *PHYC* gene, and that upon hybridization and polyploidization, this gene was contributed from each diploid to the allotetraploid ancestor of *G. hirsutum* and *G. barbadense*.

For comparison with the other phytochromes, we also analyzed a portion of the *PHYC* alignment corresponding to the hinge region only. This alignment was 296 nucleotide pairs in length, with pairwise sequence similarity of 99.0%, 290 sites (98.0%) identical across all taxa, with no InDels. Although it encompassed fewer variable nucleotides, NJ analysis of the hinge region alone could be used to differentiate the *PHYC.A* and *PHYC.D* clades (100% bootstrap support) and to infer the composition and evolutionary inheritance of the *PHYC* gene family in cottons (data not shown).

Our failure to obtain *PHYC* hinge amplification with several sets of both universal (e.g. PHYdeg-F/PHYdeg-R) and rosid specific primers was entirely due to substantial nucleotide differentiation in *PHYC*, particularly within the hinge region. For example, the 24 nt long PHYdeg-R primer had six nucleotide mismatches with the cotton *PHYC* genes, including three transitions and three transversions. Five of the six mismatches occurred at what are considered to be invariant (e.g. non-degenerate) nucleotide positions. It should be noted that these divergent nucleotides in the conserved primer-binding site did not alter the amino acid sequence (PFPLRYAC).

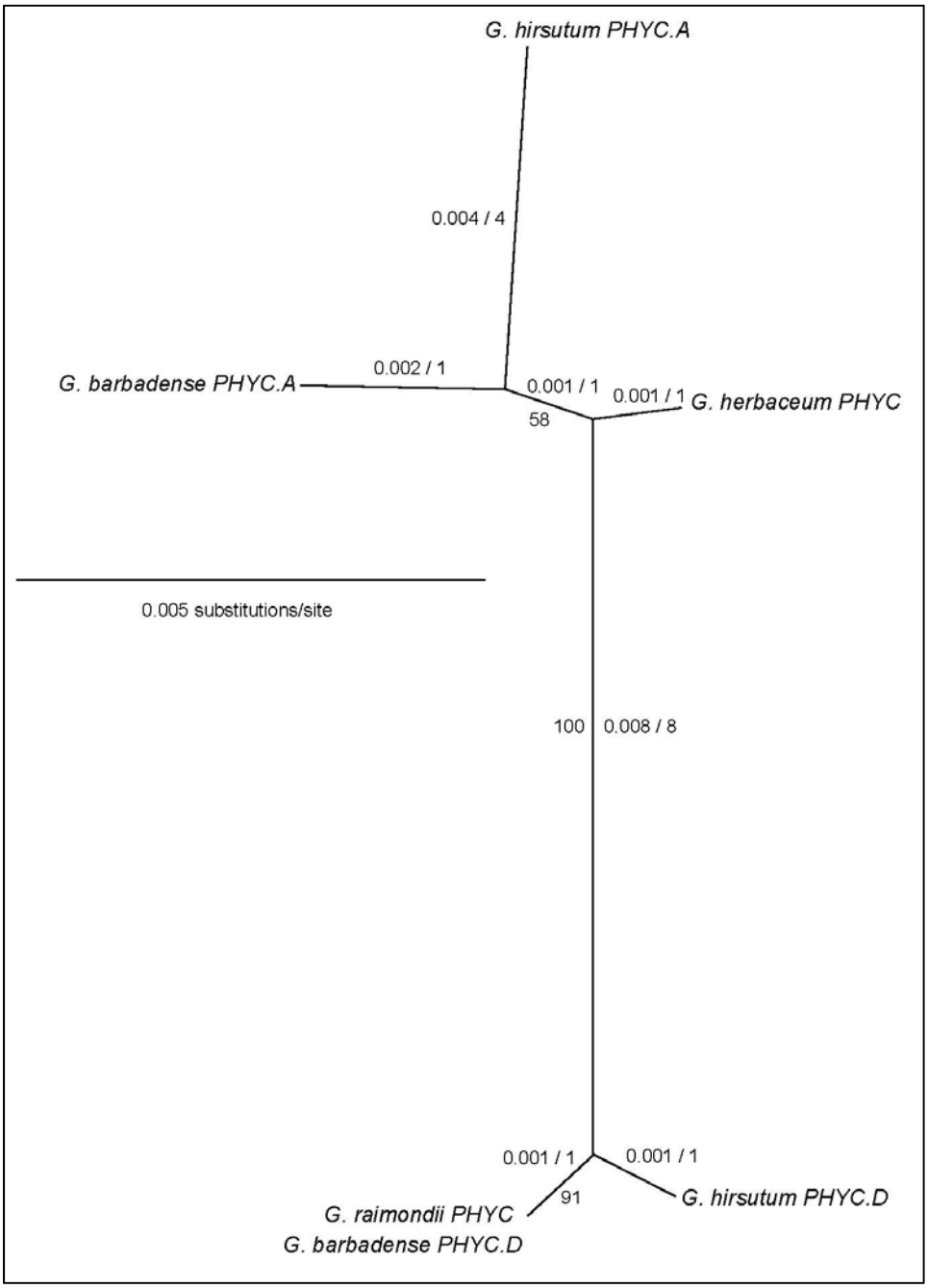


Figure 9 Phylogenetic Divergence of *Phytochrome C* in Cotton

The *PHYE* Gene Sub-Family

PHYE hinge region consensus contigs from our study taxa formed a 270 bp alignment with an average pairwise similarity of 98.9%, with 264 (97.8%) invariant sites, no InDels, and no stop codons in any taxa. The consensus of the aligned *PHYE* sequences had 80% nucleotide similarity to the corresponding fragment of the Arabidopsis *PHYE* gene. Based on maximum parsimony, nucleotide diversity in the cotton *PHYE* hinge sequences could be explained by a minimum of six nucleotide changes, all of which were synonymous. NJ analysis of the cotton *PHYE* hinge region showed two distinct clades (97% bootstrap support) corresponding to the A- and D-genome derived orthologs (designated *PHYE.A* and *PHYE.D*), a finding consistent with a hypothesis in which each diploid ancestor contributed a single *PHYE* ortholog to the allotetraploid lineage (Figure 10). Interestingly, while two distinct *PHYE* contigs were obtained from *G. hirsutum*, only a single contig, which grouped with the D-genome clade, was obtained from *G. barbadense.* Available EST sequences indicated that at least one *PHYE* locus is expressed in *G. hirsutum* (Supplemental Table 1).

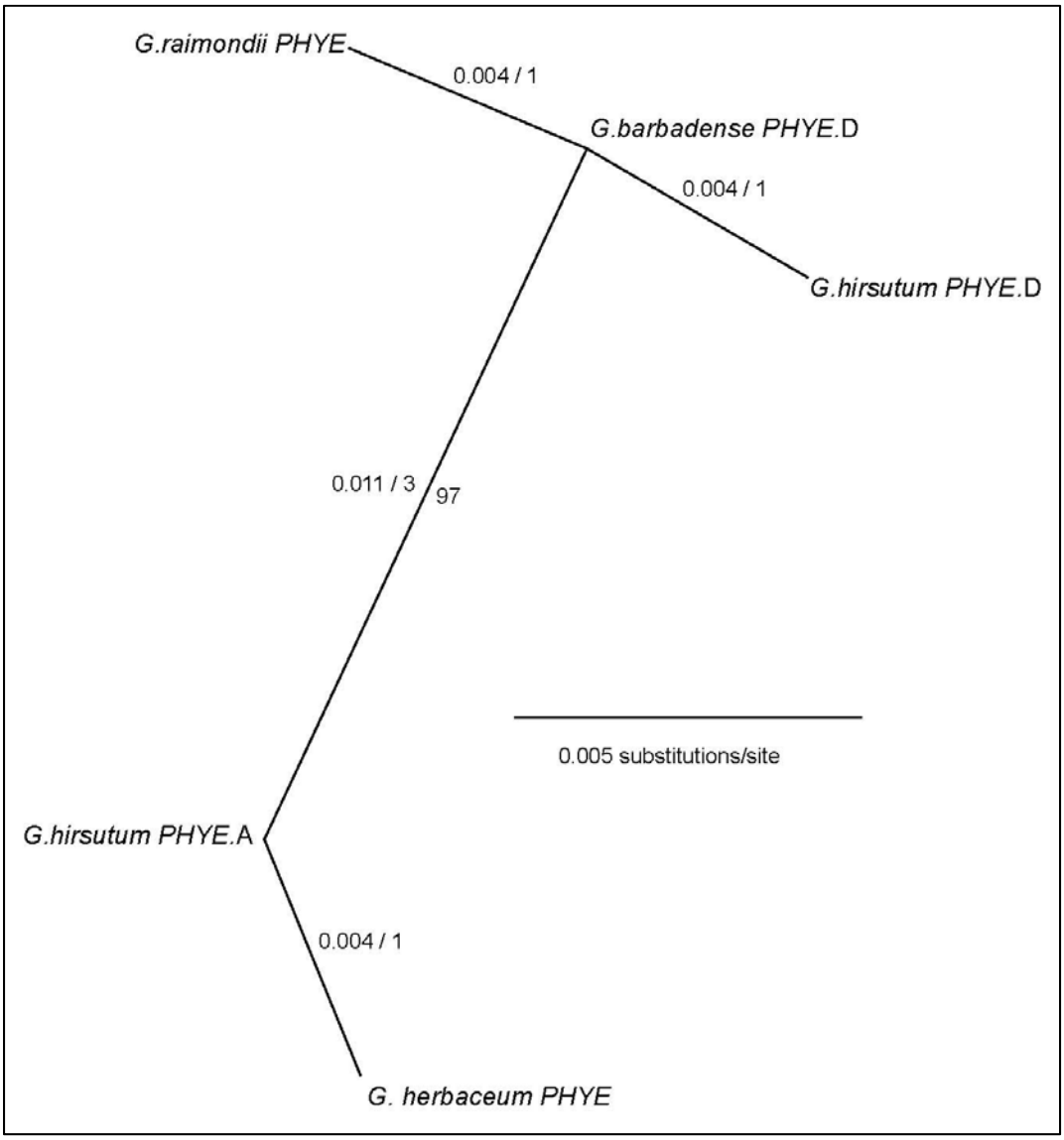


Figure 10 Phylogenetic Divergence of *Phytochrome E* in Cotton

A Global Hinge-Based Alignment of Arabidopsis and Cotton Phytochromes

PHYA, *PHYB*, *PHYC* and *PHYE* hinge regions from Arabidopsis and *Gossypium spp.* were aligned to create a global phytochrome alignment 358 nucleotides in length, with an average pairwise similarity of 69.4% and 123 identical sites (34.4%). The gene phylogeny generated from this alignment (Figure 11) reflected divergence of *PHYA*, *PHYB*, *PHYC* and *PHYE* as a result of speciation (nodes 1A, 1B, 1C and 1E, respectively) and gene duplication (nodes 2 and 3). The level of nucleotide divergence of each of the gene sub-families after nodes 1A, 1B, 1C and 1E (Kimura 2-parameter distances) was similar, with a mean of 0.297 ± 0.21 nucleotide substitutions per site. However, the synonymous (K_S) and non-synonymous (K_A) substitution rates were both significantly more variable among the various gene sub-families defined by nodes 1A, 1B, 1C and 1D than were simple nucleotide distances (Table 4). Despite this variation, all sub-families showed a K_A/K_S ratio < 0.1, implying that each remains under purifying selection for function. Further, excessively long branch-lengths, which are often found in pseudogenes, were not observed. In the *PHYB, PHYC* and *PHYE* clades, the branch lengths leading to the Arabidopsis orthologs, which have known biological functions, were longer than the branches leading to their respective cotton orthologs. Considered together, these lines of evidence indicate that each of the phytochrome sub-families retains some biological function in *Gossypium,* as they do in Arabidopsis [63, 64, 68-70, 103, 182, 184-194]. Further, our topology supports the conclusion that *PHYD* is the result of a relatively recent gene duplication that may be exclusive to the *Brassicaceae* family [182].

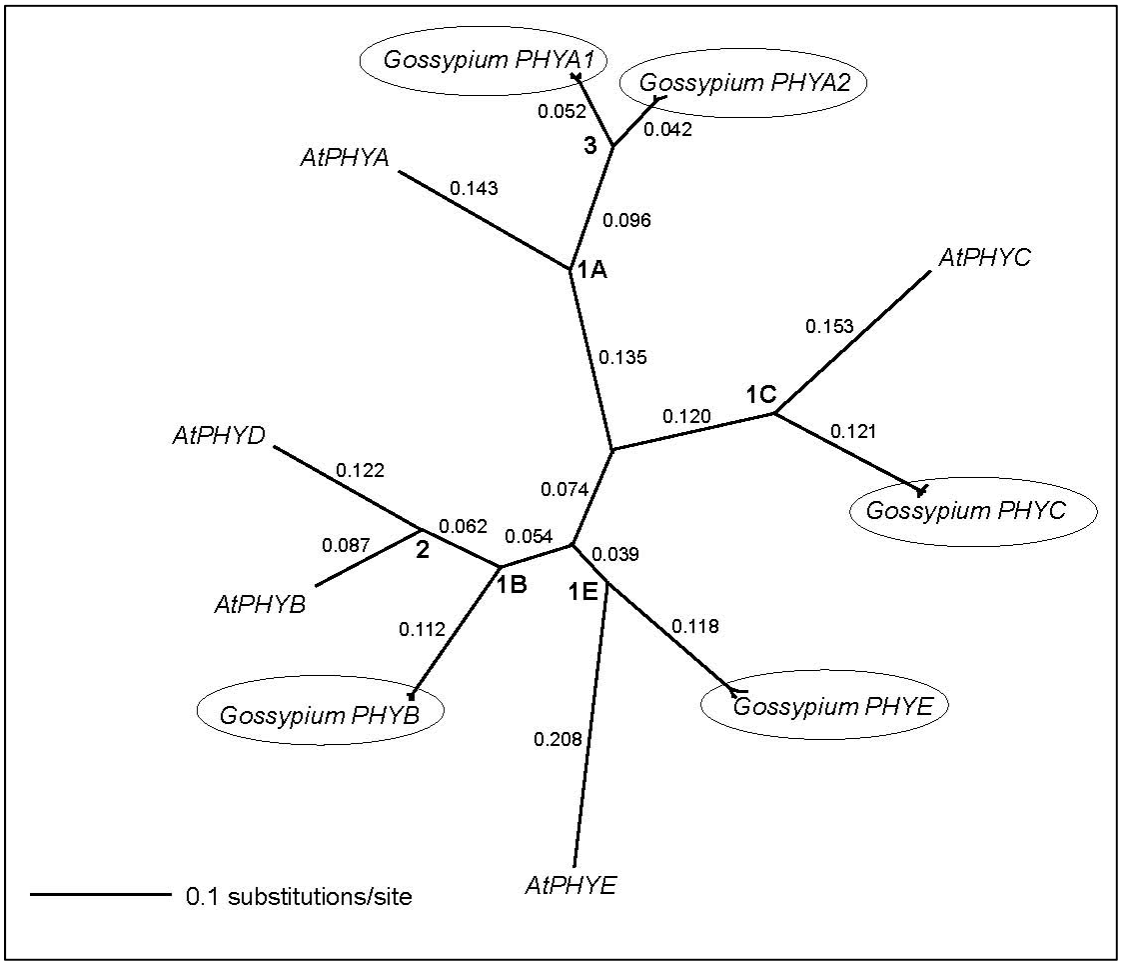


Figure 11 Phylogenetic Divergence of All Phytochromes in Cotton and Arabidopsis

Discussion

Resolution of the Phytochrome Gene Family

In three out of four cases, we were able to successfully resolve the inventory and evolutionary relationships of the phytochrome genes in diploid and allotetraploid cottons using the hinge region only. This finding supports the general utility of employing the hinge region for identifying GSTs for phytochromes. In only one case (*PHYB*) was additional gene sequence required for sufficient phylogenetic resolution. In another case (*PHYC*), nucleotide divergence at a commonly used primer-binding site prevented the characterization of the hinge region by the typical strategy of using primers based on conserved flanking peptides HYPATDIP and PFPLRYAC. However, nucleotide diversity within the *PHYC* hinge region itself was sufficiently informative to resolve the pattern of evolutionary inheritance through allotetraploidization event.

The sequencing of phytochrome gene fragments from A- and D-genome diploids, as well as from AD allotetraploid taxa, provides an essential foundation for all subsequent analysis of phytochrome function and evolution in *Gossypium*. The sequenced fragments provide sufficient information (at least two diagnostic nucleotide characters) to unequivocally identify or 'tag' various orthologs, homeologs and paralogs, as well as monitor their patterns of nucleotide divergence, and trace their evolutionary inheritance through the allopolyploidization event. This information will serve as a foundation for further sequence assembly and annotation, and will be used to design locus-specific primer sets for quantitative RT-PCR assays that will measure transcript levels for each gene family member. In some cases (e.g. *PHYA1* vs. *PHYA2*) levels of

sequence divergence are high enough to support studies of gene function using RNAi or amiRNA approaches to create gene-specific knockouts [170]. The use of well characterized 'candidate genes' of agronomic interest is becoming an integral component of marker-assisted selection efforts in plants [207]. Several SNP-based molecular markers [208, 209] are now being developed using the diagnostic nucleotide characters identified in this study, and are being mapped in experimental cotton populations that show segregation of phytochrome-controlled traits such as fiber length and flowering time.

The Ancestral Phytochrome Gene Complement of the *Malvales* and *Brassicales*

Our study indicated that the diploid ancestors to the world's major fiber crops (*G. hirsutum* and *G. barbadense*) had a complement of phytochrome apoprotein genes that was very similar to that of the model plant *Arabidopsis thaliana*. This was not entirely unexpected given the relatively close phylogenetic relationship of the two lineages [204, 205]. The most-simple evolutionary scenario is that the last common ancestor of Arabidopsis and cotton, possibly an arborescent species in the late Cretaceous period [205], had a phytochrome gene complement consisting of one functional gene in each of the *PHYA*, *PHYB/D*, *PHYC* and *PHYE* subfamilies.

PHYA* Duplication in *Gossypium

After the divergence of the *Malvales* and *Brassicales*, the ancestral *PHYA* gene underwent duplication resulting in the observed *PHYA-1* and *PHYA-2* paralogs of modern *Gossypium spp.* As the A- and D-genome diploids have both paralogs, the duplication event occurred prior to the divergence of the A- and D-genome lineages.

Using 85 MYA (range 68 MYA to 96 MYA) as a rough estimate of the time of divergence of the *Malvales* and *Brassicales* [204, 210], along with our observed K_s of 1.82 in the *PHYA* hinge region in this time interval, we can derive a crude estimate of 0.011 substitutions / synonymous-site / million years, and an estimate of the time of *PHYA* duplication of ~14 MYA. This estimate places the duplication well within the crown group of *Malvales* and the *Malvaceae* family [205]. Given our time estimate, the *PHYA* duplication may be exclusive to the genus *Gossypium*, but would have occurred prior to the estimated time of divergence of the A and D genome groups [203]. As neither we nor others [40, 203, 210] have observed evidence of additional nuclear gene duplications or chromosomal duplications in this time period, the *PHYA* event was likely a tandem or segmental duplication, rather than whole genome duplication.

After a gene duplication event, one of the two newly duplicated genes is theoretically unconstrained by selection for function, and is thus free to accumulate mutations leading to a pseudo-gene fate, sub-functionalization, or neo-functionalization [211-215]. Although we did not obtain definitive evidence of pseudo-genic sequences in any of the phytochromes or taxa studied (e.g. no stop codons or frame shift mutations), we did observe significant variation in K_A/K_s ratios in pairwise interspecific comparisons (discussed below), leaving open the possibility of pseudo-gene outcomes. Alternatively, one of the duplicated genes may undergo positive selection to gain a novel function (neo-functionalization). Further, duplicated gene-pairs may subdivide the function of ancestral gene (sub-functionalization). Perhaps the most intriguing fate, which has been observed empirically, but not yet explained in theory, is the situation in which both gene

copies may be retained for a lengthy period under what appears to be purifying or negative selection [216, 217]. One approach to understanding the evolutionary fates of duplicated genes is through an analysis of the signature of natural selection on amino acid encoding sequences.

Although the hinge regions of phytochromes display relatively high levels of nucleotide diversity [218], they do not evolve under neutrality. The hinge region participates in inter-domain communication in phytochrome molecules [219]. For example, phosphorylation of a serine residue in the *PHYA* hinge plays a likely role in regulating protein-protein interactions between phytochrome and downstream signal-transducing molecules [220]. Compared to wild-type, a mutation in the hinge region of Arabidopsis *PHYB* is deficient in localization into distinct nuclear bodies [221]. Further, a single nucleotide polymorphism (SNP) in the hinge of one of two *PHYB* genes in Aspen (*Populus tremula*, *Salicaceae*) was associated with natural geographic variation in the timing of bud-set [222].

In comparisons between cotton and Arabidopsis (Table 5), the K_A/K_s ratio for the *PHYA* hinge region was 0.068 — a value that is typical for genes under purifying selection [223]. In contrast, the K_A/K_s ratio for *PHYA* after gene duplication (node 3) was 0.163, or ~2.4-fold higher. This value is also ~2.1-fold greater than the mean K_A/K_s ratio of all phytochrome hinge regions (corresponding to nodes 1A, 1B, 1C, and 1D in figure 6) of approximately 0.079 ± 0.014. This significantly elevated K_A/K_s ratio after the *PHYA* duplication could be attributed to a relaxation of stabilizing selection and/or

subfunctionalization of the nascent *PHYA* paralogs (these two alternative possibilities are remarkably difficult to distinguish on the basis of sequence information alone).

Table 5 Nucleotide Divergence in Phytochrome Genes in Comparisons of Arabidopsis and Cotton

	K-2P	S Dif	K_S	NS Dif	K_A	K_A/K_S
Node 1A	0.291	46.5	1.82	27.4	0.123	0.068
Node 1B	0.296	36.5	1	17.5	0.086	0.09
Node 1C	0.274	41	1.55	30.3	0.147	0.095
Node 1E	0.326	49	>2.0	23	0.122	<0.061
Node 3	0.094	17	0.309	11.8	0.05	0.163

Nodes refer to the NJ tree in Figure 6. K-2P indicates the mean Kimura 2-parameter distances between Arabidopsis and cotton gene sequences.

The possible functional divergence of *PHYA1* and *PHYA2* may be more pronounced after the separation of the A- and D-genome lineages (Table 4). A

comparison of *PHYA2* in the two diploids yields a K_A/K_s ratio of ~8.2, primarily due to amino acid substitutions in *PHYA2.D*, while *PHYA1* has a K_A/K_s ratio of 0.000 in the same taxonomic comparisons. Although this difference is suggestive of possible differential rates of functional evolution in the paralogs, it is not statistically significant in Fisher's exact test ($P = 0.2485$). It will be of interest to determine whether the cotton *PHYA* paralogs have distinct functions. Experiments are underway to determine the respective biological functions of each *PHYA-1* and *PHYA-2* in *G. hirsutum* and *G. barbadense* using paralog-specific RT-PCR, RNAi gene knockout, and tests for genetic associations between phytochrome-controlled phenotypic traits and *PHYA-1* and *PHYA-2* specific molecular markers. A 'candidate gene' approach has recently been used in soy (*Glycine max*) to uncover a genetic linkage between the photoperiod insensitivity locus *E4* and one of the two the *PHYA* genes, designated *GmphyA1* and *GmphyA2* [115]. Loss of photoperiodic flowering is associated with a *Ty1/copia*-like retrotransposon insertion into exon 1 of *GmphyA2*. The authors argue that gene duplication and partial redundancy of the *PHYA* genes may have facilitated the loss of photoperiod sensitivity by allowing the *GmphyA2 (E4)* mutant to avoid the major deleterious phenotypic effects that would have been caused by complete deficiency of *PHYA* gene function.

Persistence and Loss of Phytochrome Paralogs after Allopolyploidization

All phytochromes underwent gene duplication by polyploidization at the time of formation of the AD allotetraploids, on the order of 0.5-2.0 MYA [35, 38, 201, 224]. For example, in *G. hirsutum*, we detected a minimum set of ten distinct phytochrome genes, including four *PHYA* genes. In order to assess the evolutionary trajectory of these

recently duplicated genes, we examined the synonymous and non-synonymous divergence rates of A- and D-genome phytochrome orthologs and homeologs (Table 4) in pairwise comparisons of 1) diploids with diploids (D-D), 2) diploids with tetraploids (D-T), and 3) tetraploids with tetraploids (T-T). Given that the allotetraploid cottons had both A- and D-genome derived copies of each gene on the order of hundreds of thousands of years, we hypothesized that there may be a relaxation of selection in the allotetraploids, as one of the two copies should no longer be evolutionarily constrained. However, in comparisons of A- vs. D-genome derived orthologs or homeologs for six GSTs (Table 3), we did not observe dramatic differences in K_A/K_s between diploid and allotetraploids in any GST except the hinge region of *PHYA2* (in this case, the observed K_A/K_s ratio was actually ~30-fold higher in the extant diploids than in the allotetraploids). Because of low levels of nucleotide divergence, we employed Fisher's exact test [225] and found no significant differences in the patterns of nucleotide evolution in allotetraploids vs. diploids. Thus, there was no broad evidence of dramatic relaxation of natural selection on gene function after gene duplication by allotetraploidization. Further, the generally low K_A/K_s ratios across all genes and taxa support a model in which that the phytochrome homeologs are largely evolving independently by a birth-and-death model rather than concerted evolution [226].

The coding sequences of the *PHYB* 2.1 kb fragment also appeared be evolving under stabilizing selection in both the diploids (K_A/K_s= 0.251) and allotetraploids (K_A/K_s= 0.300) reflecting continued selective constraint on coding sequence evolution after polyploidization. However, there was a significant excess of non-synonymous

substitutions in both diploids and allotetraploids ($P = 0.01$ and $P = 0.004$, respectively, in Fisher's exact test) indicating a partial relaxation of negative selection and/or functional divergence of the *PHYB* homeologs.

In the allotetraploid cottons, both *PHYC.A* and *PHYC.D* are also evolving in a pattern consistent with purifying selection ($K_A/K_s = 0.184$ over 340 codons). However, it should be noted that the *PHYC.D* clade appears to be evolving at distinctly faster rate (8 parsimonious substitutions, including 6 non-synonymous) than the *PHYC.A* clade (2 parsimonious substitutions, both synonymous). This suggests either a relaxation of purifying selection in, or functional divergence of *PHYC.D*. In a similar study of phytochromes in cultivated sorghum (*Sorghum bicolor*) and its wild congeneric relatives [172], *PHYC* was undergoing faster amino acid evolution than *PHYA* or *PHYB*. In the both the *PHYB* and *PHYC* gene subfamilies of cotton, the sequences of the C-terminal signaling domain had higher K_A/K_s ratios than the corresponding hinge region alone. This may reflect the co-evolution of protein-protein interactions with downstream signaling partners, which are mediated by the C-terminal 'signal transduction' domain [162-164, 173-175].

While *PHYE*-related contigs had low K_A/K_s values (0.000 to 0.071), indicating purifying selection, no contig corresponding to an expected *G. barbadense PHYE.A* ortholog was observed. This may have been due to under-sampling of *G. barbadense* clones for sequencing, or due to nucleotide divergence in primer sites (as observed in *PHYC*). Of the 16 *PHYE*-like clone sequences obtained from *G. barbadense*, all were in the D-genome derived clade, which would be a unlikely result ($P < 0.005$, chi-square

test) assuming equal amplification efficiencies for *PHYE.A* and *PHYE.D*. Alternatively, the apparent lack of a *PHYE.A* ortholog in *G. barbadense* could be explained by concerted evolution, gene conversion, or by PCR-mediated recombination [224, 227]. Overall, the *PHYE* genes, like the other cotton phytochromes, had more synonymous than non-synonymous nucleotide substitutions, favoring a birth-and-death model of gene evolution.

Conclusions

Our preliminary efforts to obtain an inventory of the cotton phytochrome gene family (based largely on 'hinge' region) indicated that diploid A- and D-genome diploid cottons have two paralogous *PHYA* genes (designated *PHYA1* and *PHYA2)*, and one each of *PHYB*, *PHYC*, and *PHYE* gene sub-families. Coding sequence evolution in *PHYA2* was significantly elevated, suggesting loss of selection for function, or incipient sub-functionalization. Other than this duplication and the lack of a separate *PHYD* gene, the phytochrome complement of diploid cottons was very similar to that observed in the closely related model plant *Arabidopsis thaliana*, which greatly facilitates cross-species comparisons.

Whole genome duplication via allopolyploidization (~0.5-2.0 MYA) resulted in additive amalgamation of phytochrome genes within a single nucleus in the allotetraploid, retaining complete gene complements of at least four *PHYA* genes, two genes of each *PHYB*, *PHYC* and *PHYE* in AD-genome *G. hirsutum. G. barbadense* may lack the *PHYE* gene contributed by the A-genome ancestor. Strong purifying selection on nearly all of the phytochrome genes suggests some level of conservation of function

of each of the genes after polyploidization. With the possible exception of one of the *PHYE.A* homeologs in *G. barbadense*, we did not see evidence of gene loss. We did not observe any convincing evidence of concerted evolution by gene conversion. Rather, the genes duplicated by allopolyploidy appear to be largely retained, and evolving independently as observed in 48 other nuclear genes in allotetraploid cottons [228].

These results further our understanding of the evolutionary fates of duplicate genes following allopolyploidization. Information on key evolutionary events (such as duplications), as well as rates and patterns of evolutionary change, are an important component of the functional annotation of genes and genomes [229]. These data provide the foundation for more comprehensive studies of the biological functions of each of the cotton paralogs and homeologs. The development of phytochrome 'candidate gene' markers based on the GSTs identified here may prove useful in the mobilization of valuable genes from photoperiodic wild and primitive cottons into elite cotton varieties, in order to improve stress tolerance, disease resistance, fiber quality, and other traits.

Methods

Plant Materials

To simplify the assignment of sequences to orthologous or paralogous phytochrome loci (as opposed to alternative alleles at a single locus) we employed diploid and allotetraploid strains that were highly homozygous. Diploid cotton species (*G. raimondii* Ulbr, and *G. herbaceum* L.) were obtained from the cotton germplasm collection at the Institute of Genetics and Plant Experimental Biology, Tashkent, Uzbekistan. These lines had been maintained by selfing for multiple generations.

Genetic standard genotypes *G. hirsutum* L. cv. TM-1 and *G. barbadense* L. cv. 3-79 were obtained from the USDA-ARS Cotton Germplasm Unit, at College Station, Texas, USA. *G. hirsutum* cv. TM-1 [230] is a highly inbred line (>40 generations of selfing). *G. barbadense* cv. 3-79 is a doubled-haploid line [231].

Genomic DNA Isolation and PCR Amplification

Genomic DNAs were isolated from fresh leaf tissue of individual plants from each taxon using the method described by Dellaporta et al. [232]. The primers used in this study (Table 3) were designed using sequences from phytochromes of dicotyledonous plants obtained from the GenBank database (http://www.ncbi.nlm.nih.gov) and aligned using CLUSTALX software [233]. These included the degenerate primer pair PHYdeg-F/PHYdeg-R, which was designed to amplify the hinge region of the entire phytochrome gene family, and primer pairs PHYABnondeg-F/PHYAdeg-R and PHYABnondeg-F /PHYBdeg-R, designed to amplify the hinge regions of the *PHYA* and *PHYB/D* subfamilies, respectively. In order to amplify additional regions of several the cotton phytochrome genes, degenerate primers that amplify amplicons downstream of the hinge region (in the C-terminal domain) were also designed using this approach. Conserved regions that had approximately 40-55% G+C content were used for primer design. The primer design criteria have been described [234].

PCR reactions were performed in a Robocycler thermocycler (Agilent, USA) with an initial denaturation cycle at 94°C for 3 min., followed by 45 cycles of 94°C for 1 min., 55° C for 1 min. (annealing) and 72°C for 2 min. (extension), followed by a single

5 min. extension at 72°C. A manual 'hot start' cycling protocol was performed through the addition of *Thermus aquaticus* (*Taq*) DNA polymerase in the annealing step of first cycle.

DNA Sequence Analyses

PCR products were cloned into the vector pCR4-TOPO and transformed into *E. coli* TOP10 cells according to manufacturer's instructions (Invitrogen, USA). Cloning was necessary to resolve sequences of duplicated genes. Recombinant plasmids were purified by mini-prep (Qiagen, USA) and sequenced using Big-Dye DNA version 1 cycle sequencing chemistry (Applied Biosystems, USA) along with vector-specific forward and reverse primers. As native *Taq* polymerase has an appreciable nucleotide substitution error rate [235], at least 10 clones were sequenced for each amplicon from each diploid taxon, and 20 clones were sequenced from each allotetraploid taxon. Unincorporated dye-labeled terminators were removed from the extension products by Bio-gel P-30 spin column purification (Bio-Rad, USA). Extension products were sequenced using the ABI 310 and ABI3130 Genetic Analyzers (Applied Biosystems, USA).

Data Analyses

Double-stranded, finished sequences for each clone were assembled with Sequencher 4.8 software (Gene Codes, USA). After trimming of vector and amplification primers, sequences were searched against GenBank databases using BLASTN [236]. Searches of the non-redundant nucleotide database (nr) and the *Arabidopsis thaliana* database (Taxid: 3702) were performed using the "discontinuous

megablast" method as implemented by the NCBI database [237]. Alignments of clones obtained from each amplicon/taxon combination were performed using ClustalX. Within each taxon, clone sequences were grouped into contigs on the basis of (in all cases) at least two shared diagnostic SNPs and (if present) shared indel polymorphisms. When a single clone differed from other clones in the same consensus contig at a single nucleotide position, these sporadic differences were assumed to be products of *Taq* polymerase substitution error [235].

Consensus sequences were then aligned across all taxa and used for phylogenetic analyses. Distance-based phylogenetic trees were generated using neighbor-joining [238], using a minimum evolution objective, with gaps (indels) ignored, and either uncorrected "p" distances or Kimura two-parameter distances [239], as noted in the figure legends. Parsimony analysis was performed by an exhaustive search implemented by the PAUP software package version 4.0b10 [240]. The robustness of each phylogenetic tree was evaluated by bootstrap replication [241]. Estimates of synonymous substitution rate K_S and non-synonymous substitution rate K_A were based the Jukes-Cantor correction [242] and calculated by the method of Nei and Gojobori [243] as implemented by the DnaSP ver. 5 software package [244]. The significance of differences in K_A and K_S were determined by Fisher's exact test [225]. Sequence alignments were scanned for possible recombination using the software package RDP3, employs a suite of recombination detection and analysis methods [206]. Phytochrome ESTs from *Gossypium spp.* were identified in GenBank by searching non-human, non-

mouse ESTs (est_others) and *Gossypium* (Taxid: 3633) using the "discontinuous megablast" method as implemented by the NCBI database [237].

CHAPTER III

USE OF ROCHE 454 AMPLICON PYROSEQUENCING TO IDENTIFY ORTHOLOGS, PARALOGS AND SNPS OF CANDIDATE GENES IN DIPLOID AND TETRAPLOID COTTONS *(GOSSYPIUM SPP.)*

Comparative SNP Diversity among Diploid and Tetraploid Cottons *(Gossypium spp.)* for Candidate Genes from the Floral Network, Circadian Clock, and Photoreceptor Biosynthetic Pathways

Overview

Overview Rationale and Objectives

The genomes of cultivated cottons were found to be large, complex, incompletely characterized and they included both diploids (2N = 2X = 26) and allotetraploids (2N = 4X = 52). The use of single nucleotide polymorphisms (SNPs) for genetic analyses, such as QTL mapping and association mapping, has been made complicated by the presence of multiple orthologs and paralogs. With the emergence of long-read next-generation sequencing, like the Roche 454, it has become possible to sort out nucleotide differences between different orthologs, paralogs, and alleles in the absence of complete genome sequences, cytogenetic studies, or complete linkage data. It has been hypothesized that differences in floral initiation between cultivated allotetraploid 'AD' cottons and the wild allotetraploid 'AD' relatives were due to genetic variation within the floral regulatory pathway. This study was an exploratory measure in floral gene regulation.

Overview Methods

This study postulated that SNP polymorphisms associated with 38 candidate floral regulatory genes were identified from *Arabidopsis thaliana*. In a partial Roche GS-FLX run (1/8 gasket), 56 gene amplicons representing 38 genes in the flowering pathway were sequenced from eight taxa including *Gossypium raimondii* (D5), *Gossypium herbaceum* (A1), *Gossypium barbadense* (AD2), *Gossypium hirsutum* (AD1), and the out-group *Gossypium incanum* (E4). Each of the taxa was barcoded using a novel Y-adapter strategy. From a dataset of 104,230 reads, we were able to parse out the 'A' and 'D' genome orthologs of candidate genes, and polymorphisms between the diploid orthologs and the allotetraploid paralogs.

Overview Results and Conclusions

This study characterized polymorphism levels (including exonic and intronic) of 38 candidate genes in three pathway categories: photoreceptors, circadian clock genes, and floral regulators. These sequences showed high similarity to orthologs in the model plant *Arabidopsis thaliana*, allowing for the various cotton orthologous and paralogous loci to be identified, and nucleotide polymorphisms within each locus to be easily characterized between the eight cotton taxa. Our findings implied that despite duplications in allotetraploids, informative genetic changes in candidate genes can be identified and used in subsequent experiments to correlate candidate genes with phenotypic differences in photoperiodic flowering.

Overview Keywords

Cotton, Duplicate Gene Evolution, Gene Conversion, *Gossypium*, Polyploidy, Linkage Disequilibrium, Candidate Gene, Flowering, Photoperiodism, Photoperiod, SNP, Orthologs, Circadian Clock, Paralogs

Background

Flowering time was identified as a vital trait in domestication and agronomy of higher plants. Loss of the ability to sense photoperiodic cues to initiate flowering was essential to the dissemination and diversification of many crops to different longitudinal ranges throughout the world [245-247]. Flowering time has been highly influenced by the plant's ability to discern how many hours of light have passed during a day. For the ability to spread crops throughout the globe, this phenomenon has been studied extensively in many low latitude (tropical and sub-tropical) organisms, such as *Zea Maes, Solanum lycopersicum*, *Oryza sativa,* and *Sorghum bicolor,* that are of tropical or sub-tropical origin, but have spread to higher latitudes [17, 22, 41-44, 80, 246, 248-255]. Historically, agrarian societies have disseminated these tropical origin crops throughout the world, as their societies expanded into new frontiers at differing latitudes.

During the past 10,000 years, wild plants and animals have been converted into domesticated species by the introduction of farming in agrarian cultures [12-14, 17-20, 23, 36, 44, 107, 246, 247, 256-263]. Through agricultural advancements, farmers propelled these plants and animals forward into elite lines by new techniques in crop breeding and animal husbandry [7-10, 17, 45, 107, 172, 208, 261, 264-275]. During this agricultural domestication process, wild plants and animals were selected based on

phenotypic and genetic changes caused by rapid evolutionary responses. The phenotypic and genetic changes caused by domestication provide exemplary models to study [5, 12, 13, 42, 161, 219, 225, 260, 261, 276]. One phenotypic change, resulting during the birth of agrarian societies, was a plant's ability to discern the amount of daylight hours.

Photoperiodism (the perception of the amount of day and night hours) plays a key role in the domestication process of several sub-tropical and tropical plants [19-21, 166, 223, 246, 247, 253, 254, 277]. When photoperiodism does not affect a plant's ability to flower, the plant was deemed day-neutral (photoperiod insensitive) [12, 24, 45, 75, 80, 97, 98, 105, 107, 147, 150, 165, 172, 247, 249, 250, 252, 254, 255, 258, 259, 267, 277-286]. Typically, day-neutral plants flower according to certain developmental ages or environmental cues other than day length [50]. While undomesticated varieties of the cotton species, *Gossypium barbadense* and *Gossypium hirsutum,* were short day (SD) plants, modern cultivars of these species display day-neutrality [48, 49]. As the days grow shorter, short day plants were cued to flower when a certain amount of sunlight hours has been reached. In undomesticated cotton, this occurred when a reduction of daylight hours were equal to ten [32, 33].

It has been hypothesized that humans began inadvertently selecting for early flowering cotton, as a result of poly-cultural harvest practices, approximately 5000 years ago in both the old and new worlds [5, 29, 31, 33-36, 201, 202, 225, 287-290]. In the western hemisphere, Meso-Americans were thought to have begun domesticating cotton (*Gossypium hirsutum* L. and *Gossypium barbadense* L.) during their proto-agriculture phase [5, 30, 31, 33, 287]. The various species in these mixed agricultural plots were

harvested simultaneously; thus, while indigenous people gathered *Teosinte* (primitive corn), *Canvalia* (beans), and other crops for food, they unintentionally selected cotton that produced bolls at an earlier time, rather than the ancestral photoperiod-sensitive cotton, which naturally flowered later [5, 11, 32].

Selection for day-neutrality and other traits during domestication of several tropical and sub-tropical crop species has resulted in severe genetic bottlenecks [5, 12-21, 23, 25-27, 36, 42, 44, 107, 246, 247, 249, 256, 257, 260-263, 277]. The limited genetic diversity in modern cultivated cotton has stifled crop improvement. Therefore, the transfer of valuable traits, such as tolerance to biotic and abiotic stresses from wild relatives, has emerged as a key strategy in cotton improvement [291]. These wild genetic resources have valuable assets that have not been utilized and should be incorporated into traditional breeding programs [6]. Unfortunately, these wild relatives have been hampered by photoperiod sensitive flowering, so traditional breeding programs were impaired [11].

Most commercial cotton-producing areas in the world have not provided day length conditions that allow wild cotton species to flower in the span of a growing season [6, 27, 32, 35, 36, 48, 268, 292-296]. These 'wild cotton taxa' required the shortening of red light to nearly eleven hours in order to flower [11, 48, 169, 263, 268, 297]. The site of this study has been in College Station, TX (30° N latitude), where flowering in photoperiod dependent cotton occurs during late October.

Modern breeding techniques have used marker-assisted selection (MAS) to rapidly integrate desirable traits into elite cultivars [7-10, 32, 169, 231, 270, 297-301].

Achieving markers representing desirable traits from wild relatives has been difficult, but straightforward. An optimal MAS strategy would be to use markers to select for desirable traits, such as biotic or abiotic stress tolerance, and simultaneously select against SD photoperiodic flowering [11, 48, 268, 297].

To identify molecular markers, one approach utilized tightly linked to photoperiodic flowering to discern polymorphisms in the actual genes underlying the phenotypic variation between 'wild' and cultivated cotton floral initiation. With an improved understanding of the molecular-genetic determents behind day-neutrality in modern cotton and SD primitive cotton, new strategies to introgress valuable genetic traits from wild Germplasm for crop improvement can be applied [48, 49]. This approach identified many candidate genes that might be implicated in the photoperiod-independent evolution of flowering time during domestication of cultivated cotton (*Gossypium barbadense*).

Candidate Gene Approach

For the past decade, controversy has erupted among scientists on whether to use a genome-wide study using association mapping (GWAS) or a hypothesis-driven study using candidate genes [302]. The difference is that genome-wide studies look for anonymous polymorphisms throughout the genome, while candidate gene approaches are based on genes that are involved in the pathway that is influencing a phenotypic trait.

The primary reasons for utilizing the candidate gene approach in this study were: 1) well characterized pathways in *Arabidopsis thaliana* (floral regulatory network, photoreceptor pathway, and circadian clock pathway), and 2) the phylogenetic

relatedness between Arabidopsis and cotton (*Malvaceae*) [12, 24, 45, 46, 75, 105, 107, 113, 147, 165, 169, 172, 247, 249, 252, 254, 255, 258, 259, 267, 277-286, 303]. Our candidate gene approach incorporated genes from the well-characterized flower developmental network in *Arabidopsis thaliana* [12, 24, 45, 47, 51, 69, 70, 72, 75, 80, 87, 88, 90, 94, 97, 98, 101, 105-107, 113, 121, 122, 132, 136-138, 140, 144, 147, 149, 150, 152, 159, 165, 167, 172, 184, 192, 247, 249, 250, 252, 254, 255, 258, 259, 267, 277-286, 303-316]. *Arabidopsis thaliana* was chosen as a model for cotton because of well-documented studies of photoperiod-independence and minimal evolutionary divergence being located in the same phylogenetic clade, Eurosid II. [46] Outside of *Brassicales*, cotton is the closest mapped agnate to *Arabidopsis thaliana* [317, 318].

Another reason for choosing the candidate gene approach, rather than GWAS, was the large extent of linkage disequilibrium (*LD*) in cotton [273, 291, 319]. Abdurakhmonov et al. found cotton's *LD* blocks to be approximately 5 to 6 cM in size, corresponding to 2.5 to 3.0 Mb [291]. With GWAS finding anonymous markers with very close physical and genetic linkage to photoperiodic flowering would be a daunting task because the very large linkage blocks, therefore going with candidate genes seemed more effective being closely linked with photoperiodic flowering.

A reference genome was not available for our search for cotton orthologs of our Arabidopsis candidate genes. Only recently were the *Gossypium raimondii* D genome scaffolds released with partial annotations (January 6, 2012, http://www.phytozome.net/cotton.php). Corrections to the 2012 PLoS ONE article, by Blenda et al., updated the *Gossypium raimondii* D genome scaffolds to version 2.1,

which realigned some chromosomes and scaffold orientations [320]. Another draft genome of *Gossypium raimondii* was released in August 2012 from the Beijing Genomics Institute (BGI) without annotations [321]. At the time this experiment was started, the resources available were expressed sequence tags (EST) libraries in GenBank at the National Center for Biotechnology Information (NCBI) and tentative consensus sequence (TCs) libraries in the Gene Index Project at Computational Biology and Functional Genomics of the Dana-Faber Cancer Institute (DFCI). However, a vast majority of those *Gossypium* EST and TC libraries were made from post-anthesis tissues (after flowering, fiber only). Therefore, finding genes expressed during other plant developmental stages (such as floral initiation) was difficult.

The specific objective of this project was to identify SNP, Single Insertion/Deletion (SID), and Insertions/Deletions (InDel) polymorphisms in genes of the floral developmental network genes in cotton. Thirty-eight genes of interest were selected based on the Arabidopsis floral regulatory literature [24, 45, 47, 54, 61, 63, 65-76, 78, 79, 81, 85-90, 94, 95, 97, 98, 101-104, 106, 114, 121-123, 125, 126, 128-133, 135-140, 143, 144, 148, 150-152, 159, 161, 165, 166, 168, 173, 176, 179, 182, 184-186, 188-192, 194, 196, 198, 200, 219, 221, 222, 247, 250, 252, 253, 255, 258, 259, 276, 280-282, 286, 303-306, 309-316, 322-337]. Priority was assigned to those genes with a known influence for photoperiod regulation of floral initiation. Cotton orthologs of these genes of interest were PCR amplified from eight different species of cotton [*Gossypium raimondii* (D5)*, Gossypium herbaceum* (A1)*, Gossypium barbadense* 3-79 (AD2 - genetic standard)*, Gossypium barbadense* PS-6 (AD2)*, Gossypium hirsutum* TM-1 (AD1

– genetic standard)*, and Gossypium incanum* (E4)] and two photoperiod sensitive accessions [*Gossypium barbadense* K-46 (AD2) *and Gossypium hirsutum* TX-231 (AD1)] and sequenced using Roche 454 pyrosequencing. Our research included genes from photoreceptors, circadian clock genes, and transcription factors known to act as floral integrators. The goal was to isolate sequence differences between photoperiodic and non-photoperiodic lines, in order that these differences could be tested for genetic linkage to photoperiodic flowering and later used for marker assisted selection [7-10].

Results

Discovery of Genes within Cotton Divergence through Evolution

In an effort to encompass a comprehensive characterization of SNPs in the *Gossypium* taxa, samples across an evolutionary time span had to be taken. Samples in this study spanned across diploid evolutionary relatives of the 'A' and 'D' sub-genome, the uncultivated tetraploid 'AD' *G. barbadense* and *G. hirsutum* relatives, the modern cultivated lines of the tetraploid 'AD' *G. barbadense* and *G. hirsutum*, and, finally, a distant evolutionary out-group of the 'E' sub-genome (Table 6). By breaking down the sequences by sub-genomes, the confidence of distinguishing SNPs favoring significant changes between uncultivated and cultivated taxa can be teased out, while excluding the 'A'/'D' evolutionary changes occurring ten million years ago (MYA).

Table 6 List of Cotton Used

Latin Name	Variety	Designation	Ploidy	Floral Cue	Origin
Gossypium raimondii	**D5**	D5	Diploid	Photoperiodic	Peru
Gossypium herbaceum	**A1**	A1	Diploid	Photoperiodic	Africa-Asia
Gossypium barbadense	**3-79**	AD	Allotetraploid	Photoperiod Independent	Genetic Standard
Gossypium hirsutum	**TM-1**	AD1	Allotetraploid	Photoperiod Independent	Genetic Standard
Gossypium hirsutum	**TX-231**	AD1	Allotetraploid	Photoperiodic	Texas
Gossypium barbadense	**K-46**	AD2	Allotetraploid	Photoperiodic	Guadeloupe
Gossypium barbadense	**PS-6**	AD1	Allotetraploid	Photoperiod Independent	Arizona
Gossypium incanum	**E4**	E4	Diploid	Photoperiodic	Afro-Arabian

Fifty-six primer pairs representing thirty-eight genes were tested across the eight taxa. Most primer pairs appeared to give single banded PCR products in the ‘D’ and ‘A’ sub-genome. Frequently, the tetraploid species had two PCR bands present. This represented the slight changes between the ‘A’ and ‘D’ homologs within the tetraploid species. All primer pairs worked efficiently across all taxa (data not shown).

In this study, expressed sequence tag (EST) sequences and tentative consensus sequences (TCs) from public databases [GenBank at National Center for Biotechnology Information (NCBI) and Gene Index Project at Computational Biology and Functional Genomics of the Dana-Faber Cancer Institute (DFCI)] were used to design primers for more than fifty gene fragments in the floral regulatory pathway. These primers were

used in the amplification of potential genes in *G. raimondii* D5. The products were then sequenced through traditional Sanger sequencing [338]. For PCR products with more than one band, Blunt-ended Topo kits were used to clone out different bands. Those clones were then sequenced, and PCR primers were refined. For some short ESTs, gene walking was done to identify the unknown regions flanking the known DNA region. All primers were refined to be specific for the *Gossypium raimondii* sequence and to span across the exon-intron regions. BLAST analysis showed that all of the sequenced potential gene regions corresponded back to the orthologous Arabidopsis gene [236, 237, 339, 340].

DNA vs. Amino Acid Substitutions within Coding Sequences

The sequenced exonic regions of cotton had high similarity to that of the orthologous Arabidopsis exon regions, while the intronic similarity varied greatly against that of the model plant. The translated exonic amino acid sequences were usually similar, if not identical, to those in Arabidopsis (Figure 12)**.** Occasionally, some exonic regions showed higher levels of change within the translated amino acid sequences from Arabidopsis to cotton, like the upstream hinge region of *Phytochrome E (PHYE)* (Figure 12). As expected, the levels of amino acid changes and nucleotide changes between *Gossypium raimondii* and *Gossypium herbaceum* were low in comparison with those between Arabidopsis and cotton (Table 7 a-b).

Genes	Amino Acid Sequence
ATGRP7 AT CDS	G G L A W A T D D R A L E T A F A Q Y G D V I D S K I I N D R E T G R S R G F G F V T F K D E K A M
ATGRP7 D5	G G L A W A T D D R A L E E A F S A F G E I V E S K I I N D R E T G R S R G F G F V T F R D E K A M

Genes	Amino Acid Sequence
LHY AT CDS	L R L Y G R A W Q R I E E H I G T K T A
LHY1 D5	L K L Y G R A W Q R I E E H I G T K T A

Genes	Amino Acid Sequences
PHYE AT CDS	F R I L G L S D N S S D F L G L L S L P S T S H S G E F D K V K G L I G I D
PHYE A1 Upper Hinge	F R I I G Y S E N C F G L L G L D L D S E D E I K G V - - - - K G L I G I D
PHYE D5 Upper Hinge	F R I I G Y S E N C F G L L G L D L D S E D E I K G V - - - - K S L I G I D

Figure 12 Examples of Arabidopsis versus Cotton Amino Acid Diversity within Candidate Genes

Table 7 DNA and Amino Acid Exonic Substitution Comparison

a) Arabidopsis vs. Cotton

Gene	DNA Changes	AA Changes	DNA Change Ratio	AA Change Ratio
AGL16	4 : 10	2 : 3	0.4000	0.6667
AGL3_SEP4	61 : 238	28 : 79	0.2563	0.3544
AGL30	32 : 122	11 : 40	0.2623	0.2750
AGL32	1 : 9	0 : 3	0.1111	0.0000
AGL6	10 : 89	1 : 29	0.1124	0.0345
AP1	18 : 90	5 : 30	0.2000	0.1667
ATGRP7	42 : 154	9 : 50	0.2727	0.1800
COL4	98 : 216	43 : 72	0.4537	0.5972
COL5	102 : 319	45 : 106	0.3197	0.4245
COP1	24 : 91	6 : 29	0.2637	0.2069
CRY1 A	24 : 87	6 : 29	0.2759	0.2069
CRY1 B	184 : 910	44 : 303	0.2022	0.1452
CRY2 A	63 : 362	15 : 120	0.1740	0.1250
CRY2 B	84 : 359	46 : 119	0.2340	0.3866
CRY3	74 : 241	25 : 80	0.3071	0.3125
DET1	1 : 12	0 : 4	0.0833	0.0000
ELF3	1 : 25	0 : 8	0.0400	0.0000
FD	109 : 514	64 : 170	0.2121	0.3765
FKF1_ADO3	50 : 257	21 : 85	0.1946	0.2471
GI A	50 : 324	32 : 108	0.1543	0.2963
GI B	81 : 522	43 : 173	0.1552	0.2486
HY6	130 : 387	48 : 128	0.3359	0.3750
LHY 1	8 : 60	1 : 20	0.1333	0.0500
LHY 2	7 : 61	4 : 20	0.1148	0.2000
PFT1	21 : 97	5 : 32	0.2165	0.1563
PHYA 1	113 : 536	29 : 178	0.2108	0.1629
PHYA 2	184 : 870	85 : 289	0.2115	0.2941
PHYB	144 : 524	48 : 174	0.2748	0.2759
PHYC	275 : 1086	133 : 362	0.2532	0.3674
PHYE	148 : 603	88 : 200	0.2454	0.4400
PRR5	3 : 20	0 : 6	0.1500	0.0000
PRR7 A	74 : 192	33 : 63	0.3854	0.5238
PRR7 B	72 : 189	28 : 62	0.3810	0.4516
SPA4	17 : 104	4 : 34	0.1635	0.1176
TOC1	47 : 173	20 : 57	0.2717	0.3509

Table 7 Continued.

b) *G. raimondii* vs. *G. herbaceum*

Gene	DNA Changes	AA Changes	DNA Change Ratio	AA Change Ratio
AGL16	0 : 10	0 : 3	0.0000	0.0000
AGL3_SEP4	2 : 238	1 : 79	0.0084	0.0127
AGL30	0 : 122	0 : 40	0.0000	0.0000
AGL32	0 : 9	0 : 3	0.0000	0.0000
AGL6	0 : 89	0 : 29	0.0000	0.0000
AP1	2 : 90	1 : 30	0.0222	0.0333
ATGRP7	1 : 154	0 : 50	0.0065	0.0000
COL4	3 : 216	3 : 72	0.0139	0.0417
COL5	11 : 319	3 : 106	0.0345	0.0283
COP1	0 : 91	0 : 29	0.0000	0.0000
CRY1 A	1 : 87	1 : 29	0.0115	0.0345
CRY1 B	9 : 910	3 : 303	0.0099	0.0099
CRY2 A	1 : 362	0 : 120	0.0028	0.0000
CRY2 B	6 : 359	4 : 119	0.0167	0.0336
CRY3	5 : 241	3 : 80	0.0207	0.0375
DET1	0 : 12	0 : 4	0.0000	0.0000
ELF3	0 : 25	0 : 8	0.0000	0.0000
FD	11 : 514	8 : 170	0.0214	0.0471
FKF1_ADO3	3 : 257	1 : 85	0.0117	0.0118
GI A	0 : 324	0 : 108	0.0000	0.0000
GI B	5 : 522	4 : 173	0.0096	0.0231
HY6	8 : 387	5 : 128	0.0207	0.0391
LHY 1	0 : 60	0 : 20	0.0000	0.0000
LHY 2	1 : 61	1 : 20	0.0164	0.0500
PFT1	0 : 97	0 : 32	0.0000	0.0000
PHYA 1	7 : 536	3 : 178	0.0131	0.0169
PHYA 2	8 : 870	4 : 289	0.0092	0.0138
PHYB	7 : 524	2 : 174	0.0134	0.0115
PHYC	13 : 1086	5 : 362	0.0120	0.0138
PHYE	10 : 603	7 : 200	0.0166	0.0350
PRR5	0 : 20	0 : 6	0.0000	0.0000
PRR7 A	5 : 192	3 : 63	0.0260	0.0476
PRR7 B	8 : 189	4 : 62	0.0423	0.0645
SPA4	1 : 104	0 : 34	0.0096	0.0000
TOC1	2 : 173	0 : 57	0.0116	0.0000

Coding regions from genes within gene families (e.g. paralogs), such as *Cryptochrome 1B* (*CRY1 B*), *Cryptochrome 2B* (*CRY2 B)*, and *Phytochrome A2* (*PHYA 2)* above, showed more substitutions in amino acids and nucleotides (Table 7 a-b). Furthermore, *De-etiolated 1* (*DET1)*, *Early Flowering 4* (*ELF4)*, and *Long Hypocotyl* (*LHY 1)* showed no divergence (0%) from Arabidopsis, while *Pseudo Response Regulator 7 A* and *B* (*PRR7 A* and *PRR7 B)* had 38% nucleotide difference and a 45% to 52% amino acid difference from Arabidopsis. One ortholog, *Flowering Locus D (FD)*, had a higher level of DNA and amino acid substitutions at 21% and 37%, than other non-paralogous orthologs (Table 7 a).

454 Pyrosequencing and Reference Assembly

The multiplex identifiers (MID) used for each taxon allowed all gene reactions to be pooled into one sample. All gene-specific primers permitted the identification of how each fragment was correlated to each taxon. The initial 454 run consisted of 104,230 sequences with an average length of ~400 bp. In this multiplexed sample, 96% of the sequence reads (99,699 sequences) were separated by their MID into 8 taxa (Table 8).

Table 8 Mid Divergence Statistics for Roche 454 Run

ID Tag	Barcode	Mid Identifier Designation	Variety		Floral Cue	MID Parsed Sequences
AF1	ACGAGTGCGT	*Gossypium raimondii*	**D5**	Diploid	Photoperiodic	22693
AF2	ACGCTCGACA	*Gossypium herbaceum*	**A1**	Diploid	Photoperiodic	11825
AF3	AGACGCACTC	*Gossypium barbadense*	**3-79**	Allotetraploid	Photoperiod Independent	23166
AF4	AGCACTGTAG	*Gossypium hirsutum*	**TM-1**	Allotetraploid	Photoperiod Independent	19141
AF5	ATCAGACACG	*Gossypium hirsutum*	**TX-231**	Allotetraploid	Photoperiodic	8152
AF6	ATATCGCGAG	*Gossypium barbadense*	**K-46**	Allotetraploid	Photoperiodic	7010
AF7	CGTGTCTCTA	*Gossypium barbadense*	**PS-6**	Allotetraploid	Photoperiod Independent	2920
AF8	CTCGCGTGTC	*Gossypium incanum*	**E4**	Diploid	Photoperiodic	4792
		Ungrouped				4531

Some of the initial genes and miRNA were discarded from 454 results because those sequences were not well amplified in the pyrosequencing reaction. Some initially low yielding genes were re-pyrosequenced in the SNP validation run. Alternatively, the D5 reads were mapped back to the original sequence created through cloning and Sanger sequencing. These D5 Sanger sequences formed the basic scaffold on which the D5 454 sequencing reads could be aligned, so that a consensus sequence could be created. These newly aligned D5 454 consensus sequences of the candidate genes were then used as a reference for mapping the A2 genes. Once the A2 sequences were aligned, modified consensus sequences representing the A2 nucleotide changes were created. These new consensus sequences from A2 were then used as the A2 reference sequence. For a consensus change to take place, a threshold of sequence similarity had to represent 65% of all A2 sequences for that fragment (Table 9 and Figure 13).

Table 9 Amplicon Coverage Distribution of 'A' and 'D' Genome Sequences by Ortholog

Genes (orthologs)	D5 Sequences	A1 Sequences
AGL1	0	1
AGL16	54	6
AGL2	1	4
AGL3	22	5
AGL30	13	56
AGL32	1116	83
AGL6	62	84
AGL65	170	60
AGL9	2	0
AP1	20	131
AP3	4	15
ATGRP7	87	30
COL3	103	4
COL5	192	61
COP1	74	56
CRY1A / CRY1B	1065	1250
CRY2	55	100
CRY3	36	19
DET1	180	160
ELF3	126	47
FD	27	4
FKF1	39	34
FT	300	15
GI	3544	1487
HY6	308	299
LHY	139	77
miRNA172c	8	0
PFT1	170	96
PHYA1 / PHYA2	338	144
PHYB	1655	1004
PHYC	120	52
PHYE_Upper_hinge	204	247
PI A / PI B	671	56
PRR5	106	23
PRR7 A/PRR7 B	76	105
SPA4	815	513
TOC1	2705	1121
ZTL	3	0

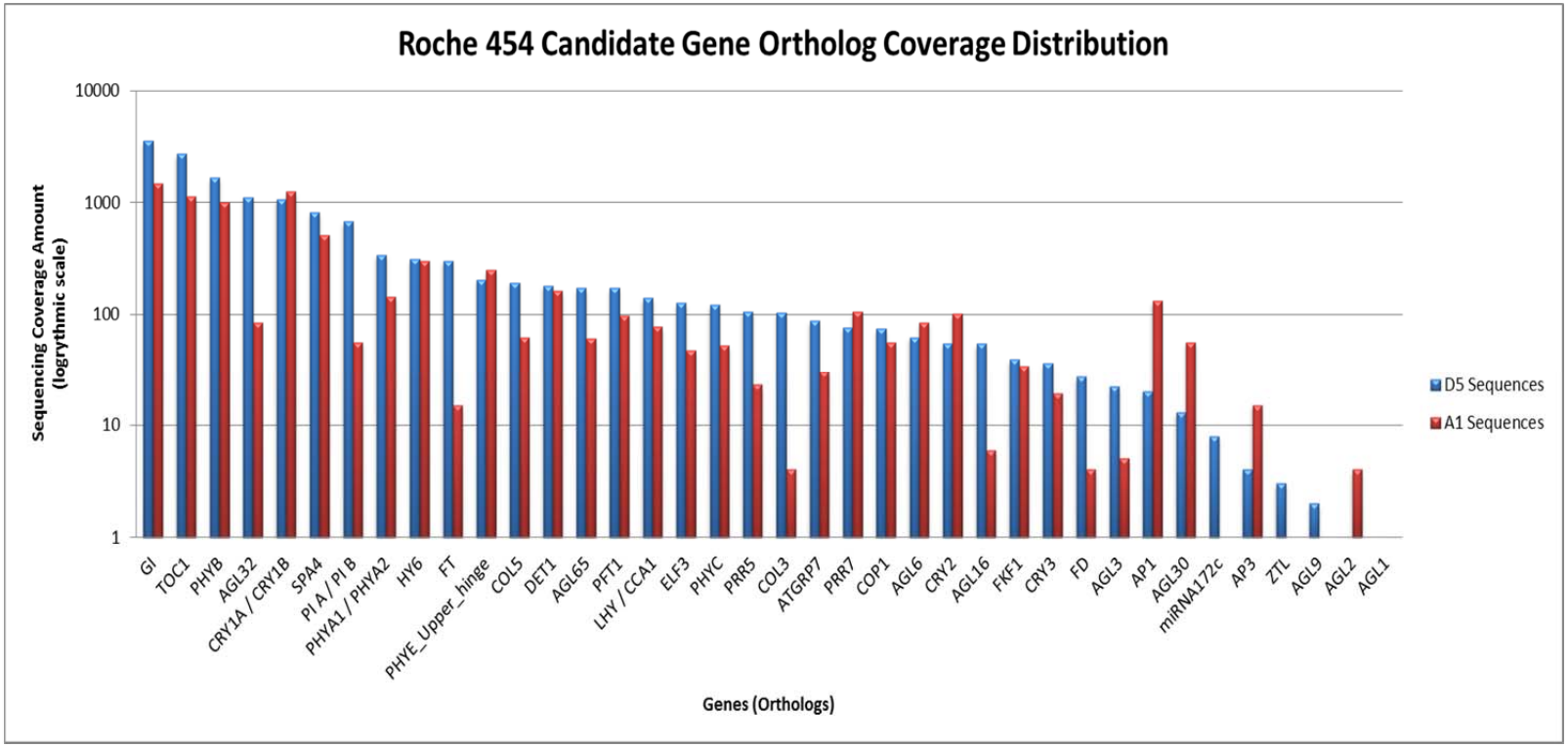


Figure 13 Roche 454 Candidate Gene Ortholog Coverage Distribution between 'A' and 'D' Genomes

The remaining sequences for all the other taxa were mapped to both the 'A' and 'D' 454 reference consensus sequences in CLCBio®. The ability to map-to-reference two sub-genomes at the same time on a taxon was imperative in separating the sub-genome reads by taxa in the tetraploid sequences. The consensus sequences of each taxon were exported, representing both the 'A' and 'D' sub-genomes, into Geneious®. A software platform change from CLCBio® to Geneious® was necessary due to the lack of alignment manipulation within CLCBio®. Geneious® has the ability to create multiple alignments. Nucleotides in those sequence alignments were visually shifted around to account for SNPs, Single Insertion/Deletions (SIDs), and Insertions/Deletions (InDels) within a sequence.

SNP, SID, and INDEL Detection

A total of 762 SNPs were uncovered in the eight taxa. Of these SNPs, 466 (296 without *Gossypium incanum*) were found in exonic regions (Table 10), while 678 (594 without *Gossypium incanum*) were discerned in the intronic regions (Table 11). A single SNP was found every ~ 27 bp in the intron regions, while in it was ~ 37 bp in the exon regions (not including *Gossypium incanum*) (Table 12).

Table 10 SNPs in Exonic Regions

Gene	A/D genome	*G. hirsutum*/ *G. barbadense*	Cultivated/ Wild	*G. incanum*	Single SNP
AGL16	0	0	0	1	0
AGL3_SEP4	0	0	2	0	2
AGL30	0	0	0	0	1
AGL32	0	0	0	0	0
AGL6	0	0	0	0	0
AGL65	3	0	0	1	2
AP1	2	0	0	1	0
ATGRP7	1	0	0	1	1
COL4	8	0	2	2	3
COL5	9	2	0	0	2
COP1	0	0	0	1	0
CRY1 A	1	0	0	0	0
CRY1 B	11	1	2	58	0
CRY2 A	1	1	2	2	0
CRY2 B	3	2	2	21	4
CRY3	4	0	0	2	1
DET1	0	0	0	0	0
ELF3	0	0	0	1	0
FD	12	0	1	4	4
FKF1_ADO3	3	2	0	1	0
GI Ex9to10_ A	0	0	0	0	0
GI Ex10to11_ A	0	0	0	1	0
GI Ex10to11_ B	0	0	0	0	0
GI Ex11to12_ A	0	2	0	0	0
GI Ex11to12_ B	6	0	1	1	1
HY6	12	2	11	2	3
LHY 1	0	0	0	0	0
LHY 2	1	0	0	1	0
PFT1	0	0	0	0	0
PHYA 1	15	3	3	7	3
PHYA 2	6	6	2	26	17
PHYB	8	1	1	1	3
PHYC	6	17	12	19	15
PHYE	5	3	0	6	6
PI A	0	0	0	0	0
PI B	0	0	0	0	0
PRR5	1	0	1	2	1
PRR7 A	6	0	0	3	1
PRR7 B	7	3	2	3	3
SPA4	1	0	0	2	0
TOC1	2	0	0	0	0
Exon Totals	134	45	44	170	73

Table 11 SNPs in Intronic Regions

Gene	A/D genome	*G. hirsutum*/ *G. barbadense*	Cultivated/ Wild	*G. incanum*	Single SNP
AGL16	5	2	0	3	7
AGL3_SEP4	2	0	4	0	8
AGL30	15	3	4	3	4
AGL32	1	0	0	3	1
AGL6	7	0	0	0	7
AGL65	22	3	3	6	9
AP1	10	2	0	7	0
ATGRP7	4	0	0	1	1
COL4	1	1	0	2	1
COL5	0	2	0	0	1
COP1	14	2	0	12	3
CRY1 A	14	2	5	0	6
CRY1 B	24	2	1	9	2
CRY2 A	2	1	0	3	3
CRY2 B	0	0	4	3	3
CRY3	8	0	2	7	8
DET1	1	0	0	0	0
ELF3	11	4	3	11	3
FD	1	2	1	2	0
FKF1_ADO3	14	5	2	11	4
GI Ex9to10_ A	9	1	0	7	13
GI Ex10to11_ A	5	2	3	3	0
GI Ex10to11_ B	0	2	1	0	2
GI Ex11to12_ A	14	2	1	4	3
GI Ex11to12_ B	3	0	2	4	3
HY6	17	2	0	11	5
LHY 1	5	2	7	1	0
LHY 2	4	0	0	1	2
PFT1	4	0	0	5	1
PHYA 1	1	0	0	0	1
PHYA 2	1	0	0	0	0
PHYB	15	2	3	1	4
PHYC	19	0	0	10	4
PHYE	0	0	0	0	0
PI A	11	4	3	11	4
PI B	3	1	2	0	3
PRR5	14	0	1	10	7
PRR7 A	4	0	1	1	2
PRR7 B	4	0	1	2	0
SPA4	1	0	1	0	0
TOC1	4	1	0	0	0
Intron Totals	294	50	55	154	125

Table 12 Overall Amplicon Sizes Split Into Intronic and Exonic Regions

Gene	Total bp Amplicon Coverage	Intron Region	Exon Region
AGL16	533	523	10
AGL3_SEP4	661	432	229
AGL30	737	620	117
AGL32	106	96	10
AGL6	274	185	89
AGL65	1,135	914	221
AP1	424	343	81
ATGRP7	430	276	154
COL4	383	141	242
COL5	448	144	304
COP1	724	633	91
CRY1 A	710	623	87
CRY1 B	1508	657	851
CRY2 A	526	170	356
CRY2 B	462	103	359
CRY3	645	406	239
DET1	93	81	12
ELF3	582	557	25
FD	518	81	437
FKF1_ADO3	979	722	257
GI Ex9to10_ A	557	494	63
GI Ex10to11_ A	419	202	217
GI Ex10to11_ B	249	188	61
GI Ex11to12_ A	664	574	90
GI Ex11to12_ B	423	152	271
HY6	1326	927	399
LHY 1	262	202	60
LHY 2	241	180	61
PFT1	459	347	112
PHYA 1	1664	66	1598
PHYA 2	1228	170	1058
PHYB	924	388	536
PHYC	1598	516	1082
PHYE	591	0	591
PI A	573	570	3
PI B	469	466	3
PRR5	751	696	55
PRR7 A	298	116	182
PRR7 B	272	92	180
SPA4	235	131	104
TOC1	285	112	173
Total bp Coverage	25450	14296	11154

As expected, the exonic region exposed more interspecific SNPs between *Gossypium hirsutum* and *Gossypium barbadense* than intraspecific SNPs between the cultivated and wild *Gossypium* lines (Table 13). Interestingly, however, there were more intronic SNPs observed between the cultivated and wild *Gossypium* lines than between *Gossypium hirsutum* and *Gossypium barbadense* (Table 13).

Table 13 Average SNP Estimates

SNP Ratio					
A/D Genome					
	Exon:	1	:	83	bp
	Intron:	1	:	49	bp
Intraspecific (Cultivated vs. Wild)					
	Exon:	1	:	252	bp
	Intron:	1	:	260	bp
Interspecfic (*G. hirsutum* vs. *G. barbadense*)					
	Exon:	1	:	246	bp
	Intron:	1	:	286	bp
E Genome					
	Exon:	1	:	65	bp
	Intron:	1	:	93	bp

There were three predicted genes that had very high levels of SNP diversity: *Cryptochrome 1B (CRY1 B)*, *Cryptochrome 2B (CRY2 B)*, and *Phytochrome A2 (PHYA 2)*. With the removal of these predicted genes the total number of detected SNPs was lowered from 1,144 to 934 (Figure 14). The revised SNP count increased the average SNP per intron to every ~ 27 bp, while decreasing the average SNP per exon to every ~ 36 bp (Supplemental Tables 2 and 3).

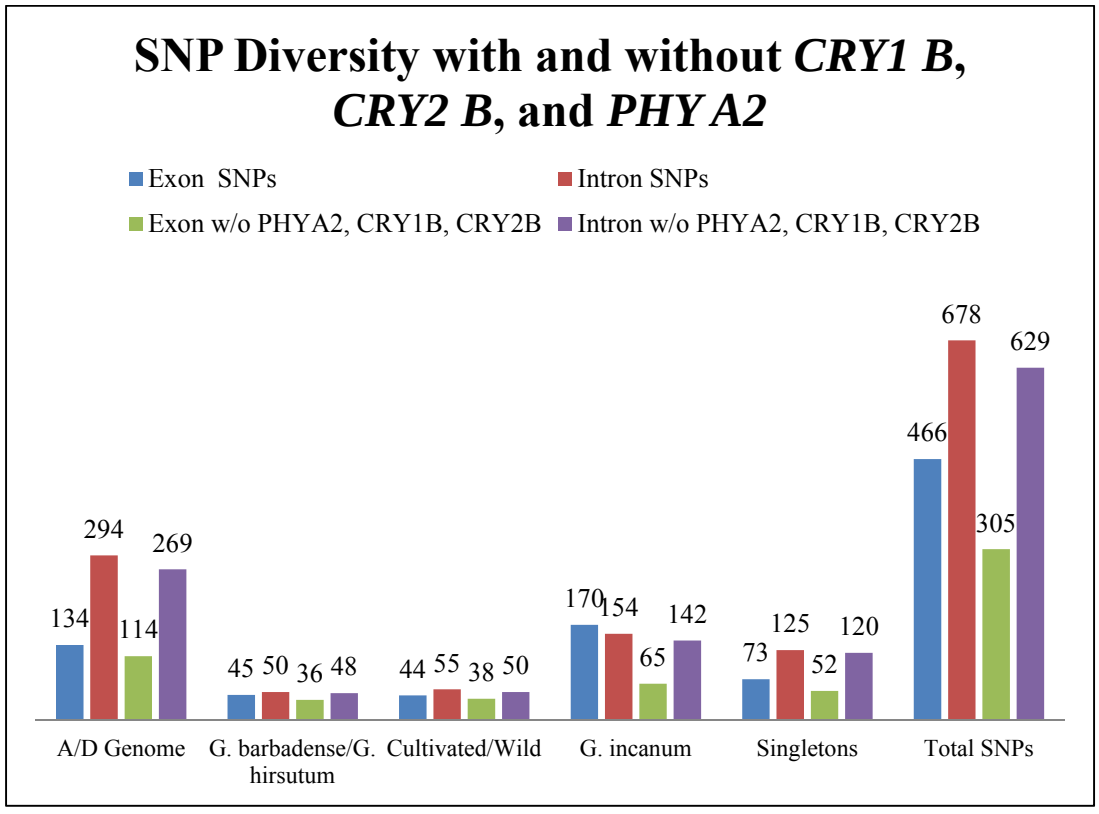


Figure 14 SNP Diversity with and without *CRY1 B*, *CRY2 B*, and *PHY A2*

A total of 233 SIDs were revealed within the eight taxa. SIDs were found to be more common in the intronic regions of the amplicons, every ~ 87 bp, than in the exonic regions, every ~ 326 bp. Notably, more SIDs were observed between wild species versus their cultivated counterparts, than between *Gossypium hirsutum* and *Gossypium barbadense* (Table 14). Those SID changes between *Gossypium hirsutum* and *Gossypium barbadense* occurred less frequently than changes between the ancestral sub-genomes 'A' and 'D' and the out-group E4 genome.

Table 14 Average SID Estimates

SID Ratio					
A/D Genome					
	Exon:	1	:	2214	bp
	Intron:	1	:	270	bp
Intraspecific (Cultivated vs. Wild)					
	Exon:	1	:	651	bp
	Intron:	1	:	227	bp
Interspecfic (*G. hirsutum* vs. *G. barbadense*)					
	Exon:	1	:	3690	bp
	Intron:	1	:	953	bp
E Genome					
	Exon:	1	:	1006	bp
	Intron:	1	:	681	bp

A total of 105 InDels were discovered in the candidate genes of the eight taxa. The InDels were found to be more common in the intronic regions of the amplicons, every ~ 234 bp, than in the exonic regions, every ~ 583 bp. Remarkably, less intronic InDels were seen between *Gossypium hirsutum* and *Gossypium barbadense,* than the ancestral sub-genomes 'A' and 'D', cultivated versus wild counterparts, and *Gossypium incanum* (Table 15).

Table 15 Average InDel Estimates

InDel Ratio					
A/D Genome					
	Exon:	1	:	2768	bp
	Intron:	1	:	549	bp
Intraspecific (Cultivated vs. Wild)					
	Exon:	1	:	1845	bp
	Intron:	1	:	841	bp
Interspecfic (*G. hirsutum* vs. *G. barbadense*)					
	Exon:	1	:	5535	bp
	Intron:	1	:	3574	bp
E Genome					
	Exon:	1	:	2768	bp
	Intron:	1	:	681	bp

Frequencies of SNPs across Different Pathway Categories

Genes holding key positions of the floral regulatory network may be influenced by strong purifying selection; therefore the relative amounts of SNPs may be fewer. Moreover, genes that have redundant functions may have increased SNP frequency. Frequency procedures (Freq procedure) and generalized linear mixed model (GLIMMIX) with Poisson distributions were applied to determine the effects of SNP frequencies by gene copy number (redundancy) and placement within: 1) the circadian clock pathway, 2) the photoreceptor pathway, and 3) the floral regulatory network.

Further analyses showed that the introns of the photoreceptor and circadian clock pathway categories have significantly more SNPs than do those of the floral network (Table 16). The photoreceptor pathway category has a higher frequency of A/D, interspecific and intraspecific SNPs in exonic regions, than the circadian clock pathway category and the floral network category (Table 17). In the coding regions, interspecific changes occurred in every ~220 bp, excluding the floral network pathway category. The floral network appeared to have high conservation between *Gossypium hirsutum* and *Gossypium barbadense*. This resulted in significantly fewer SNPs occurring in the exonic regions of the floral network and the circadian clock in comparison to the photoreceptor pathway categories (Table 16 and Figure 15). Outlier orthologs in the photoreceptor pathway category of the exonic region caused the photoreceptor pathway category mean to be above the average, so a Poisson distribution was done to normalize the data. No other pathway category combination differed significantly from each other.

Table 16 Frequency Procedures and Generalized Linear Mixed Model with Poisson Distributions for SNPs in Different Pathways

The Freq Procedure

Table of Pathway by Type			
Pathway	**Type**		
Frequency **Percent** **Row Pct** **Col Pct**	**Exon**	**Intron**	**Total**
Clock	56 4.87 19.86 11.84	226 19.64 80.14 33.33	282 24.50
Photorec	346 30.06 62.79 73.15	205 17.81 37.21 30.24	551 47.87
Flower	71 6.17 22.33 15.01	247 21.46 77.67 36.43	318 27.63
Total	473 41.09	678 58.91	1151 100.00

Statistics for Table of pathway by type

Statistic	DF	Value	Prob
Chi-Square	2	205.9956	<.0001
Likelihood Ratio Chi-Square	2	212.7148	<.0001
Mantel-Haenszel Chi-Square	1	0.0003	0.9864
Phi Coefficient		0.4230	
Contingency Coefficient		0.3896	
Cramer's V		0.4230	

Table 16 Continued.

Sample Size = 1151

The GLIMMIX Procedure

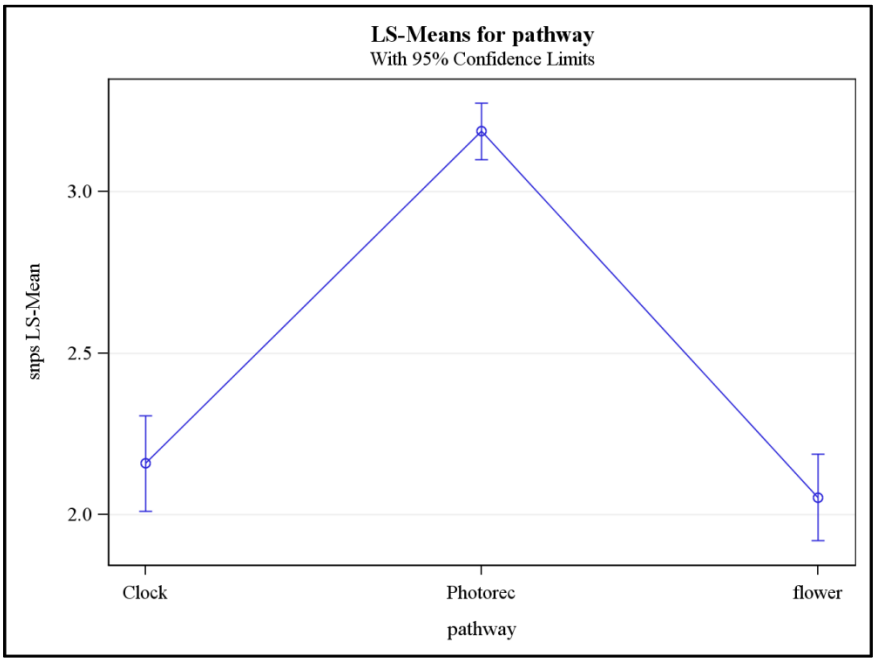
Model Information	
Data Set	CJ.PATHWAYS
Response Variable	snps
Response Distribution	Poisson
Link Function	Log
Variance Function	Default
Variance Matrix	Diagonal
Estimation Technique	Maximum Likelihood
Degrees of Freedom Method	Residual

Pathway Least Squares Means					
Pathway	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Clock	2.1580	0.07464	76	28.91	<.0001
Photorec	3.1868	0.04407	76	72.31	<.0001
Flower	2.0528	0.06733	76	30.49	<.0001

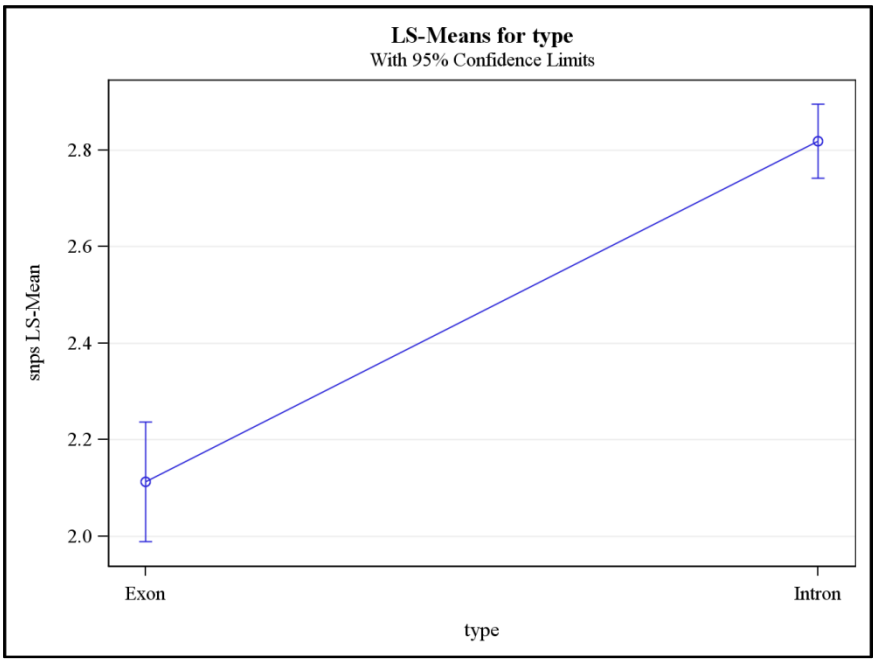
Table 16 Continued.

Type Least Squares Means					
type	Estimate	Standard Error	DF	t Value	Pr > |t|
Exon	2.1128	0.06221	76	33.96	<.0001
Intron	2.8190	0.03852	76	73.19	<.0001

Pathway*Type Least Squares Means						
Pathway	Type	Estimate	Standard Error	DF	t Value	Pr > |t|
Clock	Exon	1.4604	0.1336	76	10.93	<.0001
Clock	Intron	2.8556	0.06652	76	42.93	<.0001
Photorec	Exon	3.4485	0.05376	76	64.15	<.0001
Photorec	Intron	2.9251	0.06984	76	41.88	<.0001
Flower	Exon	1.4295	0.1187	76	12.04	<.0001
Flower	Intron	2.6762	0.06363	76	42.06	<.0001

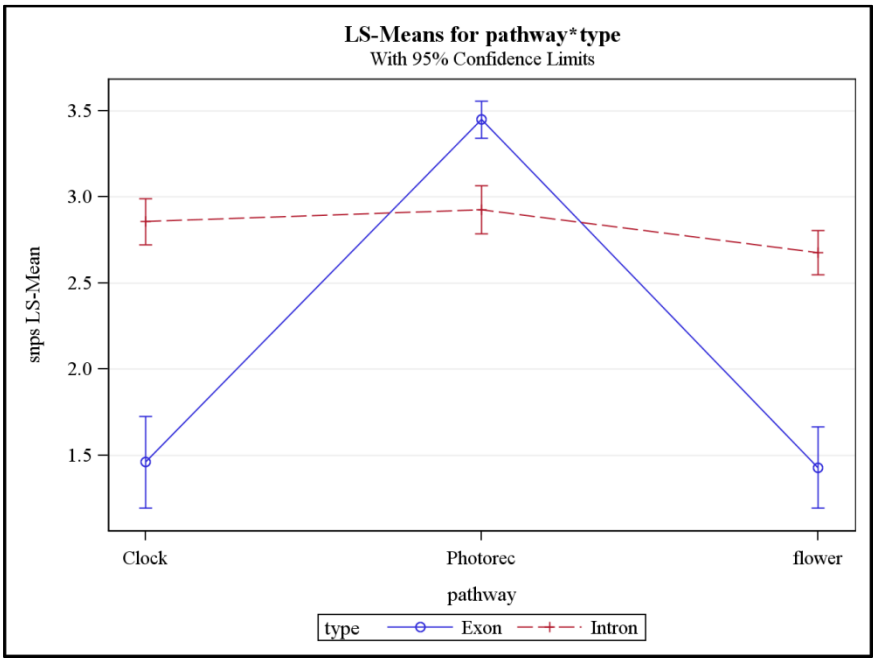


a) Pathway Least Square Means

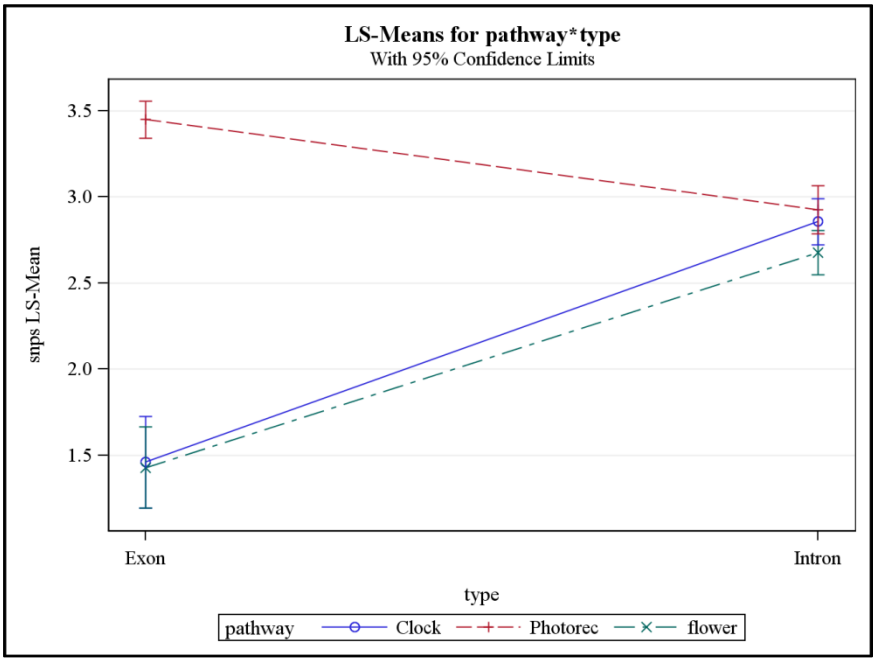


b) Type Least Square Means

Figure 15 Least Square Means Interaction Plots for SNPs



c) Least Square Means for Pathway by Type Vertical



d) Least Square Means for Pathway by Type Horizontal

Figure 15 Continued.

Table 17 SNP Ratios in Different Pathway Categories

		Floral Network	Circadian Clock	Photoreceptors without *PHYA2*, *CRY1 B*, and *CRY2 B*	Photoreceptors
Intronic	A/D genome	59	61	41	40
	Intraspecific (w/c)	332	330	310	268
	Interspecific (b/h)	272	268	442	447
Exonic	A/D genome	62	65	96	101
	Intraspecific (w/c)	444	424	122	121
	Interspecific (b/h)	1110	242	188	199

No significant differences were found between A/D homeolog gene copies in the photoreceptor pathway category. The photoreceptor pathway category had an increased level of A/D genomic SNPs in the intronic region compared to the circadian clock pathway category and the floral network category. Interestingly, the floral network pathway category had more intraspecific (cultivated vs. wild) SNPs in the exons, than interspecific (*G. barbadense* vs. *G. hirsutum*) SNPs (Table 20), while the floral network category had fewer intraspecific SNPs in the introns and more interspecific SNPS (Table 20). The breakdown of SNPs within each pathway by intronic and exonic values is shown in Tables 18, 19, and 20.

Table 18 SNPs in the Photoreceptors

a) Exonic SNPs

Gene	A/D genomes	*G. hirsutum*/ *G. barbadense*	Cultivated/ Wild	*G. incanum*	Single SNP
CRY1 A	1	0	0	0	0
CRY1 B	11	1	2	58	0
CRY2 A	1	1	2	2	0
CRY2 B	3	2	2	21	4
CRY3	4	0	0	2	1
PHYB	8	1	2	2	1
PHYA 1	15	2	3	7	4
PHYA 2	6	7	15	25	2
PHYC	6	17	22	19	15
PHYE	4	3	0	6	6
HY6	12	2	11	2	3
Exon Totals	71	36	59	144	36

b) Intronic SNPs

Gene	A/D genomes	*G. hirsutum*/ *G. barbadense*	Cultivated/ Wild	*G. incanum*	Single SNP
CRY1 A	14	2	5	0	6
CRY1 B	24	2	1	9	2
CRY2 A	2	1	0	3	3
CRY2 B	0	0	4	3	3
CRY3	8	0	2	7	8
PHYB	15	2	3	1	4
PHYA 1	1	0	0	0	1
PHYA 2	1	0	0	0	0
PHYC	19	0	0	10	4
PHYE	0	0	0	0	0
HY6	17	2	0	11	5
Intron Totals	101	9	15	44	36

Table 19 SNPs in the Circadian Clock

a) Exonic SNPs

Gene	A/D genomes	*G. hirsutum/ G. barbadense*	Cultivated/ Wild	*G. incanum*	Single SNP
ELF3	0	0	0	1	0
FKF1_ADO3	3	2	0	1	0
GI Ex9to10_ A	0	0	0	0	0
GI Ex10to11_ A	0	0	0	1	0
GI Ex10to11_ B	0	0	0	0	0
GI Ex11to12_ A	0	2	0	0	0
GI Ex11to12_ B	6	0	1	1	1
PRR5	1	0	1	2	1
PRR7 A	6	0	0	3	1
PRR7 B	7	3	2	3	3
LHY 1	0	0	0	0	0
LHY 2	1	0	0	1	0
TOC1	2	0	0	0	0
Exon Totals	26	7	4	13	6

b) Intronic SNPs

Gene	A/D genomes	*G. hirsutum/ G. barbadense*	Cultivated/ Wild	*G. incanum*	Single SNP
ELF3	11	4	3	11	3
FKF1_ADO3	14	5	2	11	4
GI Ex9to10_ A	9	1	0	7	13
GI Ex10to11_ A	5	2	3	3	0
GI Ex10to11_ B	0	2	1	0	2
GI Ex11to12_ A	14	2	1	4	3
GI Ex11to12_ B	3	0	2	4	3
PRR5	14	0	1	10	7
PRR7 A	4	0	1	1	2
PRR7 B	4	0	1	2	0
LHY 1	5	2	7	1	0
LHY 2	4	0	0	1	2
TOC1	4	1	0	0	0
Intron Totals	70	16	13	55	35

Table 20 SNPs in the Floral Network

a) Exonic SNPs

Gene	A/D genomes	*G. hirsutum/ G. barbadense*	Cultivated/ Wild	*G. incanum*	Single SNP
AGL16	0	0	0	1	0
AGL3_SEP4	0	0	2	0	2
AGL30	0	0	0	0	1
AGL32	0	0	0	0	0
AGL6	0	0	0	0	0
AGL65	3	0	0	1	2
AP1	2	0	0	1	0
ATGRP7	1	0	0	1	1
COL4	8	0	2	2	3
COL5	9	2	0	0	2
COP1	0	0	0	1	0
DET1	0	0	0	0	0
FD	12	0	1	4	4
PFT1	0	0	0	0	0
PI A	0	0	0	0	0
PI B	0	0	0	0	0
SPA4	1	0	0	2	0
Exon Totals	36	2	5	13	15

Table 20 Continued.

b) Intronic SNPs

Gene	A/D genomes	*G. hirsutum*/ *G. barbadense*	Cultivated/ Wild	*G. incanum*	Single SNP
AGL16	5	2	0	3	7
AGL3_SEP4	2	0	4	0	8
AGL30	15	3	4	3	4
AGL32	1	0	0	3	1
AGL6	7	0	0	0	7
AGL65	22	3	3	6	9
AP1	10	2	0	7	0
ATGRP7	4	0	0	1	1
COL4	1	1	0	2	1
COL5	0	2	0	0	1
COP1	14	2	0	12	3
DET1	1	0	0	0	0
FD	1	2	1	2	0
PFT1	4	0	0	5	1
PI A	11	4	3	11	4
PI B	3	1	2	0	3
SPA4	1	0	1	0	0
Intron Totals	102	22	18	55	50

Synonymous and Non-synonymous Sites across *Gossypium*

Utilizing Geneious®, we were able to align all exonic fragments from the eight taxa and compare those with the Arabidopsis coding DNA sequence (CDS). Geneious® allowed us to export those alignments as .meg files, which easily were integrated into Mega 5.2.1[341, 342]. These coding regions allowed us to calculate the rates of

synonymous (dS) and non-synonymous (dN) divergence within the candidate genes between species and in the different pathway categories. Synonymous rates told us the rate of nucleotide substitutions resulting in coding for the same amino acid. The non-synonymous rates showed us the nucleotide substitutions that caused changes in the coding for amino acids. The combined ratio of dN/dS provided the protein composition selection pressures. The dN/dS ratio >1 indicated possible positive selection.

Across the Candidate Genes

As shown in Table 21, we estimated the dN/dS for each candidate gene in pairwise comparisons across the following categories: *Arabidopsis thaliana* vs. *Gossypium raimondii* (D5); *Gossypium raimondii* vs. *Gossypium herbaceum* (A1); *Gossypium barbadense* (cultivated vs. wild); *Gossypium hirsutum* (cultivated vs. wild); and *Gossypium barbadense* (3-79) vs. *Gossypium hirsutum* (TM-1). Some candidate genes were not represented due to no changes in the exonic region or too few amino acid sites. Our total averages for dN/dS and dS/dN (Table 21 a-e) were based on candidate genes where nucleotide changes occurred.

The mean dN/dS ratio difference between *Arabidopsis thaliana* and *Gossypium raimondii* for 35 candidate genes was 0.324 for the 3,266 sites. The mean dN/dS ratio between *Gossypium herbaceum* and *Gossypium raimondii* was 0.204 for the 3,266 sites. Within *Gossypium hirsutum* (cultivated versus wild), the 'A' sub-genome dN/dS ratio was 0.003, while the 'D' sub-genome dN/dS ratio was 0.041. For the intraspecific *Gossypium barbadense* (cultivated versus wild), the 'A' sub-genome had a dN/dS ratio of 0.124, while the 'D' sub-genome had a dN/dS ratio of 0.029. Finally, the mean dN/dS ratio between *Gossypium hirsutum* (TM-1) and *Gossypium barbadense* (3-79) was 0.061 for the 3,266 sites.

Very few genes showed mean dN/dS of greater than one. *Flowering Locus D (FD)* displayed positive selection for A/D divergence, while *PRR7 A* appeared to be close to positive selection for A/D divergence (Table 21b). *PHYA 2* exhibited positive selection for divergence within 'A' sub-genome of *Gossypium barbadense* (Table 21c). *PHYC-like* had positive selection for divergence within the 'A' sub-genome between *Gossypium hirsutum* and *Gossypium barbadense* (Table 21e).

Table 21 Candidate Genes dN/dS and dS/dN Ratios

a) *Arabidopsis thaliana* (At) vs. *Gossypium raimondii* (D5)

Gene	dN/dS Ratio D5 vs. At	dS/dN Ratio D5 vs. At	Total Sites
AGL16	0.490	2.042	3
AGL3	0.771	1.297	79
AGL30	0.084	11.891	40
AGL6	0.070	14.353	29
AP1	0.089	11.200	30
COL4	0.558	1.792	72
COL5	0.244	4.105	106
CRY1B	0.098	10.202	303
CRY2 A	0.073	13.784	120
CRY2 B	0.531	1.884	119
CRY3	0.164	6.110	80
FD	1.054	0.949	170
FKF1	0.332	3.013	85
GI A	1.380	0.724	108
GI B	1.423	0.703	173
HY6	0.215	4.658	128
LHY 1	0.029	34.818	20
LHY 2	0.349	2.866	20
PFT1	0.057	17.547	32
PHYA 1	0.074	13.436	178
PHYA 2	0.279	3.583	289
PHYB	0.118	8.469	174
PHYC	0.382	2.618	362
PHYE	1.354	0.739	200
PRR7 A	0.389	2.570	63
PRR7 B	0.380	2.629	62
SPA4	0.064	15.550	34
TOC1	0.290	3.444	57
Total Averages	**0.405**	**7.035**	**3136**

Table 21 Continued.

b) *Gossypium raimondii* (D5) vs. *Gossypium herbaceum* (A1)

Gene	Dn/Ds Ratio D5 vs. A1	Ds/Dn Ratio D5 vs. A1	Total Sites
AGL3	0.750	1.333	79
AP1	0.266	3.765	30
COL5	0.119	8.429	106
CRY1 B	0.161	6.200	303
CRY2 B	0.517	1.933	119
CRY3	0.425	2.353	80
FD	1.381	0.724	170
FKF1	0.162	6.167	85
HY6	0.514	1.944	128
PHYA 1	0.200	5.000	178
PHYA 2	0.353	2.833	289
PHYB	0.106	9.400	174
PHYC	0.222	4.500	362
PHYE	0.708	1.412	200
PRR7 A	0.958	1.043	63
PRR7 B	0.286	3.500	62
Total Averages	**0.446**	**3.784**	**2428**

c) *Gossypium barbadense* (Cultivated vs. Wild)

	dN/dS Ratio *G. barbadense* Cultivated vs. Wild	dN/dS Ratio *G. barbadense* Cultivated vs. Wild	dS/dN Ratio *G. barbadense* Cultivated vs. Wild	dS/dN Ratio *G. barbadense* Cultivated vs. Wild	Total Sites
Gene	**A sub-genome**	**D sub-genome**	**A sub-genome**	**D sub-genome**	
CRY1 B		0.071		14.083	303
HY6	0.104	0.304	9.625	3.292	128
PHYA 2	3.500		0.286		289
PHYB		0.333		3.000	174
PHYC	0.727	0.308	1.375	3.250	362
Total Averages	**1.444**	**0.254**	**3.762**	**5.906**	**1256**

Table 21 Continued.

d) *Gossypium hirsutum* (Cultivated vs. Wild)

	dN/dS Ratio *G. hirsutum* Cultivated vs. Wild	dN/dS Ratio *G. hirsutum* Cultivated vs. Wild	dS/dN Ratio *G. hirsutum* Cultivated vs. Wild	dS/dN Ratio *G. hirsutum* Cultivated vs. Wild	Total Sites
Gene	**A sub-genome**	**D sub-genome**	**A sub-genome**	**D sub-genome**	
CRY2 A		0.308		3.250	120
CRY2 B	0.093		10.750		119
GI A		0.286		3.500	108
HY6		0.286		3.500	128
PHYC		0.556		1.800	362
Total Averages	**0.093**	**0.359**	**10.750**	**3.013**	**837**

e) *Gossypium barbadense* (3-79) *vs. Gossypium hirsutum* (TM-1)

	dN/dS Ratio 3-79 vs. TM-1	dN/dS Ratio 3-79 vs. TM-1	dS/dN Ratio 3-79 vs. TM-1	dS/dN Ratio 3-79 vs. TM-1	Total Sites
Gene	**A sub-genome**	**D sub-genome**	**A sub-genome**	**D sub-genome**	
CRY1 B		0.600		1.667	303
CRY2 B		0.286		3.500	119
GI A	0.440		2.273		108
HY6		0.286		3.500	128
PHYC	1.500	0.517	0.667	1.933	362
PHYE		0.625		1.600	200
Total Averages	**0.970**	**0.463**	**1.470**	**2.440**	**1220**

Across Different Pathway Categories

The dN/dS ratio across different pathway categories was calculated for different pathway categories the following groups*: Arabidopsis thaliana* vs. *Gossypium raimondii* and *Gossypium raimondii* vs. *Gossypium herbaceum*. *Gigantea (GI) A-like and B-like* appeared to be diverging more rapidly from the other clock genes when comparing Arabidopsis versus *Gossypium raimondii* D5 (Table 22b). Respectively, both *FD-like* and *PHYE-like* also appeared to be diverging more rapidly from Arabidopsis in their corresponding pathways (Table 22 a, c).

Table 22 *Arabidopsis thaliana* vs. *Gossypium raimondii* dN/dS and dS/dN Ratio by Pathway Category

a) Photoreceptor Pathway Category

Gene	dN/dS Ratio D5 vs. At	dS/dN Ratio D5 vs. At	Total Sites
CRY1 B	0.098	10.202	303
CRY2 A	0.073	13.784	120
CRY2 B	0.531	1.884	119
CRY3	0.164	6.110	80
HY6	0.215	4.658	128
PHYA 1	0.074	13.436	178
PHYA 2	0.279	3.583	289
PHYB	0.118	8.469	174
PHYC	0.382	2.618	362
PHYE	1.354	0.739	200

Table 22 Continued.

b) Circadian Clock Pathway Category

Gene	dN/dS Ratio D5 vs. At	dS/dN Ratio D5 vs. At	Total Sites
FKF1	0.332	3.013	85
GI A	1.380	0.724	108
GI B	1.423	0.703	173
LHY 1	0.029	34.818	20
LHY 2	0.349	2.866	20
PRR7 A	0.389	2.570	63
PRR7 B	0.380	2.629	62
TOC1	0.290	3.444	57

c) Floral Network Pathway Category

Gene	dN/dS Ratio D5 vs. At	dS/dN Ratio D5 vs. At	Total Sites
AGL3	0.771	1.297	79
AGL30	0.084	11.891	40
AGL6	0.070	14.353	29
AP1	0.089	11.200	30
COL4	0.558	1.792	72
COL5	0.244	4.105	106
FD	1.054	0.949	170
PFT1	0.057	17.547	32
SPA4	0.064	15.550	34

There were fewer divergent candidate genes showing a ratio dN/dS ratio greater than zero between *Gossypium raimondii* and *Gossypium herbaceum*, than between *Gossypium raimondii* and Arabidopsis in the different pathway categories. In the floral network pathway category, *FD-like* indicated positive selection for divergence between *G. raimondii* and *G. herbaceum* (Table 23c). Although both the photoreceptor and circadian clock pathway categories did not have any candidate genes with a dN/dS ratio above one, *PRR7 A–like* had a 0.958 dN/dS ratio indicating a strong likelihood positive selection may take place (Table 23 b).

Table 23 *Gossypium raimondii* (D5) vs. *Gossypium herbaceum* (A1) dN/dS and dS/dN Ratio by Pathway Category

a) Photoreceptor Pathway Category

Gene	dN/dS Ratio D5 vs. A1	dS/dN Ratio D5 vs. A1	Total Sites
HY6	0.514	1.944	128
CRY1 B	0.161	6.200	303
CRY2 B	0.517	1.933	119
CRY3	0.425	2.353	80
PHYA 1	0.200	5.000	178
PHYA 2	0.353	2.833	289
PHYB	0.106	9.400	174
PHYC	0.222	4.500	362
PHYE	0.708	1.412	200

Table 23 Continued.

b) Circadian Clock Pathway Category

Gene	dN/dS Ratio D5 vs. A1	dS/dN Ratio D5 vs. A1	Total Sites
FKF1	0.162	6.167	85
PRR7 A	0.958	1.043	63
PRR7 B	0.286	3.500	62

c) Floral Network Pathway Category

Gene	dN/dS Ratio D5 vs. A1	dS/dN Ratio D5 vs. A1	Total Sites
FD	1.381	0.724	170
AGL3	0.750	1.333	79
AP1	0.266	3.765	30
COL5	0.119	8.429	106

Overall, the dN/dS ratio within both the floral network and the circadian clock was lower than the photoreceptor pathway category. All dN/dS ratios were from the photoreceptor pathway category within *Gossypium barbadense* cultivated versus wild. In both sets of *Gossypium hirsutum* cultivated versus wild and *Gossypium barbadense* (3-79) versus *Gossypium hirsutum* (TM-1), all candidate genes representing divergence were from the photoreceptor pathway category except for *GI A-like* in the circadian clock pathway category (Table 24 a-c). Within *Gossypium barbadense*, *PHYA 2-like* was diverging more quickly, than any other candidate gene. Within *Gossypium hirsutum*, no candidate genes were under positive selection. No positive selection was seen between *Gossypium barbadense* (3-79) and *Gossypium hirsutum* (TM-1) for divergence.

Table 24 The dN/dS and dS/dN Ratios within *Gossypium barbadense* and *Gossypium hirsutum*

a) Intraspecific dN/dS and dS/dN of *Gossypium barbadense*

Gene	dN/dS Ratio Cult. vs Wild	dN/dS Ratio Cult. vs Wild	dS/dN Ratio Cult. vs Wild	dS/dN Ratio Cult. vs Wild	Total Sites
	A sub-genome	D sub-genome	A sub-genome	D sub-genome	
CRY1B		0.071		14.083	303
HY6	0.104	0.304	9.625	3.292	128
PHYA 2	3.500		0.286		289
PHYB		0.333		3.000	174
PHYC	0.727	0.308	1.375	3.250	362

Table 24 Continued.

b) Intraspecific dN/dS and dS/dN of *Gossypium hirsutum*

Gene	dN/dS Ratio Cult. vs Wild	dN/dS Ratio Cult. vs Wild	dS/dN Ratio Cult. vs Wild	dS/dN Ratio Cult. vs Wild	Total Sites
	A sub-genome	**D sub-genome**	**A sub-genome**	**D sub-genome**	
CRY2 A		0.308		3.250	120
CRY2 B	0.093		10.750		119
GI_A		0.286		3.500	108
HY6		0.286		3.500	128
PHYC		0.556		1.800	362

c) Interspecific dN/dS and dS/dN of *Gossypium barbadense* (3-79) and *Gossypium hirsutum* (TM-1)

Gene	dN/dS Ratio 3-79 vs TM-1	dN/dS Ratio 3-79 vs TM-1	dS/dN Ratio 3-79 vs TM-1	dS/dN Ratio 3-79 vs TM-1	Total Sites
	A sub-genome	**D sub-genome**	**A sub-genome**	**D sub-genome**	
CRY1 B		0.600		1.667	303
CRY2 B		0.286		3.500	119
GI A	0.440		2.273		108
HY6		0.286		3.500	128
PHYC	1.500	0.517	0.667	1.933	362
PHYE		0.625		1.600	200

Location of Genes on the D5 Genome

The characterized thirty-eight homologous candidate gene fragments were queried against the *Gossypium raimondii* draft genome in Geneious® to determine locations based on sequence. These fragments were then aligned to the draft genome (Figures 16 a-c). Also, other proposed candidate genes known in *Arabidopsis thaliana* for the floral development network were positioned onto the draft genome. Possible linkage groupings of close homologous candidate genes were ascertained by their alignment to the *Gossypium raimondii* draft genome.

a)

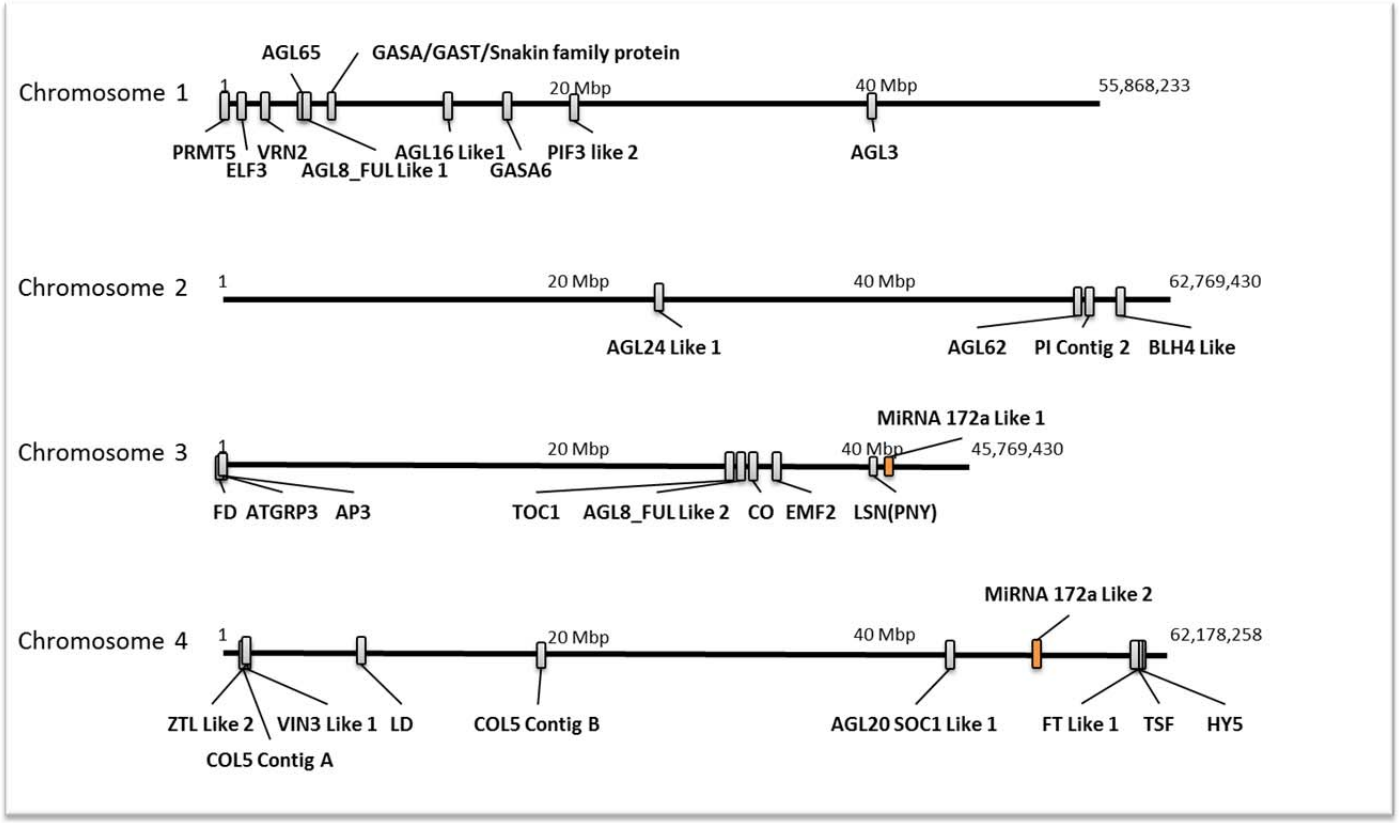


Figure 16 Alignment of Photoreceptors, Circadian Clock, and Floral Regulatory Network Genes upon the *Gossypium raimondii* Draft Genome

b)

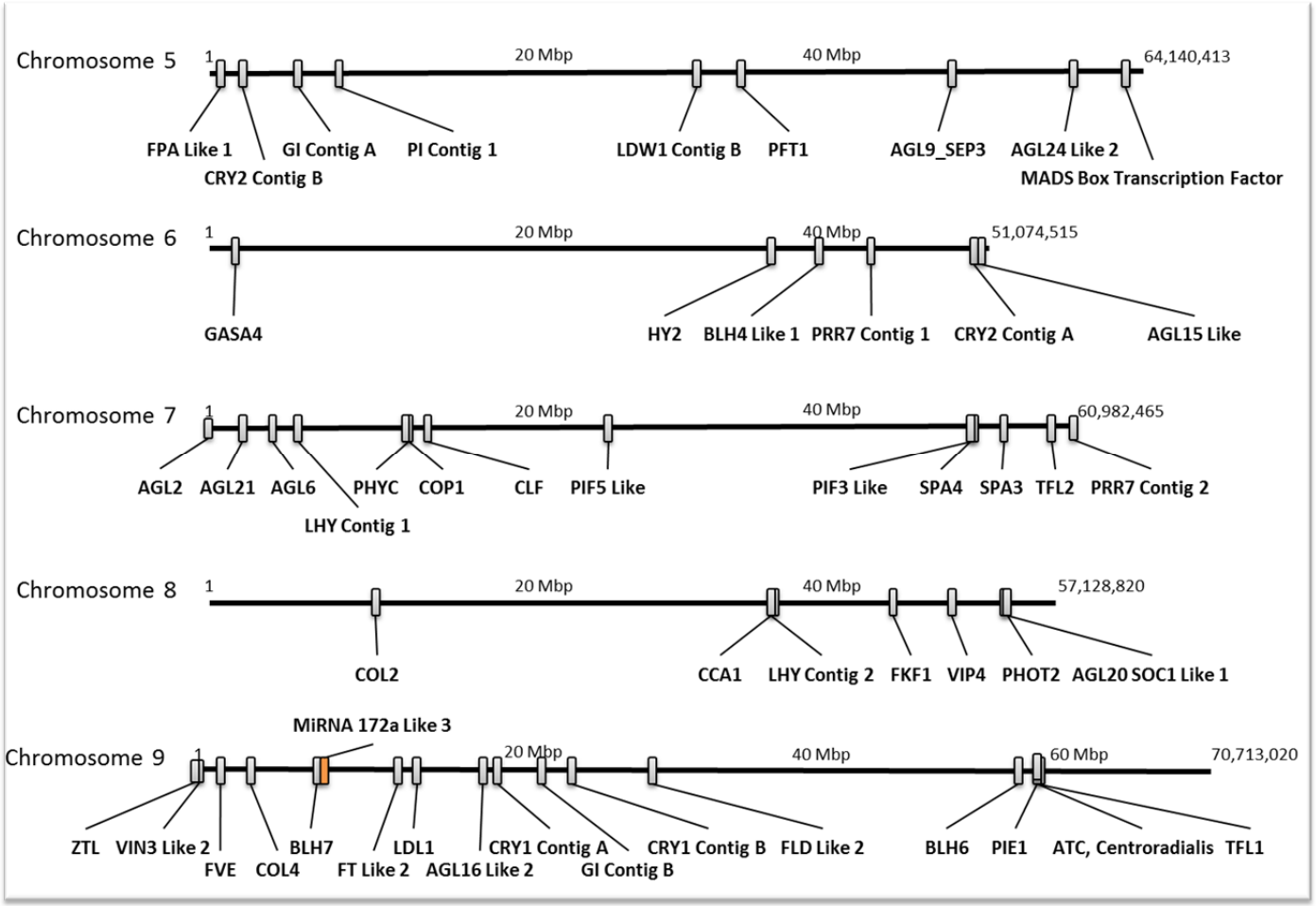


Figure 16 Continued.

c)

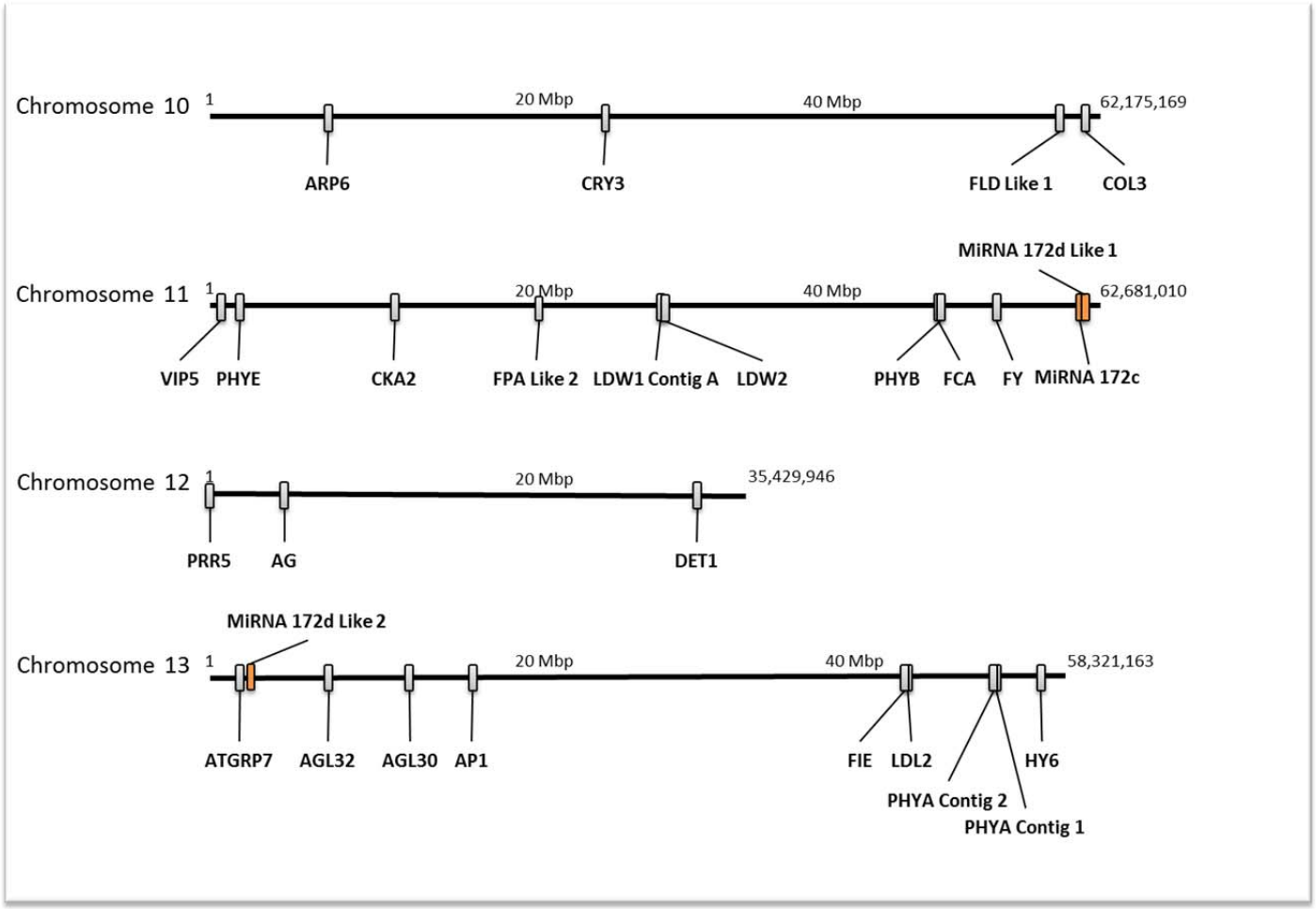


Figure 16 Continued.

From the original candidate genes initially thought to be involved in photoperiod independence, two candidate genes were not found when they were queried against the *Gossypium raimondii* draft genome. The two undiscovered genes were: *Frigida (FRI)* and *Flowering Locus C (FLC)*. Recent searches on NCBI were done to verify no *FRI* and *Frigida-Like 2 (FRL2)* complements were found in the non-redundant databases for *Gossypium* (taxid id: 3633), the uni-gene sequences, or the EST databases. On May 7th, 2013*, Frigida-Like 1 (FRL1)*(Table 25) and *FLC* (Tables 26, 27, 28) did have complements shown in the NCBI search, but the reciprocal blast approach did not yield associated complements to those genes when queried against The Arabidopsis Information Resource (TAIR10) version 10 [339]. The *Frigida-Like 1* complement reciprocal blast showed no significant results with the highest E-value at 0.061, while the *FLC* complement reciprocal blast mapped to the K-box region of the *Agamous-like* family.

Table 25 *Frigida-Like 1 (FRI1)* Blast Result from NCBI

>gi|48803701|gb|CO105015.1|CO105015 GR__Eb0034P10.r GR__Eb *Gossypium raimondii* cDNA clone GR__Eb0034P10 3', mRNA sequence
TTTAATTGTTAGGGATTTAAAAAAACATTAGTCCAGAATTTTAACAATTCAGTAGTTGATCTCTC CAAAACCAAAAGGAATGTCATCGGAAAATCTCTCGGCGACGCCGTCGCTGAGTAGTATTGAATC GGCAATCAAACTTATTGACGTCAAAAAGGACGAACTCAAGAGAGCGTTCGATGATCTCCAAGCC AACTCTCAGCGCCTCTCTTCTTTCTCTATCTCCTGGTCCGACCTCGACTCTCACTTCACCGCTCTT CAAAATTCCGTCACTCAACGCTTCCGCATCCTCCTGTCTCGCGAAGCCATTCACAACCCAATTGA TCACGTCCCCACTCTACCACTCGCCTCTAAACAGGGAGACCCATCTGCTCCACACCCGTTAACCG AACAAGACCCAGTTGACTCGGTGGTTAATCAGCCATTAACTCATCTTTTACCGTTGGATAGTGAT AAACCATCGTCTTCAAATTCGTTGAGCGTGCAGTCCGATGGTTCGGTTCCGCCGTCAGGCAACAT TGACTCGGTGGTGACTCGACCAGAACTGAAGGAATTTTGTGAGAGAATGGATGGCAAGGGGTT GAGGAAATACATAAACGACCACGTAAAAGAGCGGGAAGCTATCCGAATGGAGCTTCCCGATGC ACTGCAGAGTGCTGCGGACCCTGGGGCAATGGTTTTGGATGCAATGGAGGGGTTTTACGCTGAG AATTTGCATTCTAAAGGTGAGAAAGACCCTGAATTGCTCGGATTAAGGAGGGTTTGTGTGGTTT TATTGGAGCAGTTAATGGAAACTGGTTTAAGTTTCAGT

Table 26 *FLC* Blast Result from NCBI - TAIR Reciprocal Blast to *Agamous (AG)*

>gi|122938398|gb|EF202091.1| *Gossypium hirsutum* MADS-box protein MADS7 mRNA, complete cds, alternatively spliced
TTGGAATTCAAGAGCTTAGCCTTTCTCATAGTTCAATTCCTTGCCAGCTTTGAACCATGGAGTTCCC TAATCTAGACCCAGAAAGCTCTTCCCAGAAAAAAATGGGAAGAGGCAAAATCGAGATTAAGCGG ATCGAGAACACCACTAATCGACAAGTTACCTTCTGCAAGCGCCGCAATGGACTGCTCAAAAAGGC CTATGAATTATCTGTTCTTTGTGATGCTGAGGTGGCTCTCATAGTCTTCTCCAGCCGTGGCCGTCTC TATGAATATGCTAACAACAGTGTGAGAGCAACAATTGAGAGATACAAGAAAGCATGCTCAGATG CAACAACTCCAGGGTCTGTGGCTGAAGCCAACATTCAGTTCTACCAGCAAGAAGCCACCAAACTC CGACGACAAATTCGTGATGTTCAGAACATGAACAGGCATATCCTTGGAGAGGCTCTGAGCTCATT GACCTTTAAGGAACTCAAGAACCTTGAAGGCAGGCTGGAGAAAGGCATTTGTAGAATCAGATCCA AAAAGAACGAATTACTGTTTGCAGAAATTGGATTCATGCAAAAGAGGGAAGTTGAACTGCAGAA CGATAACATGTACCTGCGAGCAAAAATAGCTGAAAATGAAAGAGCGCAACAACAATCAAACCAG CTGATGCAAGCAGCCTCCTCCTACAATCGCAACTTTCTGCCAGTAAACCTGCTGGAACCCAGTAAT AATGATTACTCCAACCAAGACCAAACTCCTCTTCAACTTGTATAAGCTTCTCTCTTCCACCTCCATC TATTCAAGCTGACATGCAAGCAGCAGCCTCTAATGTTCTACCAACTTTATCTCTGGTTAAGCATAT ATATATATCTTATAAGCCATGCAAGTATATATATATATATTTATATATGGTTTATAAACAGTTGAA AGCCTAGTTGTTGATTATGATGAATGTAAGGAACTTAGGTTTCAGCACTGGACTCCTTTTTAACTG TTTTATTGCAGGAAAGGTTTATTTTTGTATTATATATGTTTCTTAAAACTCTACATATAGTTTTCAT CCCTTCTCTCAAAAAAAAAAAAAAAAA

Table 27 *FLC* Blast Result from NCBI - TAIR Reciprocal Blast to *Agamous-Like 6 (AGL6)*

>gi|212525793|gb|FJ409870.1| *Gossypium hirsutum* MADS-13 mRNA, complete cds
GATCACCATTTTTGTATAACACCAAAACCCCACCTCTCAGACACCACAGCCATTGTTCTAAAACCA
AAAAAAATGGGGAGAGGAAGAGTGGAGCTGAAGAGGATAGAAAACAAGATCAACAGACAAGTG
ACCTTCTCTAAGAGAAGAAATGGTTTGCTTAAGAAAGCTTATGAGCTTTCTGTTCTTTGTGATGCTG
AGGTTGCTCTTATCATCTTCTCTAGTCGTGGCAAGCTCTACGAGTTTGGCAGTTCAGGTATGAGCA
AGACCCTTGAGCGATACCAGCGTTGCTGCTTTACTCCTCAAGACAACAGCCTTGAACGCGAAACAC
AGAATTGGTACCAAGAGGTAACCAAGTTAAAGGCAAAATATGAAGCACTGCAACGCACTCAAAG
GCATTTGCTTGGAGAAGATCTTGGACCATTGAATGTTAAGGAGCTGCAAAACCTTGAGAAACAGC
TTGAAGGAGCTCTTGCACTGGCTAGACAAAGGAAGACACAGATCATGATAGAACAAATGGAAGAC
CTCCGTAAAAAGGAGCGTGAGCTTGGAGACCTTAACAAACAGCTGAAAATCAAGCTAGAGGCAGA
AGGACAAAACCTCAAAACAATCCAAGGTTTATGGAGTAGTGGTGCAGCAGCTGAAACTAGCAACT
TTCCTCTGCATCCTTCTCACCCACATCCTATGGATTGTGATCATGAACCTGTTCTGCAAATAGGGTA
CCATCACTTTGTTCAGGCTGAAGGATCTTCAGTCCCAAAAAGCATGGCCGGTGAGACCAACTTCAT
CCACGGATGGGTCATTTGAGCCCTCTCCTAAAAGCAACACAGCTACATATATAATATTTTTGTGAT
TTTTGTCTCTTGTTTTTTGTTTGGGATTTGTAATATTGCCATCATATATATATATAGAGACAGCTTGT
TCAAGTGTACAACATAAGAAAACATGCATGGATCTTAAGGAGCATTTCTCCTCTATTGTGATATAT
ACTGTTGCCTATATATATACAGCTTTTTACTGTTATATTTCAATTTCAGGCACTATTACCATTTGAG
CCTACAAAAAAAAAAAAAAAAAAAAA

Table 28 *FLC* Blast Result from NCBI - TAIR Reciprocal Blast to *Agamous-Like 11 (AGL11-STK)*

>gi|122938394|gb|EF190548.1| *Gossypium hirsutum* MADS-box protein MADS5 mRNA, complete cds
TAGACCCATTTGAAAGGTGAGATTTATGATGAAATTGGAGTGCAAATGCAAGCTATATAGTTTTGA
ATAGGGACTGATTTTCTTTGCAATAGCTGGCCTCACATCTTCAACCAAAAATCCAAACACTTTGGC
CTTCTTTGAATCTGTAATTTTCCGAAAGCTGAAGTGGCCGGAACTGTGGCCTAACCTCAGTAGGAA
TATATAAAAGGGTGAAGATGGGAAGAGGAAAGATAGAGATAAAGAGGATCGAGAACACCACGAA
TCGTCAGGTTACGTTTTGCAAGCGCAGGAATGGGCTGCTGAAGAAAGCATACGAGCTGTCAGTCCT
TTGTGATGCAGAAGTTGCTCTTATCGTCTTCTCCACCCGTGGCCGCCTCTACGAGTACTCCAACAAC
AACATAAGATCAACAATAGAGAGGTACAAGAAGGCATGTTCCGGTACCTCAAACACAAATACCGT
CACGGAAATCAATGCTCAGTATTATCAACAAGAATCAGCCAAGTTGAGGCAGCAGATTCAAATGC
TACAGAATTCTAGCAGGCATCTAATGGGAGATTCGTTGAGTTCCTTGACAGTGAAGGAGTTAAAGC
AGCTAGAGAATAGACTTGAACGAGGGATTACTAGAATCAGATCAAAGAAGCACGAAATGCTGCTA
GCTGAAATTGAATATTTTCAGAAAAGGGAAGTTGAGCTGGAAAATGAAAGTGTATGTCTCCGAGC
TAAGATTGCAGAGATAGAGAGGGTTGAGGAAGCAAACATGGTAACAGGAGCAGAGCTGAATGCT
ATTCAAGCCTTGGCATCTCGCAATTTCTTTACTCCAAATGTGATTGAGAGAGGAACTCCCACTCCCT
ACTCCCACCATGACAAGAAGATTCTCCATCTTGGGTAGAGAGAGTTGGAGAGAACAAATCTGAAA
AGGGTGTTGTGATATTATGAGATTAAATAAGGATGCATTTCAACCATATGTACTCTACATATATTA
TGAAGCTGTCATGTAATTTGTTACTTGTTTTGTTCTGTTATATTGTCTGAAAACCTATACTAGTGTA
ACGTGTTTAATTTTGTGTGTTAAAAAAAAAAAAAAAAA

A few longer sequences from the *FLC* complements were correlated past the K-box region to: 1) *Agamous (AG)* with an E-value of $4e^{-16}$, 2) *Agamous-like 6 (AGL6)* with an E-value of $2e^{-54}$, and 3) *Agamous-like 11 (AGL11-STK)* with an E-value of $7e^{-30}$ (Tables 26, 27, 28). The K-box region in the *Agamous-like* family is very similar, so distinguishing a gene based solely off that region is difficult.

FLC has not been discovered outside the *Brassicales* species. One study reported the discovery of an *FLC* homolog in *Cichorium intybus (CiMFL)*, but upon closer evaluation this fragment was found to have a 99% maximum identity match to *Arabidopsis thaliana* on the NCBI blast query [343]. Not even close relatives in the brassica family had this high level of identical matching nucleotide. Looking into this anomaly, we found identical matches to Arabidopsis for the *CiMFL* GenBank accession numbers: FJ347972 and FJ347973. Interpreting this data, we concluded that *FLC* has not been discovered outside the *Brassicales* family.

Discussion

Similarity between Arabidopsis and Cotton

In the introduction, we discussed how flowering time is a vital function in a crop's reproduction. Basic knowledge to understand both the genetic and evolutionary history of the different pathway categories behind flowering (the floral network, the circadian clock, and the photoreceptors) was needed. Our study supported the idea that using candidate genes obtained from well-defined pathways in a closely related model plant species, *Arabidopsis thaliana*, was feasible because of the low percentages in nucleotide and amino acid substitutions (Table 7 a-b). The coding regions had high similarity to the homologous Arabidopsis regions. The highest level of substitution for coding sequences was 59% in *Constans-Like 4 (COL4)* and 52% in *Pseudo Response Regulator 7 A (PRR7 A)*. Since both of these orthologs are in multi-gene families with some redundant functions, the ability to diverge evolutionary might be due to a relaxed selective pressure or sub-functionalization.

Importance of Characterizing the A and D Genome Orthologs

Including both the 'A' and 'D' diploid sequence was imperative to this study because of the difficulty in distinguishing which paralogs within the allotetraploid cottons related back to the seven to eight (MYA) evolutionary divergences. Without having the 'A' diploid sequence, we would have mistaken certain nucleotide polymorphisms to be singletons in candidate gene orthologs that did not have significant coverage across all allotetraploids. In retrospect, Sanger sequencing out both the *G. raimondii* D5 and *G. herbaceum* 'A' orthologs would have increased the speed in

evaluating our data. Due to the high similarity between the 'A' and 'D' diploid sequences (Table 7 b), we utilized the 'D' genome sequence as a reference for the 'A1' Roche 454 sequences. Only a 6.5% difference in amino acid substitution levels was shown between the A' and 'D' diploid candidate gene orthologous sequences at the highest ratio (Table 7 b).

SNP Density Comparison

As expected, our study found higher levels of SNPs in the non-coding intronic regions, than those of the coding exonic regions. In exploring the differences between the 'A' and 'D' changes, we found that every 83 bp in an exonic region there would be a SNP indicating the evolutionary split seven to eight MYA. This was less frequent than the changes occurring in cotton EST sequences as reported by Flagel et al. in 2012 [344]. Our 'A' and 'D' changes occurring in the intronic region was every 49 bp.

Our data results agreed with Flagel et al. that frequency for *G. hirsutum* and *G. barbadense* changes was considerably less frequent than the 'A' and 'D' changes occurring within the sequences [344]. We reported that approximately every 255 bp, a *G. barbadense/G. hirsutum* change occurred in both the intronic and exonic regions. In our results, we discovered that every 246 bp changes occurred between the cultivated cotton lines and wild accessions in the exonic region. This was surprising because it was slightly more frequent than the *G. barbadense/G. hirsutum* changes. It suggests hybridization or introgression of *G. barbadense* and *G. hirsutum* during or since domestication, as has been suggested in the literature [5, 6, 27, 31, 33, 36, 40, 48, 201, 263, 265, 273, 287, 292, 294, 295, 345-348].

The out-group, *G. incanum* E4 genome group, was used to root all our SNP results. Moreover, it helped to verify the divergence estimate of the different taxa seven to eight MYA. The E genome had a higher frequency of changes in the in coding regions, when compared to that of the 'A' and 'D' genomes. This might indicate that the E genome diverged farther back than eight MYA. Notably, some changes occurring in the PS-6 line appeared to be more similar to *G. incanum* than to 3-79. This might suggest that *G. incanum* or a related taxon could have been introduced into the pedigree of PS-6 at some point.

Genes occurring only once within the genome (single genes) were under higher purifying selection. Multi-gene families with redundant functions did tend to have more SNPs, but no significant differences were found in SNP ratios when comparing genes and possible pseudo-genes. Fewer SNPs were observed in the single copy genes than genes in multi-gene families or having redundant functions.

Synonymous and Non-synonymous Divergence within the Candidate Genes

Our data was slightly lower than Flagel et al. for the ratio for dN/dS for evolution between *Gossypium raimondii* (D5) and *Gossypium herbaceum* (A1)[344]. In 2012, Flagel et al. reported at dN/dS ratio of 0.308 for D5 versus A1[344]. This study reported a dN/dS ratio of 0.204. These results still fall within the standard error reported by Flagel et al., but our study has comparatively fewer sites[344]. The results from both studies were statistically similar and gave credence to the divergence of the A1 genome and the D5 genome seven to eight MYA.

Comparatively, our dN/dS ratio for evolutionary changes within *Gossypium hirsutum* and within *Gossypium barbadense* was much lower, than that of Flagel et al. [344]. Although our results differ, it was not unexpected. This study focuses on a highly conserved set of pathways in all angiosperms. Our lower dN/dS ratio was likely due to three things: 1) the highly conserved nature of the photoreceptor, circadian clock, and floral network pathway candidate genes; 2) the fewer amount of synonymous and non-synonymous sites; and 3) a smaller number of genes analyzed.

Synonymous and Non-synonymous Divergence of Different Pathway Categories

Interestingly, this study found that the most divergent pathway category within cotton was the photoreceptor pathway. Both the floral network pathway category and the circadian clock pathway category showed fewer evolutionary divergences. Only, *PRR7 A* and *FD* suggested positive selection was occurring between the *Gossypium raimondii* (D5) and *Gossypium herbaceum* (A1) sequences.

Within the intraspecific comparisons for both *Gossypium barbadense* and *Gossypium hirsutum*, nine of the ten genes showing evolutionary divergence were from the photoreceptor pathway category. Many of these photoreceptor genes had multiple orthologous copies within cotton. These copies increased the dN/dS ratio discovered within our averages. Ortholog *GI A-like* was the only circadian clock pathway category gene to show evolutionary divergence within *Gossypium hirsutum* and between the interspecific comparison of *Gossypium barbadense* (3-79) and *Gossypium hirsutum* (TM-1). A greater likelihood that the changes behind photoperiodism during domestication might be located in the pathway categories before the onset of the floral network pathway *CO* and *FT* because of the higher divergence ratios occurring within the photoreceptor and circadian clock pathway categories.

Conclusion

Before novel traits from undomesticated species could be integrated into cultivated elite lines, the genetic basis of the loss of photoperiodism cotton in domesticated cotton must be understood. In cotton, there was a paucity of information on how different pathways interconnect to influence flowering time. This study characterized polymorphism differences in thirty-eight homologs of genes within the flowering time network, including photoreceptors, light dependent transcripts, circadian clock regulators, and floral integrators. This research asked if these genes were sound candidates for understanding the relationship between photoperiodism and flowering in cotton. Overall, we discovered appreciable SNP diversity within the candidate gene orthologs, including SNPs differentiating cultivated and wild *Gossypium barbadense* and *Gossypium hirsutum*. We found 36 SNPs within *Gossypium hirsutum* and 53 SNPs within *Gossypium barbadense*. These SNPs can be utilized as markers for integrating in new traits from wild accessions and to explore the molecular evolution of different genomes.

Methods

Plant Growth and Collection

3-79, D5, TM-1, PS-6, A1, K-46, TX-231 and E4

We took DNA from the postulated modern species A and D genomes (*Gossypium herbaceum* A1 and *Gossypium raimondii* D5) that are most closely related to the ancestral A and D to create a base reference to compare evolutionary changes between A/D genome differentiation occurring seven to eight (MYA) and changes post natural hybridization /genome duplication event occurring one to two MYA [39]. Tissue from photoperiod independent (PI) cultivars and photoperiod dependent (PD) varieties/accessions were collected to postulate the nucleotide differences between cultivated PI cotton and PD cotton varieties/accessions. To make accurate evolutionary comparisons between PI/PD *Gossypium hirsutum* and *Gossypium barbadense*, DNA from out-group, *Gossypium incanum* (E4) (Accession Number: PI-530984; Saudi Arabia), outside the A/D genome was selected. The major cultivar/lines of *Gossypium hirsutum* and *Gossypium barbadense* are TM-1 (2005; Accession Number: 05-PI-607172; Texas, USA), 3-79 (a double-haploid line), and PS-6 (2005; Accession Number: 05-6745).

One photoperiod insensitive plant from the *Gossypium hirsutum* TX-231(2005; Accession Number: 05-PI-163725; Guatemala) had been grown at the USDA Southern Plains Agricultural Research Station (USDA-ARS-SPARC) from 2005 to 2008 in the greenhouse facilities. This plant was transferred to Dr. Alan Pepper's laboratory at the Texas A&M University Institute for Plant Genomics and Biotechnology (IPGB). There

the plant was maintained in the greenhouse and outside conditions through December of 2009. New seeds from TX-231 have recently been grown using the seed transplants to greenhouse conditions for back up tissue. The second photoperiod insensitive line was from *Gossypium barbadense* K-46 (1988; Accession Number: 88-PI-528313; Guadeloupe). The accession ID information on K46 will be added, once information from the USDA is obtained from Richard Percy.

The experiment contained collected leaf tissues from: *Gossypium raimondii* D5 (1989; Accession Number: 03-PI-408785, Peru); *Gossypium herbaceum* A1 (2003; Accession Number: 89-PI-530898); *Gossypium barbadense* lines 3-79, PS-6, and K-46; and *Gossypium hirsutum* line TM-1and TX-231. All seeds for these taxa were taken from the USDA Southern Plains storage facility in College Station, TX. Seeds were imbibed in distilled water in the -20ºC refrigerator overnight. Falcon petri dishes (150x20) mm had two autoclaved growth paper pieces cut to fit them. One circle of paper was placed in the bottom of the petri dish. Seed coats were removed from the cotton seeds. Eight to ten seeds were placed on the paper circle. A second paper circle was placed on top of the seeds. The sheets were dampened with distilled water. This petri dish lids were them placed on top and parafilm was placed around the petri dishes. These dishes were left out on bench tops under growth lamps for 3 to 5 days. Once cotyledons had sprouted, the seedlings were transferred to a soil mixture.

The transferred seedlings were grown in a mixture of metro mix 500, vermiculite, and perlite. The ratio consisted of 2 parts metro mix to 1 part vermiculite and 1 part perlite. Specimens were grown in a long day growth chamber at 78 ºC with 10

hours of light and 14 hours of darkness. They were grown in these chambers for a period of 2 months. The plants were transplanted to larger pots and moved to Borlaug greenhouse chamber 16 for 4 more months of growth. DNA was collected from new leaf tissues during phases in the greenhouse. Tissue samples were directly frozen in liquid nitrogen and stored in a -80ºC freezer.

DNA Extraction

Sample tubes were removed tubes from the -80ºC freezer and placed into a liquid nitrogen container until ready to grind. Tissue samples were transferred by VWR Disposable Spatulas (North American Catalog Number 80081-188) into new 1.5mL tubes. The samples were then ground with blue pestles in liquid nitrogen to form a powdered state. Warmed extraction buffer and RNA Plant Isolation Aid (Ambion™ AM9690) were added to each sample to enhance DNA recovery. Liquid was eluted and moved into new tubes containing100% Isopropyl alcohol (IPA) to condense the DNA into a pellet. Supernatant was removed, and pellets were allowed to dry for 30 minutes. Super TE (50mM Tris-EDTA, pH 7.5) was added to each pellet and gently ground with a blue pestle until it was re-suspended.

After re-suspension, un-dissolved solids were removed from the samples by vortexing and centrifuging for one minute at full speed. Supernatant was then transferred into new tubes, where 3M NaOAc pH 5.2 and IPA were added to precipitate out the DNA. The condensed DNA was deposited upon the bottom of the tube in a pellet formation after centrifugation occurred. Before the DNA was washed, the IPA liquid mixture was removed and replaced with 80% ethanol (EtOH). The ethanol was removed,

and samples were allowed to dry for 30 minutes. Each DNA was re-suspended in 50-100µl of 0.5x TE (Tris-EDTA, pH 8) with RNAse and stored at -20ºC.

Candidate Gene Selection and Primer Design

Nucleic acid sequences from well annotated genes involved in floral development, circadian clock, and related photoreceptor pathways were obtained from the *Arabidopsis* genome at The Arabidopsis Information Resource (TAIR) website (http://www.arabidopsis.org) [24, 45, 47, 54, 62, 65-73, 75, 76, 78, 79, 81, 85-90, 94, 95, 97-99, 101-104, 106, 114, 120-123, 125, 126, 128-133, 135-140, 143-146, 148, 150-153, 157, 159-161, 165, 166, 168, 173, 184, 188, 192, 194, 200, 247, 250, 252, 253, 255, 258, 259, 276, 280-282, 286, 303-306, 309-316, 322-328, 330-334, 336, 349, 350]. The CDS sequences (coding sequences) of the *Arabidopsis* candidate genes used to search correlated cotton sequences using BlastX from the available *Gossypium hirsutum* and *Gossypium raimondii* ESTs and TCs. The cotton UniGene sequences were acquired from NCBI (ftp://ftp.ncbi.nih.gov/repository/UniGene/). The UniGene sequences were then utilized in the Stand-Alone BLAST program (blast-2.2.18-ia32-win32.exe) from NCBI (ftp://ftp.ncbi.nih.gov/blast/executables/LATEST). The *Arabidopsis* candidate genes were then queried against the UniGene sequences. These genes of interest were also queried against the DFCI Gene Index Project for the cotton TC sequences (http://compbio.dfci.harvard.edu/tgi/plant.html).

To verify that we had obtained the closest known cotton ortholog to a particular Arabidopsis gene, we used a reciprocal Blast approach [339]. We queried the cotton sequences against the *Arabidopsis* sequences at the TAIR website (http://www.arabidopsis.org/Blast/index.jsp) to confirm those gene identities. The cotton fragments procured from the TAIR query were aligned with the *Arabidopsis thaliana* cDNA and CDS of each candidate genes in Sequencher v. 4.2 - 4.8 (Gene Codes Corp., Ann Arbor, MI) and in Geneious® Pro v 3.0.5 – 6.1.5 (Biomatters Ltd.).

The candidate gene sequeneces were annotated using the known *Arabidopsis thaliana* gene model on TAIR10. In order to attain greater diversity and occurrence of SNPs, the primers were designed across exon-intron spanning regions (Table 29). Each primer was built containing 21 bp to 30 bp. The standard format for creating possible primers was to make sure that the sequence had a GC clamp on the end with 35% to 60% GC content and 63ºC to 65ºC melting temperature (salt adjusted) [351].

In some cases Genome Walking, via the GenomeWalker® Universal Kit (Clonetech Laboratories), was used to obtain intronic sequences, which could not be amplified by PCR. After sequencing the walked sections, a precise end primer was made from the walked fragment.

Through the amalgamation of sequence information that was derived, consensus sequences for the exon-intron spanning regions for each gene were determined. From the original amplification sequences, a second set of gene-specific primers were made resulting in 56 amplicons that ranged in sizes from 400 bp to 800 bp across 38 genes (Table 29).

Table 29 454 Candidate Gene Primers

Primer Name	Sequence	bp	Tm	%GC
AGL1_RintW	CGACCACGCGTGCCCTATAGT	21	64.5	62%
AGL1_W1_DFCI	ATCGACAAGTTACCTTCTGCAAGC	24	61.6	46%
AGL2_RintW	CTCAAAACAGTTCCAAGCAATTAGGAC	27	61.2	41%
AGL2_W1_DFCI	GGAACGGGTTGTTGAAGAAGGCT	23	63.4	52%
AGL3_RintW	GGCTGGTAAAGTCATTGGAGGTG	23	61.4	52%
AGL3_W1_DFCI	GGGTTGCTCAAGAAAGCTTATGAG	24	60.1	46%
AGL6_RintW	CGATCGACACGTTTCATCTCATATC	25	59.9	44%
AGL6_W1_NCBI	CAGACAAGTGACCTTCTCTAAGAG	24	58.3	46%
AGL9_SEP3_2F_DFCI	GGTACCAAAAATGCAACTATGGAGC	25	60.9	44%
AGL9_SEP3_2R_DFCI	AAGTTGTTTCTCAAGTGACTCGAGC	25	61.7	44%
AGL16_F2_DFCI	AAATGAGCCTTCGCGGTGTTCG	22	64.2	55%
AGL16_R2_DFCI	TACTTGTATTTCTTCGATTAACATTTGATC	30	57.6	27%
AGL30_RintW	CCACTACTAATTCGCATCATTTGAGC	26	60.8	42%
AGL30_W1_DFCI	AACACAAATGGCCGTCAGGCGA	22	65.5	55%
AGL32_F1_DFCI	TGTCCTCAAAGTCCGAGAGCG	21	62.1	57%
AGL32_R1_DFCI	GATTATCAAGCTGTTGCTGCAAGAG	25	61	44%
AGL65_RintW	CACCTAGAAAGTCATGTATGTTTAGATC	28	58.1	36%
AGL65_RintW2	CAAATCCAGAGATCAAAATGGCAACC	26	61.3	42%
AGL65_W1_DFCI	GACAAGTTACATATTCGAAACGCAG	25	59	40%
AGL65_W3_DFCI	CTCGAGAATGATTTTTACCTACTGCAG	27	60.3	41%
AP1_2F_NCBI	GAAGATCCTTGAACGCTATGAAAGG	25	60	44%
AP1_2R_NCBI	CTCTGGTTTCTCTCCAAAAGCTC	23	59.2	48%
AP3_RintW	CCTTACTATAGGGCACGCGTG	21	60.6	57%
AP3_W2_DFCI	CACTGTTCTTTGTGACGCCAAGG	23	62.6	52%
ATGRP7_F1_DFCI	ATGTCGAGTTCCGGTGCTTCG	21	63	57%
ATGRP7_R1_DFCI	GTTCATTCCTTCGATCGCGTCC	22	61.9	55%
COL3_R	TCTTTCTCTTCTCTCTGTACCTCAG	25	59.1	44%
COL3_W1	CTCCTCTAACCCTCTCGCTCG	21	61.4	62%
COL5_1R_DFCI	CTGTTCTTTCTCTTCTCTCTGTACC	25	58.6	44%
COL5_Fint	GACATGGATCCGTTTATTGATTTTGAG	27	59.4	37%
COP1_1Fint	CCATGTTAGTCAGAGAAAAGATGCCT	26	61.1	42%
COP1_1R_NCBI	CGGATGCTTGGAGATGATTCTAGC	24	61.4	50%
CRY1_1F_NCBI	GACGTCGCTTTTGCAGGTAACAC	23	62.9	52%
CRY1_1R_NCBI	CTGCGTTGAATGATCGAACCGC	22	63	55%
CRY1_3F_NCBI	AATGCCTTAACATGCCTTTTGACCCT	26	63.6	42%
CRY1_3R_NCBI	AAGCTTTATCTGCATTGCTCCACC	24	61.9	46%
CRY1_4F_NCBI	AGGGTGGAGCAATGCAGATAAAGC	24	63.5	50%
CRY1_4R_NCBI	ACGGCCATCCAGAAGAGTACC	21	61.6	57%
CRY2_1R_NCBI	AGCTCCGGCAGCCATTGCCT	20	₮7.2	65%
CRY2_2F	TATCCGTTCACTCATGAGAGATCG	24	59.8	46%
CRY3_FW2_NCBI	TACGACCCGTGTTCAAATTACGGA	24	62.3	46%
CRY3_Rint_W	TTGCGTAACCCGGAGTGTCTC		62.4	57%
DET1_F1a_DFCI	TAATATGGAGACAACTGAAATTGTTGCA	28	60.1	32%

Table 29 Continued.

DET1_R1_DFCI	CAAATAATTGGTAAAGTTCATCTGCTGC	28	60.2	36%
ELF3_1R_NCBI	CTTAATCAGCCTATGTAACTCAAACAC	27	58.6	37%
ELF3_FintW	TGCGACAAAAGCATGTGGCTTCAG	24	64.8	50%
FD_FintW	AGTATGGAACGACATCACCCTCG	23	62	52%
FD_R1a_DFCI	GATGAGCAACTTCAAGTTCTAGCTC	25	59.9	44%
FKF1_ADO3_1F_DFCI	GACTTCTTTCGTTGTTTCCGATGC	24	61.1	46%
FKF1_ADO3_1R_DFCI	TTCCATCATCATCACGTATAGGTGC	25	60.8	44%
FKF1_Fint	ATACATAGGCGTCAACTGAATTGGAG	26	61.2	42%
FKF1_Rint	CAATTCAGTTGACGCCTATGTATCC	25	60.1	44%
FT_3F_NCBI	GTTACTGATATTCCAGCCACAACTG	25	59.9	44%
FT_3R_NCBI	AAAGTCCCTAGTGTTGAAGTTTTGG	25	59.6	40%
GI_1F_NCBI	CTGCCCAACAGGGATGGAGAC	21	62.7	62%
GI_1R_NCBI	CGGAGATGCTGATACGACATTGCA	24	63.3	50%
GI_2F_NCBI	CGGCAAAAGCAGCAACTGCAG	21	63.7	57%
GI_3F_NCBI	GTGCCACTGATGGAATGCTCG	21	61.9	57%
GI_3R_NCBI	CTGGCTGAACTGCTCTAGCTG	21	60.7	57%
GI_4R	ACGAGCATTCCATCAGTGGCAC	22	63.4	55%
HY6_F1a_DFCI	GGTTCGTGGCTATGAAATTGCATAC	25	61	44%
HY6_F2_DFCI	GGATCCTCAAGCGTTCATATGCCA	24	63.1	50%
HY6_R2_DFCI	GTTCTTCTCGTCTCTAGTCCAGC	23	60.2	52%
HY6_Rint	TGGCATAACCTTGCTCTTCGAACC	24	63.4	50%
LHY_F1_NCBI	GAGCATAATAGGTTCCTAGAGGCT	24	59.2	46%
LHY_R1_NCBI	GAGCATGACTCCTGATCTGCAC	22	61.1	55%
MiRNA172c_FW1_DFCI	ACCACCGTCCATCAACAGATGTG	23	63	52%
MiRNA172c_R	GCCCGGGCTGGTATATGAATATG	23	61	52%
PFT1_2F_PGDB	AAGGGCAGCCGGTCTTTATCAC	22	62.9	55%
PFT1_2R_PGDB	GAGATATAAGACGAACTATTTGCATGG	27	58.3	37%
PFT1_3F_PGDB	TGCAAATAGTTCGTCTTATATCTCAGG	27	59.1	37%
PFT1_3R_PGDB	AAAGCATCAACGTCTGTGAAGGCA	24	63.5	46%
PHYA_1F_Contig1	GTTCCATTTCCCCTCAGGTATGC	23	61.3	52%
PHYA_1F_NCBI	GTTCCATTCCCACTCAGGTATGC	23	61.2	52%
PHYA_1R_NCBI	CTTGTTGCAGCTCATGGCTTGC	22	63.4	55%
PHYA_2Fint	CATATTCCAACATAGTTACGAGTGCT	26	59.3	39%
PHYA_2Rint	GTCCATGACAATCTTCTGAGCTG	23	59.4	48%
PHYA_3Rint	CATAGTCCTTCCATGGCAAACTCC	24	61.5	50%
PHYA_F2_NCBI	ACTTCATGTCCAGCGATTAACAGAG	25	61.1	44%
PHYA_Fint	CATGGAAGGACTATGAAATGGATGC	25	60	44%
PHYA_R2_NCBI	GATACGTTTTCCATTGCTCGTCAC	24	60.7	46%
PHYB_F2_IBR	GTTGTTTGTCATCATACTTCTGCACG	26	61.3	42%
PHYB_F3_IBR	AACAGGACTCTCAGTTGAGGAAGC	24	62.3	50%
PHYB_F4_IBR	GGTGCAAAGCATCATCCAGAGGA	23	63.1	52%
PHYB_R3_IBR	TGAGAGTCCTGTTGATTCTGCAGC	24	62.9	50%
PHYB_R_IBR	CCTTACTAGAGCAAGCGTTCACCA	24	62.6	50%
PHYC_1F	TTGGTACCGAGCTCGGATCCA	21	63	57%

Table 29 Continued.

PHYC_1R_DFCI	GCAGTCTCAATCAGGCGGACC	21	63.3	62%
PHYC_2F	AGTATGGGATCAATTGCATCTCTTGTG	27	61.6	41%
PHYC_FW2	GGTCCGCCTGATTGAGACTGC	21	63.3	62%
PHYC_R_NCBI	TCTCGACTACAACATGCATTAACAACC	27	61.9	41%
PHYC_Rint	CAAGAACTGAAGCACCTGGATAGC	24	61.7	50%
PHYE_2F_NCBI	GTATGATTGCTGTTGAAGAACCCAG	25	60.4	44%
PHYE_R_IBR_hinge	AACCGGATTTGCATGGCAGTCAC	23	64.3	52%
PI_F1b_DFCI	AAGAAACTATGGGATGCTAAACATGAG	27	59.7	37%
PI_R1_DFCI	TTTCTTGATTCTATCTATTTCATTGCTGAG	30	58.8	30%
PRR5_F1	CCACCAGGCAGATTATATCTGCTC	24	60.9	50%
PRR5_Rint	TTGAACCACAATCCGTAACCATGTC	25	61.8	44%
PRR7_F3	GACTGTAGTTAGGGATGAGCGGA	23	61.2	52%
PRR7_R3	CAATGTTATTGCTACCGACATTTGAGC	27	61.6	41%
SPA4_F2_DFCI	ACACGAGAAACGCGTATGGTCC	22	63.1	55%
SPA4_R2_DFCI	CGCAGCAAACATTGGCCTTTGTC	23	64	52%
TOC1_F1_DFCI	TCTGCTAGACAGGTCATTGATGC	23	60.4	48%
TOC1_R1_DFCI	CAAGATAGTCAGCAGCACCAAGC	23	62	52%
ZTL_1F_NCBI	GGTTACCGGTTATCGCGCCGA	21	65.1	62%
ZTL_Rint	AGAAAATTTATGGGTGGCAAGGTGCA	26	63.9	42%

Primer Validation and Amplification

Primers were verified on the diploid D5 *Gossypium raimondii* genome using GoTaq® Green Master Mix (Promega™) for PCR on the Tetrad™ PTC-225 Thermocycler (MJ Research). The amplicons were then checked using gel electrophoresis. The amplification products were then noted to be single or multi-banded. Amplified single banded products indicated a single gene representing only a single locus with no polymorphisms, while multi-banded products amplified either represented amplification of multiple loci or a wide double-band represented

heterozygosity at the locus. These products were amplified in D5 *Gossypium raimondii* because it was the smallest primitive diploid ancestor to the tetraploid species and was being completely sequenced. Amplicons that appeared as single bands were re-amplified using *Phusion*® High-Fidelity *PCR* Master Mix (New England Biolabs), instead of GoTaq®. To check whether the amplification had worked, the single banded Phusion products were re-run on gels. They were then purified using either Qiaquick® PCR Purification (Qiagen™) or BayGene purification columns. Next, the ¼ ABI Big Dye Terminator Cycle reactions were prepared for sequencing the amplicon products. These reactions were then cleaned to remove the terminator dye by Performa® DTR Gel Filtration Cartridges (EdgeBio™). Afterward, these fragments were sequenced using the ABI model 3730. Finally, in Sequencher and/or Geneious®, DNA fragments were aligned giving genetic data from exon to exon.

For the multi-banded reactions, these amplicons were rerun using the Phusion® High-Fidelity PCR Master Mix (New England Biolabs). Thereafter, the PCR products were taken through blunt end Topo cloning PCR using Zero Blunt® Topo® Cloning Kit (Invitrogen™). The reactions were plated out on LB Agar plates plus kanamycin. Colonies were grown overnight at 37ºC. The next day, the colonies were picked using pipette tips and PCR amplified with GoTaq®. The colony PCR products were then purified using an ExcelaPure 96 well plate (EdgeBio™). After gel evaluation, they were then re-arrayed for ¼ ABI Big Dye Terminator Cycle reactions. To remove the terminator dye from the plate, a Performa DTR Ultra-96-well plate was used.

Once cotton sequences for candidate genes were obtained, they were aligned by employing Sequencher® and Geneious® software. Chromatograms were scanned for possible sequencing errors from dye blobs and poor reads. Consensus sequences were then made.

Multiplexed Amplicon Sequencing for the GS-FLX Roche 454

Amplification

The 56 amplicons from eight taxa were amplified using Phusion reactions in four and half 96 well plates (American Standard™). These reactions were then purified using 96 well ExcelaPure plates (EdgeBio™). Next, the samples needed to be quantified for double-stranded (DS) DNA. A new product called AccuBlue (Biotium™) was shown to quantify DS DNA. AccuBlue High Sensitivity Kit (Biotium™) was used to quantify DS DNA by taking the samples and measuring them on Victor™ *X3* Multilabel Plate Reader (PerkinElmer™) with 485nm of excitation and 535 nm emission levels. Afterward, they were measured against a set of control standards. This was a light sensitive reaction, so reactions were made in a black 96 well flat bottom plate (Corning™ CLS3916). This was to reduce light refraction from other wells. After quantification, the amplicon PCR products were diluted with 2X TE (Tris-EDTA, pH 8). In new plates, the amplicon products were transferred, so normalization at 5.5 x 10^8 molecules/µl was maintained. All 56 amplicons were then multiplexed together by taxa, using 10µl per reaction, into one tube. Once reactions were multiplexed, library preparation for each taxon on the GS-FLX® Titanium Roche 454 (Roche Life Sciences™) was ready to begin.

Library Adapter Preparations

The amplicon sequencing via Roche 454™ approach utilizes the Y-Primer technology. This Y-Primer technology allows partial double-stranded adapters attached by ligation to the amplicon template with an end-repair reaction of a 3' adenine overhang [352]. To accomplish high throughput sequencing with amplicons, the sequencing primers were modified with Primer 'A' and 'B' keys and MIDs (multiplex identifiers). These primers were ligated to the original amplification primers, in order that the dual barcoding strategy would work with the GS-FLX Titanium emPCR Kits (Lib-L)[353-355]. On the Roche 454™ website, fourteen MID (multiplex identifier) adapters were listed. From the available adapters, this experiment selected eight MIDs to create a barcoding system to help identify the eight taxa [356, 357]. The system of dual barcoding gave the ability to identify the taxon by MID and the gene by original PCR primers.

To ensure that the Y-Primer would ligate to the original PCR amplicon, the Y-Primer needed to be modified. This Y-Primer modification on the 'A' strand oligonucleotides was created with a phosphate on the 5' end to allow for an overhang that attached to the amplicon fragment. Next, the Y-Primer A and B strands needed to be annealed before ligation to the PCR amplicon could occur. By diluting with STE (Sodium-Tris-EDTA), the nseqA-F primer and nseqB-R0 primer allowed molecules to anneal stably to each other [358]. Next, an annealing reaction between the forward and reverse primers were performed at 95º C and allowed to decrease to room temperature.

Dilutions of the nseqAB annealed adapter reactions were created at a 5pmol/µl concentration.

Amplicon Pooling, Phosphorylation, Adenylation, and Ligation

After the completion of the amplicon library adapters, the next step was to prepare the pooled samples by taxa for ligation with the 'A' and 'B' tagged primers (Figure 17). The pooled samples were quantified by AccuBlue (Biotium™) to $5x10^8$ molecules/µl. A phosphorylation reaction added a phosphate to the 5' end of the pooled amplicons and removed any unwanted phosphoryl groups on the 3' end. The phosphorylation step was imperative because the PCR amplicons were blunt ended via the Phusion reaction.

In the phosphorylation reaction, pooled PCR amplicon reactions were heated to 70º C, so that the end of the fragment would begin to unwind. By immediately plunging these amplicons into an ice bath, the amplicon ends were locked into an open position. The phosphorylation used the T4 Polynucleotide Kinase (PNK), the NEBNext® End Repair Buffer (New England Biolabs), and 25mM $MgCl_2$ to account for the TE (Tris-EDTA, pH 8) added in the normalization step. These samples were the purified with the MinElute kit (Qiagen™).

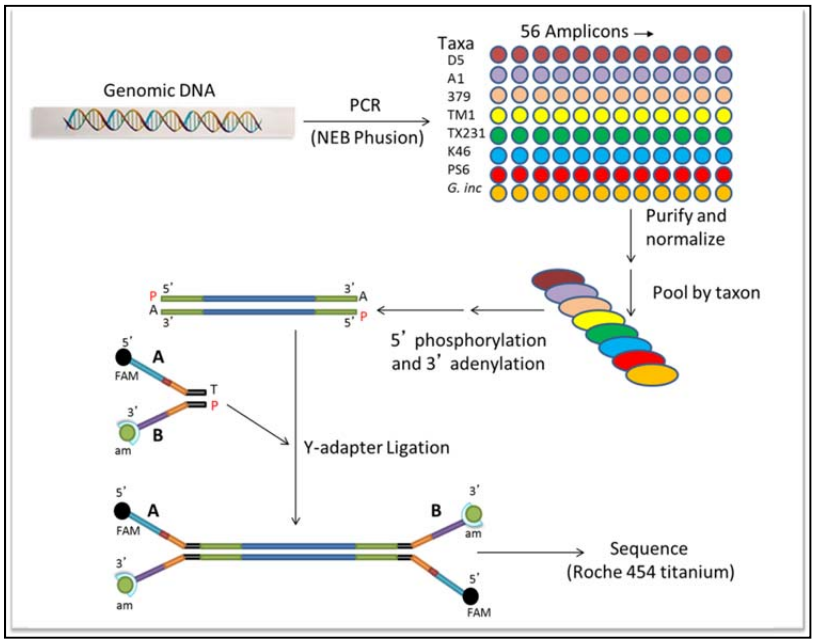


Figure 17 Amplicon Library Pooling and Preparation for Roche 454

Consequently, a 3' adenylation step was completed on the pooled samples with the NEBNext® dA Tailing Module Klenow Fragment (New England Biolabs). This was to make sure that the 3' end of the blunted pooled PCR fragments would contain a free adenosine that the Y-Primer could ligate on to. Thereon, the samples were purified with the MinElute kit (Qiagen™). With AccuBlue (Biotium™), the pooled samples were quantified, so the proper ligation adapter concentration could be calculated. This ligation adapter concentration was .5 pmol/µl.

Using the NEBNext® Quick Ligation kit (New England Biolabs), the diluted A and B annealed adapters were ligated on to the ends of the pooled PCR amplicons. Once the different annealed adapters, coding for different MIDs, were ligated onto the separate taxon pooled samples, all eight taxon ligations reactions were pooled together into a single reaction. The sample was purified using the MinElute kit (Qiagen™), and the sample was re-purified again using the AMPure® XP tube protocol (Agencourt™). Then, the sample was quantified by Fluorescein (FAM: 6-FAM phosphoramidite), a fluorophore that emits light upon excitation, on the Perkin Elmer Victor 3 in a black 384 flat-bottom plate. The concentration of the clean ligated pooled sample was 3.23×10^8 molecules/µl. The sample was diluted to 1×10^6 and 2×10^6 with TE [354]. Finally, the samples were given to the Laboratory for Genome Technology where Lib-A kit (Roche Life Sciences™) was used to perform EmPCR. Pyrosequencing was completed on the GS-FLX® Titanium Roche 454 (Roche Life Sciences™).

Analysis of Roche 454 data through CLC Genomics Workbench and Geneious®

Once pyrosequencing was completed, the ".sff" files were imported into CLC® Genomics Workbench (CLCBio™) where they were analyzed. The sequences were parsed into the eight taxa by the multiplexing option to process tagged sequences. To parse the sequences by barcode (Table 8), the barcodes and 16 base-linker (GTGACTCCGACTAGGT) were entered into the splitting parameters and sequence length was left up to 1000 bp. The minimum number of sequences per group was lowered, so that nine groups could be created. This gave one extra group that did not

have the MID barcodes identified. All sequences that had a non-functional barcode were discarded (4.4%)

In CLC® Workbench, the D5 *Gossypium raimondii* reference sequences from the ABI files in Sequencher® 4.8 and Geneious® were imported to make a consensus alignment for each gene that correlated with the new 454 D5 sequences. The reads for D5 *Gossypium raimondii* were then mapped to the reference sequence on CLC® Workbench. A new consensus read was created with the ABI and 454 data for the candidate genes in the D5 *Gossypium raimondii* taxa.

These new consensus fragments were used as the D base reference for mapping the reads from other tetraploid taxa for each gene. The A1 base reference sequence was made from making the correlating the new D5 base reference sequence with the A1 454 sequences. The correlation was set to 85%, so that it could pick up the differences between the 'A' and 'D' genomes.

Using the 'A' and 'D' reference sequences, the each taxon of 454 sequences were mapped to both reference sequences at the same time on CLC® Workbench. A consensus fragment representing the A and D strands for each amplicon was created for each taxon. The consensus sequences were kept at high levels of confidence with at least 80% of reads required to confirm them.

Each consensus sequence that was constructed for both the A and D strands of the eight taxa was exported and imported into Geneious® version 5.1 [359]. The sequences were assembled into contigs. Contigs represent the alignment of several DNA fragments that overlapped to create a single DNA consensus sequence. They were then

evaluated, tagged and scored for differences. Geneious® gave the ability to move, annotate, and align sequences within a contig more easily than CLC® Workbench version 3.3.5[360]. Sequencing errors within the consensus amplicons were easily spotted and verified against the hundreds of basal 454 sequences in CLC® used to form the original consensus sequence. This was imperative due to CLC® Workbench version 3.3.5 map to reference alignment problems. The final cleaned contigs of the eight taxa were then analyzed for SNPs, SIDs, and INDELs between the taxa, the A/D strands, and the wild versus the cultivated lines.

SNP, SID, and INDEL Detection, Analysis and Verification

Geneious® software version 5.1 to version 6.1.5 (Biomatters Ltd.) was used to discover single nucleotide polymorphisms, single nucleotide variations, nucleotide insertions (>2 bp), and nucleotide deletions (>2 bp) in the amplicon contig assemblies for each gene in every taxa. In areas that were well covered, these changes were noted showing the A/D changes, the wild and cultivated changes, and the *Gossypium hirsutum* versus the *Gossypium barbadense* lines.

This information was analyzed by calculating the number of variants in both the exon and intron spanning regions. These variants were categorized into groups by SNPs, SIDs, and INDELs. The SNPs, SIDs, and INDELs were averaged into ratios by groups of introns, exons, intron-exon spanning regions in Microsoft© Office 10 Excel. From this data, procedure frequencies and generalized linear mixed models were used to correlate SNPs within the different pathways and by type using the Base SAS© 9.3 Procedures Guide (Table 16 and Figure 15) [361]. The procedure frequency and generalized linear

mixed model for this paper was generated using SAS© software, Version 9.3 of the SAS System for Windows 64 bit (see appendix for code in SAS). Copyright © 2011 SAS© Institute Inc. SAS© and all other SAS© Institute Inc. product or service names are registered trademarks or trademarks of SAS© Institute Inc., Cary, NC, USA.

Synonymous and Non- synonymous Sites in Coding Regions

From the amplicon contig assemblies for each candidate gene in Geneious® version 5.1 to version 6.1.5 (Biomatters Ltd.)[228], exonic sequence alignments from each gene were extracted to analyze non-synonymous variants to synonymous variants in the exonic regions. Also exonic regions of Arabidopsis were assembled with the cotton exonic sequence alignments in Geneious® version 5.1 to version 6.1.5 (Biomatters Ltd.)[228]. These files were exported as .meg files and imported into Mega 5.2.1 [341, 342]. Mega 5.2.1 was used to calculate nucleotide changes, amino acid changes, non-synonymous variants, and to synonymous variants. The number of synonymous substitutions per synonymous site was calculated from averaging over all sequence pairs with their standard error estimates. This data was obtained by a bootstrap procedure of 500 replicates using the Nei-Gojobori model [243]. The number of non-synonymous substitutions per non-synonymous site was calculated from averaging over all sequence pairs with their standard error estimates. Again, this data was obtained by a bootstrap procedure of 500 replicates using the Nei-Gojobori model [243]. All ambiguous positions were removed for each sequence pair in both analyses.

CHAPTER IV

COMPARISON OF CANDIDATE GENE-BASED AND GENOTYPING-BY-SEQUENCING (GBS) APPROACHES TO TRAIT MAPPING IN *GOSSYPIUM BARBADENSE* L.

A Comparison of Candidate Gene-based and Genotyping-By-Sequencing (GBS) Approaches to Trait Mapping in *Gossypium barbadense* L.

Overview

Overview Rationale and Objectives

The use of marker assisted selection (MAS) has been a recent part of modern breeding techniques. With the advent of next-generation sequencing, we are now able to effectively help breeders reduce the amount of time it takes to bring in new traits of interest into a cultivated line. This study facilitated the use of genotyping-by-sequencing (GBS) to associate SNP alleles that with our trait of interest, photoperiod independence. This was done by comparing SNPs that present in various orthologs, homeologs and paralogs within a fully introgressed wild cotton Germplasm line.

Overview Methods

We employed a genotyping-by-sequencing (GBS) approach to narrow target SNPs associated with a major locus that contributes to photoperiodic flowering in *Gossypium barbadense*. We used a simplified restriction digestion protocol to achieve reduced representation for sequencing. We utilized Targeted GBS (TGBS) to show linkage with photoperiod independence in a segregating BC4F2 wild cotton conversion

line (PS6 x wild). This was an ideal way to correlate phenotype with genotype in a breeding population.

Overview Results and Conclusions

We identified SNPs between cultivated AD2 *G. barbadense* and wild AD2 *G. barbadense* relatives. We correlated this data to the *Gossypium raimondii* reference genome with the known floral regulatory genes from *Arabidopsis thaliana*. We showed segregation of loci within the segregating population. This schema showed potential for identifying potential time and money savings in the development of new lines not only reduces time in breeding strategies and to bring in beneficial traits into elite cotton cultivars more rapidly.

Overview Keywords

Genotype by Sequencing, Cotton, *Gossypium*, Reduced Representation, Marker Assisted Selection, Loci, Linkage Disequilibrium, Photoperiodism, Photoperiod, Flowering, Wild Germplasm Introgression, GBS, Targeted GBS

Background

Economically, cotton is the fourth most important crop in the world [284]. Therefore, it has high economic value in the world's trade markets. Many countries' gross domestic products (GDP) rely on the production of goods sold at fair trade values [362]. In agricultural business, the goal is to produce the best product at minimal cost [363]. Therefore, producing the best quality cotton at the lowest cost helps to maintain the crop's economic importance in the world. Commercially, there are two main types of marketable cotton: Diploids (A genome) and Allotetraploids (AD Genome). The diploids

(*Gossypium herbaceum* and *Gossypium arboreum*) are primarily grown for trait introgression into the allotetraploids, but accrue value for non-textiles (bed stuffing) [293]. Diploid cotton is small leafed, lanky, and shrub-like in appearance and is highly distinguishable from allotetraploid cottons in fiber and form [31, 33, 287, 293]. Diploid cultivated cotton species have short, weak fibers[293]. During spinning, these weak fibers are blended with other cotton fibers to achieve specific Micronaire values[293]. Cotton's Micronaire values indicate the fiber's fineness (linear density) and maturity (cell wall degree of development) [364]. In short, this value accounts for air permeability in the fiber for spinning. Micronaire is the one of the most important fiber characteristics for commercial value [364]. Therefore, it is important to achieve world-wide industry standards for Micronaire values, so criteria for commodity trade values can be established.

The other types of marketable cotton are the allotetraploids. There are two main types of allotetraploid cotton (*Gossypium barbadense* and *Gossypium hirsutum*). *Gossypium barbadense* and *Gossypium hirsutum* are marketed for fiber in textiles, fiber in medical supplies, seed-oil content in vegetable oils, cottonseed meal for animal feed and fertilizer [294]. *Gossypium hirsutum*, known as upland cotton, is a three to five foot shrub with cordate leaves of three to five lobes that produces strong fiber around ¾ of an inch to one inch [294]. While *Gossypium barbadense*, known as extra-long staple (ELS) cotton, is a shrub-like tree with three to five deep lobed cordate leaves and produces fiber from 1 and 3/8 of an inch or longer[33, 295, 348]. This ELS cotton is softer and longer, than *Gossypium hirsutum*. Therein, the commercial price for *Gossypium*

barbadense is much higher [295, 346]. Both *Gossypium barbadense* and *Gossypium hirsutum* are the major contributors to the global economic commodity production, imports, and exports of cotton world-wide. Discovering ways to improve these allotetraploid commodities benefits the national GDP of cotton producing countries. Hence, it is vital to have an understanding about the genetic diversity within cotton.

Modern allotetraploid cultivated cotton has limited genetic diversity because of an evolutionary bottleneck caused by Mesoamerican proto-agricultural tribes selecting for early flowering during the domestication process [5, 287]. The oldest discovered cotton textiles in the New World were of *Gossypium barbadense* L. in Northern Peru with the Huaca Prieta civilization circa 2400 B.C. [32]. The other modern cultivated allotetraploid cotton, *Gossypium hirsutum* L. was first domesticated near the Yucatán peninsula [27]. The domestication process of cotton bottlenecked because modern allotetraploid cotton cultivars originated only from these two domestication sites. Eventually a plateau for breeding new cotton cultivars will be reached. Therefore, developing practical traits from wild relatives is needed. These untapped wild genetic resources have valuable assets which should be incorporated into traditional breeding programs [6].

Cotton breeders are asked to increase their products value by bringing in more desirable traits from wild Germplasm. Often this achieved by searching for desirable traits from a sub-standard stock and integrating those traits into elite lines through breeding schemas [9, 10, 271]. Desirable traits which breeders might acquire from

undomesticated cotton are: 1) higher disease tolerance, 2) pest tolerance, 3) heat tolerance, 4) salt tolerance, 5) fiber length, 6) fiber strength, and 7) stalk vigor.

However, bringing in desirable traits is not a simple process because a key trait, 'photoperiod sensitivity', was lost during the allotetraploid cotton domestication [11, 245-247, 262]. Most commercial cotton producing areas in the world do not undergo light conditions allowing wild cotton species to flower during a growing season's span following today's current cultivation practices. Using specialized light conditions in greenhouses/growth chambers, undomesticated and cultivated species can be bred together, but this cross results in offspring exhibiting photoperiod sensitivity. Therefore, these offspring are rendered useless for commercial production.

Our study focused on *Gossypium barbadense. Gossypium barbadense's* genetic diversity further narrowed through breeding strategies for habitat and day neutrality. The colonial shipping industry had increased the trade commodity that was easily grown in similar habitat with day neutrality for increased yield [31, 33]. The long strand staple cotton, 'Sea Island Cotton' [*Gossypium barbadense* SI], was the primary cotton developed throughout the Caribbean basin into late 18th century [348]. In 19th century, Egyptians developed elite cotton lines from the introduced 'Sea Island Cotton'[348, 365]. Germplasm from some Egyptian cultivars were utilized in the development of the original 'Pima' cultivars [295, 366-369]. The elite cotton cultivars used in this study are Pima S-5 (PS-5) and Pima S-6 (PS-6) [369-371].

Our overarching theory was that a single gene controls and influences photoperiodism in *Gossypium barbadense's* flowering pathway. It was supported by the

significant 3:1 non-flowering to flowering phenotypic ratio of the *Gossypium barbadense PS-6 x* wild *Gossypium barbadense* field trial (material provided by Richard Percy, USDA-ARS) This BC4F2 segregating field demonstrated a that a recessive mutation led to photoperiod independent flowering in *Gossypium barbadense*. The BC4F2 segregating population was composed of *Gossypium barbadense PS-6* crossed with eight wild *Gossypium barbadense, where the wild Gossypium barbadense* was the recurrent parent. This effort was to retain a majority of the wild Germplasm alleles, while still segregating for photoperiod independence.

This project's specific aim was to narrow down heritable markers for photoperiod regulation in cotton. GBS was used as a tool to find closely linked SNPs for photoperiod independence in a fully introgressed wild population. The SNP differences were correlated back to the genetic donor parent. Our hypothesis is that the inherited *Gossypium barbadense* PS-5 marker loci retained in both fully introgressed lines (*Gossypium barbadense* PS-5 x *Gossypium barbadense* PI-435242 and *Gossypium barbadense PS-5 x Gossypium barbadense* K-56) were closely linked with the photoperiod independence gene. These loci containing SNP differentiation can then be used as markers in MAS breeding schemes.

This cotton Genotype-By-Sequencing (GBS) study promises to benefit worldwide cotton production by identifying the loci near the photoperiod independence trait, an integral part of cotton domestication. Since cotton has a highly complex genome containing 26 chromosome pairs, a marker assisted selection (MAS) breeding program needs a large efficient number of SNPs for genomic mapping, quantitative trait loci

(QTLs), and accruing unique traits from primitive Germplasm [7-10, 271, 372, 373]. Therefore, discovering linked markers to a segregating loci/trait would be highly valuable in a MAS breeding schema. These loci will ease the ability to integrate genetically diverse desirable traits without attaining photoperiod dependency from wild stocks. Unfortunately, desirable characteristics may be linked with photoperiodism, so linkage between photoperiodism and the desirable trait must be disrupted. For example, there is compelling evidence shown for genetic variability in exotic *Gossypium hirsutum* L. for heat tolerance [297]. If heat tolerance was closely linked to photoperiodism, then this would be an ideal trait to break apart from photoperiod dependency genes.

Genotype-By-Sequencing

Genotype-By-Sequencing (GBS), a novel approach to sequencing principal fragments of differing organisms within a multiplexed reaction, has been extremely valuable to research studies because of the ease in segmenting DNA with targeted restriction enzymes. In species with larger genomes, GBS has employed specific restriction enzymes to select the same corresponding fragments from several taxa, so they can be compared and overlapped. This method has characterized many traits in various populations and diverse taxa. GBS has been practiced as a complementary method for: (1) association mapping studies, (2) candidate gene studies, and (3) mapping populations. Using GBS, fully sequenced and annotated organisms are not required for polymorphic change comparison. This cost effective method has worked well in breeding populations by identifying the key nucleotide differences correlated to phenotypic data and mapping those polymorphisms back to a specific parent.

Complexity reduction has been adopted in GBS in several ways: (1) two enzyme reduction and, (2) single enzyme size selection reduction. The first method has been used to demonstrate reduction by a combination with two restriction enzymes. In the first case, a rare base cutter and a frequent base cutter would be applied to the sample organism's DNA. When used in conjunction, sample reduction takes place by the selection of the rare base cutter to elicit out the key DNA fragments. Next, a more frequent base cutter would be used to decrease the key fragments' sizes. High levels of stringency and precision to reduce complexity would be required to obtain sufficient quantities, without sacrificing too much of the limiting factor, the sample DNA. When used, this method will not select against highly repetitive regions. The second method has been used to establish complexity reduction through the use of an efficient single restriction enzyme and a gel extraction. This reduction method selected DNA fragments which excluded highly repetitive regions in the organism's DNA, such as chloroplast and mitochondria. When selecting for restriction enzymes in both methods, two elements have been considered: (1) GC rich enzymes allow for more gene rich regions, and (2) higher AT content that increases upstream and downstream elements

The GBS method has been applied to a variety of organisms such as maize, wheat, spotted gar, sorghum, and many others [374-379]. Elshire et al. provides a model case study for single enzyme reduction GBS to sequence out differences in maize breeding populations. The maize nucleotide differences were compared with the fully sequenced maize genome. Thus, this data helped reorder the maize genome and select out many new SNP markers in several populations [375].

Until this study, GBS had not been applied to cotton research. Our approach utilized the single enzyme size selection reduction method. The key differences between our experiment and Elshire et al. was: 1) the restriction enzyme used [ApeKI (Elshire), HinP1I/BsrGI (Logan-Young)], and 2) reduced representation through chloroplast elimination. We did this experiment as a complementary experiment to our candidate gene method in order to narrow down the amount of candidate genes. (Logan-Young, Chapter 3) We used two enzymes in our research to fragment cotton DNA for sequencing on the Illumina GAII®:*HinP1I* and *BsrGI*.

The first enzyme we used was *HinP1I*. This enzyme cut DNA at highly GC rich areas, usually exonic gene regions. The DNA region avoiding chloroplast DNA contamination utilizing *HinP1I* was approximately 20bp. This narrow window minimized the region for gel extraction causing difficulty to cut without chloroplast contamination. Since *HinP1I* selected highly GC rich areas, less polymorphisms were discovered in exonic gene regions. Although the discovery of any SNPs was positive, it would have been more beneficial to receive more SNP data on the intronic regions (upstream and downstream elements). Lastly, the *HinP1I* fragment was methylated sensitive, greatly reducing the amount of reads received.

Our second GBS enzyme was *BsrGI*. This enzyme was chosen because it had a good efficiency rate in NEBuffer® 1, 2, 3, and 4. It was also fairly cost efficient and cut the *Gossypium hirsutum* chloroplast less frequently in silico. Since the narrow bp window caused difficulties *HinP1I*, the next enzyme *BsrGI* allowed for a 100bp region of DNA to be extracted without contamination of chloroplast genes. This enzyme was

methylated insensitive to increase the amount of sequences with possible polymorphisms.

Several factors were considered while utilizing GBS. In order to rule out single nucleotide polymorphisms (SNPs) present since the ancestral divergence event 12 MYA, differentiation between the two diploid lineages (A/D paralogs) was imperative to identify. When a viewed fragment shows a probable SNP, one must make a determination of whether the SNP occurrence was (1) a sequencing error, (2) a true variance, (3) a homeolog, (4) a heterozygote, or (5) a paralog. Depending on the organism, methylated sensitive restriction enzymes have been shown to reduce the DNA complexity within an organism [380-382]. In polyploids, one paralog may have been methylated; so therefore an important consideration when selecting a restriction enzyme is whether it is methyl sensitive [382, 383]. If sequencing both paralogs was essential, then using a methylated insensitive restriction enzyme would be more beneficial. In this experiment, both diploid and allotetraploid cotton were used; hence complexity reduction at the individual level was requisite within the multiplexed reaction.

Our experiment looked for SNPs correlated to the fully integrated phenotypic trait of early flowering into an undomesticated cotton species from a cultivated cotton species. Through the use of flowering independent crossed lines by Dr. Richard Percy at the United States Department of Agriculture Southern Plains Agricultural Research Station (USDA-ARS-SPARC), we keyed out markers to be utilized for trait integration from uncultivated species without the detrimental effect of photoperiod dependency. In future wild introgression breeding schemas, these markers can be used as identifiers for

plants carrying photoperiod flowering dependence in a segregating F1 population. Plants with other beneficial traits will be selected, while rejecting those with photoperiod dependency. Through the use of replicates of different non-domesticated cotton species crossed with the same cultivated species and the selection of the phenotypic trait of early flowering, the overlapping regions should contain the elusive region for photoperiod independence.

Results

Illumina GAII Single End 76bp GBS Sequencing

Multiplex identifiers (MID) barcodes were used for identification of each taxon. This allowed all taxa to be pooled into one sample. Reduced representation of *BsrGI* and *HinP1I* sites allowed for the same fragment sizes to be captured. The *HinP1I* GBS run consisted of 24,591,611 sequences with a length of 76 bp. In this multiplexed sample, 99.8% sequence reads (24,561,522 sequences) were separated by their MID barcodes into ten taxa. Of these sequences, 18,810,285 represented eight cotton taxa (Table 30). *HinP1I* gave significantly more sequences, while *BsrGI* gave fewer sequences (24,591,611: 6,320,415). The *BsrGI* GBS run had 6,320,415 sequences at a length of 76 bp. This multiplexed sample had a MID barcode separation rate of 99.5% sequence reads (6,287,347 sequences) into ten cotton taxa (Table 31).

Table 30 MID Divergence Statistics for *HinP1I*

***HinP1I* Illumina Results**	
MID Barcodes	**# Sequences**
A1-1_PS-5_ATCACG	8539793
A1-2_PI-435242_CGATGT	1591530
A1-3_K-56_TTAGGC	3796931
A1-4_TM-1_TGACCA	32005
A1-5_3-79_ACAGTG	498485
A1-6_TX-231_GCCAAT	623384
A1-7_4127_PR2_CAGATC	1807966
A1-8_4024_PR1_ACTTGA	1920191
A1-9_CAA_GATCAG	4089583
A1-10_CAB_TAGCTT	1661654
Unknown	30089
Total # Sequences	***24591611***

Table 31 MID Divergence Statistics for *BsrGI*

***BsrGI* Illumina Results**	
MID Barcodes	**# Sequences**
B1-1_PS-5_ATCACG	434088
B1-2_PI-435242_CGATGT	237219
B1-3_K-56_TTAGGC	954486
B1-4_3-79_TGACCA	576652
B1-5_TM-1_ACAGTG	407220
B1-6_TX-231_GCCAAT	1067995
B1-7_4127_PR2_CAGATC	819355
B1-8_4024_PR1_ACTTGA	881371
B1-9_D5_GATCAG	295879
B1-10_A1_TAGCTT	613082
Unknown	33068
Total # Sequences	***6320415***

STACKs Loci Grouping

After sequences were trimmed to 66 bp for quality, they were sorted using the process_radtags command in the STACKs program. Tagged reads were grouped into separate folders for organization and different experimental conditions with parental crosses and offspring. Stacks of individual samples and the number of SNPs found within those samples were created, before catalogs referencing crosses were done. Very few SNPs were discovered within the GC rich regions of individual taxa of the *HinP1I* sequences (Table 32). Lower numbers (>100) of SNPs were expected due to sequencing of GC rich genic regions. Inefficient sequencing depth accounted for low numbers (>10) of SNPs (ie. TM-1). More individual SNPs were found within PS-5 of the *HinP1I* experiment, due to slight pipetting inaccuracies when diluting for sample concentration. *BsrGI* individual stacks contained more variability within an individual sample (Table 33). This was expected due to sequencing upstream and downstream elements from GC rich gene centers.

Catalogs, for different crosses (parental, wild/cultivated, and hirsutum/barbadense) and those crosses with offspring, were created, so SNPs could be analyzed between those crosses. These catalogs were built by using the map_denovo command in STACKs program. Similar sequence reads were assembled together, counted, and collapsed into a single library loci entry in each catalog.

Table 32 Average *HinP1I* SNPs by Taxa

Taxa	Unique Stacks	SNPS (within Sample)	Barcode
379	1548	90	ACAGTG
TM1	120	5	TGACCA
TX231	1819	99	GCCAAT
PS5	30165	1742	ATCACG
PI435242	8001	219	CGATGT
K56	18030	793	TTAGGC
PR1	4073	264	ACTTGA
PR2	7546	291	CAGATC

Table 33 Average *HinP1I* SNPs by Taxa

Taxa	Unique Stacks	SNPs (within Sample)	Barcode
379	12996	1114	ACAGTG
TM1	9880	864	TGACCA
TX231	16356	1482	GCCAAT
PS5	8463	751	ATCACG
PI435242	7109	625	CGATGT
K56	14970	1388	TTAGGC
PR1	15339	1327	ACTTGA
PR2	15367	1431	CAGATC
A1	11771	1135	TAGCCT
D5	6237	365	GATCAG

Loci Associated with Photoperiod Independence in BC4F5

In the methyl sensitive *HinP1I* experiment, interspecific crosses between the genetic standards *Gossypium barbadense* 3-79 and *Gossypium hirsutum* TM-1 were evaluated. Fewer loci involving TM-1 were discovered because of the reduced number of sequences. In this analysis, 1,505 loci of similar sequences were found between the genetic standards. In effort to see the difference between interspecific crosses (*Gossypium barbadense* and *Gossypium hirsutum)*, 3-79 was examined against TX-231. The total loci amount increased to 2,688. Another difference evaluated was between cultivated and wild *Gossypium spp.* In *Gossypium hirsutum*, intraspecific cross (cultivated TM-1 and wild TX-231) sequences were surveyed against one another and found to have 1800 similar loci.

In *Gossypium barbadense*, two intraspecific crosses (cultivated PS-5 x wild K-56 and PS-5 x wild PI-435242) with progeny were reviewed for similar sequences. The cross (PS-5 x K-56) showed 39, 561 loci, while the cross (PS-5 x PI-435242) showed 34,108 loci. From these created catalogs, a survey locating one to three SNPs in a single locus, between taxa, was done. The crosses containing the largest number of SNPs between a locus was in the two intraspecific crosses with PS-5. This resulted in 129 (PS-5 x K-56) and 135 (PS-5 x PI-435242) sequences with one to three SNPs, as seen in Table 34.

Table 34 *HinP1I* SNPs between Intraspecific Crosses and Interspecific Crosses

		1 to 3 SNPs	1 to 3 SNPs		
HinP1I	**Total Loci**	**Loci 2 parents**	**Loci 2 parents w/ progeny**	**PS-5 Progeny**	**Undomesticated Progeny**
379/TM1	1505	5	n/a	n/a	n/a
379/TX231	2688	76	n/a	n/a	n/a
PS5/K56	39651	129	42	12	21
PS5/PI435242	34108	135	36	17	10
TM1/TX231	1800	2	n/a	n/a	n/a

Table 35 *BsrGI* SNPs between Intraspecific Crosses and Interspecific Crosses

		1 to 3 SNPs	1 to 3 SNPs		
BsrGI	**Total Loci**	**Loci 2 parents**	**Loci 2 parents w/ progeny**	**PS-5 Progeny**	**Undomesticated Progeny**
379/TM1	15314	2407	n/a	n/a	n/a
379/TX231	19247	3273	n/a	n/a	n/a
PS5/K56	15027	1035	464	123	341
PS5/PI435242	10848	902	397	211	186
TM1/TX231	17215	1221	n/a	n/a	n/a
A1/D5	16545	346	n/a	n/a	n/a

From these loci, 42 (PS-5 x K-56) and 36 (PS-5 x PI-435242) loci were linked to their offspring (PR-1_4024 and PR-2_4127). Out of the 42 loci from the PS-5 x K-56 cross, 12 progeny (PR-4024) loci with SNPs had an ancestral relationship to the PS-5 parent, while 21 progeny (PR-4024) loci with SNPs were linked to the K-56 parent. Data analysis revealed 17 progeny (PR-4024) loci out of the 36 loci from the PS-5 x K-56 cross had SNPs linked to the PS-5 parent, while only ten progeny (PR-4024) loci shared SNP linkage to the K-56 parent.

In the *BsrGI* experiment, loci between the interspecific and intraspecific crosses were well distributed. In this analysis, the interspecific (3-79 x TM-1) genetic standard divergence showed 15,314 loci of similar sequences. Although TM-1 had sufficient sequence amounts in the *BsrGI* experiment, a comparison of 3-79 against TX-231 was done to show the interspecific differences between *Gossypium barbadense* and *Gossypium hirsutum* of the two experiments: *BsrGI* and *HinP1I*. The total loci amount increased to 19,247. Next, an evaluation between the cultivated and wild *Gossypium spp.* was done. In *Gossypium hirsutum*, cultivated TM-1 and wild TX-231 sequences were surveyed against one another and found to have 17,215 similar loci. In *Gossypium barbadense*, again the two intraspecific crosses (cultivated PS-5 x wild K-56 and PS-5 x wild PI-435242) with progeny were reviewed for similar sequences. Cross (PS-5 x K-56) showed 15, 027 loci, while the cross (PS-5 x PI-435242) showed 10,848 loci. A control factor, the inter-genomic cross of the ancestral lines *Gossypium herbaceum* A1 and *Gossypium raimondii* D5 uncovered 16,545 loci of similar sequences.

Between taxa matches, a survey to locate one to three SNPs in a single locus was done. The largest number of SNPs within a locus was contained in the interspecific crosses of *Gossypium barbadense* and *Gossypium hirsutum* (3-79 x TM-1 and 3-79 x TX-231). This resulted in 2,407 (3-79 x TM-1) and 3,273 (3-79 x TX-231) sequences with one to three SNPs, as seen in Table 35. The perceived SNP amount in the intraspecific crosses (wild and cultivated *Gossypium spp.)* was lower than the interspecific crosses. In *Gossypium hirsutum,* intraspecific cross (cultivated TM-1 and wild TX-231), 1,221 loci shared one to three SNPs. In *Gossypium barbadense*, the intraspecific crosses with PS-5 had 1,035 (PS-5 x K-56) and 902 (PS-5 x PI-435242) SNPs within their comparative cataloged sequences, as seen in Table 35. The inter-genomic cross (A1 x D5) was comprised of 346 loci sharing one to three SNPs.

Data ascertained from the *Gossypium barbadense* intraspecific SNPs revealed linkage to each individual parental genotype. From these loci, 464 (PS-5 x K-56) and 397 (PS-5 x PI-435242) loci were linked to their offspring (PR-1_4024 and PR-2_4127). Out of the 464 loci from the PS-5 x K-56 cross, 123 progeny (PR-4024) loci with SNPs had complete ancestral linkage to the PS-5 parent, while 341 progeny (PR-4024) loci with SNPs were linked to the K-56 parent. From the 397 loci in the PS-5 x PI-435242 cross, 211 progeny (PR-4127) loci with SNPs revealed linkage to the PS-5 parent, while 186 progeny (PR-4127) loci shared SNPs linked to the PI-435242 parent.

Comparison between Intraspecific *Gossypium barbadense* lines of related SNP loci

After the discovery of SNPs correlated to specific parental loci, a comparison of loci representing PS-5 in both crosses was done. In *HinP1I*, no overlap between loci linked to PS-5 was seen. Also, overlap of the uncultivated *Gossypium barbadense* (K-56 or PI-435242) uncovered no overlap in loci. The number of sequences in the *HinP1I* parental loci did not effectively meet minimal coverage. In the *BsrGI* experiment, correlation was found between the two intraspecific *Gossypium barbadense* crosses.

The identical loci in both populations representing the same SNPs mapping with the same parents were indicated by the phrase overlapping (O). The loci, represented in only one intraspecific cross, were designated as singletons (S). The SNP data from the BsrGI experiment established ten overlapping loci representing linkage to cultivated *Gossypium barbadense* PS-5, while it uncovered 47 overlapping loci linked to both uncultivated 'wild' *Gossypium barbadense* (K-56 and PI-435242). There were 272 singleton loci in one of the two intraspecific crosses representing PS-5, whereas 391 singleton loci revealed association to either K-56 or PI-435242 (Table 36). There were 42 sites of disagreement, where one cross linked the loci to the cultivated PS-5 parent and the other cross showed loci linkage to the wild parent.

Table 36 Informative SNPs with Correlating Progeny

Heat Map

Loci Relatedness to Parental Genotype

	Highest ***Overlapping*** **PS-5**	**High** ***Singleton*** **PS-5**	**Low** ***Singleton*** **Wild**	**Lowest** ***Overlapping*** **Wild**	**Disagreement** ***Overlapping*** **PS-5 / Wild**
Number of Progeny/Parent Loci	10	272	391	47	42

The statistical probability of the BC4F5 population containing PS-5 genomic of all alleles being un-linked was 1.9% to 3.1% based on the possible breeding strategy one (Table 37a). The second possible breeding strategy required maintaining a F1 as the recurrent parent in the backcross schema over many years (Table 37b). These breeding strategies were based on alleles in random mating without linkage to set a baseline for heredity. Both the PS-5 x K-56 population and the PS-5 x PI-435242 chi-squares in Table 38 show that our populations are not randomly mated and there is linkage. The chi-square implicated other factors additive, dominance, and environmental factors were likely to be involved in genomic heredity of the breeding populations (Equation 1). After analyzing the number of overlapping loci in Table 36, we used Bayes Theorem to determine the probability of those two crosses having identical informative SNPs at the same loci (Equation 2 and Table 39). The likelihood of these two populations containing 10 overlapping loci that map back to the cultivated PS-5 genome was 11.4%.

Table 37 Baseline for Heridity with Random Mating and Non-Linkage

a) Possible Breeding Strategy 1

Generation	Cross	Progeny	Cross Probability	Selection For Early Flowering	Probability of Cultivated Genome	Probability of Wild Genome
F1	aa (cult P1) x Aa (wild P2)	Aa or aa	0.5	aa	0.5	0.5
BC1	aa (F1) x Aa (P2)	Aa or aa	0.5	aa	0.25	0.75
BC2	aa (BC1) x Aa (P2)	Aa or aa	0.5	aa	0.125	0.875
BC3	aa (BC2) x Aa (P2)	Aa or aa	0.5	aa	0.0625	0.9375
BC4	aa (BC4) x Aa (P2)	Aa or aa	0.5	aa	0.03125	0.96875

Selection Pressure for Keeping Mostly Wild Traits for Conversion Lines

Generation	Cross	Cross Probability	Probability of Cultivated Genome	Probability of Wild Genome
BC4F2	BC4 x BC4	(0.5 to 1) x .03125	0.015625 to 0.03125	0.96875 to 0.984348
BC4F3	..F2 x ..F2	(0.5 to 1) x (0.015625 to 0.03125)	0.0078125 to 0.03125	0.96875 to 0.9921875
BC4F4	..F3 x ..F3	(0.5 to 1) x (0.0078125 to 0.03125)	0.00390625 to 0.03125	0.96875 to 0.99609375
BC4F5	..F4 x ..F4	(0.5 to 1) x (0.00390625 to 0.03125)	0.001953125 to 0.03125	0.96875 to 0.998046875

Table 37 Continued.

b) Possible Breeding Strategy 2

Generation	Cross	Progeny	Cross Probability	Selection For Early Flowering	Probability of Cultivated Genome	Probability of Wild Genome
F1	aa (cult P1) x AA (wild P2)	Aa	1	N/A	0.5	0.5
F2	Aa (F1) x Aa (F1)	AA, Aa, aa	0.25	aa	0.25	0.75
BC1	aa (F2) x Aa (F1)	Aa or aa	0.5	aa	0.125	0.875
BC2	aa (P1) x Aa (F1)	Aa or aa	0.5	aa	0.0625	0.9375
BC3	aa (P1) x Aa (F1)	Aa or aa	0.5	aa	0.03125	0.96875
BC4	aa (P1) x Aa (F1)	Aa or aa	0.5	aa	0.015625	0.98348

Selection Pressure for Keeping Mostly Wild Traits for Conversion Lines

Generation	Cross	Cross Probability	Probability of Cultivated Genome	Probability of Wild Genome
BC4F2	BC4 x BC4	(0.5 to 1) x .0.015625	0.0078125 to 0.015625	0.984348 to 0.9921875
BC4F3	..F2 x ..F2	(0.5 to 1) x (0.0078125 to 0.015625)	0.00390625 to 0.015625	0.984348 to 0.99609375
BC4F4	..F3 x ..F3	(0.5 to 1) x (0. 00390625 to 0.015625)	0.001953125 to 0.015625	0.984348 to 0.998046875
BC4F5	..F4 x ..F4	(0.5 to 1) x (0.001953125 to 0.015625)	0.0009765625 to 0.015625	0.984348 to 0.9990234375

Equation 1 Heredity with Population Mean with Additive, Dominance, and Environmental Effects

$Y_{ijk} = \mu + a_i + \beta_{ij} + e_{ij}$

Population mean = μ; Additive Effects = a; Dominace effects = β; *and Environmental Effects = e*

Table 38 Chi Square for Random Mating and Linkage

BsrGI	PS-5 Progeny	Undomesticated Progeny	Row Totals
PS5/K56	123	341	**464**
PS5/PI435242	211	186	**397**
Column Totals	**334**	**527**	**861**

BsrGI	PS-5 Expected	Undomesticated Expected	Row Totals
PS5/K56	14	450	**464**
PS5/PI435242	12	385	**397**
Column Totals	**334**	**527**	**861**

ChiSquare	4277.99
Degrees of Freedom	3
Number of Classes	4

	Probability				
Degrees of Freedom	**0.9**	**0.5**	**0.1**	**0.05**	**0.01**
1	0.02	0.46	2.71	3.84	6.64
2	0.21	1.39	4.61	5.99	9.21
3	0.58	2.37	6.25	7.82	11.35
4	1.06	3.36	7.78	9.49	13.28
5	1.61	4.35	9.24	11.07	15.09

HO: Random mating and no linkage

HA: Linkage and non-random mating

We found that random mating without linkage is not occurring P>0.05.

Equation 2 Bayes Theorm

$$P(A|B) = (P(B|A)P(A)) / P(B)$$

Table 39 Bayes Theorm for Identical Loci

	Prior Probability PS5		Prior Probability wild	
PS5/K56	0.2651	PS-5	0.7349	K56
PS5/PI435242	0.5315	PS-5	0.4685	PI435242

	Probability
Double Overlap (10 loci)	0.0299

Event A1: We had 10 matching sequences K56
Event A2: We did not have 10 matching sequences
Event B1: We had 10 matching sequences PI435242
Event B2: We did not have 10 matching sequences

Event	Probability
P(A1)	0.2651
P(A2)	0.9701
P(B1)	0.5315
P(B2)	0.4685

$$P(A_1 | B_1) = \frac{P(A_1)\,P(B_1 | A_1)}{P(A_1)\,P(B_1 | A_1) + P(A_2)\,P(B_1 | A_2) + P(A_1)\,P(B_2 | A_1) + P(A_2)\,P(B_2 | A_2)}$$

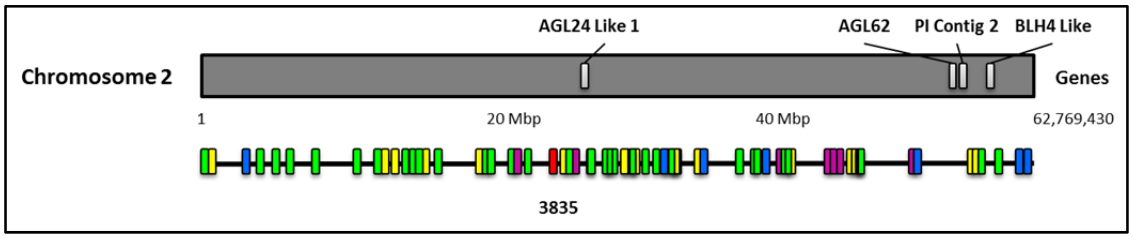
$$P(A_1 | B_1) = (0.2651)(0.5315) \;/\; [\,(0.2651)(0.5315) + (0.9701)(0.5315) + (0.2651)(0.4685) + (0.9701)(0.4685)\,]$$

$$P(A_1 | B_1) = 0.114071122$$

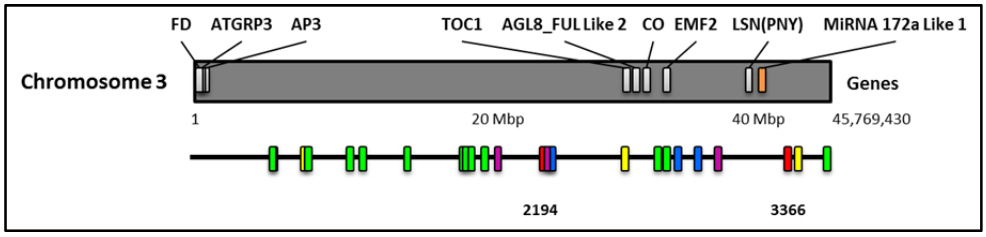
Mapping with Candidate Genes and GBS markers to *Gossypium raimondii* Draft Genome

The related loci represented in the heat map (Table 36) were queried against the *Gossypium raimondii* draft genome in Geneious® to ascertain locations based on sequence. These loci fragments were then aligned to the draft genome. The proposed candidate genes known in *Arabidopsis thaliana* for the floral development network were overlaid onto the draft genome. This allowed for correlation between loci representing the intraspecific cross to be illustrated with proposed homologous candidate genes on the *Gossypium raimondii* draft genome. The focused alignment was placed on overlapping loci related to the cultivated *Gossypium barbadense* PS-5. Locus 3835 was aligned near *AGL24-Like 1* on chromosome two of the *Gossypium raimondii* draft genome (Figure 18 a). Loci 2194 and 3366 were aligned to chromosome three (Figure 18 b). Locus 8037 was aligned near *AGL9_SEP3-Like* on chromosome five (Figure 18 c). Loci 2165, 2187, and 3241 were aligned on chromosome nine near *PIE1-Like*, *TFL1-Like*, and *ATC_Centroradialis-Like* (Figure 18 d). Loci 2193 and 3240 were aligned to chromosome eleven. Some loci near genes were *FY-Like*, *LDW1-Like 1*, *LDW2-Like*, *miRNA 172c-Like*, and *miRNA 172d-Like* (Figure 18 e). Locus 4429 was aligned to chromosome thirteen (Figure 18 f).

a)



b)



c)

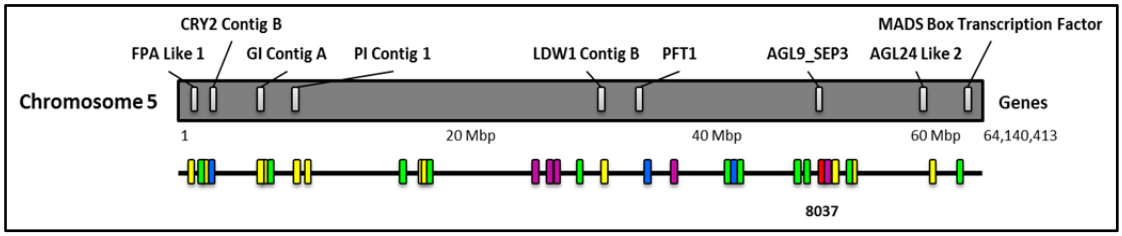
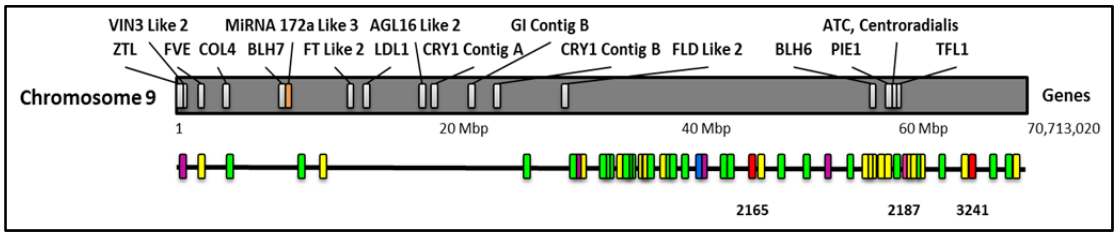
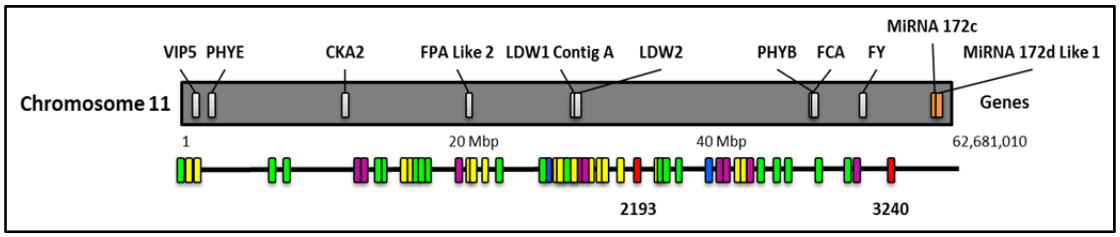


Figure 18 Alignment of Informative SNPs to *G. raimondii* Draft Genome

d)



e)



f)

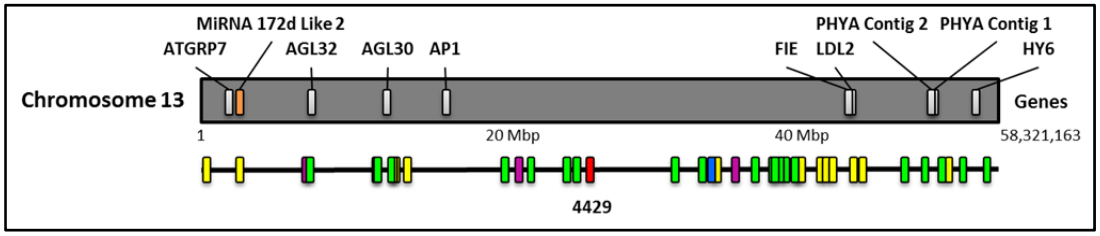


Figure 18 Continued.

Targeted GBS

Ten gene specific forward primers correlated the SNPs within overlapping loci were adapted to the Illumina® TruSeq Indexed Adaptor P5. The Illumina® TruSeq Indexed Adaptor P5 and P7 allowed for a population to be run on a single plate with both forward and reverse in-line barcodes. 'Reduced representation' was achieved by the combination of the *BsrGI* restriction enzyme and gene specific primers. This Targeted GBS (TGBS) on the Illumina® HiSeq 2500 gave 18,656,276 total sequences (Table 40) from splitting the reverse barcodes. The sequences were then imported into Geneious® where they were split using the forward barcode (Table 41). The forward and reverse barcodes indicated which individual was being sequenced.

Table 40 TGBS Reverese Illumina TruSeq Barcode Sequences Split

TGBS Reverse Barcodes	
Reverse P7 Barcodes	**# Sequences**
P7 TruSeq 1	9,410,359
P7 TruSeq 2	1,244,203
P7 TruSeq 3	1,013,110
P7 TruSeq 4	938,926
P7 TruSeq 5	1,084,218
P7 TruSeq 6	2,555,050
P7 TruSeq 8	2,410,410
Total Number of Sequences	***18,656,276***

Table 41 TGBS Forward Illumina TruSeq Barcode Sequences Split

TGBS Forward Barcodes	
Total Forward P5 Barcodes	**# Sequences**
P5 TruSeq 8	4,690,515
P5 TruSeq 9	4,428,201
P5 TruSeq 10	3,444,071
P5 TruSeq 11	2,855,232
P5 TruSeq 12	2,074,695
P5 TruSeq 1 -Unknown	1,163,622
Total P7 TruSeq 1 Sequences	***18,656,336***

Within the PS-6 and K-46 individuals, we looked for our targeted loci primers. First, we correlated our original informative loci back to *G. barbadense* 3-79, a double haploid from our original sequences, to see if more than one copy existed (Figure 19). Sequence 2187 had two similar loci in *G. barbadense* 3-79, while sequence 8037 had two different loci only on *Gossypium raimondii* D5 and no matching reference in *G. barbadense* 3-79. Reference sequences from *G. barbadense* 3-79 were queried against PS-6 and K-46. Noticeably, not all sequences had 100% matches, so the best possible match was taken. A new set of reference sequences from the on-target PS-6 and K-46 sequences were created (Table 42). Each individual in the segregating population was mapped back to these references. Many individuals had only a few loci results.

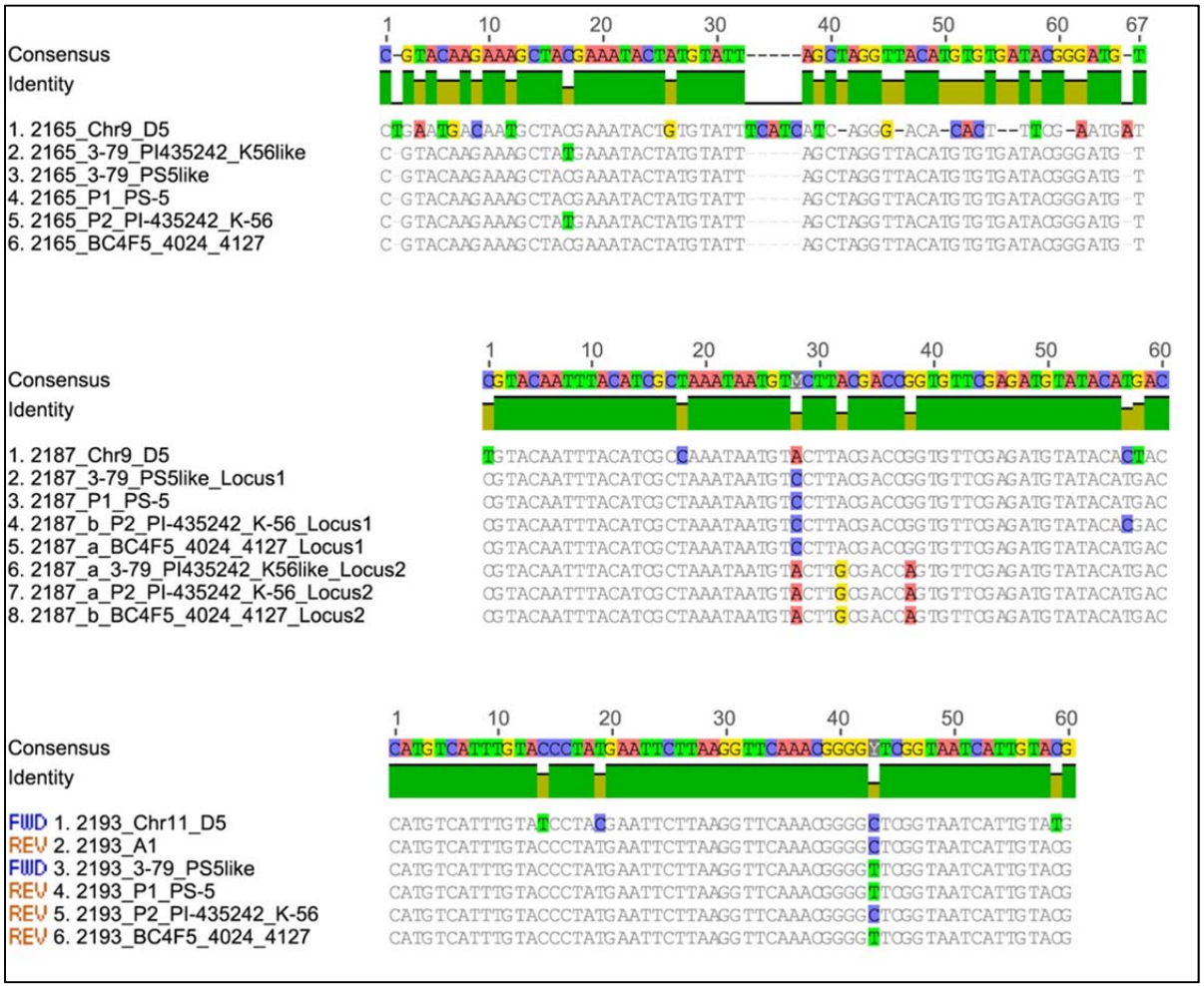


Figure 19 Alignment of 10 Loci with D5, PS-5, K-56/PI-435242, 379, and A1*(*if available)

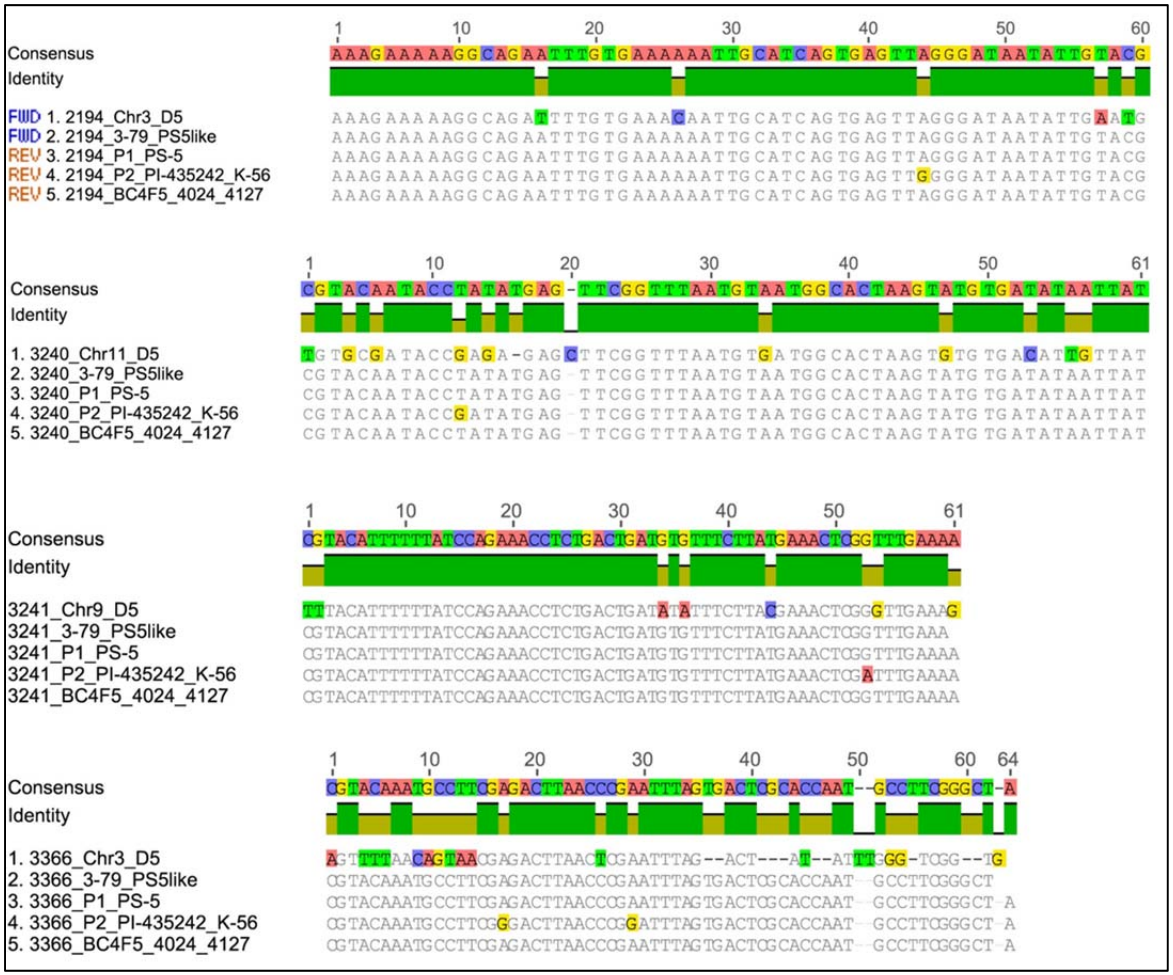


Figure 19 Continued.

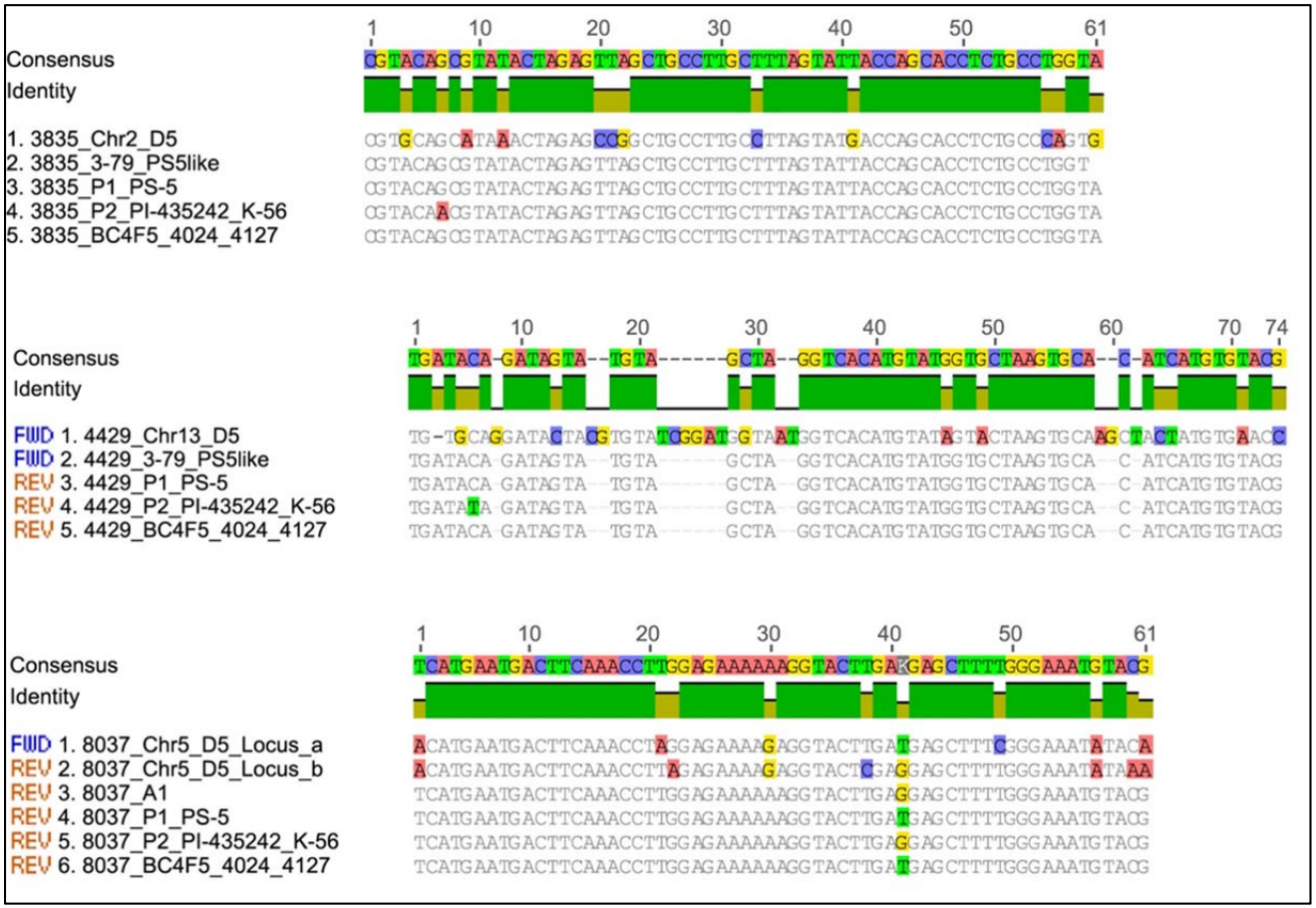


Figure 19 Continued.

Table 42 On-target Reference Sequences

Reference Name	Sequence
2187a on PS6 Locus1	CGTACAATTTACATCGCTAAATAATGTCCTTACGACCGGTGTTCGAGATGTATACA
3241 on PS6	CGTACATTTTTTATCCAGAAACCTCTGACTGATGTGTTTCTTATGAAACTCGGTTTGAA
3241 on K46	CGTACATTTTTTATCCAGAAACCTCTGACTGATATGCTTCTTTTGGAACTCAATTTGGAAGA
3366 on PS6 Het A	CGTACAAATGCCTTCGAGACTTAACCCGGATTTAGTGACTCGCACCAATGCCTTCGGGCT
3366 on PS6 Het_B	CGTACAAATGCCTTCGGGACTTAACCCGAATTTAGTGACTCGCACCAATGCCTTCGGGCT
3366 on K46	CGTACAAATGCCTTCGAGACTTAACCCGGATTTAGTAACTCGCACCAATGCCTTCGGGCT
4429 on PS6	TGATACAGATAGTATGTAGCTAGGTCACATGTATGGTGCTAAGTGCG
4429 on K46	TGATACAGATAGTATGTAGCTAGGTCACATGTATGGTGCTAAGTGCG

We found that only four of the initial primers targeted the correct fragment. Several sequence paralogs were discovered from our initial reference mapping. The TGBS sequences for thirty-three individuals of the segregating BC4F2 wild conversion lines were queried for identical sequences. Interestingly, it appeared that only 2187 seemed to be highly correlated with flowering (Table 43). In table 43, the sequence fragment 4429 showed strong linkage with the C nucleotide, while 3366 appeared to be heterozygous. Also table 43 illustrated that the SNPs for 3366 did not show linkage to flowering for the nucleotides of G/A at position 17 and G/A at the 29th position. While 3241 may have linkage with flowering, too few sequences were obtained from the TGBS data to make an accurate analysis of co-segregation with the photoperiod independence locus.

Table 43 TGBS Results

TS Fwd	*TS Rev*	*Plant ID*	*Pheno*	*PS-6 2187*	*PS-6 3241*		*K-46 3241*		*PS-6 3366 Het A*				*PS-6 3366 Het B*			*K-46 3366*				*PS-6 4429*		*K-46 4429*	
Guadeloupe - BC4F2 - K46				*N/A*	*A/G 54*				*A/G 17 G 29*				*A/G 17 A 29*			*A/G 17 G 29*				*C/T 6*		*C/T 6*	
					A	G	A	G	AA	AG	GA	GG	AA	AG	GA	AA	AG	GA	GG	C	T	C	T
1	8	4341	F	5	0	2	0	2	0	16	1	21	0	0	0	23	631	0	0	10	0	736	19
1	9	4349	NF	1	0	0	0	1	0	1	0	0	0	0	0	0	0	14	266	582	21	39	3
1	10	4357	NF	1	0	0	0	0	0	7	0	12	0	0	0	5	224	15	327	0	0	18	0
1	11	4365	F	0	0	1	0	0	0	11	0	0	0	0	0	0	219	0	0	321	0	19	0
1	12	PS-6	F	1	0	1	0	1	0	10	0	0	0	0	10	9	215	24	389	32	0	379	7
2	8	4342	NF	1	0	0	0	0	0	1	0	2	0	0	3	0	63	0	75	98	0	4	0
2	9	4350	NF	0	0	0	0	0	0	2	0	0	0	1	1	0	25	0	48	0	0	10	1
2	10	4358	NF	1	0	0	0	0	0	4	0	3	0	0	3	2	14	3	35	18	1	15	2
2	11	4366	NF	0	0	0	0	0	1	2	0	8	0	0	1	2	20	2	41	49	3	2	0
2	12	4374	NF	1	0	0	0	0	1	20	0	26	0	0	3	2	26	2	55	35	1	5	0
3	8	4343	F	1	0	0	0	0	0	0	1	0	0	0	3	0	45	0	77	59	0	16	0
3	9	4351	NF	0	0	0	0	0	0	2	0	2	1	0	3	0	24	0	44	44	2	8	1
3	10	4359	F	1	0	0	0	0	0	0	0	0	0	0	2	0	21	0	26	10	1	22	1
3	11	4371	NF	0	0	0	0	0	0	1	1	1	0	0	3	0	10	0	39	30	0	1	0
3	12	4375	F	0	0	0	0	0	0	1	0	0	0	0	2	0	37	0	60	43	0	1	0
4	8	4344	NF	1	0	1	0	0	0	1	0	5	0	0	3	4	57	0	0	7	0	87	0
4	9	4352	NF	0	0	0	0	0	0	0	0	1	0	0	3	0	20	0	37	3	0	48	0
4	10	4360	NF	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	26	1	11	0
4	11	4368	F	0	0	0	0	0	0	1	0	3	0	0	1	0	17	0	27	344	1	27	1
4	12	4376	NF	0	0	0	0	0	0	2	0	2	0	0	1	0	21	0	50	49	0	2	0

Table 43 Continued.

5	8	4345	F	0	1	0	0	0	0	5	0	5	0	0	4	1	57	5	62	92	3	92	3
5	9	4353	NF	0	0	0	0	0	0	0	0	2	0	0	0	0	31	0	44	3	0	3	0
5	10	4361	NF	0	0	0	0	0	0	1	0	2	0	0	1	0	15	0	31	58	0	58	0
5	11	4369	NF	1	0	0	0	0	0	1	0	2	0	0	4	0	19	0	44	26	2	26	2
5	12	K-46	NF	0	0	0	0	1	0	0	0	0	0	0	3	0	34	0	43	49	1	49	1
6	8	4346	NF	0	0	2	0	0	0	6	0	0	0	0	0	0	137	0	0	14	0	14	0
6	9	4355	NF	0	0	0	0	0	0	2	0	2	0	0	0	0	48	0	90	9	1	9	1
6	10	4363	NF	0	0	0	0	0	0	2	0	1	0	0	0	0	32	0	51	38	0	38	0
6	11	4370	F	1	0	0	0	0	0	0	0	0	0	0	6	0	39	0	76	71	0	71	0
6	12	4378	NF	0	0	0	0	1	0	4	0	4	0	0	5	0	39	0	98	88	3	88	3
8	8	4348	NF	0	0	1	0	0	0	5	0	5	0	0	0	0	90	0	156	124	5	124	5
8	9	4356	F	1	0	0	0	0	0	1	0	1	0	0	0	2	26	0	33	58	0	58	0
8	10	4364	NF	0	0	0	0	0	1	1	0	0	0	0	0	0	25	0	48	30	1	30	1
8	11	4372	NF	0	0	0	0	0	0	3	0	2	0	0	0	0	35	4	54	2	0	2	0
8	12	4380	NF	0	0	0	0	0	0	0	0	0	0	0	0	0	25	4	66	2	0	2	0

Discussion

Efficiency of Reduced Representation and Analysis

Using two different enzymes for GBS taught us the differences between having highly conserved coding regions and more variability in the upstream and downstream elements. As expected, we found more diversity in the *BsrGI* sequences, but were surprised by the extreme lack of informative SNPs within the *HinP1I* sequences. The low levels of polymorphism within the Hinp1I sequences were expected due to the fact *HinP1I* GC richness selected more coding region fragments. These extremely low levels may also indicate that one homeolog is methylated.

Using STACKs software helped us to easily and efficiently sort through the data, but as any program goes there were some errors. We were able to sort through these minor errors by evaluating the data with queries in Microsoft© Access. We found 10 loci from the progeny of two crosses that showed linkage to the cultivated photoperiod independent parent *Gossypium barbadense* L. Pima S-5 (PS-5). These loci mapped back to the *Gossypium raimondii* draft genome groups: 2, 3, 5, 9, 11, and 13. These loci were correlated with the genes involved in flowering (Logan-Young, Chapter 3).

Significance of Having the 'A' and 'D' Genome for Comparison

The significant loci were mapped to the released 'D' genome with little trouble, but deciphering whether these were A/D SNPs or SNPs between cultivars was difficult without a fully sequenced 'A' reference genome. Sequencing the A1 and D5 cotton samples in the BsrGI did help to place some loci on the 'A' or 'D' strand, but more

sequencing depth of A1 *Gossypium herbaceum* and D5 *Gossypium raimondii* would be advisable to future studies.

Despite not having fully sequenced parents, we were still able to attain the same loci using selective enzyme digestion and reduced representation by size selection. We compared both the 'A' and the 'D' sequences to PS-5 to find loci that matched our significant loci.

GBS Loci Discovery Putatively Linked to Photoperiod Independence

This research has led to the discovery of markers linked with photoperiod independence from the cultivated *Gossypium barbadense* PS-5 parent. These loci will allow for a quick and efficient way to narrow down the candidate genes implicated behind photoperiod independence.

These loci were tested in a segregating population (BC4F2 - PS-6 x K-56) using a new method called Targeted GBS. We found that loci 2187, 3241, 3366, and 4429 worked well within the segregating population using the Targeted GBS protocol, while loci 2165, 2193, 2194, 3240, 3835, and 8037[1] gave poor results. One reason was due to our design for SNPs located to close to the *BsrGI* cut site. Another reason appears to be a problem with GBS on the Illumina® HiSeq® 2500. There seems to be a phenomenon of an all or nothing response to sequencing on the Illumina® HiSeq® 2500. In the future, we may overcome the difficulties found with paralogs by designing primers that will select against other paralogs.

[1] No results were seen with 8037.

From this data, we were able to correlate which loci were seen in photoperiod dependent and photoperiod independent *Gossypium barbadense* from the segregating population (BC4F2 - PS-6 x K-56). We showed that one fragment segregated with photoperiod independence. That fragment was 2187. Another fragment, 3241, might have showed linkage with photoperiod independence, but too few fragments were obtained.

Conclusion

Cotton produces one generation per year in most production regions. A normal backcross breeding program would take five years. Next, the breeder would want to cross this plant with another plant, of the recurrent parent line, and then self the progeny for five to seven generations.

In traditional breeding programs, this process would take 10 to 20 years to bring in a new trait from the wild Germplasm. One way to alleviate this long and arduous process is to select only those plants carrying the trait of interest after a F1 cross, without bringing photoperiodic sensitivity by utilizing marker assisted selection (MAS) breeding. Our study provided one closely linked marker to the trait photoperiodism that breeders can utilize in MAS breeding programs.

Methods

Plant Growth

Seeds from TM-1, 3-79, PS-5, PI-435242, K-56, A2, and D5 were imbibed in distilled water in the +4ºC refrigerator overnight. Falcon petri dishes (150 x 20 mm) had two autoclaved growth paper pieces cut to fit them (Table 44). One circle of paper was placed in the bottom of the petri dish. Seed coats were removed from the cotton seeds. Eight to ten seeds were positioned independently upon the paper circle. Another paper circle was placed over them. The sheets were dampened with distilled water. The petri dish lids were then fixed on top and parafilm was placed around the petri dishes. Next, these dishes were set out upon bench tops under growth lamps for five to eight days. Once the cotyledons were present, the whole plant was deposited into a labeled 1.5mL tube with a hole punched through the lid. The tube was then cast into liquid nitrogen to preserve the DNA and placed into a -80ºC freezer until DNA extractions could be performed.

Table 44 Cotton Seed Information

Year	Scientific Name	Name	Alternate no.	PI No.	Type	Origin
2003	*G. herbaceum*			03-PI408785	D5-1	Peru
1989	*G. raimondii*			89-PI530898	A1-27	
1988	*G. barbadense*	K-46	AE-CRC-88229	88-PI528313		Guadeloupe, St. Francois
2005	*G. barbadense*	PS-6	May-45			
2006	*G. barbadense*	PS-5	May-44			
1984	*G. barbadense*	K-56	AE-CRC-8429	84-PI274514		Peru, Piura, Sinchao Chico
1985	*G. barbadense*	PI-435242		85-PI4352342		
2005	*G. hirsutum*	GHOP 04-05	SA-2269	05-PI607172		USA, Texas, College Station
2002	*G. barbadense*	3-79 2:10:12	GB-1585			
2005	*G. hirsutum*	TX-231		05-PI163725		Guatemala, Zacapaa, Zacapa

Barbadense Field Trial

In 2008, a field trial created by Dr. Richard Percy at the United States Department of Agriculture Southern Plains Agricultural Research Station (USDA-ARS-SPARC) in College Station, TX had a BC4F5 population of eight lines consisting of five samples each. While most of the wild traits were retained in the offspring, photoperiod independent early flowering from the established PS-5 cultivar was observed in the wild cultivar offspring. Finally, leaf tissue was taken from these plants and put into labeled

1.5mL tubes with holes punched through their lids. After being dipped in liquid nitrogen, they were then placed into a -80ºC freezer until DNA extractions were performed.

DNA Samples

DNA was taken from genetic standards *Gossypium hirsutum* TM-1 and *Gossypium barbadense* 3-79, the parental lines *Gossypium barbadense* lines (PS-5, PI-435242, and K-56), the ancestral lines *Gossypium herbaceum* A1 and *Gossypium raimondii* D5 (only used in *BsrGI* experiment), photoperiod dependent *Gossypium hirsutum* TX-231, and fully integrated photoperiod independent offspring *Gossypium barbadense* BC4F5 4024 and 4127. High quality DNA was extracted using the Pepper Lab Cotton DNA Extraction protocol (Logan-Young, 2013, Chapter 3).

Restriction Enzymes and Design of Adaptor

Restriction Endonuclease (RE) selection based on frequent base pair (bp) cutting and high efficiency was imperative for eliminating sequence elements containing repetitive regions. The first RE selected was *HinP1I*, a 4bp GC rich cutter (G^CGC) having 100% activity in NEBuffer® 1, 2, 3, and 4(New England Biolabs). This enzyme was selected because of its methylation sensitivity. Methylation sensitivity has been used as a method of genomic reduction [375, 379, 380]. The high GC concentration allowed for selection of gene rich coding regions rather than upstream and downstream elements. This RE left a 1bp overhang of C to where the adaptor can ligate.

The *HinP1I* experiment consisted of two adaptors being ligated to the cut DNA. First, two oligonucleotide sequences were constructed with the Illumina® paired end (PE) Adaptor A1, a specific 6bp barcode, and the HpaII site overhang. The HpaII site

was complementary to the *HinP1I* site. The first sequence of the oligonucleotide for PE Adaptor A1 was the top strand 5'bACACTCTTTCCCTACACGACGCTCTTCCGATCxxxxxxC, where xxxxxx represented the top barcode. On the bottom strand of the PE Adaptor A1, the oligonucleotide sequence was 5'CGGyyyyyyAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT, where yyyyyy represented the bottom barcode. (Figure 20) Second, two additional oligonucleotide sequences consisted of the Illumina® PE Adaptor A2, and the HpaII site overhang. The top strand sequence for PE Adaptor A2 was 5'bCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTC. The bottom strand was 5'CGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG. These two adaptors were ordered and prepared during the next-generation library sequencing preparation. (Figure 20)

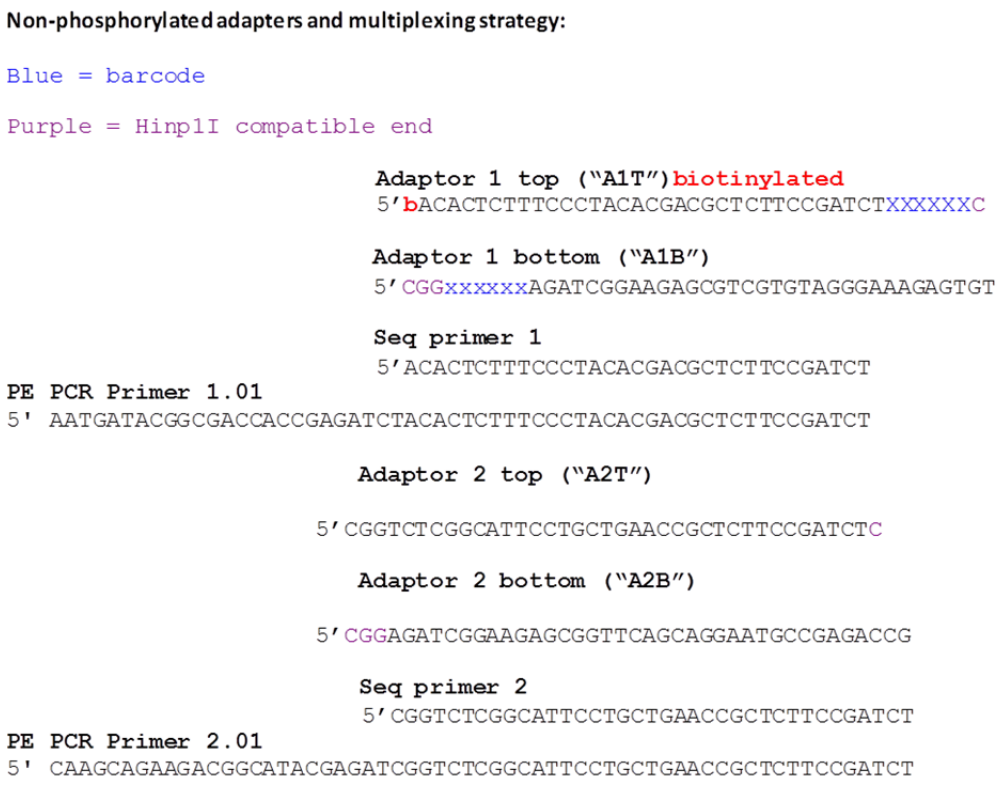


Figure 20 *HinP1I* Adaptor Strategy

The *BsrGI* experiment consisted of two adaptors with the same PE Adaptor A1, a 6bp barcode and a PE Adaptor A2 site, but contained a *BsrGI* overhang. The top and bottom strands were complementary.

Next-Generation Library Sequencing Preparation

For each oligonucleotide pair,10µl of 100mM stock of each of the top and bottom strands were diluted in 20µl of S^2TE. Each pair was annealed in a water bath at 95ºC under foil for five minutes, then the water bath was turned off and adaptors were allowed to cool to room temperature. Each 40µl annealing reaction was then diluted with 60µl of STE (Sodium-Tris-EDTA) to create a working stock at 10pmol/µl.

Digestion-Ligation

Each DNA (250ng), restriction endonuclease, 10xNEB2® buffer (New England Biolabs), DiH_2O, spermadine, and BSA was loaded into strip tube wells in 20µl reactions. They were then digested for two hours at 37ºC (Figure 21). Then adaptors were ligated in a competitive reaction process. Adaptors A1 and A2 (10 pmol/µl) with T4 DNA ligase, DiH_2O and 10xNEB® T4 ligase buffer (New England Biolabs) were added to the digestion reaction. This was incubated at 22ºC for one hour and heated to 37ºC for 30 minutes. The ligation reaction was then heated to 65ºC for 20 minutes to inactivate the T4 DNA ligase (Figure 22). Once completed, these reactions were then pooled together into one 1.5mL tube.

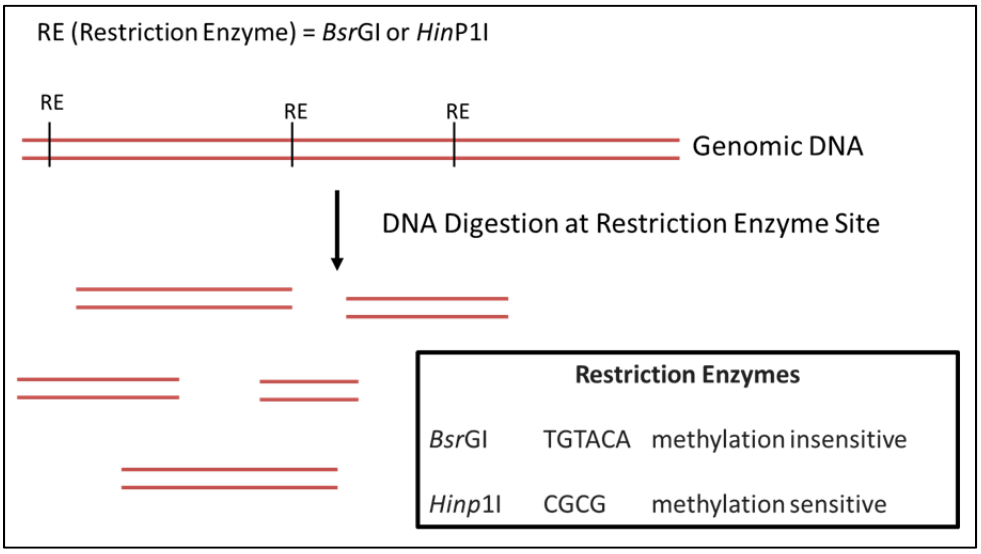


Figure 21 Digestion

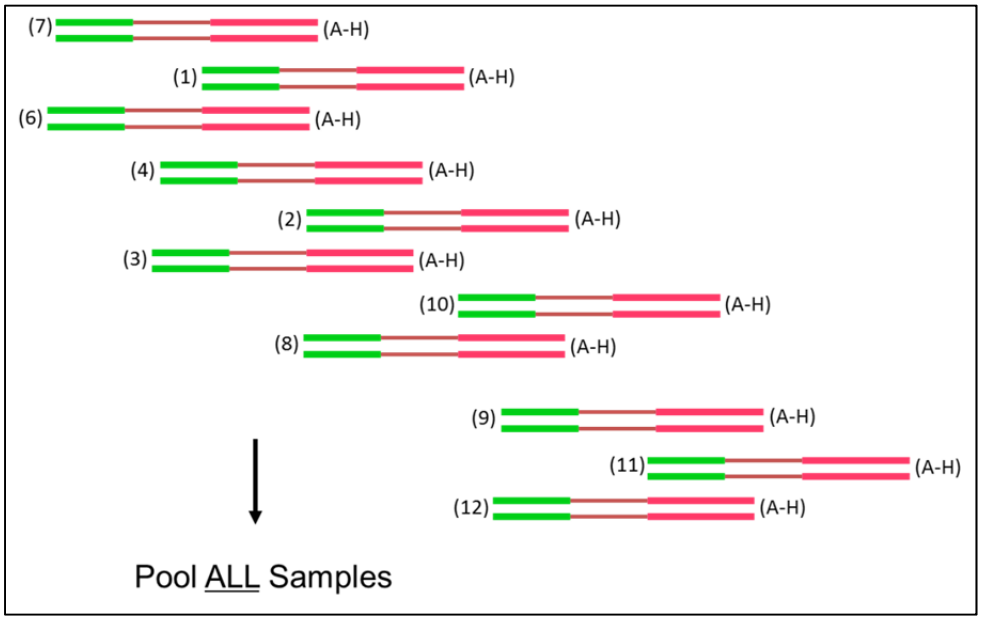


Figure 22 Ligation Reaction

Gel Extraction

Complexity reduction to avoid chloroplast contamination was performed by a two hour gel electrophoresis of adaptor ligated DNA fragments run across a Gel Green® 2.5% 1xTBE gel (*BsrGI*: Gel Green® 2% 1xTAE gel). DNA was extracted from 270bp to 290bp using the X-tracta® tool (LabGadget LLC) by moving the gel selection into a new recess on the dark reader. Size was selected via a 100bp ladder and lambda-280 fragment run as controls. The 270bp to 290bp fragment was then run into the Recochip® (TaKaRa 9039) through 30 minutes of gel electrophoresis at 110 volts. The samples were then purified to remove Gel Green® (Biotium) dye.

In the *BsrGI* experiment, using the X-tracta® tool was not needed because the size of the region to avoid chloroplast contamination was 150bp. Sizes were verified via a 1Kb+ ladder and lambda-230-350 mix run as controls. Hence, two Recochips® were used to isolate the region between 230bp to 350bp. The first chip blocked larger fragments from entering the isolation zone as gel electrophoresis proceeded (Figure 23). The second chip captured the intended fragment after 30 minutes of gel electrophoresis at 110 volts (Figure 24). When samples were run on a TAE gel, then a MinElute PCR Purification (Qiagen™) was performed to change buffer solutions to TE.

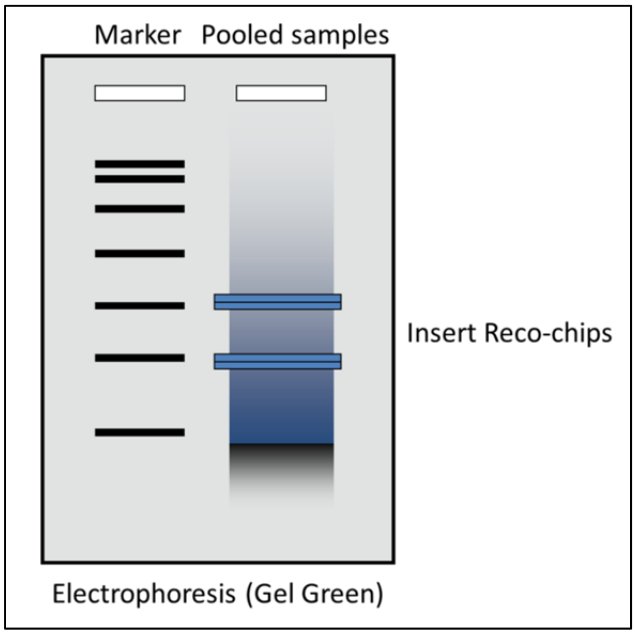


Figure 23 Recochip Complexity Reduction

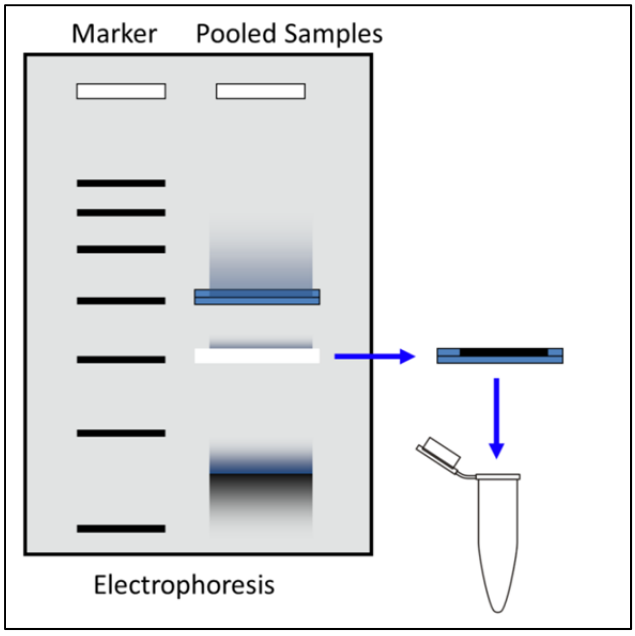


Figure 24 Recochip Capture

Fill-in and ssDNA Isolation

After complexity reduction, the samples needed to have their nicks filled in via the NEBNext® Fill-in and ssDNA Isolation Module (New England Biolabs). During the ligation process, nicks were present because the restriction enzyme fragments lacked a 5' phosphate on the bottom strand. The Bst DNA polymerase recognizes these sites and repairs the missing bases derived from the complementary DNA fragment and Adaptor A1 [384].The single stranded isolation uses a wash step where the un-biotinylated fragments will be washed away and the double biotinylated fragments will not be eluted. Thus, DNA fragments having one biotinylated end were kept for amplification.

PCR Amplification

The adaptor-ligated DNA fragments were amplified using two partial complementary PCR primers. The paired end PCR primers were recognized by the Illumina® GAII. These primers were PE_PCR_Primer_1.01 and PE_PCR_Primer_2.01 (Table 45).

Table 45 Paired End Primers

PE_PCR_Primer_1.01
5'-AATGATACGGGACCACCGAGATCTACACTCTTTCCCTACACGACGACGCTCTTCCGATCT

PE_PCR_Primer_2.01
5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

PCR cycling conditions consisted of 98ºC for 30 seconds followed by 30 cycles (35 cycles *BsrGI*) of 98ºC for 12 seconds, 65ºC for 30 seconds, 72ºC for 30 seconds, then with a final extension time of one minute at 72ºC. These amplified pooled samples made up the library. The samples were then cleaned and purified using the MinElute PCR Purification kit (Qiagen®) and the AMPure XP Tube (Agencourt®) protocol[2].

Quantification

Samples were quantified by nanodrop. This was used to get a general approximation of the sample's ng/µl concentration. Next, the sample was quantified to a specific concentration of double-stranded DNA with the High Sensitivity AccuBlue

[2] Second gel extraction was done on the *BsrGI* fragment to get a smaller cleaner PCR product.

Quantification Kit (Biotium™) on the Victor™ *X3* Multilabel Plate Reader (PerkinElmer™). The quality control check was done on the Bio-Analyzer (Agilent™) with a pico-chip to make sure the fragment contained good quality DNA and correct size selection. The GBS samples were then diluted to a 10nM concentration to run on the Illumina. The 10nM GBS sample concentration was calculated using an equation (Equation 3). The 10nM sample was given to the TAMU AgriLife Genomics & Bioinformatics Services to run one lane on the Illumina. The *BsrGI* sample was spiked into another runs lane.

Equation 3 Calculation for Conversion of an ng/µl Solution to 10nM Concentration

(((## ng/µl * 1x10^6 µl/L)* 1 bp mol/660g)* 1/fragment size bp) = ## nM

Filtering Raw Reads

The unfiltered fastq 76bp paired end reads were imported into Geneious® to trim reads for quality. The region being trimmed consisted of a greater than five percent error per base. The maximum length after trimming was 66bp. The fastq reads after being trimmed to 66bp were exported from Geneious® and imported to CLCbio® to remove reads under 66bp. When bringing the fastq sequences into CLCbio® from Geneious®,

the data was imported as Sanger Sequence fastq quality scores. This step was imperative in making sure the quality scores associated with the fasta sequences were correct.

During the *BsrGI* GBS experiment, two runs were completed on the same data. The first run had a technical problem with the Illumina due to tiling errors, but some reads were useable. Using Geneious®, both runs were grouped together into a single fastq list for single run for STACKs. Reads had to be exactly the same length before running STACKs, thus catalog creation errors were avoided.

To get approximate values per barcode, reads were separated via barcode. Geneious® was used over other software versions because it allowed for single base pair mismatches in the barcodes and customized specific barcode sets. Also, the end-adapter filtered out some biased ends. Geneious® streamlined the data for easy taxa separated barcode visualization.

Building STACKs

The trimmed 66bp full unsorted fastq sequencing data was transferred to the linux server with a compiled STACKs program version 0.998. The renz.h file was changed to add the *BsrGI* and *HinP1I* barcodes. The sequences were filtered using the process radtags command. These barcode sorted sequence .fq files were then compared in pairwise combinations with and without progeny using the denovo_map.pl command. The data was loaded into several MySQL tables and was accessible via the STACKs web-interface.

Discovery of SNPs and Creation of Probable Markers

Cataloged sequences were annotated by SNP changes between: loci two parents, loci two parents and progeny, cultivated loci on progeny, wild loci on progeny, informative heterozygotes two parents with progeny, non-informative two parents with progeny, informative heterozygotes two parents, and non-informative two parents. The two tables STACKs generated, genotypes and observed haplotypes, were transferred to Microsoft® Access, so data could be easily filtered through SQL query scripts. Significant photoperiod independent SNP markers in both fully integrated offspring were exported into fasta format and moved into Geneious®.

Alignment to *Gossypium raimondii*

The significant photoperiod independent SNP markers were then queried against the released D5 *Gossypium raimondii* genome tentative contigs (A. Patterson, August 2012 release) in a localized Geneious® database. These reads were placed upon a scaffold map of the D5 tenative contigs. The candidate genes were also queried and aligned to the D5 scaffold map.

Verification of Significant SNPs in BC4F2 Segregating Populations

DNA was extracted from segregating BC4F2 (*Gossypium barbadense* PS-6 x *Gossypium barbadense* K-46) population using the Pepper Lab Cotton DNA Extraction protocol (Logan-Young, 2013, Chapter 3). Phenotypic data from the segregating population was taken during DNA sampling in the 2008 USDA-ARS-SPARC field trials. The entire PS-6xK-46 DNA population was diluted to a 30ng/ul concentration level. Due to multiplexing constraints 33 samples from the population of 40 individuals

were used in the Targeted GBS experiment with two additional parental DNA samples (PS-6 and K-46). The diluted 35 samples were arranged into a 96-well plate format: A-H rows (skipping row G) x 8-12 columns.

Probability Statistical Methods for Gene Linkage

A simple probability chart (Table 37) was created based on the probable heritability of a genome introgression in a natural population using random mating without linkage. This was important to establish because it laid the foundation for what the maximum threshold would be for a cultivated plant's genome to be incorporated into a wild genome without linkage disequilibrium (*LD*). Using a chi squared statistic, we showed that our population is not randomly mating and contains significantly more of the cultivated genome. This correlated to knowledge that some other traits besides photoperiod independence were being selected for unknowingly by the breeder during the creation of the wild Germplasm line. Finally, we were able to then use Bayes theorem (Equation 2) to calculate the comparative data from the *BsrGI* experiment in Table 39. Our equation in Table 39 estimated the probability of the two BC4F2 populations containing the same overlapping sequences.

Targeted GBS

Ten specific forward primers correlating to SNPs found in the over-lapping loci were created. These were adapted to the Illumina® TruSeq Indexed Adaptor P5 (Figure 25 and Tables 46-48). Two oligonucleotide sequences were constructed with the Illumina® TruSeq Indexed Adaptor P5 site, a specific 6bp barcode, *BsrGI* site, and a gene specific primer. The forward oligonucleotide sequence was

5’CTACACGACGCTCTTCCGATCTACTTGACTTCCGATCT- xxxxxx-CGTACA-qqqqqqqqq, where xxxxxx represented the forward barcode and qqqqqqqqq represented the gene specific primer. On the reverse strand, sequences were constructed with the Illumina® TruSeq Indexed Adaptor P7 site, a specific 6bp barcode, and a partial *BsrGI* site. The reverse oligonucleotide sequence was 5’ TGTACG-AGATCGGAAGAGCACACGTCTGAACTCCAGTCACAyyyyyyTCTCGTATGCCG TCTTCTGCTTG [Btn], where yyyyyy represented the index barcode (Figure 25 and Tables 46- 48).

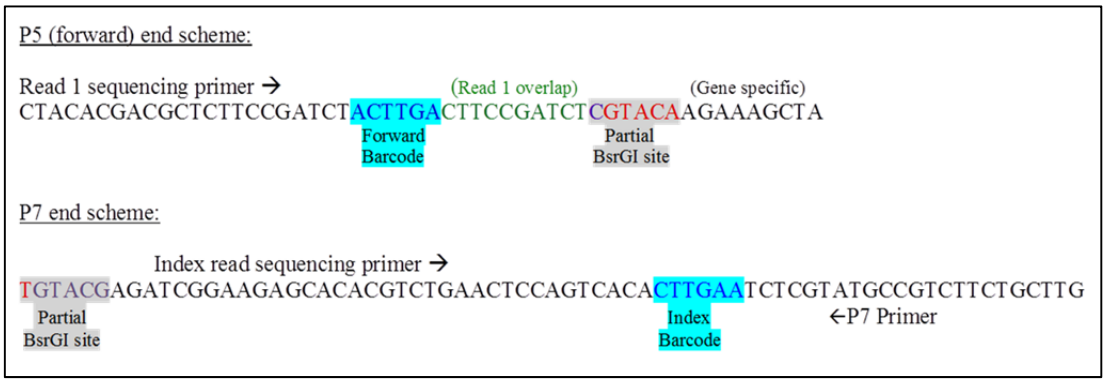


Figure 25 P5 and P7 Adaptor Schema

Table 46 P7 End Adaptors

P7 End Adapters	
UP7-IR-T-A	[BTN]CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTC
UP7-IR-T-B	[BTN]CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTC
UP7-IR-T-C	[BTN]CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTC
UP7-IR-T-D	[BTN]CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTC
UP7-IR-T-E	[BTN]CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTC
UP7-IR-T-F	[BTN]CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTC
UP7-IR-T-G	[BTN]CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTC
UP7-IR-T-H	[BTN]CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTC

Table 47 *BsrGI* P7-side Adapters for Index Read Multiplexing on Illumina (Hi-seq)

BP7 series	**BsrGI P7-side adapters for index read multiplexing on illumina (Hi-seq)**
BP7-IR-B-A	GTACGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG
BP7-IR-B-B	GTACGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG
BP7-IR-B-C	GTACGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG
BP7-IR-B-D	GTACGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG
BP7-IR-B-E	GTACGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG
BP7-IR-B-F	GTACGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG
BP7-IR-B-G	GTACGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG
BP7-IR-B-H	GTACGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTG

Table 48 P5 Adapter Forward Barcodes

P5 Adapter Forward Barcodes		TruSeq	
2165-8	CTACACGACGCTCTTCCGATCTACTTGACTTCCGATCTCGTACAAGAAAGCTA	ACTTGA	8
2165-9	CTACACGACGCTCTTCCGATCTGATCAGCTTCCGATCTCGTACAAGAAAGCTA	GATCAG	9
2165-10	CTACACGACGCTCTTCCGATCTTAGCTTCTTCCGATCTCGTACAAGAAAGCTA	TAGCTT	10
2165-11	CTACACGACGCTCTTCCGATCTGGCTACCTTCCGATCTCGTACAAGAAAGCTA	GGCTAC	11
2165-12	CCTACACGACGCTCTTCCGATCTCTTGTACTTCCGATCTCGTACAAGAAAGCTA	CTTGTA	12
3366-8	CTACACGACGCTCTTCCGATCTACTTGACGTACAAATGCCTTCGRGACTTAAC	ACTTGA	8
3366-9	CTACACGACGCTCTTCCGATCTGATCAGCGTACAAATGCCTTCGRGACTTAAC	GATCAG	9
3366-10	CTACACGACGCTCTTCCGATCTTAGCTTCGTACAAATGCCTTCGRGACTTAAC	TAGCTT	10
3366-11	CTACACGACGCTCTTCCGATCTGGCTACCGTACAAATGCCTTCGRGACTTAAC	GGCTAC	11
3366-12	CTACACGACGCTCTTCCGATCTCTTGTACGTACAAATGCCTTCGRGACTTAAC	CTTGTA	12
2187-8	CTACACGACGCTCTTCCGATCTACTTGAGATCTCGTACAATTTACATCGCTAAATAATG	ACTTGA	8
2187-9	CTACACGACGCTCTTCCGATCTGATCAGGATCTCGTACAATTTACATCGCTAAATAATG	GATCAG	9
2187-10	CTACACGACGCTCTTCCGATCTTAGCTTGATCTCGTACAATTTACATCGCTAAATAATG	TAGCTT	10
2187-11	CTACACGACGCTCTTCCGATCTGGCTACGATCTCGTACAATTTACATCGCTAAATAATG	GGCTAC	11
2187-12	CTACACGACGCTCTTCCGATCTCTTGTAGATCTCGTACAATTTACATCGCTAAATAATG	CTTGTA	12
2193-8	CTACACGACGCTCTTCCGATCTACTTGATTCCGATCTCGTACAATGATTACCGA	ACTTGA	8
2193-9	CTACACGACGCTCTTCCGATCTGATCAGTTCCGATCTCGTACAATGATTACCGA	GATCAG	9
2193-10	CTACACGACGCTCTTCCGATCTTAGCTTTTCCGATCTCGTACAATGATTACCGA	TAGCTT	10
2193-11	CTACACGACGCTCTTCCGATCTGGCTACTTCCGATCTCGTACAATGATTACCGA	GGCTAC	11
2193-12	CTACACGACGCTCTTCCGATCTCTTGTATTCCGATCTCGTACAATGATTACCGA	CTTGTA	12
2194-8	CTACACGACGCTCTTCCGATCTACTTGACTTCCGATCTCGTACAATATTATCCC	ACTTGA	8
2194-9	CTACACGACGCTCTTCCGATCTGATCAGCTTCCGATCTCGTACAATATTATCCC	GATCAG	9
2194-10	CTACACGACGCTCTTCCGATCTTAGCTTCTTCCGATCTCGTACAATATTATCCC	TAGCTT	10
2194-11	CTACACGACGCTCTTCCGATCTGGCTACCTTCCGATCTCGTACAATATTATCCC	GGCTAC	11
2194-12	CTACACGACGCTCTTCCGATCTCTTGTACTTCCGATCTCGTACAATATTATCCC	CTTGTA	12
3240-8	CTACACGACGCTCTTCCGATCTACTTGAGCTCTTCCGATCTCGTACAATACC	ACTTGA	8
3240-9	CTACACGACGCTCTTCCGATCTGATCAGGCTCTTCCGATCTCGTACAATACC	GATCAG	9
3240-10	CTACACGACGCTCTTCCGATCTTAGCTTGCTCTTCCGATCTCGTACAATACC	TAGCTT	10
3240-11	CTACACGACGCTCTTCCGATCTGGCTACGCTCTTCCGATCTCGTACAATACC	GGCTAC	11
3240-12	CTACACGACGCTCTTCCGATCTCTTGTAGCTCTTCCGATCTCGTACAATACC	CTTGTA	12

Table 48 Continued.

3241-8	CTACACGACGCTCTTCCGATCTACTTGAGTACATTTTTTATCCAGAAACCTCTGACTG	ACTTGA	8
3241-9	CTACACGACGCTCTTCCGATCTGATCAGGTACATTTTTTATCCAGAAACCTCTGACTG	GATCAG	9
3241-10	CTACACGACGCTCTTCCGATCTTAGCTTGTACATTTTTTATCCAGAAACCTCTGACTG	TAGCTT	10
3241-11	CTACACGACGCTCTTCCGATCTGGCTACGTACATTTTTTATCCAGAAACCTCTGACTG	GGCTAC	11
3241-12	CTACACGACGCTCTTCCGATCTCTTGTAGTACATTTTTTATCCAGAAACCTCTGACTG	CTTGTA	12
4429-8	CTACACGACGCTCTTCCGATCTACTTGAGCACTTAGCACCATACATGTGACC	ACTTGA	8
4429-9	CTACACGACGCTCTTCCGATCTGATCAGGCACTTAGCACCATACATGTGACC	GATCAG	9
4429-10	CTACACGACGCTCTTCCGATCTTAGCTTGCACTTAGCACCATACATGTGACC	TAGCTT	10
4429-11	CTACACGACGCTCTTCCGATCTGGCTACGCACTTAGCACCATACATGTGACC	GGCTAC	11
4429-12	CTACACGACGCTCTTCCGATCTCTTGTAGCACTTAGCACCATACATGTGACC	CTTGTA	12
8037a-8	CTACACGACGCTCTTCCGATCTACTTGAGATCTCGTACATTTCCCAAAAGCTC	ACTTGA	8
8037a-9	CTACACGACGCTCTTCCGATCTGATCAGGATCTCGTACATTTCCCAAAAGCTC	GATCAG	9
8037a-10	CTACACGACGCTCTTCCGATCTTAGCTTGATCTCGTACATTTCCCAAAAGCTC	TAGCTT	10
8037a-11	CTACACGACGCTCTTCCGATCTGGCTACGATCTCGTACATTTCCCAAAAGCTC	GGCTAC	11
8037a-12	CTACACGACGCTCTTCCGATCTCTTGTAGATCTCGTACATTTCCCAAAAGCTC	CTTGTA	12

The PS-6xK46 DNA population was digested with *BsrGI* in the 96-well plate dilution format, as stated above. Once digested, an adaptor BP7A through BP7-H (skipping BP7-G) was ligated to the samples by corresponding row (Table 48). Next, the samples were pooled by column to create 5 pooled samples. These pooled samples were purified using a MinElute PCR Purification kit (Qiagen®) to remove primers and the digestion/ligation mix.

Primer mixes to multiplex out *BsrGI* with the additional specific Targeted GBS primer were developed. They were made into 5 primer mixes to correspond to ten

specific loci for each column. The primer mixes were diluted to 10 pmol/ul concentrations.

Using vent (exo-) polymerase, a primer extension was done to add a primer mix (8 through 12) and Tufts_PCR_Primer_2.1_B [Btn] (Tufts primer with biotin) to the PS-6 xK46 DNA population (Table 38). The vent polymerase PCR conditions consisted of 98ºC for one minute followed by 15 cycles of 98ºC for 30 seconds, 58ºC for 30 seconds, 72ºC for one minute, then with a final extension time of two minutes at 72ºC. All five vent samples were pooled together into one microfuge tube. The pooled vent reaction was then cleaned using a MinElute PCR Purification kit (Qiagen™).

In the effort to remove fragments without biotin after the vent PCR, the NEBNext® Fill-in and ssDNA Isolation Module (New England Biolabs) was used to remove the non-biotinylated fragments. The protocol was modified by skipping the fill-in reaction and the second 1X Bead Wash after the fill-in step. The sample was cleaned over a Qiaquick PCR Purification column (Qiagen®) and eluted in 25ul.

Final amplification for the TGBS samples using PE_PCR_Primer_1.2 and Tufts_PCR_Primer_2.1 was done to amplify the genetic library fragments (Table 49-50). The PCR conditions consisted of 98ºC for 30 seconds followed by 15 cycles of 98ºC for 12 seconds, 65ºC for 30 seconds, 72ºC for 30 seconds, then with a final extension time of one minute at 72ºC.

Table 49 Tufts Paired End PCR Primer 2.1

Tufts_PCR_Primer_2.1
5'- CAAGCAGAAGACGGCATACGAG

Table 50 Paired End PCR Primer 1.2

PE_PCR_Primer_1.2
5'- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAT

The library sample was then purified using the AMPure XP Tube (Agencourt®) with a 0.7ul sample reaction: 1ul bead ratio. The sample was eluted with 50 ul of elution buffer. The sample was quantified to ~2ng double-stranded DNA with the High Sensitivity AccuBlue Quantification Kit (Biotium™) on the Victor™ *X3* Multilabel Plate Reader (PerkinElmer™). A quality control check was done on the Bio-Analyzer (Agilent™) with a pico-chip to make sure the fragment was good quality DNA and correct size selection. The TGBS samples were then diluted to a 5nM concentration to run on the Illumina HiSeq 2500. The 5nM sample was given to the TAMU AgriLife Genomics & Bioinformatics Services to run one lane on the Illumina HiSeq 2500.

From the TAMU AgriLife Genomics & Bioinformatics Services, data files sorted by reverse barcode were received. The .fastq files were uploaded into Geneious. Next, the Illumina data was sorted by our forward. The next step was to identify the individual sequences for PS-6 and K-46, which were the parents to the segregating population for photoperiodism.

An alignment of the original loci from PS-5 and the wild accessions (PI-435242 and K-56) to *Gossypium barbadense* 3-79 was done to see if each fragment had multiple loci. The new 3-79 consensus sequences for each loci was then blasted against the PS-6 and K-46 sequences. After obtaining the closest sequence to each locus in PS-6 and K-46, a reference sequence set was made for both PS-6 and K-46. The other individuals in the segregating population were then mapped to these references.

CHAPTER V

CONCLUSION

World-wide modern cultivated cotton production has been limited by the current genetic diversity within cotton's elite cultivars, so there has been a strong need to develop practical traits from 'wild' relatives. To increase diversity, valuable assets from untapped 'wild' genetic resources were found in other studies and should be incorporated into traditional breeding programs [6].

Currently, there has been a problem with incorporating this myriad of untapped 'wild' genomic resources through traditional breeding schemas. The major problem has been that 'wild' cotton has been hampered by photoperiod sensitivity [11]. In short, this means that most commercial cotton producing areas in the world do not have the correct natural light conditions to allow 'wild' cotton species to flower in the span of a growing season under today's current cultivation practices. On the other hand, modern cultivated cotton was able to flourish all over the world because it had the ability to establish flowering under early maturation.

Wild conversion trait introgression has been difficult because 'wild' and cultivated cotton species do not flower during overlapping times. Crosses between 'wild' and cultivated cotton species were done in the past, but always resulted in offspring exhibiting photoperiod sensitivity. This rendered these progeny useless for commercial production.

The final problem that this research addressed was the length of time to bring in valuable traits from 'wild' Germplasm. In traditional breeding programs, this trait introgression process would take ten to twenty years. One way to alleviate this long and arduous process has been to select only those plants carrying the trait of interest after an F1 cross, without bringing photoperiodic sensitivity by utilizing marker assisted selection (MAS) breeding. Our research has provided markers associated with photoperiod independence.

This study asked what fundamental research was needed before novel traits from undomesticated 'wild' cotton species could be integrated into cultivated elite lines. We researched the floral transition network within cotton to overcome the major hurdle, photoperiod independence, for trait introgression from 'wild' cotton species. With limited knowledge of information existing within the floral transition network within cotton, we conducted an experiment to look at possible candidate genes for photoperiod independence. A second experiment looked for independent single nucleotide polymorphisms (SNPs) outside the candidate genes that were associated with photoperiod independence.

The candidate gene study reported SNP differences in thirty-eight homologs of genes within the floral transition network, including photoreceptors, light dependent transcripts, circadian clock regulators, and floral integrators. We uncovered appreciable SNP diversity within the candidate gene orthologs, including SNPs differentiating cultivated and 'wild' *Gossypium barbadense* and *Gossypium hirsutum*. We located 36 intraspecific SNPs within *Gossypium hirsutum* and 53 intraspecific SNPs within

Gossypium barbadense. From this research, we laid the foundation to test our intraspecific SNPs in our BC4F2 segregating 'wild' cotton conversion population.

The Genotype-By-Sequencing (GBS) research used redundancy measures to link other discrete SNP differences (unassociated with candidate genes) to photoperiodicity within cotton. This GBS study found ten overlapping loci containing intraspecific SNPs from two BC4F5 'wild' cotton conversion populations that had full introgression of photoperiod independence. From these loci, targeted GBS (TGBS) was conducted on a segregating BC4F2 'wild' cotton conversion population to show linkage of these markers with photoperiod independence. Our GBS study provided one closely linked marker to the photoperiod independence trait.

In conclusion, our research has provided markers associated with photoperiod independence. Future MAS breeding programs may incorporate our intraspecific candidate gene SNPs and our discrete GBS SNPs as markers for 'wild' trait introgression. Also, other research scientists may utilize our markers to explore the molecular evolution of different cotton genomes. From this research, our future goal will be to focus on those candidate genes near our discrete GBS SNPs associated with photoperiod independence to locate the polymorphism(s) behind photoperiod independence in *Gossypium barbadense*.

REFERENCES

1. Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, Chen X, Stelly DM, Rabinowicz PD, Town CD *et al*: **Toward Sequencing Cotton (*Gossypium*) Genomes**. *Plant Physiology* 2007, **145**(4):1303-1310.

2. Zheng J-l, Yi W-m, Wang N-n: **Bio-Oil Production from Cotton Stalk**. *Energy Conversion and Management* 2008, **49**(6):1724-1730.

3. Liu Q, Singh S, Green A: **Genetic Modification of Cotton Seed Oil Using Inverted-Repeat Gene-Silencing Techniques**. *Biochemical Society Transactions* 2000, **28**(927):927-929.

4. Mohamed OE, Satter LD, Grummer RR, Ehle FR: **Influence of Dietary Cottonseed and Soybean on Milk Production and Composition**. *Journal of Dairy Science* 1988, **71**(10):2677-2688.

5. Wendel JF, Cronn RC: **Polyploidy and the Evolutionary History of Cotton**. In: *Advances in Agronomy.* Edited by Neale DB, vol. 78. New York, NY: Academic Press; 2003: 139-186.

6. McCarty Jr JC, Percy RG: **Genes from Exotic Germplasm and Their Use in Cultivar Improvement in *Gossypium hirsutum* L. and *G. barbadense* L**. In: *Genetic Improvement of Cotton: Emerging Technologies.* Edited by Jenkins J, Saha S. Enfield, NH: Science Publishers; 2001: 65-80.

7. Xu Y, Crouch JH: **Marker-Assisted Selection in Plant Breeding: From Publications to Practice** *Crop Science* 2008, **48**(2):391-407.

8. Collard BCY, Mackill DJ: **Marker-Assisted Selection: An Approach for Precision Plant Breeding in the Twenty-First Century**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2008, **363**(1491):557-572.

9. Mohan M, Nair S, Bhagwat A, Krishna TG, Yano M, Bhatia CR, Sasaki T: **Genome Mapping, Molecular Markers and Marker-Assisted Selection in Crop Plants**. *Molecular Breeding* 1997, **3**(2):87-103.

10. Paterson AH, Tanksley SD, Sorrells ME: **DNA Markers in Plant Improvement**. In: *Advances in Agronomy.* Edited by Sparks DL, vol. 46. New York, NY: Academic Press; 1991: 39-90.

11. Kohel RJ, Richmond TR: **The Genetics of Flowering Response in Cotton. IV. Quantitative Analysis of Photoperiodism of Texas 86, *Gossypium hirsutum* Race latifolium, in a Cross with an Inbred Line of Cultivated American Upland Cotton**. *Genetics* 1962, **47**(11):1535-1542.

12. Blackman BK, Rasmussen DA, Strasburg JL, Raduski AR, Burke JM, Knapp SJ, Michaels SD, Rieseberg LH: **Contributions of Flowering Time Genes to Sunflower Domestication and Improvement**. *Genetics* 2011, **187**(1):271-287.

13. Ross-Ibarra J, Morrell PL, Gaut BS: **Plant Domestication, a Unique Opportunity to Identify the Genetic Basis of Adaptation**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(1):8641-8648.

14. Burger JC, Chapman MA, Burke JM: **Molecular Insights into the Evolution of Crop Plants**. *American Journal of Botany* 2008, **95**(2):113-122.

15. Doebley JF, Gaut BS, Smith BD: **The Molecular Genetics of Crop Domestication**. *Cell* 2006, **127**(7):1309-1321.

16. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez G. J, Buckler E, Doebley J: **A Single Domestication for Maize shown by Multilocus Microsatellite Genotyping**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(9):6080-6084.

17. Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J: **Identifying Genes of Agronomic Importance in Maize by Screening Microsatellites for Evidence of Selection During Domestication**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(15):9650-9655.

18. Dubcovsky J, Dvorak J: **Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication**. *Science* 2007, **316**(5833):1862-1866.

19. Luo MC, Yang ZL, You FM, Kawahara T, Waines JG, Dvorak J: **The Structure of Wild and Domesticated Emmer Wheat Populations, Gene Flow between Them, and the Site of Emmer Domestication**. *TAG Theoretical and Applied Genetics* 2007, **114**(6):947-959.

20. Ozkan H, Brandolini A, Pozzi C, Effgen S, Wunder J, Salamini F: **A Reconsideration of the Domestication Geography of Tetraploid Wheats**. *TAG Theoretical and Applied Genetics* 2005, **110**(6):1052-1060.

21. Simons KJ, Fellers JP, Trick HN, Zhang Z, Tai Y-S, Gill BS, Faris JD: **Molecular Characterization of the Major Wheat Domestication Gene Q**. *Genetics* 2006, **172**(1):547-555.

22. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L *et al*: **Resequencing 50 Accessions of Cultivated and Wild Rice Yields Markers for Identifying Agronomically Important Genes**. *Nature Biotechnology* 2012, **30**(1):105-111.

23. He Z, Zhai W, Wen H, Tang T, Wang Y, Lu X, Greenberg AJ, Hudson RR, Wu C-I, Shi S: **Two Evolutionary Histories in the Genome of Rice: The Roles of Domestication Genes**. *PLoS Genetics* 2011, **7**(6):1-10.

24. Yang Y, Peng Q, Chen G-X, Li X-H, Wu C-Y: ***OsELF3* is Involved in Circadian Clock Regulation for Promoting Flowering under Long-Day Conditions in Rice**. *Molecular Plant* 2012, **6**:202-215.

25. Gao L-z, Innan H: **Nonindependent Domestication of the Two Rice Subspecies, *Oryza sativa ssp. indica* and *ssp. japonica*, Demonstrated by Multilocus Microsatellites**. *Genetics* 2008, **179**(2):965-976.

26. Londo JP, Chiang Y-C, Hung K-H, Chiang T-Y, Schaal BA: **Phylogeography of Asian Wild Rice, *Oryza rufipogon*, Reveals Multiple Independent Domestications of Cultivated Rice, *Oryza sativa***. *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(25):9578-9583.

27. Brubaker CL, Wendel JF: **Reevaluating the Origin of Domesticated Cotton (*Gossypium hirsutum*; Malvaceae) Using Nuclear Restriction Fragment Length Polymorphisms (RFLPs)**. *American Journal of Botany* 1994, **81**(10):1309-1326.

28. Brubaker CL, Paterson AH, Wendel JF: **Comparative Genetic Mapping of Allotetraploid Cotton and its Diploid Progenitors**. *Genome* 1999, **42**(2):184-203.

29. Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF: **Polyploid Formation in Cotton is Not Accompanied by Rapid Genomic Changes**. *Genome* 2001, **44**(3):321-330.

30. Hutchinson JB: **Intra-Specific Differentiation in *Gossypium hirsutum***. *Heredity* 1951, **5**(2):161-193.

31. Hutchinson JB: **New Evidence on the Origin of the Old World Cottons**. *Heredity* 1954, **8**(2):225-241.

32. Hutchinson J: **The Application of Genetics to Cotton Improvement**. London, UK: Cambridge University Press, London; 1959.

33. Hutchinson JB, Silow R, Stephens S: **The Evolution of Gossypium and the Differentiation of the Cultivated Cottons**. London, UK: Empire Cotton Growing Corporation. Oxford University Press, London.; 1947.

34. Wendel JF, Percy RG: **Allozyme Diversity and Introgression in the Galapagos Islands Endemic *Gossypium darwinii* and Its Relationship to Continental *G. barbadense***. *Biochemical Systematics and Ecology* 1990, **18**(7–8):517-528.

35. Wendel JF, Brubaker CL, Percival AE: **Genetic Diversity in *Gossypium hirsutum* and the Origin of Upland Cotton**. *American Journal of Botany* 1992, **79**(11):1291-1310.

36. Wendel JF, Olson PD, Stewart JM: **Genetic Diversity, Introgression, and Independent Domestication of Old World Cultivated Cottons**. *American Journal of Botany* 1989, **76**(12):1795-1806.

37. Beasley J: **Meiotic Chromosome Behavior in Species, Species Hybrids, Haploids, and Induced Polyploids of *Gossypium***. *Genetics* 1942, **27**(1):25-54.

38. Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF: **Rate Variation Among Nuclear Genes and the Age of Polyploidy in *Gossypium***. *Molecular Biology and Evolution* 2003, **20**(4):633-643.

39. Lee JJ, Woodward AW, Chen ZJ: **Gene Expression Changes and Early Events in Cotton Fibre Development**. *Annals of Botany* 2007, **100**(7):1391-1401.

40. Endrizzi JE, Turcotte EL, Kohel RJ: **Genetics, Cytology, and Evolution of *Gossypium***. In: *Advances in Genetics*. Edited by Caspari EW, Scandalios JG, vol. 23. New York, NY: Academic Press; 1985: 271-375.

41. Wright SI, Gaut BS: **Molecular Population Genetics and the Search for Adaptive Evolution in Plants**. *Molecular Biology and Evolution* 2005, **22**(3):506-519.

42. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS: **The Effects of Artificial Selection on the Maize Genome**. *Science* 2005, **308**(5726):1310-1314.

43. Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH, Aquadro CF, Kresovich S: **Diversity and Selection in Sorghum: Simultaneous Analyses Using Simple Sequence Repeats**. *TAG Theoretical and Applied Genetics* 2005, **111**(1):23-30.

44. Hamblin MT, Casa AM, Sun H, Murray SC, Paterson AH, Aquadro CF, Kresovich S: **Challenges of Detecting Directional Selection After a Bottleneck: Lessons From *Sorghum bicolor***. *Genetics* 2006, **173**(2):953-964.

45. Michaels SD: **Flowering Time Regulation Produces Much Fruit**. *Current Opinion in Plant Biology* 2009, **12**(1):75-80.

46. APG: **An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II**. *Botanical Journal of the Linnean Society* 2003, **141**(4):399.

47. Chen H, Huang X, Gusmaroli G, Terzaghi W, Lau OS, Yanagawa Y, Zhang Y, Li J, Lee J-H, Zhu D *et al*: **Arabidopsis *Cullin 4-Damaged DNA Binding Protein 1* Interacts with *Constitutively Photomorphogenic 1-Suppressor of PHYA* Complexes to Regulate Photomorphogenesis and Flowering Time**. *The Plant Cell Online* 2010, **22**(1):108-123.

48. McCarty JC, Jenkins JN, Wu J: **Primitive Accession Derived Germplasm by Cultivar Crosses as Sources for Cotton Improvement: I. Phenotypic Values and Variance Components**. *Crop Science* 2004, **44**(4):1226-1230.

49. McCarty JC, Wu J, Jenkins JN: **Genetic Diversity for Agronomic and Fiber Traits in Day-Neutral Accessions Derived from Primitive Cotton Germplasm**. *Euphytica* 2006, **148**(3):283-293.

50. Mauseth JD: **Botany : An Introduction to Plant Biology**, 3rd ed edn. Sudbury, MA: Jones and Bartlett Learning; 2003.

51. Khan M, Khan I, Ali G: ***MPF2*-Like MADS-Box Genes Affecting Expression of *SOC1* and *MAF1* are Recruited to Control Flowering Time**. *Molecular Biotechnology* 2012, **54**(1):1-12.

52. Bünning E: **Circadian Rhythms and the Time Measurement in Photoperiodism**. In: *Cold Spring Harbor Symposia on Quantitative Biology: 1960 1960; Cold Springs Harbor, New York*: Cold Spring Harbor Laboratory Press; 1960: 249-256.

53. Pittendrigh CS, Minis DH: **The Entrainment of Circadian Oscillations by Light and Their Role as Photoperiodic Clocks**. *The American Naturalist* 1964, **98**(902):261-294.

54. Hayama R, Coupland G: **The Molecular Basis of Diversity in the Photoperiodic Flowering Responses of Arabidopsis and Rice**. *Plant Physiology* 2004, **135**(2):677-684.

55. Tournois J: **Influence de la Lumière sur la Floraison du Houblon Japonais et du Chanvre Déterminées par des Semis Haitifs**. *Comptes Rendus de l'Académie des Sciences* 1912, **155**:297-300.

56. Klebs AC: **The Historic Evolution of Variolation**. Baltimore, MD: Johns Hopkins Hospital; 1913.

57. Garner WW, Allard HA: **Effect of the Relative Length of Day and Night and Other Factors of the Environment on Growth and Reproduction in Plants**. *Monthly Weather Review* 1920, **48**:415-415.

58. Garner WW, Allard HA: **Further Studies in Photoperiodism: The Response of the Plant to Relative Length of Day and Night**. *Journal of Agricultural Research* 1923, **23**(11):1-88.

59. Rizzini L, Favory J-J, Cloix C, Faggionato D, O'Hara A, Kaiserli E, Baumeister R, Schafer E, Nagy F, Jenkins GI *et al*: **Perception of UV-B by the Arabidopsis *UVR8* Protein**. *Science* 2011, **332**(6025):103-106.

60. Quail PH: **An Emerging Molecular Map of the Phytochromes**. *Plant, Cell & Environment* 1997, **20**(6):657-665.

61. Li J, Li G, Wang H, Wang Deng X: **Phytochrome Signaling Mechanisms**. *The Arabidopsis Book* 2011, **148**:1-27.

62. Yu X, Liu H, Klejnot J, Lin C: **The Cryptochrome Blue Light Receptors**. *The Arabidopsis Book* 2010, **135**:1-27.

63. Clack T, Mathews S, Sharrock R: **The Phytochrome Apoprotein Family in Arabidopsis is Encoded by Five Genes: The Sequences and Expression of *PHYD* and *PHYE***. *Plant Molecular Biology* 1994, **25**(3):413-427.

64. Cowl J, Hartley N, Xie D, Whitelam G, Murphy G, Harberd N: **The *PHYC* Gene of Arabidopsis: Absence of the Third Intron Found in *PHYA* and *PHYB***. *Plant Physiology* 1994, **106**:813-814.

65. Somers DE, Devlin PF, Kay SA: **Phytochromes and Cryptochromes in the Entrainment of the Arabidopsis Circadian Clock**. *Science* 1998, **282**(5393):1488-1490.

66. Devlin PF, Kay SA: **Cryptochromes Are Required for Phytochrome Signaling to the Circadian Clock but Not for Rhythmicity**. *The Plant Cell Online* 2000, **12**(12):2499-2509.

67. Valverde F, Mouradov A, Soppe W, Ravenscroft D, Samach A, Coupland G: **Photoreceptor Regulation of *Constans* Protein in Photoperiodic Flowering**. *Science* 2004, **303**(5660):1003-1006.

68. Reed JW, Nagatani A, Elich TD, Fagan M, Chory J: ***Phytochrome A* and *Phytochrome B* Have Overlapping but Distinct Functions in Arabidopsis Development**. *Plant Physiology* 1994, **104**(4):1139-1149.

69. Devlin PF, Robson PRH, Patel SR, Goosey L, Sharrock RA, Whitelam GC: ***Phytochrome D* Acts in the Shade-Avoidance Syndrome in Arabidopsis by Controlling Elongation Growth and Flowering Time**. *Plant Physiology* 1999, **119**(3):909-916.

70. Franklin KA, Praekelt U, Stoddart WM, Billingham OE, Halliday KJ, Whitelam GC: ***Phytochromes B, D, and E* Act Redundantly to Control Multiple Physiological Responses in Arabidopsis**. *Plant Physiology* 2003, **131**(3):1340-1346.

71. Halliday KJ, Salter MG, Thingnaes E, Whitelam GC: **Phytochrome Control of Flowering is Temperature Sensitive and Correlates with Expression of the Floral Integrator *FT***. *The Plant Journal* 2003, **33**(5):875-885.

72. El-Din El-Assal S, Alonso-Blanco C, Peeters AJM, Raz V, Koornneef M: **A QTL for Flowering Time in Arabidopsis Reveals a Novel Allele of *CRY2***. *Nature Genetics* 2001, **29**(4):435-440.

73. Balasubramanian S, Sureshkumar S, Agrawal M, Michael TP, Wessinger C, Maloof JN, Clark R, Warthmann N, Chory J, Weigel D: **The *Phytochrome C* Photoreceptor Gene Mediates Natural Variation in Flowering and Growth Responses of *Arabidopsis thaliana***. *Nature Genetics* 2006, **38**(6):711-715.

74. Filiault DL, Wessinger CA, Dinneny JR, Lutes J, Borevitz JO, Weigel D, Chory J, Maloof JN: **Amino Acid Polymorphisms in Arabidopsis *Phytochrome B* Cause Differential Responses to Light**. *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(8):3157-3162.

75. Tsuji H, Taoka K-i, Shimamoto K: **Regulation of Flowering in Rice: Two Florigen Genes, a Complex Gene Network, and Natural Variation**. *Current Opinion in Plant Biology* 2011, **14**(1):45-52.

76. Izawa T, Takahashi Y, Yano M: **Comparative Biology Comes into Bloom: Genomic and Genetic Comparison of Flowering Pathways in Rice and Arabidopsis**. *Current Opinion in Plant Biology* 2003, **6**(2):113-120.

77. Hayama R, Yokoi S, Tamaki S, Yano M, Shimamoto K: **Adaptation of Photoperiodic Control Pathways Produces Short-Day Flowering in Rice**. *Nature* 2003, **422**(6933):719-722.

78. Doi K, Izawa T, Fuse T, Yamanouchi U, Kubo T, Shimatani Z, Yano M, Yoshimura A: ***Ehd1*, a B-Type Response Regulator in Rice, Confers Short-Day Promotion of Flowering and Controls *FT-Like* Gene expression Independently of *Hd1***. *Genes & Development* 2004, **18**(8):926-936.

79. Izawa T, Oikawa T, Sugiyama N, Tanisaka T, Yano M, Shimamoto K: **Phytochrome Mediates the External Light Signal to Repress *FT* Orthologs in Photoperiodic Flowering of Rice**. *Genes & Development* 2002, **16**(15):2006-2020.

80. Yano M, Kojima S, Takahashi Y, Lin H, Sasaki T: **Genetic Control of Flowering Time in Rice, a Short-Day Plant**. *Plant Physiology* 2001, **127**(4):1425-1429.

81. Ishikawa R, Tamaki S, Yokoi S, Inagaki N, Shinomura T, Takano M, Shimamoto K: **Suppression of the Floral Activator *Hd3a* Is the Principal Cause of the Night Break Effect in Rice**. *The Plant Cell Online* 2005, **17**(12):3326-3336.

82. Ishikawa R, Shinomura T, Takano M, Shimamoto K: **Phytochrome Dependent Quantitative Control of *Hd3a* Transcription is the Basis of the Night Break Effect in Rice Flowering**. *Genes & Genetic Systems* 2009, **84**(2):179-184.

83. Ishikawa R, Aoki M, Kurotani K-i, Yokoi S, Shinomura T, Takano M, Shimamoto K: ***Phytochrome B* Regulates *Heading date 1 (Hd1)*-Mediated Expression of Rice Florigen *Hd3a* and Critical Day Length in Rice**. *Mol Genet Genomics* 2011, **285**(6):461-470.

84. Takano M, Inagaki N, Xie X, Kiyota S, Baba-Kasai A, Tanabata T, Shinomura T: **Phytochromes are the Sole Photoreceptors for Perceiving Red/Far-Red Light in Rice**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(34):14705-14710.

85. Tamaki S, Matsuo S, Wong HL, Yokoi S, Shimamoto K: ***Hd3a* Protein Is a Mobile Flowering Signal in Rice**. *Science* 2007, **316**(5827):1033-1036.

86. Shalitin D, Yu X, Maymon M, Mockler T, Lin C: **Blue Light–Dependent in Vivo and in Vitro Phosphorylation of Arabidopsis *Cryptochrome 1***. *The Plant Cell Online* 2003, **15**(10):2421-2429.

87. Mizoguchi T, Wright L, Fujiwara S, Cremer F, Lee K, Onouchi H, Mouradov A, Fowler S, Kamada H, Putterill J *et al*: **Distinct Roles of *Gigantea* in Promoting Flowering and Regulating Circadian Rhythms in Arabidopsis**. *The Plant Cell Online* 2005, **17**(8):2255-2270.

88. Koornneef M, Alonso-Blanco C, Peeters AJM, Soppe W: **Genetic Control of Flowering time in Arabidopsis**. *Annual Review of Plant Biology* 1998, **49**(1):345-370.

89. Jang S, Marchal V, Panigrahi KC, Wenkel S, Soppe W, Deng XW, Valverde F, Coupland G: **Arabidopsis *COP1* Shapes the Temporal Pattern of *CO* Accumulation Conferring a Photoperiodic Flowering Response**. *EMBO Journal* 2008, **27**(8):1277-1288.

90. Ishikawa M, Kiba T, Chua NH: **The Arabidopsis *SPA1* Gene is Required for Circadian Clock Function and Photoperiodic Flowering**. *The Plant Journal* 2006, **46**(5):736-746.

91. Phee B-K, Kim J-I, Shin DH, Yoo J, Park K-J, Han Y-J, Kwon Y-K, Cho M-H, Jeon J-S, Bhoo SH *et al*: **A Novel Protein Phosphatase Indirectly Regulates *Phytochrome-Interacting Factor 3* Via Phytochrome**. *Biochemical Journal* 2008, **415**(2):247-255.

92. Yu J-W, Rubio V, Lee N-Y, Bai S, Lee S-Y, Kim S-S, Liu L, Zhang Y, Irigoyen ML, Sullivan JA *et al*: ***COP1* and *ELF3* Control Circadian Function and Photoperiodic Flowering by Regulating GI Stability**. *Molecular Cell* 2008, **32**(5):617-630.

93. Goto N, Katoh N, Kranz AR: **Morphogenesis of Floral Organs in Arabidopsis*:* Predominant Carpel Formation of the Pin-Formed Mutant**. *The Japanese Journal of Genetics* 1991, **66**(5):551-567.

94. Mozley D, Thomas B: **Developmental and Photobiological Factors Affecting Photoperiodic Induction in *Arabidopsis thaliana* Heynh. Landsberg erecta**. *Journal of Experimental Botany* 1995, **46**(2):173-179.

95. Bagnall D, King R, Hangarter R: **Blue-Light Promotion of Flowering is Absent in *HY4* Mutants of Arabidopsis**. *Planta* 1996, **200**(2):278-280.

96. Bruggemann E, Handwerger K, Essex C, Storz G: **Analysis of Fast Neutron-Generated Mutants at the *Arabidopsis thaliana HY4* Locus**. *The Plant Journal* 1996, **10**(4):755-760.

97. Blazquez MA, Ahn JH, Weigel D: **A Thermosensory Pathway Controlling Flowering Time in *Arabidopsis thaliana***. *Nature Genetics* 2003, **33**(2):168-171.

98. Guo H, Yang H, Mockler TC, Lin C: **Regulation of Flowering Time by Arabidopsis Photoreceptors**. *Science* 1998, **279**(5355):1360-1363.

99. Liu H, Yu X, Li K, Klejnot J, Yang H, Lisiero D, Lin C: **Photoexcited *CRY2* Interacts with *CIB1* to Regulate Transcription and Floral Initiation in Arabidopsis**. *Science* 2008, **322**:1535-1539.

100. Liu H, Liu B, Zhao C, Pepper M, Lin C: **The Action Mechanisms of Plant Cryptochromes**. *Trends in Plant Science* 2011, **16**(12):684-691.

101. Jung J-H, Seo Y-H, Seo PJ, Reyes JL, Yun J, Chua N-H, Park C-M: **The *Gigantea*-Regulated *MicroRNA172* Mediates Photoperiodic Flowering Independent of *Constans* in Arabidopsis**. *The Plant Cell Online* 2007, **19**(9):2736-2748.

102. Endo M, Nakamura S, Araki T, Mochizuki N, Nagatani A: ***Phytochrome B* in the Mesophyll Delays Flowering by Suppressing *Flowering Locus T* Expression in Arabidopsis Vascular Bundles**. *The Plant Cell Online* 2005, **17**(7):1941-1952.

103. Franklin KA, Davis SJ, Stoddart WM, Vierstra RD, Whitelam GC: **Mutant Analyses Define Multiple Roles for *Phytochrome C* in Arabidopsis Photomorphogenesis**. *The Plant Cell Online* 2003, **15**(9):1981-1989.

104. Monte E, Alonso JM, Ecker JR, Zhang Y, Li X, Young J, Austin-Phillips S, Quail PH: **Isolation and Characterization of *PHYC* Mutants in Arabidopsis Reveals Complex Crosstalk between Phytochrome Signaling Pathways**. *The Plant Cell Online* 2003, **15**(9):1962-1980.

105. Takano M, Inagaki N, Xie X, Yuzurihara N, Hihara F, Ishizuka T, Yano M, Nishimura M, Miyao A, Hirochika H *et al*: **Distinct and Cooperative Functions of *Phytochromes A, B, and C* in the Control of De-etiolation and Flowering in Rice**. *The Plant Cell Online* 2005, **17**(12):3311-3325.

106. Samis KE, Heath KD, Stinchcombe JR, Rausher M: **Discordant Longitudinal Clines in Flowering Time and *Phytochrome C* in *Arabidopsis thaliana***. *Evolution* 2008, **62**(12):2971-2983.

107. Saïdou A-A, Mariac C, Luong V, Pham J-L, Bezançon G, Vigouroux Y: **Association Studies Identify Natural Variation at *PHYC* Linked to Flowering Time and Morphological Variation in Pearl Millet**. *Genetics* 2009, **182**(3):899-910.

108. Ouedraogo M, Hubac C: **Effect of Far Red Light on Drought Resistance of Cotton**. *Plant and Cell Physiology* 1982, **23**(7):1297-1303.

109. Singh G, Garg OP: **Effect of Red, Far-Red Radiations on Germination of Cotton Seed**. *Plant and Cell Physiology* 1971, **12**(3):411-415.

110. Kasperbauer MJ: **Cotton Plant Size and Fiber Developmental Responses to FR/R Ratio Reflected from the Soil Surface**. *Physiologia Plantarum* 1994, **91**(2):317-321.

111. Kasperbauer MJ: **Cotton Fiber Length Is Affected by Far-Red Light Impinging on Developing Bolls Mention of a Trademark or Product Does Not Constitute a Guarantee or Warranty of the Product by the USDA and Does Not Imply Its Approval to the Exclusion of Other Products or Vendors that may also be Suitable**. *Crop Science* 2000, **40**(6):1673-1678.

112. Childs KL, Miller FR, Cordonnier-Pratt M-M, Pratt LH, Morgan PW, Mullet JE: **The Sorghum Photoperiod Sensitivity Gene, *Ma3*, Encodes a *Phytochrome B*[1]**. *Plant Physiology* 1997, **113**(2):611-619.

113. Hanumappa M, Pratt LH, Cordonnier-Pratt M-M, Deitzer GF: **A Photoperiod-Insensitive Barley Line Contains a Light-Labile *Phytochrome B***. *Plant Physiology* 1999, **119**(3):1033-1040.

114. Izawa T, Oikawa T, Tokutomi S, Okuno K, Shimamoto K: **Phytochromes Confer the Photoperiodic Control of Flowering in Rice (a Short-Day Plant)**. *The Plant Journal* 2000, **22**(5):391-399.

115. Liu B, Kanazawa A, Matsumura H, Takahashi R, Harada K, Abe J: **Genetic Redundancy in Soybean Photoresponses Associated With Duplication of the Phytochrome A Gene**. *Genetics* 2008, **180**(2):995-1007.

116. Devlin PF, Kay SA: **Circadian Photoperception**. *Annual Review Physiolology* 2001, **63**(1):677-694.

117. Cashmore AR: **Cryptochromes: Enabling Plants and Animals to Determine Circadian Time**. *Cell* 2003, **114**(5):537-543.

118. Sancar A: **Structure and Function of DNA Photolyase and Cryptochrome Blue-Light Photoreceptors**. *Chemical Reviews-Columbus* 2003, **103**(6):2203-2238.

119. Bell-Pedersen D, Cassone VM, Earnest DJ, Golden SS, Hardin PE, Thomas TL, Zoran MJ: **Circadian Rhythms From Multiple Oscillators: Lessons From Diverse Organisms**. *Nature Reviews Genetics* 2005, **6**(7):544-556.

120. Johnson C, Knight M, Kondo T, Masson P, Sedbrook J, Haley A, Trewavas A: **Circadian Oscillations of Cytosolic and Chloroplastic Free Calcium in Plants**. *Science* 1995, **269**(5232):1863-1865.

121. Doyle MR, Davis SJ, Bastow RM, McWatters HG, Kozma-Bognar L, Nagy F, Millar AJ, Amasino RM: **The *ELF4* Gene Controls Circadian Rhythms and Flowering Time in *Arabidopsis thaliana***. *Nature* 2002, **419**(6902):74-77.

122. Lu SX, Webb CJ, Knowles SM, Kim SHJ, Wang Z, Tobin EM: ***CCA1* and *ELF3* Interact in the Control of Hypocotyl Length and Flowering Time in Arabidopsis**. *Plant Physiology* 2012, **158**(2):1079-1088.

123. Gendron JM, Pruneda-Paz JL, Doherty CJ, Gross AM, Kang SE, Kay SA: **Arabidopsis Circadian Clock Protein, *TOC1*, is a DNA-Binding Transcription Factor**. *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(8):3167-3172.

124. Pruneda-Paz JL, Kay SA: **An Expanding Universe of Circadian Networks in Higher Plants**. *Trends in Plant Science* 2010, **15**(5):259-265.

125. Lau On S, Huang X, Charron J-B, Lee J-H, Li G, Deng Xing W: **Interaction of Arabidopsis *DET1* with *CCA1* and *LHY* in Mediating Transcriptional Repression in the Plant Circadian Clock**. *Molecular Cell* 2011, **43**(5):703-712.

126. Helfer A, Nusinow DA, Chow BY, Gehrke AR, Bulyk ML, Kay SA: ***Lux Arrhythmo* Encodes a Nighttime Repressor of Circadian Gene Expression in the Arabidopsis Core Clock**. *Current Biology* 2011, **21**(2):126-133.

127. Nusinow DA, Helfer A, Hamilton EE, King JJ, Imaizumi T, Schultz TF, Farre EM, Kay SA: **The *ELF4-ELF3-LUX Complex* Links the Circadian Clock to Diurnal Control of Hypocotyl Growth**. *Nature* 2011, **475**(7356):398-402.

128. Wang Y, Wu J-F, Nakamichi N, Sakakibara H, Nam H-G, Wu S-H: ***Light-Regulated WD1* and *Pseudo-Response Regulator 9* Form a Positive Feedback Regulatory Loop in the Arabidopsis Circadian Clock**. *The Plant Cell Online* 2011, **23**(2):486-498.

129. Farré EM, Harmer SL, Harmon FG, Yanovsky MJ, Kay SA: **Overlapping and Distinct Roles of *PRR7* and *PRR9* in the Arabidopsis Circadian Clock**. *Current Biology* 2005, **15**(1):47-54.

130. Hicks KA, Albertson TM, Wagner DR: ***Early Flowering 3* Encodes a Novel Protein That Regulates Circadian Clock Function and Flowering in Arabidopsis**. *The Plant Cell Online* 2001, **13**(6):1281-1292.

131. Kikis EA, Khanna R, Quail PH: ***ELF4* is a Phytochrome-Regulated Component of a Negative-Feedback Loop Involving the Central Oscillator Components *CCA1* and *LHY***. *The Plant Journal* 2005, **44**(2):300-313.

132. Noh B, Lee S-H, Kim H-J, Yi G, Shin E-A, Lee M, Jung K-J, Doyle MR, Amasino RM, Noh Y-S: **Divergent Roles of a Pair of Homologous Jumonji/Zinc-Finger–Class Transcription Factor Proteins in the Regulation of Arabidopsis Flowering Time**. *The Plant Cell Online* 2004, **16**(10):2601-2613.

133. Hicks KA, Millar AJ, Carré IA, Somers DE, Straume M, Meeks-Wagner DR, Kay SA: **Conditional Circadian Dysfunction of the Arabidopsis *Early Flowering 3* Mutant**. *Science* 1996, **274**(5288):790-792.

134. McWatters HG, Bastow RM, Hall A, Millar AJ: **The *ELF3zeitnehmer* Regulates Light Signalling to the Circadian Clock**. *Nature* 2000, **408**(6813):716-720.

135. Nakamichi N, Kiba T, Henriques R, Mizuno T, Chua N-H, Sakakibara H: ***Pseudo-Response Regulators 9, 7,* and *5* Are Transcriptional Repressors in the Arabidopsis Circadian Clock**. *The Plant Cell Online* 2010, **22**(3):594-605.

136. Thines B, Harmon FG: **Ambient Temperature Response Establishes *ELF3* as a Required Component of the Core Arabidopsis Circadian Clock**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(7):3257-3262.

137. Dixon LE, Knox K, Kozma-Bognar L, Southern MM, Pokhilko A, Millar AJ: **Temporal Repression of Core Circadian Genes Is Mediated through *Early Flowering 3* in Arabidopsis**. *Current Biology* 2011, **21**(2):120-125.

138. Takase T, Nishiyama Y, Tanihigashi H, Ogura Y, Miyazaki Y, Yamada Y, Kiyosue T: ***Lov Kelch Protein 2* and *Zeitlupe* Repress Arabidopsis Photoperiodic Flowering Under Non-Inductive Conditions, Dependent on *Flavin-Binding Kelch Repeat F-Box 1***. *The Plant Journal* 2011, **67**(4):608-621.

139. Corbesier L, Vincent C, Jang S, Fornara F, Fan Q, Searle I, Giakountis A, Farrona S, Gissot L, Turnbull C *et al*: ***FT*** **Protein Movement Contributes to Long-Distance Signaling in Floral Induction of Arabidopsis**. *Science* 2007, **316**(5827):1030-1033.

140. Jaeger KE, Wigge PA: ***FT*** **Protein Acts as a Long-Range Signal in Arabidopsis**. *Current Biology* 2007, **17**(12):1050-1054.

141. Shalit A, Rozman A, Goldshmidt A, Alvarez JP, Bowman JL, Eshed Y, Lifschitz E: **The Flowering Hormone Florigen Functions as a General Systemic Regulator of Growth and Termination**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(20):8392-8397.

142. Abe M, Kobayashi Y, Yamamoto S, Daimon Y, Yamaguchi A, Ikeda Y, Ichinoki H, Notaguchi M, Goto K, Araki T: ***FD*****, a bZIP Protein Mediating Signals from the Floral Pathway Integrator** ***FT*** **at the Shoot Apex**. *Science* 2005, **309**(5737):1052-1056.

143. Wigge PA, Kim MC, Jaeger KE, Busch W, Schmid M, Lohmann JU, Weigel D: **Integration of Spatial and Temporal Information During Floral Induction in Arabidopsis**. *Science* 2005, **309**(5737):1056-1059.

144. Yoo SK, Chung KS, Kim J, Lee JH, Hong SM, Yoo SJ, Yoo SY, Lee JS, Ahn JH: ***Constans*** **Activates** ***Suppressor of Overexpression of Constans 1*** **Through** ***Flowering Locus T*** **to Promote Flowering in Arabidopsis**. *Plant Physiology* 2005, **139**(2):770-778.

145. Shannon S, Meeks-Wagner DR: **Genetic Interactions That Regulate Inflorescence Development in Arabidopsis**. *The Plant Cell Online* 1993, **5**(6):639-655.

146. Gregis V, Sessa A, Dorca-Fornell C, Kater MM: **The Arabidopsis Floral Meristem Identity Genes** ***AP1*****,** ***AGL24*** **and** ***SVP*** **Directly Repress Class B and C Floral Homeotic Genes**. *The Plant Journal* 2009, **60**(4):626-637.

147. Liu C, Xi W, Shen L, Tan C, Yu H: **Regulation of Floral Patterning by Flowering Time Genes**. *Developmental Cell* 2009, **16**(8):711-722.

148. Liu C, Thong Z, Yu H: **Coming Into Bloom: The Specification of Floral Meristems**. *Development* 2009, **136**(20):3379-3391.

149. Lee J, Lee I: **Regulation and Function of** ***SOC1*****, a Flowering Pathway Integrator**. *Journal of Experimental Botany* 2010, **61**(9):2247-2254.

150. Hanano S, Goto K: **Arabidopsis *Terminal Flower 1* is Involved in the Regulation of Flowering Time and Inflorescence Development through Transcriptional Repression**. *The Plant Cell Online* 2011, **23**:3172-3184.

151. Wu R-M, Walton EF, Richardson AC, Wood M, Hellens RP, Varkonyi-Gasic E: **Conservation and Divergence of Four Kiwifruit *SVP-like* MADS-Box Genes Suggest Distinct Roles in Kiwifruit Bud Dormancy and Flowering**. *Journal of Experimental Botany* 2012, **63**(2):797-807.

152. Aubert D, Chen L, Moon Y-H, Martin D, Castle LA, Yang C-H, Sung ZR: ***EMF1*, A Novel Protein Involved in the Control of Shoot Architecture and Flowering in Arabidopsis**. *The Plant Cell Online* 2001, **13**(8):1865-1875.

153. Kim SY, Lee J, Eshed-Williams L, Zilberman D, Sung ZR: ***EMF1* and *PRC2* Cooperate to Repress Key Regulators of Arabidopsis Development**. *PLoS Genetics* 2012, **8**(3):1-16.

154. Litt A, Kramer EM: **The ABC Model and the Diversification of Floral Organ Identity**. *Seminars in Cell & Developmental Biology* 2010, **21**(1):129-137.

155. Honma T, Goto K: **Complexes of MADS-Box Proteins are Sufficient to Convert Leaves into Floral Organs**. *Nature* 2001, **409**(6819):525-529.

156. Pelaz S, Ditta GS, Baumann E, Wisman E, Yanofsky MF: **B and C Floral Organ Identity Functions Require *Sepallata* MADS-box Genes**. *Nature* 2000, **405**(6783):200-203.

157. Smyth DR, Bowman JL, Meyerowitz EM: **Early Flower Development in Arabidopsis**. *The Plant Cell Online* 1990, **2**(8):755-767.

158. Pelaz S, Tapia-López R, Alvarez-Buylla ER, Yanofsky MF: **Conversion of Leaves into Petals in Arabidopsis**. *Current Biology* 2001, **11**(3):182-184.

159. Moon J, Suh S-S, Lee H, Choi K-R, Hong CB, Paek N-C, Kim S-G, Lee I: **The *SOC1* MADS-Box Gene Integrates Vernalization and Gibberellin Signals for Flowering in Arabidopsis**. *The Plant Journal* 2003, **35**(5):613-623.

160. Sánchez-Corrales Y-E, Álvarez-Buylla ER, Mendoza L: **The *Arabidopsis thaliana* Flower Organ Specification Gene Regulatory Network Determines a Robust Differentiation Process**. *Journal of Theoretical Biology* 2010, **264**(3):971-983.

161. Becker A, Theissen G: **The Major Clades of MADS-Box Genes and Their Role in the Development and Evolution of Flowering Plants**. *Molecular Phylogenetics and Evolution* 2003, **29**(3):464-489.

162. Kendrick RE: **Photomorphogenesis in Plants**, 2nd edn. Dordrecht, The Netherlands: Kluwer Academic Publishers; 1994.

163. Quail PH: **Photosensory Perception and Signal Transduction in Plants**. *Current Opinion in Genetics & Development* 1994, **4**(5):652-661.

164. Smith H: **Physiological and Ecological Function within the Phytochrome Family**. *Annual Review of Plant Physiology and Plant Molecular Biology* 1995, **46**(1):289-315.

165. Mathews S: **Phytochrome Evolution in Green and Nongreen Plants**. *Journal of Heredity* 2005, **96**(3):197-204.

166. Suárez-López P, Wheatley K, Robson F, Onouchi H, Valverde F, Coupland G: ***Constans* Mediates between the Circadian Clock and the Control of Flowering in Arabidopsis**. *Nature* 2001, **410**(6832):1116-1120.

167. Takano M, Hirochika H, Miyao A: **Control of Plant Flowering Time by Regulation of *Phytochrome C* Expression**. In: *US7566815 B2*. Edited by USPTO, vol. US7566815 B2, C12N15/82, A01H5/10, C12N15/29, A01H5/00 edn. Japan: National Institute of Agrobiological Sciences and National Agricultural and Bio-oriented Research Organization; 2009: 1-17.

168. Balasubramanian S, Sureshkumar S, Lempe J, Weigel D: **Potent Induction of *Arabidopsis thaliana* Flowering by Elevated Growth Temperature**. *PLoS Genetics* 2006, **2**(7):980-989.

169. Abdurakhmonov I, Buriev Z, Logan-Young C, Abdukarimov A, Pepper A: **Duplication, Divergence and Persistence in the Phytochrome Photoreceptor Gene Family of Cottons (*Gossypium spp.*)**. *BMC Plant Biology* 2010, **10**(1):1-18.

170. Udall JA, Swanson JM, Haller K, Rapp RA, Sparks ME, Hatfield J, Yu Y, Wu Y, Dowd C, Arpat AB *et al*: **A Global Assembly of Cotton ESTs**. *Genome Research* 2006, **16**(3):441-450.

171. GCC: **Sequencher 4.8**. In., 4.8 edn. Ann Arbor, MI: Gene Codes Corporation; 2007: Aligns of DNA sequences.

172. White GM, Hamblin MT, Kresovich S: **Molecular Evolution of the Phytochrome Gene Family in Sorghum: Changing Rates of Synonymous and Replacement Evolution**. *Molecular Biology and Evolution* 2004, **21**(4):716-723.

173. Chun L, Kawakami A, Christopher DA: ***Phytochrome A* Mediates Blue Light and UV-A-Dependent Chloroplast Gene Transcription in Green Leaves**. *Plant Physiology* 2001, **125**(4):1957-1966.

174. Quail PH: **Phytochrome: A Light-Activated Molecular Switch that Regulates Plant Gene Expression**. *Annual Review of Genetics* 1991, **25**(1):389-409.

175. Rockwell NC, Su Y-S, Lagarias JC: **Phytochrome Structure And Signaling Mechanisms**. *Annual Review of Plant Biology* 2006, **57**(1):837-858.

176. Sharrock RA, Quail PH: **Novel Phytochrome Sequences in *Arabidopsis thaliana*: Structure, Evolution, and Differential Expression of a Plant Regulatory Photoreceptor Family**. *Genes & Development* 1989, **3**(11):1745-1757.

177. Schneider-Poetsch HAW, Marx S, Kolukisaoglu HÜ, Hanelt S, Braun B: **Phytochrome Evolution: Phytochrome Genes in Ferns and Mosses**. *Physiologia Plantarum* 1994, **91**(2):241-250.

178. Kolukisaoglu HÜ, Marx S, Wiegmann C, Hanelt S, Schneider-Poetsch HAW: **Divergence of the Phytochrome Gene Family Predates Angiosperm Evolution and Suggests that *Selaginella* and *Equisetum* Arose prior to *Psilotum***. *Journal of Molecular Evolution* 1995, **41**(3):329-337.

179. Mathews S, Lavin M, Sharrock RA: **Evolution of the Phytochrome Gene Family and Its Utility for Phylogenetic Analyses of Angiosperms**. *Annals of the Missouri Botanical Garden* 1995, **82**(2):296-321.

180. Wada M, Kanegae T, Nozue K, Fukuda S: **Cryptogam Phytochromes**. *Plant, Cell & Environment* 1997, **20**(6):685-690.

181. Wu SH, Lagarias JC: **The Phytochrome Photoreceptor in the Green Alga *Mesotaenium caldariorum*: Implication for a Conserved Mechanism of Phytochrome Action**. *Plant, Cell & Environment* 1997, **20**(6):691-699.

182. Mathews S, McBreen K: **Phylogenetic Relationships of B-Related Phytochromes in the Brassicaceae: Redundancy and the Persistence of *Phytochrome D***. *Molecular Phylogenetics and Evolution* 2008, **49**(2):411-423.

183. Pratt LH: **Phytochromes: Differential Properties, Expression Patterns and Molecular Evolution**. *Photochemistry and Photobiology* 1995, **61**(1):10-21.

184. Devlin PF, Patel SR, Whitelam GC: ***Phytochrome E* Influences Internode Elongation and Flowering Time in Arabidopsis**. *The Plant Cell Online* 1998, **10**(9):1479-1487.

185. Shinomura T, Nagatani A, Chory J, Furuya M: **The Induction of Seed Germination in *Arabidopsis thaliana* Is Regulated Principally by *Phytochrome B* and Secondarily by *Phytochrome A***. *Plant Physiology* 1994, **104**(2):363-371.

186. Botto JF, Sanchez RA, Whitelam GC, Casal JJ: ***Phytochrome A* Mediates the Promotion of Seed Germination by Very Low Fluences of Light and Canopy Shade Light in Arabidopsis**. *Plant Physiology* 1996, **110**(2):439-444.

187. Furuya M, Schäfer E: **Photoperception and Signalling of Induction Reactions by Different Phytochromes**. *Trends in Plant Science* 1996, **1**(9):301-307.

188. Casal JJ, Sanchez RA, Yanovsky MJ: **The Function of *Phytochrome A***. *Plant, Cell & Environment* 1997, **20**(6):813-819.

189. Smith H, Xu Y, Quail PH: **Antagonistic but Complementary Actions of *Phytochromes A* and *B* Allow Optimum Seedling De-Etiolation**. *Plant Physiology* 1997, **114**(2):637-641.

190. Qin M, Kuhn R, Moran S, Quail PH: **Overexpressed *Phytochrome C* has Similar Photosensory Specificity to *Phytochrome B* but a Distinctive Capacity to Enhance Primary Leaf Expansion**. *The Plant Journal* 1997, **12**(5):1163-1172.

191. Whitelam GC, Devlin PF: **Roles of Different Phytochromes in Arabidopsis Photomorphogenesis**. *Plant, Cell & Environment* 1997, **20**(6):752-758.

192. Neff MM, Chory J: **Genetic Interactions Between *Phytochrome A*, *Phytochrome B*, and *Cryptochrome 1* During Arabidopsis Development**. *Plant Physiology* 1998, **118**(1):27-35.

193. Smith HB: **Photoreceptors in Signal Transduction: Pathways of Enlightenment**. *The Plant Cell Online* 2000, **12**(1):1-3.

194. Franklin KA, Whitelam GC: **Light Signals, Phytochromes and Cross-Talk with Other Environmental Cues**. *Journal of Experimental Botany* 2004, **55**(395):271-276.

195. Josse E-M, Foreman J, Halliday KJ: **Paths Through the Phytochrome Network**. *Plant, Cell & Environment* 2008, **31**(5):667-678.

196. Heschel MS, Butler CM, Barua D, Chiang GCK, Wheeler A, Sharrock RA, Whitelam GC, Donohue K: **New Roles of Phytochromes during Seed Germination**. *International Journal of Plant Sciences* 2008, **169**(4):531-540.

197. Hauser B, Cordonnier-Pratt M-M, Daniel-Vedele F, Pratt L: **The Phytochrome Gene Family in Tomato Includes a Novel Subfamily**. *Plant Molecular Biology* 1995, **29**(6):1143-1155.

198. Pratt L, Cordonnier-Pratt M-M, Hauser B, Caboche M: **Tomato Contains Two Differentially Expressed Genes Encoding B-Type Phytochromes, neither of which can be Considered an Ortholog of Arabidopsis *Phytochrome B***. *Planta* 1995, **197**(1):203-206.

199. Pratt LH, Cordonnier-Pratt MM, Kelmenson PM, Lazarova GI, Kubota T, Alba RM: **The Phytochrome Gene Family in Tomato (*Solarium lycopersicum* L.)**. *Plant, Cell & Environment* 1997, **20**(6):672-677.

200. Howe GT, Bucciaglia PA, Hackett WP, Furnier GR, Cordonnier-Pratt MM, Gardner G: **Evidence that the Phytochrome Gene Family in Black Cottonwood has One *PHYA* Locus and Two *PHYB* Loci but Lacks Members of the *PHYC/F* and *PHYE* Subfamilies**. *Molecular Biology and Evolution* 1998, **15**(2):160-175.

201. Wendel JF: **New World Tetraploid Cottons Contain Old World Cytoplasm**. *Proceedings of the National Academy of Sciences of the United States of America* 1989, **86**(11):4132-4136.

202. Wendel JF, Albert VA: **Phylogenetics of the Cotton Genus (Gossypium): Character-State Weighted Parsimony Analysis of Chloroplast-DNA Restriction Site Data and Its Systematic and Biogeographic Implications**. *Systematic Botany* 1992, **17**(1):115-143.

203. Cronn R, Small R, Haselkorn, Wendel J: **Rapid Diversification of the Cotton Genus (*Gossypium:* Malvaceae) Revealed by Analysis of Sixteen Nuclear and Chloroplast Genes**. *American Journal of Botany* 2002, **89**:707-725.

204. Soltis PS, Soltis DE, Chase MW: **Angiosperm Phylogeny Inferred from Multiple Genes as a Tool for Comparative Biology**. *Nature* 1999, **402**(6760):402-404.

205. Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE *et al*: **Rosid Radiation and the Rapid Rise of Angiosperm-Dominated Forests**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**(10):3853-3858.

206. Cronn R, Cedroni M, Haselkorn T, Grover C, Wendel JF: **PCR-Mediated Recombination in Amplification Products Derived from Polyploid Cotton**. *TAG Theoretical and Applied Genetics* 2002, **104**(2-3):482-489.

207. Ossowski S, Schwab R, Weigel D: **Gene Silencing in Plants Using Artificial microRNAs and Other Small RNAs**. *The Plant Journal* 2008, **53**(4):674-690.

208. Moose SP, Mumm RH: **Molecular Plant Breeding as the Foundation for 21st Century Crop Improvement**. *Plant Physiology* 2008, **147**(3):969-977.

209. Konieczny A, Ausubel FM: **A Procedure for Mapping Arabidopsis Mutations using Co-Dominant Ecotype-Specific PCR-Based Markers**. *The Plant Journal* 1993, **4**(2):403-410.

210. Neff MM, Neff JD, Chory J, Pepper AE: **dCAPS, a Simple Technique for the Genetic Analysis of Single Nucleotide Polymorphisms: Experimental Applications in *Arabidopsis thaliana* Genetics**. *The Plant Journal* 1998, **14**(3):387-392.

211. Wikström N, Savolainen V, Chase MW: **Evolution of the Angiosperms: Calibrating the Family Tree**. *Proceedings of the Royal Society B: Biological Sciences* 2001, **268**(1482):2211-2220.

212. Seelanan T, Schnabel A, Wendel JF: **Congruence and Consensus in the Cotton Tribe (Malvaceae)**. *Systematic Botany* 1997, **22**(2):259-290.

213. Ohno S: **Evolution by Gene Duplication**. New York: Springer-Verlag; 1970.

214. Force A, Lynch M, Pickett FB, Amores A, Yan Y-l, Postlethwait J: **Preservation of Duplicate Genes by Complementary, Degenerative Mutations**. *Genetics* 1999, **151**(4):1531-1545.

215. Lynch M, Force A: **The Probability of Duplicate Gene Preservation by Subfunctionalization**. *Genetics* 2000, **154**(1):459-473.

216. Wagner A: **Birth and Death of Duplicated Genes in Completely Sequenced Eukaryotes**. *Trends in Genetics* 2001, **17**(5):237-239.

217. Lynch M, Conery JS: **The Evolutionary Fate and Consequences of Duplicate Genes**. *Science* 2000, **290**(5494):1151-1155.

218. Moore RC, Purugganan MD: **The Evolutionary Dynamics of Plant Duplicate Genes**. *Current Opinion in Plant Biology* 2005, **8**(2):122-128.

219. Alba R, Kelmenson PM, Cordonnier-Pratt M-M, Pratt LH: **The Phytochrome Gene Family in Tomato and the Rapid Differential Evolution of this Family in Angiosperms**. *Molecular Biology and Evolution* 2000, **17**(3):362-373.

220. Park C-M, Bhoo S-H, Song P-S: **Inter-Domain Crosstalk in the Phytochrome Molecules**. *Seminars in Cell & Developmental Biology* 2000, **11**(6):449-456.

221. Kim J-I, Shen Y, Han Y-J, Park J-E, Kirchenbauer D, Soh M-S, Nagy F, Schäfer E, Song P-S: **Phytochrome Phosphorylation Modulates Light Signaling by Influencing the Protein–Protein Interaction**. *The Plant Cell Online* 2004, **16**(10):2629-2640.

222. Chen M, Schwab R, Chory J: **Characterization of the Requirements for Localization of *Phytochrome B* to Nuclear Bodies**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(24):14493-14498.

223. Ingvarsson PK, Garcia MV, Luquez V, Hall D, Jansson S: **Nucleotide Polymorphism and Phenotypic Associations Within and Around the *Phytochrome B2 Locus* in European Aspen (*Populus tremula*, Salicaceae)**. *Genetics* 2008, **178**(4):2217-2226.

224. Nekrutenko A, Makova K, Li W: **The K(A)/K(S) Ratio Test for Assessing the Protein-Coding Potential of Genomic Regions: An Empirical and Simulation Study**. *Genome Research* 2002, **12**:198-202.

225. Cronn RC, Small RL, Wendel JF: **Duplicated Genes Evolve Independently after Polyploid Formation in Cotton**. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**(25):14406-14411.

226. Fisher R: **On the Interpretation of Chi^2 from Contingency Tables, and the Calculation of P**. *Journal of Royal Statistical Society* 1922, **85**:87-94.

227. Nei M, Rogozin IB, Piontkivska H: **Purifying Selection and Birth-And-Death Evolution in the Ubiquitin Gene Family**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(20):10866-10871.

228. Ltd. B: **Geneious**. In., vol. Pro, R6 edn. Auckland, NZ: Biomatters Ltd.; 2012: Geneious is a DNA, RNA and protein sequence alignment, assembly and analysis software platform, integrating bioinformatic and molecular biology tools into a simple interface.

229. Borevitz JO, Ecker JR: **Plant Genomics: The Third Wave**. *Annual Review of Genomics & Human Genetics* 2004, **5**(1):443-477.

230. Kohel R, Richmond T, Lewis C: **Texas Marker-1. A Description of a Genetic Standard for *Gossypium hirsutum* L**. *Crop Science* 1970, **10**:670-671.

231. Kohel R: **Molecular Mapping and Characterization of Traits Controlling Fiber Quality in Cotton**. *Euphytica* 2001, **121**(2):163-172.

232. Dellaporta S, Wood J, Hicks J: **A Plant DNA Mini-preparation: Version II**. *Plant Molecular Biology Reporter* 1983, **1**(4):19-21.

233. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools**. *Nucleic Acids Research* 1997, **25**(24):4876-4882.

234. Reddy O, Pepper A, Abdurakhmonov I, Saha S, Jenkins J, Brooks T, Bolek Y, El-Zik K: **New Dinucleotide and Trinucleotide Microsatellite Marker Resources for Cotton Genome Research**. *Journal of Cotton Science* 2001, **5**:103-113.

235. Eckert KA, Kunkel TA: **High Fidelity DNA Synthesis by the *Thermus aquaticus* DNA Polymerase**. *Nucleic Acids Research* 1990, **18**(13):3739-3744.

236. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs**. *Nucleic Acids Research* 1997, **25**(17):3389-3402.

237. Ma B, Tromp J, Li M: **PatternHunter: Faster and More Sensitive Homology Search**. *Bioinformatics* 2002, **18**(3):440-445.

238. Saitou N, Nei M: **The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees**. *Molecular Biology and Evolution* 1987, **4**(4):406-425.

239. Kimura M: **A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences**. *Journal of Molecular Evolution* 1980, **16**(2):111-120.

240. Swofford D: **PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)**. In., 4 edn. Sunderland, MA: Sinauer Associates; 2003.

241. Felsenstein J: **Confidence Limits on Phylogenies: An Approach Using the Bootstrap**. *Evolution* 1985, **39**(4):783-791.

242. Jukes TH, Cantor CR: **Evolution of Protein Molecules**. In: *Mammalian Protein Metabolism.* Edited by Munro M. New York, NY: Academic Press; 1969: 21-123.

243. Nei M, Gojobori T: **Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions**. *Molecular Biology and Evolution* 1986, **3**(5):418-426.

244. Librado P, Rozas J: **DnaSP v5: A software for Comprehensive Analysis of DNA Polymorphism Data**. *Bioinformatics* 2009, **25**(11):1451-1452.

245. Colledge S, Conolly J: **The Origins and Spread of Domestic Plants in Southwest Asia and Europe**. Walnut Creek, CA: Left Coast Press; 2007.

246. Fuller DQ: **Contrasting Patterns in Crop Domestication and Domestication Rates: Recent Archaeobotanical Insights from the Old World**. *Annals of Botany* 2007, **100**(5):903-924.

247. Izawa T: **Adaptation of Flowering-Time by Natural and Artificial Selection in Arabidopsis and Rice**. *Journal of Experimental Botany* 2007, **58**(12):3091-3097.

248. Hirschhorn JN, Daly MJ: **Genome-Wide Association Studies for Common Diseases and Complex Traits**. *Nature Reviews Genetics* 2005, **6**(2):95-108.

249. Jiménez-Gómez JMA-B, Carlos; Borja, Alicia; Anastasio, Germán; Angosto, Trinidad; Lozano, Rafael; Martínez-Zapatera, José M.: **Quantitative Genetic Analysis of Flowering Time in Tomato**. *Genome* 2007, **50**(3):303–315.

250. Amasino R: **Seasonal and Developmental Timing of Flowering**. *The Plant Journal* 2010, **61**(6):1001-1013.

251. Bäurle I, Dean C: **The Timing of Developmental Transitions in Plants**. *Cell* 2006, **125**(4):655-664.

252. Kobayashi Y, Weigel D: **Move On Up, It's Time For Change—Mobile Signals Controlling Photoperiod-Dependent Flowering**. *Genes & Development* 2007, **21**(19):2371-2384.

253. Greenup A, Peacock WJ, Dennis ES, Trevaskis B: **The Molecular Biology of Seasonal Flowering-Responses in Arabidopsis and the Cereals**. *Annals of Botany* 2009, **103**(8):1165-1172.

254. Sasani S, Hemming MN, Oliver SN, Greenup A, Tavakkol-Afshari R, Mahfoozi S, Poustini K, Sharifi H-R, Dennis ES, Peacock WJ *et al*: **The Influence of Vernalization and Daylength on Expression of Flowering-Time Genes in the Shoot Apex and Leaves of Barley (*Hordeum vulgare*)**. *Journal of Experimental Botany* 2009, **60**(7):2169-2178.

255. Matsubara K, Ogiso-Tanaka E, Hori K, Ebana K, Ando T, Yano M: **Natural Variation in *Hd17*, a Homolog of Arabidopsis *ELF3* that is Involved in Rice Photoperiodic Flowering**. *Plant and Cell Physiology* 2012, **53**(4):709-716.

256. Henry RJ: **Next-Generation Sequencing for Understanding and Accelerating Crop Domestication**. *Briefings in Functional Genomics* 2012, **11**(1):51-56.

257. Yang C-c, Kawahara Y, Mizuno H, Wu J, Matsumoto T, Itoh T: **Independent Domestication of Asian Rice Followed by Gene Flow from japonica to indica**. *Molecular Biology and Evolution* 2012, **29**(5):1471-1479.

258. Skøt L, Sanderson R, Thomas A, Skøt K, Thorogood D, Latypova G, Asp T, Armstead I: **Allelic Variation in the Perennial Ryegrass *Flowering Locus T* Gene is Associated with Changes in Flowering Time Across a Range of Populations**. *Plant Physiology* 2011, **155**(2):1013-1022.

259. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC *et al*: **The Genetic Architecture of Maize Flowering Time**. *Science* 2009, **325**(5941):714-718.

260. Amaral AJ: **Genome-Wide Footprints of Pig Domestication and Selection Revealed Through Massive Parallel Sequencing of Pooled DNA**. *PLoS ONE* 2011, **6**(4):1-12.

261. Rubin C: **Whole-Genome Resequencing Reveals Loci Under Selection During Chicken Domestication**. *Nature* 2010, **464**:587-591.

262. Purugganan MD, Fuller DQ: **The Nature of Selection During Plant Domestication**. *Nature* 2009, **457**(7231):843-848.

263. Iqbal MJ: **A Genetic Bottleneck in the 'Evolution Under Domestication' of Upland Cotton *Gossypium hirsutum* L. Examined Using DNA Fingerprinting**. *TAG Theoretical and Applied Genetics* 2001, **103**(4):547-554.

264. Thomson M, Zhao K, Wright M, McNally K, Rey J, Tung C-W, Reynolds A, Scheffler B, Eizenga G, McClung A *et al*: **High-Throughput Single Nucleotide Polymorphism Genotyping for Breeding Applications in Rice Using the BeadXpress Platform**. *Molecular Breeding* 2012, **29**(4):875-886.

265. Hinze LL, Dever JK, Percy RG: **Molecular Variation Among and Within Improved Cultivars in the U.S. Cotton Germplasm Collection**. *Crop Science* 2012, **52**(1):222-230.

266. Mammadov JA: **Development of Highly Polymorphic SNP Markers from the Complexity Reduced Portion of Maize [*Zea mays* L.] Genome for Use in Marker-Assisted Breeding**. *TAG Theoretical and Applied Genetics* 2010, **121**:577-588.

267. Jung C, Müller AE: **Flowering Time Control and Applications in Plant Breeding**. *Trends in Plant Science* 2009, **14**(10):563-573.

268. Abdurakhmonov IY, Kushanov FN, Djaniqulov F, Buriev ZT, Pepper AE, Fayzieva N, Mavlonov GT, Saha S, Jenkins JN, Abdukarimov A: **The Role of Induced Mutation in Conversion of Photoperiod Dependence in Cotton**. *Journal of Heredity* 2007, **98**(3):258-266.

269. Draye X, Chee P, Jiang C-X, Decanini L, Delmonte TA, Bredhauer R, Smith CW, Paterson AH: **Molecular Dissection of Interspecific Variation Between *Gossypium hirsutum* and *G. barbadense* (Cotton) by a Backcross-Self Approach: II. Fiber Fineness**. *TAG Theoretical and Applied Genetics* 2005, **111**(4):764-771.

270. Lacape J-M, Nguyen T-B, Courtois B, Belot J-L, Giband M, Gourlot J-P, Gawryziak G, Roques S, Hau B: **QTL Analysis of Cotton Fiber Quality Using Multiple *Gossypium hirsutum* X *Gossypium barbadense* Backcross Generations**. *Crop Science* 2005, **45**(1):123-140.

271. Bernardo R: **Breeding for Quantitative Traits in Plants**. Woodbury, MN: Stemma Press; 2002.

272. Wang GL, Dong JM, Paterson AH: **The Distribution of *Gossypium hirsutum* Chromatin in *G. barbadense* Germ Plasm: Molecular Analysis of Introgressive Plant Breeding**. *TAG Theoretical and Applied Genetics* 1995, **91**(6-7):1153-1161.

273. Miller PA, Rawlings JO: **Breakup of Initial Linkage Blocks Through Intermating in a Cotton Breeding Population**. *Crop Science* 1967, **7**(3):199-204.

274. Esteve-Codina A: **Partial Short-Read Sequencing of a Highly Inbred Iberian Pig and Genomics Inference Thereof**. *Heredity* 2011, **107**:256-264.

275. Ramos AM: **Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology**. *PLoS ONE* 2009, **4**(8):1-13.

276. Lagercrantz U, Axelsson T: **Rapid Evolution of the Family of *Constans Like* Genes in Plants**. *Molecular Biology and Evolution* 2000, **17**(10):1499-1507.

277. Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA: **Copy Number Variation Affecting the *Photoperiod-B1* and *Vernalization-A1* Genes Is Associated with Altered Flowering Time in Wheat (*Triticum aestivum*)**. *PLoS ONE* 2012, **7**(3):1-11.

278. Campoli C, Drosse B, Searle I, Coupland G, von Korff M: **Functional Characterisation of *HvCO1*, the Barley (*Hordeum vulgare*) Flowering Time Ortholog of *Constans***. *The Plant Journal* 2012, **69**(5):868-880.

279. Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C *et al*: **Genome-Wide Association Study of Flowering Time and Grain Yield Traits in a Worldwide Collection of Rice Germplasm**. *Nature Genetics* 2012, **44**(1):32-39.

280. Pabón-Mora N, Ambrose BA, Litt A: **Poppy Apetala 1/Fruitfull Orthologs Control Flowering Time, Branching, Perianth Identity, and Fruit Development**. *Plant Physiology* 2012, **158**(4):1685-1704.

281. Zhao J, Huang X, Ouyang X, Chen W, Du A, Zhu L, Wang S, Deng XW, Li S: ***OsELF3-1*, an Ortholog of Arabidopsis *Early Flowering 3*, Regulates Rice Circadian Rhythm and Photoperiodic Flowering**. *PLoS ONE* 2012, **7**(8):1-10.

282. Laurie RE, Diwadkar P, Jaudal M, Zhang L, Hecht V, Wen J, Tadege M, Mysore KS, Putterill J, Weller JL *et al*: **The *Medicago Flowering Locus T Homolog*, *MtFTa1*, is a Key Regulator of Flowering Time**. *Plant Physiology* 2011, **156**(4):2207-2224.

283. Lou P, Xie Q, Xu X, Edwards C, Brock M, Weinig C, McClung C: **Genetic Architecture of the Circadian Clock and Flowering Time in *Brassica rapa***. *TAG Theoretical and Applied Genetics* 2011, **123**(3):397-409.

284. Murphy RL, Klein RR, Morishige DT, Brady JA, Rooney WL, Miller FR, Dugas DV, Klein PE, Mullet JE: **Coincident Light and Clock Regulation of *Pseudoresponse Regulator Protein 37* (*PRR37*) Controls Photoperiodic Flowering in Sorghum**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(39):16469-16474.

285. Watanabe S, Hideshima R, Xia Z, Tsubokura Y, Sato S, Nakamoto Y, Yamanaka N, Takahashi R, Ishimoto M, Anai T *et al*: **Map-Based Cloning of the Gene Associated With the Soybean Maturity Locus *E3***. *Genetics* 2009, **182**(4):1251-1262.

286. Roux F, Touzet P, Cuguen J, Le Corre V: **How to be Early Flowering: An Evolutionary Perspective**. *Trends in Plant Science* 2006, **11**(8):375-381.

287. Wendel JF, Brubaker C, Alvarez I, Cronn R, Stewart JM: **Evolution and Natural History of the Cotton Genus**. In: *Genetics and Genomics of Cotton.* Edited by Paterson AH. New York, NY: Springer US; 2009: 3-22.

288. Small RL, Wendel JF: **Phylogeny, Duplication, and Intraspecific Variation of Adh Sequences in New World Diploid Cottons (*Gossypium* L., Malvaceae)**. *Molecular Phylogenetics and Evolution* 2000, **16**(1):73-84.

289. Reinisch AJ, Dong JM, Brubaker CL, Stelly DM, Wendel JF, Paterson AH: **A Detailed RFLP Map of Cotton, *Gossypium hirsutum* x *Gossypium barbadense*: Chromosome Organization and Evolution in a Disomic Polyploid Genome**. *Genetics* 1994, **138**(3):829-847.

290. DeJoode DR, Wendel JF: **Genetic Diversity and Origin of the Hawaiian Islands Cotton, *Gossypium tomentosum***. *American Journal of Botany* 1992, **79**(11):1311-1319.

291. Abdurakhmonov I, Saha S, Jenkins J, Buriev Z, Shermatov S, Scheffler B, Pepper A, Yu J, Kohel R, Abdukarimov A: **Linkage Disequilibrium Based Association Mapping of Fiber Quality Traits in *G. hirsutum* L. Variety Germplasm**. *Genetica* 2009, **136**(3):401-417.

292. Niles GA, Feaster CV: **Cotton**. In: *Breeding.* Edited by Kohel RJ, Lewis CF. Madison, WI: American Society of Agronomy, Inc.; Crop Science Society of America, Inc.; Soil Science Society of America, Inc.; 1984: 201–231.

293. Kulkarni V, Khadi B, Maralappanavar M, Deshapande L, Narayanan S: **The Worldwide Gene Pools of *Gossypium arboreum* L. and *G. herbaceum* L., and Their Improvement**. In: *Genetics and Genomics of Cotton.* Edited by Paterson AH. New York, NY: Springer US; 2009: 69-97

294. Lubbers EL, Chee PW: **The Worldwide Gene Pool of *G. hirsutum* and its Improvement**. In: *Genetics and Genomics of Cotton.* Edited by Paterson AH. New York, NY: Springer US; 2009: 23-52.

295. Percy RG: **The Worldwide Gene Pool of *Gossypium barbadense* L. and Its Improvement**. In: *Genetics and Genomics of Cotton.* Edited by Paterson AH. New York, NY: Springer US; 2009: 53-68.

296. Lee SB, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H: **The Complete Chloroplast Genome Sequence of *Gossypium hirsutum*: Organization and Phylogenetic Relationships to Other Angiosperms**. *BMC Genomics* 2006, **7**(1):61-73.

297. Quisenberry JE, Jordan WR, Roark BA, Fryrear DW: **Exotic Cottons as Genetic Sources for Drought Resistance**. *Crop Science* 1981, **21**(6):889-895.

298. Shen X, Guo W, Zhu X, Yuan Y, Yu JZ, Kohel RJ, Zhang T: **Molecular Mapping of QTLs for Fiber Qualities in Three Diverse Lines in Upland Cotton Using SSR Markers**. *Molecular Breeding* 2005, **15**(2):169-181.

299. Lin Z, He D, Zhang X, Nie Y, Guo X, Feng C, STEWART JM: **Linkage Map Construction and Mapping QTL for Cotton Fibre Quality Using SRAP, SSR and RAPD**. *Plant Breeding* 2005, **124**(2):180-187.

300. Mackay I, Powell W: **Methods for Linkage Disequilibrium Mapping in Crops**. *Trends in Plant Science* 2007, **12**(2):57-63.

301. Guo W, Cai C, Wang C, Han Z, Song X, Wang K, Niu X, Wang C, Lu K, Shi B *et al*: **A Microsatellite-Based, Gene-Rich Linkage Map Reveals Genome Structure, Function and Evolution in Gossypium**. *Genetics* 2007, **176**(1):527-541.

302. Amos W, Driscoll E, Hoffman JI: **Candidate Genes Versus Genome-Wide Associations: Which Are Better for Detecting Genetic Susceptibility to Infectious Disease?** *Proceedings of the Royal Society B: Biological Sciences* 2010, **278**:1183-1188.

303. Lee Y-S, Jeong D-H, Lee D-Y, Yi J, Ryu C-H, Kim SL, Jeong HJ, Choi SC, Jin P, Yang J *et al*: ***OsCOL4* is a Constitutive Flowering Repressor Upstream of *Ehd1* and Downstream of *OsPHYB***. *The Plant Journal* 2010, **63**(1):18-30.

304. Pazhouhandeh M, Molinier J, Berr A, Genschik P: ***MSI4/FVE* Interacts with *CUL4–DDB1* and a *PRC2-Like* Complex to Control Epigenetic Regulation of Flowering Time in Arabidopsis**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(8):3430-3435.

305. Ranjan A, Fiene G, Fackendahl P, Hoecker U: **The Arabidopsis Repressor of Light Signaling *SPA1* Acts in the Phloem to Regulate Seedling De-Etiolation, Leaf Expansion and Flowering Time**. *Development* 2011, **138**(9):1851-1862.

306. Salomé PA, Bomblies K, Laitinen RAE, Yant L, Mott R, Weigel D: **Genetic Architecture of Flowering-Time Variation in *Arabidopsis thaliana***. *Genetics* 2011, **188**(2):421-433.

307. Valverde F: ***Constans* and the Evolutionary Origin of Photoperiodic Timing of Flowering**. *Journal of Experimental Botany* 2011, **62**(8):2453-2463.

308. Tiwari SB, Shen Y, Chang H-C, Hou Y, Harris A, Ma SF, McPartland M, Hymus GJ, Adam L, Marion C *et al*: **The Flowering Time Regulator *Constans* is Recruited to the *Flowering Locus T* Promoter Via a Unique Cis-Element**. *New Phytologist* 2010, **187**(1):57-66.

309. Seo E, Lee H, Jeon J, Park H, Kim J, Noh Y-S, Lee I: **Crosstalk Between Cold Response and Flowering in Arabidopsis Is Mediated through the Flowering-Time Gene *SOC1* and Its Upstream Negative Regulator *FLC***. *The Plant Cell Online* 2009, **21**(10):3185-3197.

310. Ehrenreich IM, Hanzawa Y, Chou L, Roe JL, Kover PX, Purugganan MD: **Candidate Gene Association Mapping of Arabidopsis Flowering Time**. *Genetics* 2009, **183**(1):325-335.

311. Schwartz C, Balasubramanian S, Warthmann N, Michael TP, Lempe J, Sureshkumar S, Kobayashi Y, Maloof JN, Borevitz JO, Chory J *et al*: **Cis-regulatory Changes at *Flowering Locus T* Mediate Natural Variation in Flowering Responses of *Arabidopsis thaliana***. *Genetics* 2009, **183**:723-732.

312. Melzer S, Lens F, Gennen J, Vanneste S, Rohde A, Beeckman T: **Flowering-Time Genes Modulate Meristem Determinacy and Growth Form in *Arabidopsis thaliana***. *Nature Genetics* 2008, **40**(12):1489-1492.

313. Liu L-J, Zhang Y-C, Li Q-H, Sang Y, Mao J, Lian H-L, Wang L, Yang H-Q: ***COP1*-Mediated Ubiquitination of *Constans* Is Implicated in Cryptochrome Regulation of Flowering in Arabidopsis**. *The Plant Cell Online* 2008, **20**(2):292-306.

314. Lee JH, Yoo SJ, Park SH, Hwang I, Lee JS, Ahn JH: **Role of SVP in the Control of Flowering Time by Ambient Temperature in Arabidopsis**. *Genes & Development* 2007, **21**(4):397-402.

315. Oliverio KA, Crepy M, Martin-Tryon EL, Milich R, Harmer SL, Putterill J, Yanovsky MJ, Casal JJ: ***Gigantea* Regulates *Phytochrome A*-Mediated Photomorphogenesis Independently of Its Role in the Circadian Clock**. *Plant Physiology* 2007, **144**(1):495-502.

316. Putterill J, Robson F, Lee K, Simon R, Coupland G: **The *Constans* Gene of Arabidopsis Promotes Flowering and Encodes a Protein Showing Similarities to Zinc Finger Transcription Factors**. *Cell* 1995, **80**(6):847-857.

317. Bowers JE, Chapman BA, Rong J, Paterson AH: **Unravelling Angiosperm Genome Evolution by Phylogenetic Analysis of Chromosomal Duplication Events**. *Nature* 2003, **422**(6930):433-438.

318. Rong J, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW, Delmonte TA, Ding X, Garza JJ, Marler BS *et al*: **A 3347-Locus Genetic Recombination Map of Sequence-Tagged Sites Reveals Features of Genome Organization, Transmission and Evolution of Cotton (*Gossypium*)**. *Genetics* 2004, **166**(1):389-417.

319. Zhang J, Lu Y, Cantrell RG, Hughs E: **Molecular Marker Diversity and Field Performance in Commercial Cotton Cultivars Evaluated in the Southwestern USA**. *Crop Science* 2005, **45**(4):1483-1490.

320. Blenda A, Fang DD, Rami J-F, Garsmeur O, Luo F, Lacape J-M: **A High Density Consensus Genetic Map of Tetraploid Cotton That Integrates Multiple Component Maps through Molecular Marker Redundancy Check**. *PLoS ONE* 2012, **7**(9):1-17.

321. Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S *et al*: **The Draft Genome of a Diploid Cotton *Gossypium raimondii***. *Nature Genetics* 2012, **44**(10):1098-1103.

322. Hu W, Franklin KA, Sharrock RA, Jones MA, Harmer SL, Lagarias JC: **Unanticipated Regulatory Roles for Arabidopsis Phytochromes Revealed by Null Mutant Analysis**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(4):1542-1547.

323. Matsubara K, Yamanouchi U, Nonoue Y, Sugimoto K, Wang Z-X, Minobe Y, Yano M: ***Ehd3*, Encoding a Plant Homeodomain Finger-Containing Protein, is a Critical Promoter of Rice Flowering**. *The Plant Journal* 2011, **66**(4):603-612.

324. Kolmos E, Herrero E, Bujdoso N, Millar AJ, Tóth R, Gyula P, Nagy F, Davis SJ: **A Reduced-Function Allele Reveals That *Early Flowering 3* Repressive Action on the Circadian Clock Is Modulated by Phytochrome Signals in Arabidopsis**. *The Plant Cell Online* 2011, **23**(9):3230-3246.

325. Zeevaart JAD: **Leaf-Produced Floral Signals**. *Current Opinion in Plant Biology* 2008, **11**(5):541-547.

326. Hecht V, Foucher F, Ferrandiz C, Macknight R, Navarro C, Morin J, Vardy ME, Ellis N, Beltran JP, Rameau C *et al*: **Conservation of Arabidopsis Flowering Genes in Model Legumes**. *Plant Physiology* 2005, **137**(4):1420-1434.

327. Malcomber ST, Kellogg EA: **Sepallata Gene Diversification: Brave New Whorls**. *Trends in Plant Science* 2005, **10**(9):427-435.

328. Burn JE, Smyth DR, Peacock WJ, Dennis ES: **Genes Conferring Late Flowering in *Arabidopsis thaliana***. *Genetica* 1993, **90**(2-3):147-155.

329. Pokhilko A, Fernandez AP, Edwards KD, Southern MM, Halliday KJ, Millar AJ: **The Clock Gene Circuit in Arabidopsis Includes a Repressilator with Additional Feedback Loops**. *Molecular Systems Biology* 2012, **8**:1-13.

330. Huang W, Perez-Garcia P, Pokhilko A, Millar AJ, Antoshechkin I, Riechmann JL, Mas P: **Mapping the Core of the Arabidopsis Circadian Clock Defines the Network Structure of the Oscillator**. *Science* 2012, **336**(6077):75-79.

331. McClung CR: **The Photomorphogenic Protein, *De-Etiolated 1*, Is a Critical Transcriptional Corepressor in the Central Loop of the Arabidopsis Circadian Clock**. *Molecular Cell* 2011, **43**(5):693-694.

332. Staiger D, Green R: **RNA-Based Regulation in the Plant Circadian Clock**. *Trends in Plant Science* 2011, **16**(10):517-523.

333. Nakamichi N: **Molecular Mechanisms Underlying the Arabidopsis Circadian Clock**. *Plant and Cell Physiology* 2011, **52**(10):1709-1718.

334. McClung CR: **The Genetics of Plant Clocks**. In: *Advances in Genetics.* Edited by Stuart B, vol. 74. New York, NY: Academic Press; 2011: 105-139.

335. Castillon A, Shen H, Huq E: **Blue Light Induces Degradation of the Negative Regulator *Phytochrome Interacting Factor 1* to Promote Photomorphogenic Development of Arabidopsis Seedlings**. *Genetics* 2009, **182**(1):161-171.

336. Ahmad M, Cashmore AR: **The Blue-Light Receptor *Cryptochrome 1* Shows Functional Dependence on *Phytochrome A* or *Phytochrome B* in *Arabidopsis thaliana***. *The Plant Journal* 1997, **11**(3):421-427.

337. Johnson E, Bradley M, Harberd NP, Whitelam GC: **Photoresponses of Light-Grown *PHYA* Mutants of Arabidopsis (*Phytochrome A* Is Required for the Perception of Daylength Extensions)**. *Plant Physiology* 1994, **105**(1):141-149.

338. Sanger F: **DNA Sequencing with Chain-Terminating Inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**(12):5463-5467.

339. Wall DP, Fraser HB, Hirsh AE: **Detecting Putative Orthologs**. *Bioinformatics* 2003, **19**(13):1710-1711.

340. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *Journal of Molecular Biology* 1990, **215**(3):403-410.

341. Kumar S, Stecher G, Peterson D, Tamura K: **MEGA-CC: Computing Core of Molecular Evolutionary Genetics Analysis Program for Automated and Iterative Data Analysis**. *Bioinformatics* 2012, **28**(20):2685-2686.

342. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods**. *Molecular Biology and Evolution* 2011, **28**(10):2731-2739.

343. Locascio A, Lucchin M, Varotto S: **Characterization of a *MADS Flowering Locus C-Like* (*MFL*) Sequence in *Cichorium intybus*: A Comparative Study of *CiMFL* and *AtFLC* Reveals Homologies and Divergences in Gene Function**. *New Phytologist* 2009, **182**(3):630-643.

344. Flagel L, Wendel J, Udall J: **Duplicate Gene Evolution, Homoeologous Recombination, and Transcriptome Characterization in Allopolyploid Cotton**. *BMC Genomics* 2012, **13**(1):302-315.

345. Wang F, Gong Y, Zhang C, Liu G, Wang L, Xu Z, Zhang J: **Genetic Effects of Introgression Genomic Components from Sea Island Cotton (*Gossypium barbadense* L.) on Fiber Related Traits in Upland Cotton (*G. hirsutum* L.)**. *Euphytica* 2011, **181**(1):41-53.

346. Paterson AH: **Genetics and Genomics of Cotton**, vol. 3. New York, NY: Springer US; 2009.

347. Álvarez I, Cronn R, Wendel JF: **Phylogeny of the New World Diploid Cottons (Gossypium L., Malvaceae) Based on Sequences of Three Low-Copy Nuclear Genes**. *Plant Syst Evol* 2005, **252**(3-4):199-214.

348. Stephens SG: **The Origin of Sea Island Cotton**. *Agricultural History* 1976, **50**(2):391-399.

349. Humphries JA: **Two *WD-Repeat* Genes from Cotton are Functional Homologues of the *Arabidopsis thaliana Transparent Testa Glabra1 (TTG1)* Gene**. *Plant Molecular Biology* 2005, **57**(1):67-81.

350. Mao J, Zhang Y-C, Sang Y, Li Q-H, Yang H-Q: **A Role for Arabidopsis Cryptochromes and *COP1* in the Regulation of Stomatal Opening**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(34):12270-12275.

351. Kibbe WA: **OligoCalc: An Online Oligonucleotide Properties Calculator**. *Nucleic Acids Research* 2007, **35**(2):43-46.

352. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ: **Amplification-Free Illumina Sequencing-Library Preparation Facilitates Improved Mapping and Assembly of (G+C)-Biased Genomes**. *Nature Methods* 2009, **6**(4):291-295.

353. Roche: **Rapid Library Preparation Method Manual**. In. Edited by Science RA. Mannheim, Germany: Roche Diagnostics GmbH; 2010: 1-8.

354. Roche: **Updated Titration Range for GS FLX Titanium emPCR Lib-L Kit**. In. Edited by Science RA, TCB No. 006-2010 edn. Mannheim, Germany: Roche Diagnostics GmbH; 2010: 1-3.

355. Roche: **emPCR Method Manual – Lib-L LV**. In. Edited by Science RA. Mannheim, Germany: Roche Diagnostics GmbH; 2010: 1-12.

356. Roche: **Using Multiplex Identifier (MID) Adaptors for the GS FLX Titanium Chemistry - Extended MID Set** In. Edited by Science RA, TCB No. 005-2009 edn. Mannheim, Germany: Roche Diagnostics GmbH; 2009: 1-7.

357. Roche: **Using Multiplex Identifier (MID) Adaptors for the GS FLX Titanium Chemistry - Basic MID Set**. In: *Science, Roche Applied.* Edited by Science RA, TCB No. 005-2009 edn. Mannheim, Germany: Roche Diagnostics GmbH; 2009: 1-11.

358. Santos SR: **Everything You Ever Wanted to Know Concerning Oligonucleotides, but Were Afraid to Ask**. In. Edited by Santos SR. Auburn University Website: Auburn University; 2005: 1-5.

359. Drummond AJ AB, Buxton S, Cheung M, Cooper A, Heled J, Kearse M, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A: **Geneious v5.1**. In., vol. Pro, 5.3.4 edn. Auckland, NZ: Biomatters Ltd.; 2010.

360. Bjarne Knudsen TK, Mikael Flensborg, Henrik Sandmann, Michael Heltzen, Alex Andersen, Mikkel Dickenson, Jakob Bardram, Peter J. Steffensen, Søren Mønsted, Torben Lauritzen, Roald Forsberg, Agnes Thanbichler, Jannick D. Bendtsen, Lasse Görlitz, Jane Rasmussen, David Tordrup, Morten Værum, Mikkel Nygaard Ravn, Christian Hachenberg, Esben Fisker, Patrick Dekker, Jacob Schultz, Anne-Mette K. Hein, Jonas Buur Sinding: **CLC Genomics Workbench**. In., 5.5.1 edn. Aarhus, Denmark CLC Bio; 2012: Analyze, visualize, and compare DNA, RNA, and Protein data. Run advanced workflows with large and complicated datasets.

361. Inc. SI: **Base SAS® 9.3 Procedures Guide**. Cary, NC, USA: SAS Institute Inc.; 2011.

362. Opal C, Nicholls A: **Fair Trade: Market-Driven Ethical Consumption**. Wiltshire, UK: Sage Publications Limited; 2005.

363. Schmitz A, Moss CB, Schmitz TG: **Agricultural Policy, Agribusiness, and Rent-Seeking Behaviour**. Toronto, Canada: University of Toronto Press; 2010.

364. Montalvo Jr JG: **Relationships Between Micronaire, Fineness, and Maturity. Part I. Fundamentals**. *Journal of Cotton Science* 2005, **9**(2):81-88.

365. Balls WL: **The Cotton Plant in Egypt**. London, UK: MacMillan and Company limited; 1919.

366. Smith CW, Cothren JT: **Cotton: Origin, History, Technology, and Production**, vol. 4. New York, NY: John Wiley & Sons, Inc.; 1999.

367. Feaster CV, Turcotte EL: **Registration of Pima S-2 Cotton1 (Reg. No. 57)**. *Crop Science* 1976, **16**(4):603-603.

368. Bryan WE: **Prospects for Pima S-1 Cotton**. *Progressive Agriculture in Arizona* 1955, **7**(2):6-6.

369. Silvertooth JC: **Agronomic Guidelines for Pima Cotton Production in Arizona**. In. Edited by Arizona Uo, Sciences CoAaL, vol. AZ1242. Tuscon, AZ: The University of Arizona Cooperative Extension; 2001: 1-4.

370. Feaster CV, Turcotte EL: **Registration of Pima S-6 Cotton**. *Crop Science* 1984, **24**(2):382-382.

371. Feaster CV, Turcotte EL, Young EF: **Registration of Pima S-5 Cotton01 (Reg. No. 60)**. *Crop Science* 1976, **16**(4):604-604.

372. Qureshi SN, Saha S, Kantety RV, Jenkins J: **EST-SSR: A New Class of Genetic Markers in Cotton**. 2004, **8**:112-123.

373. Park Y-H, Alabady M, Ulloa M, Sickler B, Wilkins T, Yu J, Stelly D, Kohel R, El-Shihy O, Cantrell R: **Genetic Mapping of New Cotton Fiber Loci Using EST-Derived Microsatellites in an Interspecific Recombinant Inbred Line Cotton Population**. *Mol Genet Genomics* 2005, **274**(4):428-441.

374. Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH: **Genome Evolution and Meiotic Maps by Massively Parallel DNA Sequencing: Spotted Gar, an Outgroup for the Teleost Genome Duplication**. *Genetics* 2011, **188**(4):799-808.

375. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species**. *PLoS ONE* 2011, **6**(5):1-10.

376. Ma X-F, Jensen E, Alexandrov N, Troukhan M, Zhang L, Thomas-Jones S, Farrar K, Clifton-Brown J, Donnison I, Swaller T *et al*: **High Resolution Genetic Mapping by Genome Sequencing Reveals Genome Duplication and Tetraploid Genetic Structure of the Diploid *Miscanthus sinensis***. *PLoS ONE* 2012, **7**(3):1-11.

377. Chutimanitsakun Y: **Construction and Application for QTL Analysis of a Restriction Site Associated DNA (RAD) Linkage Map in Barley**. *BMC Genomics* 2011, **12**:4-17.

378. Nelson J, Wang S, Wu Y, Li X, Antony G, White F, Yu J: **Single-Nucleotide Polymorphism Discovery by High-Throughput Sequencing in Sorghum**. *BMC Genomics* 2011, **12**(1):352-367.

379. Poland JA, Brown PJ, Sorrells ME, Jannink J-L: **Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach**. *PLoS ONE* 2012, **7**(2):1-8.

380. Laird PW: **Principles and Challenges of Genome-Wide DNA Methylation Analysis**. *Nature Reviews Genetics* 2010, **11**(3):191-203.

381. Zilberman D, Henikoff S: **Genome-Wide Analysis of DNA Methylation Patterns**. *Development* 2007, **134**(22):3959-3965.

382. Peng H, Zhang J, Wu X: **The Ploidy Effects in Plant Gene Expression: Progress, Problems and Prospects**. *Sci China Ser C-Life Sci* 2008, **51**(4):295-301.

383. Peng H, Zhang H-y, Li Y, Xu P-z, Wang X-d, Wu X-j: **Natural Homologous Triploidization and DNA Methylation in SARII-628, a Twin-seedling Line of Rice (*Oryza sativa*)**. *Rice Science* 2007, **14**(4):265-271.

384. Mead DA, McClary JA, Luckey JA, Kostichka AJ, Witney FR, Smith LM: **Bst DNA Polymerase Permits Rapid Sequence Analysis from Nanogram Amounts of Template**. *BioTechniques* 1991, **11**(1):76-87.

APPENDIX A

List of Supplemental Tables

Page

APPENDIX B

A Cotton Story

One summer, I remember gazing out the Suburban's window at the regimental lines of seeds being drawn in the fields. It was beautiful to watch the tractors tracing out these parallel rows. Quickly, I traced a line of Skittles on the back seat.

"Mom, look I'm a farmer. I am growing skittle trees."

She laughed while I smiled back happily.

My father asked, "Why are the farmers making lines of seeds?"

I contemplated this, but could not come up with a response. Quietly, I stared down at my coloring book, hoping to avoid answering the question.

Again he asked, "Well?"

I smartly replied, "It is like coloring. Staying inside a straight line is easier."

He gently replied, "That is a remarkably well thought out answer. The seeds are in a row, so they are easy to water, grow, and harvest."

Staring back out the window, I asked my mother, "What are the farmers planting?"

She replied, "Well this farm has corn, the one across the way has wheat, and that one has cotton."

With my young, inquisitive mind, I pondered what each plant produced. "Mom, I know I eat corn. What does wheat and cotton make?"

She stated, "We eat wheat, when we eat cereal and bread. We wear cotton, when we dress in t-shirts."

I am fascinated by this. I stare down at my t-shirt. I am trying to discern how a plant made these products. Not understanding, I quizzically asked, “How does cotton make my clothes?”

Instantly, she replied, “Cotton produces a fiber that when twisted together makes string. People weave thousands of strings together to create a t-shirt like yours. It is like the paper placemats you made for thanksgiving at school.”

Quietly, I took this in and was quite pleased with this new information.

In Dallas, my parents took me to the enormous farmer’s market downtown. They showed me all the different crops that farmers brought in to sell to the public, but there was one crop I did not see: cotton. On our return journey, I saw this white fluff spewing from the green bushes. Again, I asked “Mom, what are those white, fluffy things?”

Remembering that I am five, she replied, “That is cotton.”

My mom and dad then pulled the Suburban over to the road side. Out I hopped and raced to pick up some white fluff that had been blown off. We all retreated back to the vehicle, where I began to pull and twist the cotton fibers.

“What are these hard things Mom?”

She stated, “Those are seeds so that new plants can grow.”

Satisfied with the answer, I began to play with the cotton. Remembering that my mom said cotton made up my t-shirt, I hemmed and hawed over the fact that cotton was white, but my shirt was yellow. Finally, I asked “Mom, why is the cotton white, but my shirt yellow?”

My mom patiently described that factories dyed cotton different colors so that we could wear clothes in various hues and shades.

From that point on, I had a passion for plants, the environment, agriculture, and nature. I was constantly amazed at all the different products that were produced from plants. It was when I entered college at New Mexico State University that I knew I wanted to work on plants for crop improvement. I had those grandiose dreams of creating massive tomatoes that could feed hundreds of people in impoverished nations. Dr. Roy Cantrell showed me the importance of plant genetics, genetics and society, and specifically the genetics of cotton. He inspired me to go to graduate school. He had received his degree from Texas A&M University, so I decided that Texas A&M University might be the place for me to attend. This is how I arrived at this point, determined to inspire the cotton community, the world, and future students through my research.

APPENDIX C

Chapter II Supplemental Material*

Authors and Contact Information

Ibrokhim Y. Abdurakhmonov[3], Zabardast T. Buriev[3], Carla Jo Logan-Young[4],

Abdusattor Abdukarimov[3], Alan E. Pepper[4§]

Email addresses:

IYA: genomics@uzsci.net, ZTB: zabar75@yahoo.com,

CJLY: tysfira@neo.tamu.edu, AA: inst@gen.org.uz,

AEP: apepper@bio.tamu.edu

---

* Reprinted with the permission from "Duplication, Divergence and Persistence in the Phytochrome Photoreceptor Gene Family of Cottons (*Gossypium spp.*) " by Abdurakhmonov I, Buriev Z, Logan-Young C; Abdukarimov A, and Pepper A, 2010, *BMC Plant Biology*, 10(1):119

[3] Center of Genomic Technologies, Academy of Sciences of Uzbekistan. Yuqori Yuz, Qibray region Tashkent, 111226 Uzbekistan

[4] Department of Biology, Texas A&M University, College Station, Texas 77843, USA

[§]Corresponding author

Authors Contributions

IYA and AEP designed the experiment. IYA designed most of the PCR primers and cloned the *PHYA*, *PHYB* and *PHYE* gene families. ZTB performed DNA sequencing of phytochrome genes. CJLY isolated, cloned and sequenced the *PHYC* gene family and participated in the sequencing of *PHYA*, *PHYB* and *PHYE* genes. IYA, AA, CJLY, and AEP performed data interpretation and drafted the manuscript. All authors read and approved the final manuscript.

Supplemental Tables

Supplemental Table 1 GenBank EST Hits in Cotton

HSP	Taxon	ESTs	GenBank IDs	Min. Number Loci
PHYA	*G. raimondii*	4	CO117336	1
			CO092073	
			CO092074	
	G. hirsutum	5	ES822585	2
			EX168627	
			ES823391	
			DV849493	
			DW235365	
PHYB	*G. raimondii*	0	0	0
	G. hirsutum	3	DT566665	1
			ES835169	
			ES850111	
PHYC	*G. raimondii*	1	CO121409	1
	G. hirsutum	0	0	0
PHYE	*G. raimondii*	0	0	1
	G. hirsutum	2	DW478704	1
			DW506498	

APPENDIX D

Chapter III Supplemental Material

Authors and Contact Information

Carla Jo Logan-Young[5§], John Z. Yu[6], Richard G. Percy[6], Natalie B. Ware[5],

Sara E. Duke[7], Ibrokhim Y. Abdurakhmonov[8], Alan E. Pepper[5§]

Email addresses:

CJLY: tysfira@tamu.edu, JZY: john.yu@ars.usda.gov,

RGP: richard.percy@ars.usda.gov, NW: nware11@gmail.com,

IYA: genomics@uzsci.net, SED: sara.duke@ars.usda.gov,

AEP: apepper@bio.tamu.edu

[5] Department of Biology, Texas A&M University, College Station, Texas 77843, USA

[6] Cotton Germplasm Research Unit, United States Department of Agriculture, Agricultural Research Station, Southern Plains Agricultural Research Center, College Station, Texas 77845, USA

[7] Area Statistician, United States Department of Agriculture, Agricultural Research Station, Southern Plains Agricultural Research Area Office, College Station, Texas 77845, USA

[8] Center of Genomic Technologies, Academy of Sciences of Uzbekistan, Yuqori Yuz, Qibray region Tashkent, 111226 Uzbekistan

[§]Corresponding author

Authors Contributions

CJLY, IYA, AEP designed the experiment. CJLY designed most of the PCR primers and cloned all gene families. NBW helped to collect data, prep amplicons, and interpret data. RGP provided material for photoperiod dependent and independent lines. RGP provided all accession information from the USDA-ARS-SPARC Cotton Germplasm collection. SED determined the correct statistical measures using SAS. CJLY, JZY, and AEP performed data interpretation and drafted the manuscript. All authors read and approved the final manuscript.

Supplemental Table 2 SNPs in Exons without *PHY A2, CRY1 B, and CRY2 B*

Gene	A/D genome	*G. hirsutum*/ *G. barbadense*	Cultivated/ Wild	*G. incanum*	Single SNP
AGL16	0	0	0	1	0
AGL3_SEP4	0	0	2	0	2
AGL30	0	0	0	0	1
AGL32	0	0	0	0	0
AGL6	0	0	0	0	0
AGL65	3	0	0	1	2
AP1	2	0	0	1	0
ATGRP7	1	0	0	1	1
COL3	8	0	2	2	3
COL5	9	2	0	0	2
COP1	0	0	0	1	0
CRY1 A	1	0	0	0	0
CRY2 A	1	1	2	2	0
CRY3	4	0	0	2	1
DET1	0	0	0	0	0
ELF3	0	0	0	1	0
FD	12	0	1	4	4
FKF1_ADO3	3	2	0	1	0
GI Ex9to10_ A	0	0	0	0	0
GI Ex10to11_ A	0	0	0	1	0
GI Ex10to11_ B	0	0	0	0	0
GI Ex11to12_ A	6	0	1	1	1
GI Ex11to12_ B	0	2	0	0	0
HY6	12	2	11	2	3
LHY 1 /CCA1	0	0	0	0	0
LHY 2 /CCA1	1	0	0	1	0
PFT1	0	0	0	0	0
PHYA Contig 1	15	3	3	7	3
PHYB	8	1	1	1	3
PHYC	6	17	12	19	15
PHYE	5	3	0	6	6
PI 1	0	0	0	0	0
PI 2	0	0	0	0	0
PRR5	1	0	1	2	1
PRR7 Contig 1	6	0	0	3	1
PRR7 Contig 2	7	3	2	3	3
SPA4	1	0	0	2	0
TOC1	2	0	0	0	0
Exon Totals	114	36	38	65	52

Supplemental Table 3 SNPs in Introns without *PHY A2, CRY1 B, and CRY2 B*

Gene	A/D genome	*G. hirsutum/ G. barbadense*	Cultivated/ Wild	*G. incanum*	Single SNP
AGL16	5	2	0	3	7
AGL3_SEP4	2	0	4	0	8
AGL30	15	3	4	3	4
AGL32	1	0	0	3	1
AGL6	7	0	0	0	7
AGL65	22	3	3	6	9
AP1	10	2	0	7	0
ATGRP7	4	0	0	1	1
COL3	1	1	0	2	1
COL5	0	2	0	0	1
COP1	14	2	0	12	3
CRY1 A	14	2	5	0	6
CRY2 A	2	1	0	3	3
CRY3	8	0	2	7	8
DET1	1	0	0	0	0
ELF3	11	4	3	11	3
FD	1	2	1	2	0
FKF1_ADO3	14	5	2	11	4
GI Ex9to10_ A	9	1	0	7	13
GI Ex10to11_ A	5	2	3	3	0
GI Ex10to11_ B	0	2	1	0	2
GI Ex11to12_ A	3	0	2	4	3
GI Ex11to12_ B	14	2	1	4	3
HY6	17	2	0	11	5
LHY 1 /CCA1	5	2	7	1	0
LHY 2 /CCA1	4	0	0	1	2
PFT1	4	0	0	5	1
PHYA Contig 1	1	0	0	0	1
PHYB	15	2	3	1	4
PHYC	19	0	0	10	4
PHYE	0	0	0	0	0
PI 1	11	4	3	11	4
PI 2	3	1	2	0	3
PRR5	14	0	1	10	7
PRR7 Contig 1	4	0	1	1	2
PRR7 Contig 2	4	0	1	2	0
SPA4	1	0	1	0	0
TOC1	4	1	0	0	0
Intron Totals	269	48	50	142	120

Procedure Frequency, Generalized Linear Mixed Model with Poisson Regression, and Least Squared Means

```
libname cj 'C:\DATA_SPA\Carla Jo';

data cj.pathways;

   input ID $ pathway $ type $ snps;

   cards;

1  flower Exon 3

2  flower Exon 0

3  flower Exon 15

4  flower Exon 13

5  flower Exon 1

6  flower Exon 1

7  flower Exon 0

8  flower Exon 0

9  flower Exon 0

10  flower Exon 21

11  flower Exon 4

12  flower Exon 1

13  flower Exon 0

14  flower Exon 0

15  flower Exon 6

16  flower Exon 3

17  flower Exon 3

1  flower Intron 19

2  flower Intron 10
```

```
3  flower Intron 5
4  flower Intron 3
5  flower Intron 17
6  flower Intron 31
7  flower Intron 1
8  flower Intron 33
9  flower Intron 9
10  flower Intron 6
11  flower Intron 14
12  flower Intron 29
13  flower Intron 5
14  flower Intron 14
15  flower Intron 43
16  flower Intron 2
17  flower Intron 6
18 Clock Exon 1
19 Clock Exon 6
20 Clock Exon 0
21 Clock Exon 1
22 Clock Exon 0
23 Clock Exon 2
24 Clock Exon 9
25 Clock Exon 5
26 Clock Exon 10
27 Clock Exon 18
28 Clock Exon 0
29 Clock Exon 2
```

30 Clock Exon 2

18 Clock Intron 32

19 Clock Intron 36

20 Clock Intron 30

21 Clock Intron 13

22 Clock Intron 5

23 Clock Intron 24

24 Clock Intron 12

25 Clock Intron 32

26 Clock Intron 8

27 Clock Intron 7

28 Clock Intron 15

29 Clock Intron 7

30 Clock Intron 5

31 Photoreceptor Exon 1

32 Photoreceptor Exon 72

33 Photoreceptor Exon 6

34 Photoreceptor Exon 32

35 Photoreceptor Exon 7

36 Photoreceptor Exon 14

37 Photoreceptor Exon 31

38 Photoreceptor Exon 55

39 Photoreceptor Exon 79

40 Photoreceptor Exon 19

41 Photoreceptor Exon 30

31 Photoreceptor Intron 27

32 Photoreceptor Intron 38

```
33 Photoreceptor Intron 9
34 Photoreceptor Intron 10
35 Photoreceptor Intron 25
36 Photoreceptor Intron 25
37 Photoreceptor Intron 2
38 Photoreceptor Intron 1
39 Photoreceptor Intron 33
40 Photoreceptor Intron 0
41 Photoreceptor Intron 35
;
proc freq data=cj.pathways;
table pathway*type/chisq;
weight snps;
run;
proc glimmix  data=cj.pathways;
class pathway type ;
  model snps=pathway type type*pathway /dist=Poisson;
Lsmeans pathway type type*pathway/ pdiff plot=mean(sliceby=type join
cl);
Lsmeans  type*pathway/ pdiff plot=mean(sliceby=pathway join cl);
run;
```

APPENDIX E

Chapter IV Supplemental Material

Authors and Contact Information

Carla Jo Logan-Young[9§], John Z. Yu[10], Richard G. Percy[10],

Sara E. Duke[11], Alan E. Pepper[9§]

Email addresses:

CJLY: tysfira@tamu.edu, JZY: john.yu@ars.usda.gov,

RGP: richard.percy@ars.usda.gov, SED: sara.duke@ars.usda.gov,

AEP: apepper@bio.tamu.edu

[9] Department of Biology, Texas A&M University, College Station, Texas 77843, USA

[10] Cotton Germplasm Research Unit, United States Department of Agriculture, Agricultural Research Station, Southern Plains Agricultural Research Center, College Station, Texas 77845, USA

[11] Area Statistician, United States Department of Agriculture, Agricultural Research Station, Southern Plains Agricultural Research Area Office, College Station, Texas 77845, USA

[§]Corresponding author

Authors Contributions

CJLY and AEP designed the experiment. CJLY and AEP designed the GBS primers and TGBS primers. RGP provided material for photoperiod dependent and independent lines. RGP provided all accession information from the USDA-ARS-SPARC Cotton Germplasm collection. SED determined the correct statistical measures. CJLY, JZY, and AEP performed data interpretation and drafted the manuscript. All authors read and approved the final manuscript.