GENETIC VARIATION AND EVOLUTION OF THE SIZE OF NBS-LRR-ENCODING GENE FAMILY IN COTTON AND RELATED SPECIES (*GOSSYPIUM* L.)

A Thesis

by

YEN-HSUAN WU

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

MASTER OF SCIENCE

May 2009

Major Subject: Plant Breeding

GENETIC VARIATION AND EVOLUTION OF THE SIZE OF NBS-LRR-ENCODING GENE FAMILY IN COTTON AND RELATED SPECIES (*GOSSYPIUM* L.)

A Thesis

by

YEN-HSUAN WU

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

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Hongbin Zhang
Committee Members,	Wayne Smith
	Ruzong Fan
Head of Department,	David D. Baltensperger

May 2009

Major Subject: Plant Breeding

ABSTRACT

Genetic Variation and Evolution of the Size of NBS-LRR-Encoding Gene Family in Cotton and Related Species (*Gossypium* L.). (May 2009)

Yen-Hsuan Wu, B.S., National Taiwan University
Chair of Advisory Committee: Dr. Hongbin Zhang

Most of genes contained in a genome have been shown to exist in forms of families; however, little is known about their variation and evolution during the course of genome evolution. The present study shows that the numbers of the genes of the NBS-LRR-encoding gene family vary extremely significantly among different lines or cultivars of a species and among related species from the same genus. This suggests that plant genetics and evolution depend on not only gene sequence variation, but also the number of genes in multigene families. This study has further revealed that the variation of number of genes in the gene family in the *Gossypium* species is affected significantly not only by genome size variation, polyploidization and natural selection, but also by domestication/breeding. There is a positive correlation ($P \leq 0.05$) between genome size and number of genes in the family, suggesting that species with larger genomes tend to have more NBS-LRR-encoding genes. It was observed that natural polyploids have significantly larger numbers of genes in the family and larger genomes than the artificial polyploids of their putative diploid ancestors. This indicates that polyploidization, perhaps post-polyploidization as well, either led to the loss of the genes in a gene family or slowed the process of gene number increase after

polyploidization. It was shown that cultivated cottons have significantly more NBS-LRR-encoding genes than wild species at both diploid and polyploidy levels. This result indicates that plant breeding likely allows accumulation of NBS-LRR-encoding genes that potentially provide resistance to pathogens. Therefore, plant breeders have selected not only for favorable alleles and favorable allele combinations, but also for the number of genes. Finally, difference ($P \leq 0.001$) was found in number of genes in the NBS-LRR-encoding gene family among the species native to different geographical regions, suggesting that natural selection has played an important role in the variation in number of genes in the NBS-LRR-encoding gene family. The gene members that are favorable for fitness at the time are selected and accumulated in the genomes, but those that are not favorable for fitness at the time are lost in natural selection.

As this is the first study in the field, further studies remain. These include, but not limited to, the universality of the findings in plants and animals, the universality of the findings in different gene families, genetics of the gene family size variation, relationship between the gene family size variation and phenotypic variation, gene family size variation and breeding, etc. Nevertheless, the findings obtained from this study are sufficient to shed light on many fundamental questions in biology, diversity and complexity of plants and animals.

ACKNOWLEDGEMENTS

I would like to express my great appreciation to the committee chair of my graduate study, Dr. Hongbin Zhang, who gave me a good opportunity to work in this interesting research project. I am deeply grateful for his suggestions, supports, and encouragement during this research. I also want to thank the members of the committee, Dr. Wayne Smith and Dr. Ruzong Fan, for their valuable directions and supports in the study.

Special acknowledgement is also given to Dr. Meiping Zhang and Dr. Mi-Kyung Lee, for their wonderful team work that I have experienced in this research. Also, I am appreciative for all my colleagues in Dr. Hongbin Zhang's laboratory for their kind help and inspirations.

Finally, I would like to present my great appreciation to my parents, Mu-Tang Wu and Yin-Hua Chai, for their endless love and selfless support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Recent genome sequencing has revealed that most of genes contained in a genome exist in forms of families (Gibson and Muse 2004; Demuth et al. 2006) that are often defined groups of homologous genes that are likely to have similar functions. The underlying mechanism of the origin and evolution of a multigene family, especially the ribosomal RNA gene (rDNA) family, has been studied extensively at the nucleotide sequence level. Concerted evolution has been proposed and accepted widely to play an important role in the origin and evolution of a multigene family (Smith 1973; Zimmer et al. 1980; Irwin and Wilson 1990), in which the diversity and evolution of a multigene family is driven by unequal crossing over and gene conversion. Ota and Nei (1994) and Nei et al. (1997) proposed a "birth-and-death" model to explain the genetic diversity and evolution of the major histocompatibility complex (MHC) and immunoglobulin (*Ig*) gene families in the vertebrate immune system. In this model, new genes are created by repeated gene duplication. Some duplicated genes are maintained in the genome for a long time but others are deleted or become nonfunctional by deleterious mutations. Michelmore and Meyers (1998) used the birth-and-death model to explain the diversity and evolution of plant disease resistance gene clusters, especially those of the NBS-LRR-encoding gene family.

This thesis follows the style of Genetics.

Nevertheless, little is known about variation of the number of genes in a family within a species and between related species though it has been a common census that the total number of genes contained in a genome varies among diverged species. It is unknown how a multigene family evolves in term of family size, or the number of gene members in a family. Domestication, breeding, genome size variation, polyploidization, and natural selection represent major forces of crop plant genome evolution. Do and what kind of roles these activities play in the fate and evolution of a multigene family?

In this study, we have addressed these questions and provide the basic knowledge of genetic variation and evolution of the size of a multigene family using the NBS-LRR (nucleotide-binding site-leucine rich repeat)-encoding gene family, and cotton and related species (*Gossypium L.*) as experimental models. Particularly, we tested whether the NBS-LRR-encoding gene family is consistent or variable in family size among related species and populations of a species of *Gossypium*, and determined what drives the variation and evolution of the NBS-LRR-encoding gene family in cotton and related species: genome size change, polyploidization, domestication/breeding and/or natural selection.

Gossypium L. is an excellent model system for the proposed research. First, it consists of 50 species, including both diploids (2n = 26) (45) and polyploids (2n = 52) (5). Cytological studies grouped the *Gossypium* diploid species into eight genome groups, designated A through G and K, whereas all five polyploid species were shown to contain A and D genomes, designed $(AD)_1$ through $(AD)_5$. Molecular phylogenetic analyses clustered the *Gossypium* species into three major lineages: the D-genome

species lineage (13 species), the A+B+C+E+F+G+K-genome lineage (32 species) and the AD-genome species lineage (5 species) (Figure 1). These features allow studies in variation of number of genes in the family among related species and among different populations within a species and the impact of polyploidization on variation of the number of gene members of a multigene family. Second, the genome sizes of the *Gossypium* species vary dramatically (Hendrix and Stewart 2005; Figure 1). At the diploid level, the genome sizes vary from 880 Mb/1C in the D-genome species to 2,570 Mb/1C in the K-genome species, with a difference by nearly three fold. At the polyploidy level, the genome sizes vary from 2,350 to 2,490 Mb/1C with a difference of 140 Mb (5.8%). These allow studies of relationships between genome size variation and member number variation of a gene family. Third, the *Gossypium* species are geographically originated in three main continents of the globe: America, Africa-Asia and Australia (Figure 2). The D- and AD-genome species are endemic to America, especially Mexico, the A-, B-, E- and F-genome species are originated in Africa-Asia while the C-, G- and K-genome species are native to Australia. Moreover, of the 50 *Gossypium* species, 4 are cultivated and 46 are wild. At the diploid level, 2 of the 45 diploid species, *G. herbaceum* and *G. arboretum*, are cultivated while 43 are wild. At the polyploid level, 2 of the 5 polyploids, *G. hirstutm* and *G. barbadense*, are cultivated and 3 are wild. These facts provide us with a desirable tool to study the impact of domestication/breeding and natural selection on the variation of number of gene members in a multigene family.

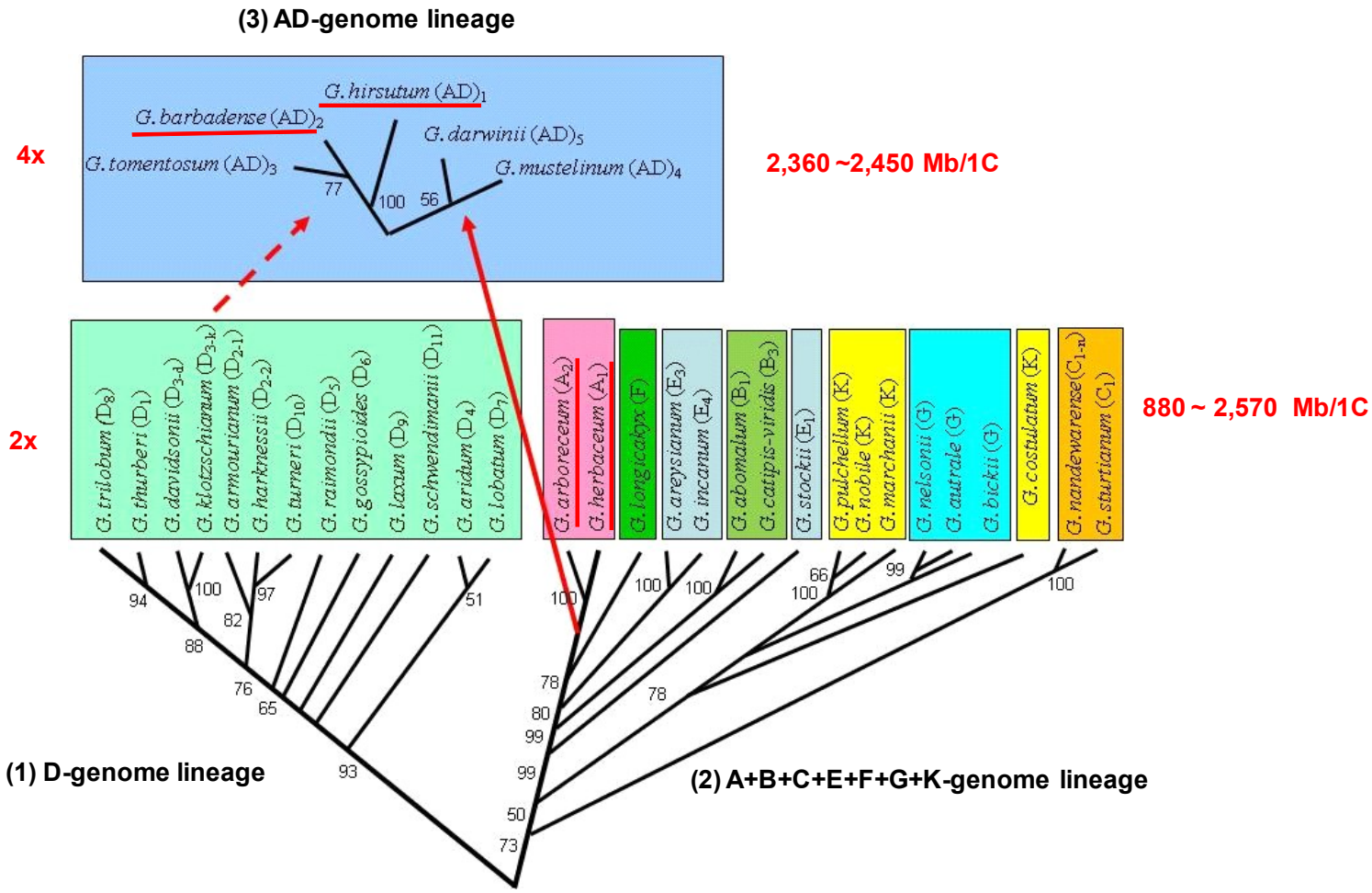

FIGURE 1.—Phylogeny and evolution of *Gossypium* species (from Ying et al. 2007). The number below each branch is the percentage of confidence calculated by bootstrap computation.

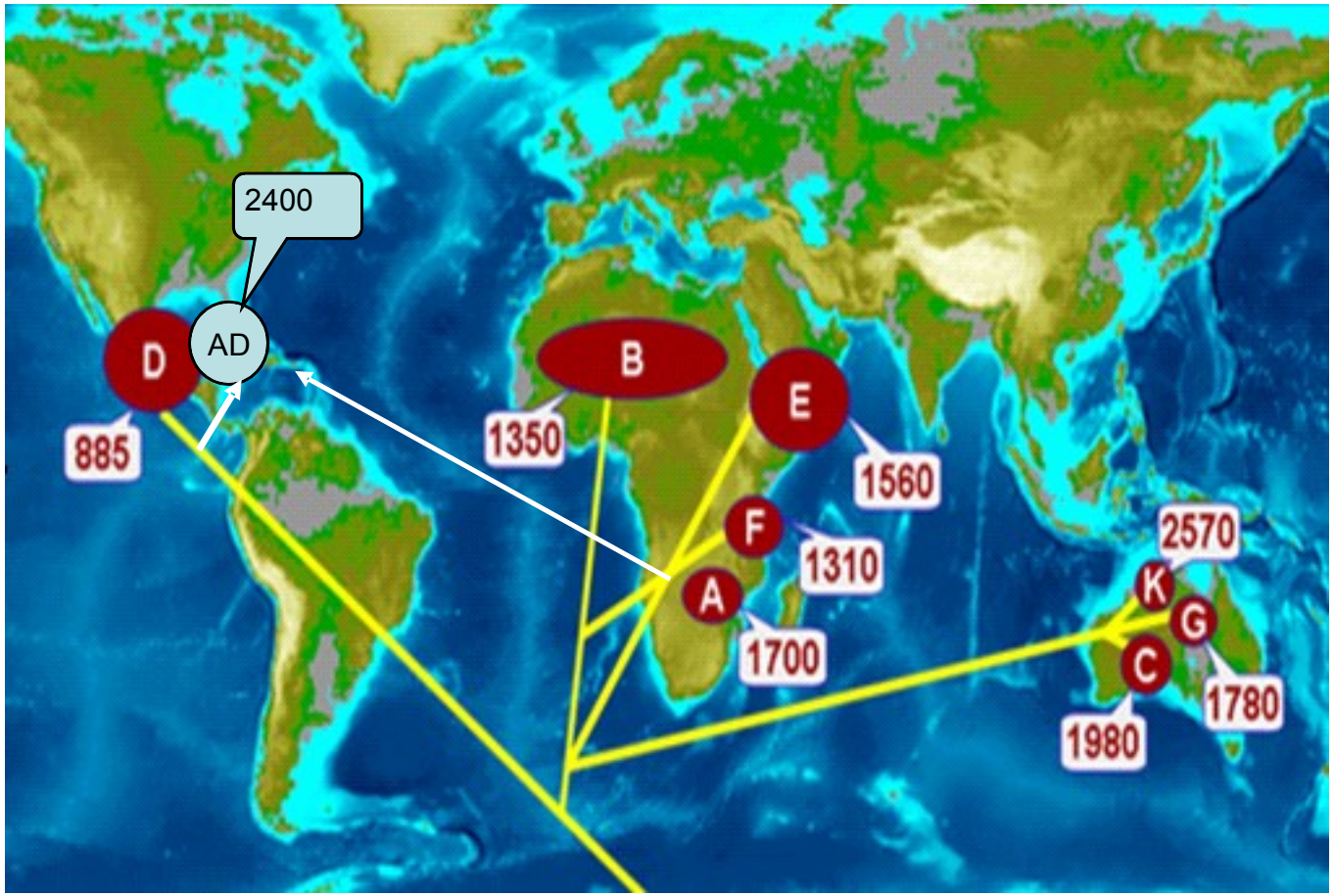

FIGURE 2.—Geographic distribution of the *Gossypium* species (down-load from http://www.eeob.iastate.edu/faculty/WendelJ/home.htm). The letters indicate the genome constitutions of the *Gossypium* species while the Arabic numbers indicate the sizes of the genomes.

The NBS-LRR-encoding gene family represents a large gene family in plants. For instance, whole-genome sequencing showed that there are ~150 NBS-LRR-encoding genes in the genome of Arabidopsis (145 Mb/1C) (Meyers et al. 2003) and ~480 NBS-LRR-encoding genes in the genome of rice (400 Mb/1C) (Zhou et al. 2004). The gene family, no matter in the genome of Arabidopsis or that of rice, all is distributed in different chromosomes and tend to cluster physically (Meyers et al. 2003; Zhou et al. 2004). It has also been shown that nearly 80% of the cloned genes conferring resistance to various plant pathogens, including bacteria, fungi, viruses and nematodes, were contributed by the NBS-LRR-encoding gene family (Takken and Joosten 2000). Therefore, the gene family has been subjected to both natural and artificial selections (domestication and breeding) during the course of crop plant evolution.

The NBS-LRR-encoding genes contained in the genome of polyploid cotton, *G. hirsutum* L., were previously cloned and characterized (He et al. 2004). It was shown that there is a large NBS-LRR-encoding gene family in the cotton genome (He et al. 2004). Phylogenetic analysis showed that the gene family has evolved rapidly and could be further categorized into 10 distinct subfamilies based on the similarities of their nucleotide sequences (Figure 3). As are the gene families in the genomes of Arabidopsis and rice (Meyers et al. 2003; Zhou et al. 2004), the NBS-LRR-encoding genes are distributed in different chromosomes and tend to cluster physically in the cotton genome (He et al. 2004). Therefore, the NBS-LRR-encoding gene family in the

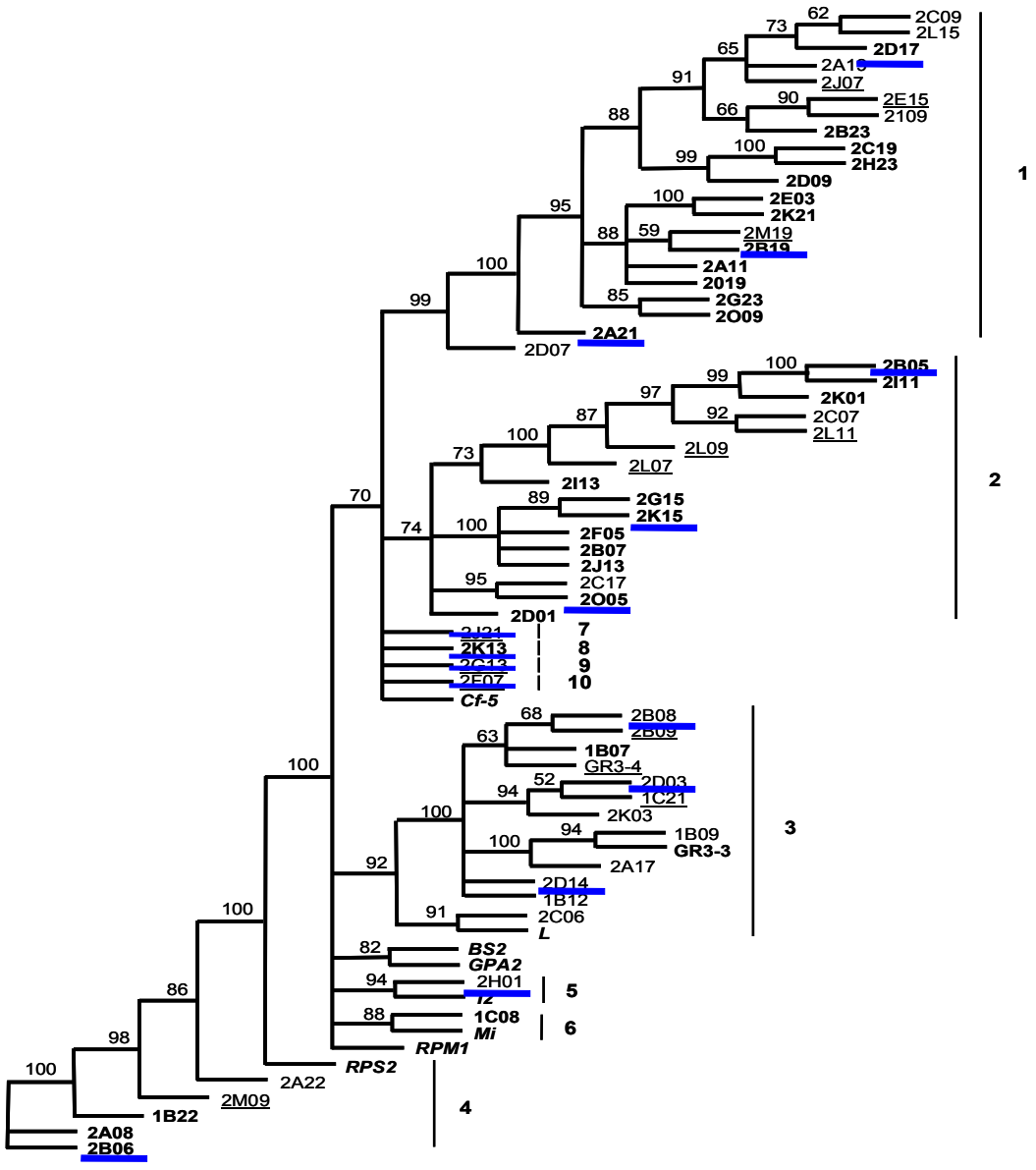

FIGURE 3.—The Phylogeny of the NBS-LRR-encoding gene family in cotton (from He et al. 2004). The numbers nearby the vertical lines indicate the subfamilies of the gene family whereas the gene members underlined indicate those used as probes to estimate the number of gene members in the gene family.

genome of cotton has provided a unique model for studies of fate and evolution of multigene families during the course of plant genome evolution.

CHAPTER II

MATERIALS AND METHODS

2.1 Plant materials and DNA isolation

Figure 4 shows that experimental flow chart of this study. A total of 96 cultivars or accessions of *Gossypium* species were used in this study (Table 1). Of the accessions or cultivars, 65 represent 30 diploid species and 31 do the five polyploid species of *Gossypium*, with each species having 1 – 12 accessions. These plant materials represent all eight genomes, all three geographical origins and all four cultivated species of the genus *Gossypium*.

The plants of each accession or cultivar were grown in a greenhouse, phenotypically verified during growth and development (Figure 5), and sampled for nuclear DNA isolation. Young leaves were collected from a single plant verified to represent its accession. Nuclear DNA was isolated with a modified CTAB (cetytrimethylammonium bromide) method that is routinely used in our laboratory. Nuclei were first isolated in the extraction buffer containing 350 mM sorbitol, 100 mM Tris, 5 mM EDTA, 0.38% (w/v) bisulfate, and then lysed to release nuclear DNA in the nuclei lysis buffer containing 0.2 M Tris.HCI, 50 mM EDTA, 2.0 M NaCI, and 2% (w/v) CTAB. The DNA was purified with a chloroform/isoamyl alcohol (24:1) mixture, collected by precipitation with iso-propanol and dissolved in TE containing 10 mM Tris.HCl and 1 mM EDTA, pH 8.0. Furthermore, RNA contaminated in the DNA was removed by treatment with RNase at 37^{o}C for 30 min, followed by extraction with an

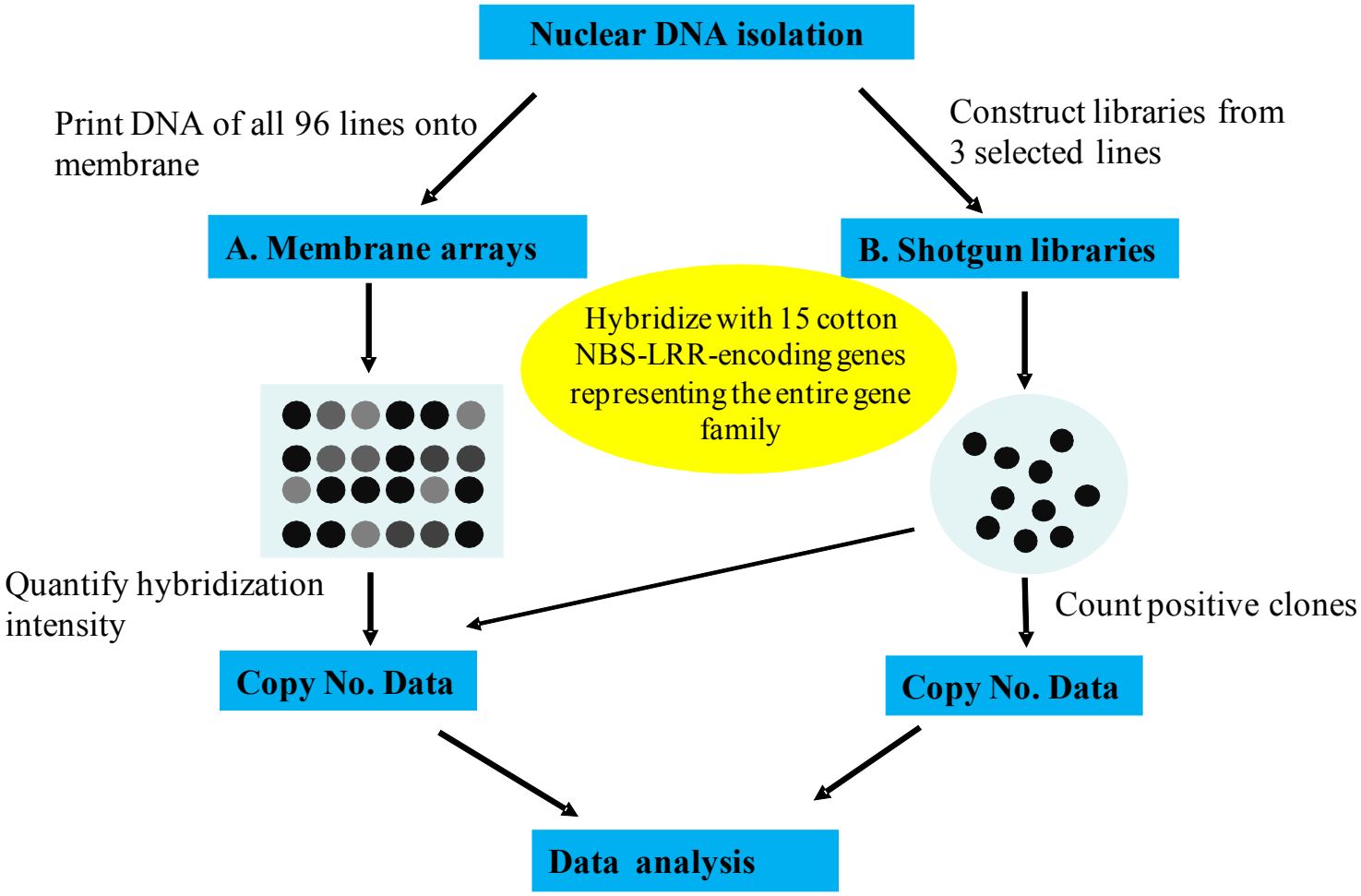

FIGURE 4.— Flow chart of the experimental plan to measure the number of genes in the NBS-LRR-encoding gene family in the genome of each accession or cultivar of the *Gossypium* species.

TABLE 1

***Gossypium* species used in this study**

No.	Species name	Genome	Accession/cultivar*	Origin
1	*G. sturtianum*	C1	C1-4	Australia
			C1-7	Australia
2	*G. nandewarense*	C1-n	C1-n-5	Australia
			C1-n-6	Australia
3	*G. costulatum*	K	C5-3	Australia
			C5-4	Australia
4	*G. nobile*	K	NWA35	Australia
5	*G. pulchellum*	K	C8-1	Australia
6	*G. marchantii*	K	NWA-6	Australia
7	*G. australe*	G	C3-1	Australia
			C3-4	Australia
8	*G. nelsonii*	G	C9-1	Australia
			C9-2	Australia
9	*G. bickii*	G1	G1-1	Australia
			G1-3	Australia
10	*G. thurberi*	D1	D1-1	Mexico
			D1-7	Mexico
11	*G. trilobum*	D8	D8-7	Mexico
			D8-8	Mexico
			D8-9	Mexico
12	*G. davidsonii*	D_{3d}	D_{3d}-1	Mexico
			D_{3d}-2	Mexico
13	*G. klotzchianum*	D_{3-k}	D_{3-k}-57	Ecuador
			D_{3-k}-58	Ecuador
			D_{3-k}-59	Ecuador
14	*G. armourianum*	D_{2-1}	D_{2-1}-7	Mexico
			D_{2-1}-9	Mexico
15	*G. harknessii*	D_{2-2}	D_{2-2}-4	Mexico
16	*G. turneri*	D10	D10-1	Mexico
			D10-2	Mexico
17	*G. aridum*	D4	D4-5	Mexico
18	*G. lobatum*	D7	D7-4	Mexico
			0208082.07	Mexico
19	*G. laxum*	D9	D9-3	Mexico
			0208021.08	Mexico
20	*G. schwendimanii*	D11	D11-1	Mexico
21	*G. gossypioides*	D6	D6-6	Mexico
			0208082.05	Mexico
22	*G. raimondii*	D5	D5-3	Peru
			D5-6	Peru
			D5-8	Peru

Table 1

(Continued)

No.	Species name	Genome	Accession/cultivar*	Origin
23	*G. herbaceum*	A1	A1-108	
			A1-111	
			A1-120	
			A1-127	
			A1-128	
			A1-129	
			A1-153	
			A1-154	
			A1-172	
			A1-180	
24	*G. arboreum*	A2	0208083.10	
			A2-142	
			A2-47	
			A2-84	
25	*G. anomalum*	B1	B1-1	Africa
			B1-7	Africa
26	*G. capitis-viridis*	B3	B3-1	Portugal
27	*G. longicakyx*	F1	F1-1	Tanzania
			F1-4	Tanzania
28	*G. stocksii*	E1	E1-3	Arabia
			E1-4	Arabia
29	*G. areysianum*	E3	E3-1	Arabia
30	*G. incanum*	E4	0208081.07	
			E4-4	
31	*G. hirsutum*	(AD)1	TM1	
			Wild Mexico Jack Jones	
			Clevewilt 6	
			Auburn 56	
			Stoneville 213	
			Coker 201	
			Coker 310	
			Deltapine 16	
			Deltapine 61	
32	*G. barbadense*	(AD)2	Pima S6	
			3-79	
			(AD)2-81	
			(AD)2-372	
			K101	
33	*G. tomentosum*	(AD)3	(AD)3-15	USA
			(AD)3-16	USA
			(AD)3-17	USA

Table 1

(Continued)

No.	Species name	Genome	Accession/cultivar*	Origin
33	*G. tomentosum*	(AD)3	(AD)3-25	USA
			0208081.05	
			(AD)3-26	
			(AD)3-1	
			(AD)3-3	
			(AD)3-4	
			(AD)3-5	
			(AD)3-7	
			(AD)3-11	
34	*G. mustelinum*	(AD)4	0208082.04	
			(AD)4-9	Brazil
			(AD)4-7	Brazil
35	*G. darwinii*	(AD)5	(AD)5-3	Ecuador
			(AD)5-7	Ecuador

* The seeds or plants were kindly provided and morphologically verified by Dr. Edward Percival, the curator of the Cotton Germplasm, USDA/ARS, College Station, Texas, or Dr. David M. Stelly, Texas A&M University, College Station, Texas.

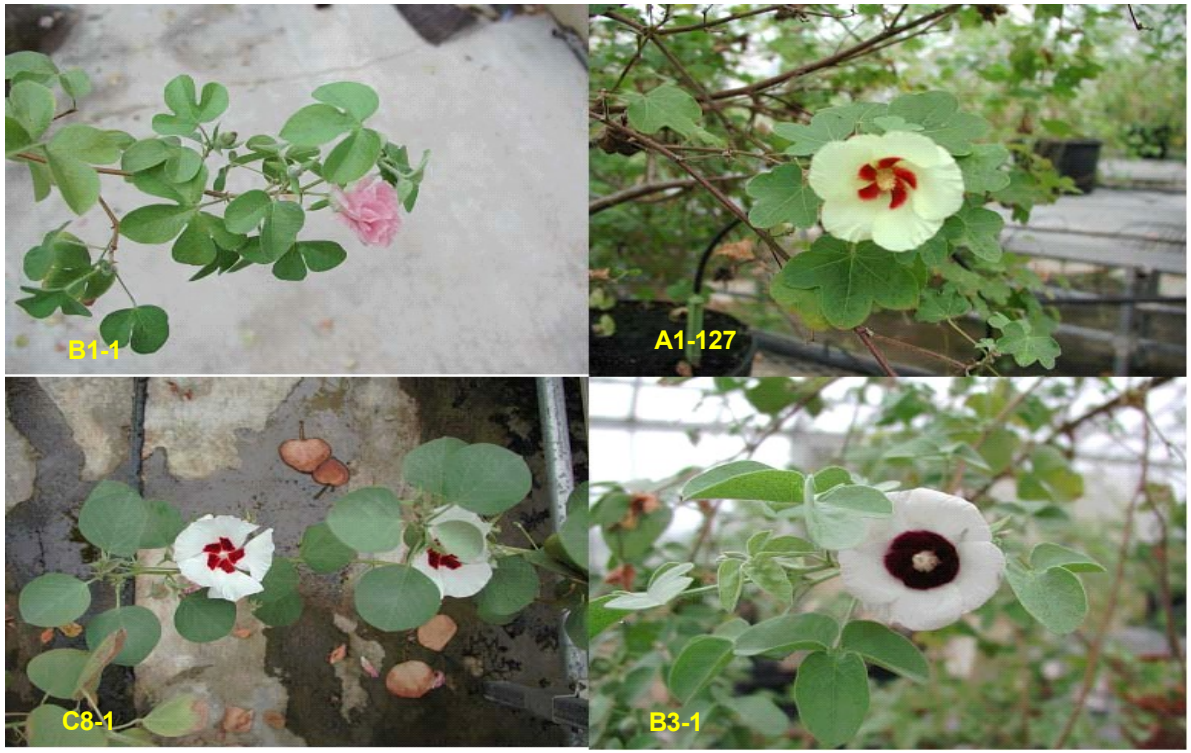

FIGURE 5.—Examples of the *Gossypium* species accessions or cultivars used in this study showing their identities.

equal volume of phenol, precipitation with ethanol and washing in 70% ethanol. DNA pellet was dried and dissolved in TE. The concentration of isolated DNA was determined with a fluorometer and by agarose gel electrophoresis.

2.2 Estimation of the number of genes in the NBS-LRR-encoding gene family in individual accession

Several methods have been used to estimate the number of genes in a multigene family in a genome. These include membrane array (e.g., Diaz et al. 2007), microarray (e.g., Chung et al. 2004), genome sequencing (e.g., Meyers et al. 2003; Zhou et al. 2004; Hawkins et al. 2006) and quantitative real-time PCR (qrtPCR) (e.g., Ferreira et al. 2006; Yi et al. 2008). For the membrane array method, arrays are fabricated by printing total genomic DNA or cDNA of target lines onto nylon supporting membrane and probed with overgos designed from target genes or their sequences. The copy number of the genes in a genome is quantified using a PhosphorImager. For the microarray method, arrays are fabricated by printing or synthesizing gene-specific oligos or large-insert DNA clones (such as bacterial artificial chromosome, BAC) on a chemically-coated glass slide and probed with total genomic DNA or cDNA. The copy number of a gene or sequence in a genome is quantified using a microarray analyzer. For the genome sequencing method, shotgun or other types of DNA libraries having certain insert sizes are constructed and all or a sample of the library clones are sequenced. The copy number of a gene family in a genome is calculated based on the number of clones sequenced, number of clones containing the target gene, clone insert

size and genome size. For the qrtPCR method, gene-specific probes and/or primers are designed and used to amplify the DNA of a target genome. The copy number of a gene in the genome is determined based on the fluoresce intensity.

Sensitivity/accuracy, capacity, reliability/reproducibility, throughput and economy are used to evaluate the methods. In comparison, it seems from previous studies (Meyers et al. 2003; Chung et al. 2004; Zhou et al. 2004; Hawkins et al. 2006; Ferreira et al. 2006; Diaz et al. 2007; Yi et al. 2008) that all four methods have a sensitivity that allows to determine a single-copy of a gene in a genome and a reasonable reliability; however, they have different advantages in other aspects. For instance, the membrane array method is readily fabricated and economical, without need for expensive instruments, and could be re-used for many times, but lower in throughput than the microarray method while it is much higher than the sequencing method. The membrane array method is also especially suited for estimation of the number of gene members in a family that has a high copy number, ranging from a single to million copies, whereas the microarray and qrtPCR methods are often used to estimate the copy number of genes that are small in the genome such as a few copies. Nevertheless, the results of all membrane array, microarray and qrtPCR methods could be significantly influenced by the nucleotide sequence identity among the members of genes and hybridization or PCR stringency, especially the qrtPCR method. The sequencing method has no such problem, but it is expensive to sequence a large number of genomes even though the new-generation high-throughput sequencing technologies are used.

Furthermore, we previously studied the reliability of the membrane array method using the DNA of two rice cultivars, Nipponbare and Teqing, with 8 replicates and five plants from each cultivar (unpublished). Statistical analysis of the results showed that no significant variation was detected in the number of genes constituting the NBS-LRR-encoding gene family among the replicates and individual plants of each cultivar, suggesting that the membrane array method is highly reliable and reproducible. Moreover, we have also modified the genome sequencing method by taking the advantages of its accuracy and high reliability in estimation of the copy number of genes, but reducing its cost by simply screening its shotgun libraries with overgos designed from target genes or their sequences as probes. Therefore, the membrane array method and the library screening method were both used in this study to estimate the number of genes of the NBS-LRR-encoding gene family in each accession of the *Gossypium* species.

2.2.1. *Membrane array method*

DNA of 0.32, 0.64 and 0.96 μg were printed per dot onto Hybond N+ membrane using a dot array blotting apparatus for D-genome species, A-, B-, C-, E-, F-, G- and K-genome species, and AD-genome species, respectively. The difference in amount of DNA printed per dot was applied to minimize the influence of genome size and/or ploidy level on the estimation of the number of genes. Moreover, to remove the potential noise background and infer the copy number of the genes, two control groups were included in the arrays. The first group contained 15 NBS-LRR-encoding genes of

known copy number as the positive controls and references to estimate the copy number of the gene family in each genome. The second group contained the printing buffer and non-homologous (salmon sperm) DNA as the negative controls to remove the noise background. Six sets of arrays were fabricated for experimental replicates (Figure 4).

Furthermore, to minimize the influence of gene member sequence variation on estimation of the gene member number of the NBS-LRR-encoding gene family in each line of the *Gossypium* species, we randomly chose 15 cotton NBS-LRR-encoding genes that represent all 10 subfamilies of the gene family, with 1 – 3 genes per subfamily depending on the sizes of subfamilies (Figure 3). Subfamily-specific primers were designed from the NBS-LRR-encoding regions of the genes and used to amplify the genes using the *G. hirsutum* genetic standard cultivar, Texas Marker-1 (TM-1), as a template. The expected PCR products (~650 bp) of the genes were purified on agarose gels and used as probes to estimate the number of gene members of the family.

The membrane array hybridization experiment was conducted with six technical replications using the combined PCR products of the 15 NBS-LRR-encoding genes as a probe. Biological replicates or multiple plants per accession was not included in the experiment because our previous study in rice (unpublished) showed that there was no significant variation in the number of gene members in the NBS-LRR-encoding gene family among different plants of a cultivar. To minimize the potential influence of the hybridization stringencies on the copy number estimation, we tested different hybridization stringencies, especially the washing stringencies (see below). The

PhosphoImager Bio-Imaging Analyzer BAS-1800II was used to quantify the hybridization intensity of each accession with the probe. The number of genes in the family in the genome of each accession was estimated by comparison with the positive controls of known copy number that were printed on the same membrane array.

2.2.2 *Library screening*

To further verify the membrane array results and to provide additional references for estimation of the number of genes in the family in the genome of each accession, we also estimated the number of genes in the family in the genomes of *Gossypium* using a second method, i.e., the library screening method. Three species were selected from the phylogenetic tree (Figure 1) to reduce the amount of the works, but they are able to represent the species of the genus. Therefore, we selected one species from each of the three major lineages of the species: *G. herbaceum* from the A+B+C+E+F+G+K-genome lineage and *G. raimondii* from the D-genome lineage at the diploid level, and *G. hirsutum* from the AD-genome lineage at the polyploid level. Two types of DNA libraries, shotgun-like and regular, were constructed from the DNA of the three genotypes to provide a reasonable representation for the genome. The shotgun-like DNA libraries were constructed by partially digested the DNA with three 4-bp, blunt-ended restriction enzymes, *Alu*I, *Hae*III and *Rsa*I, simultaneously, so that the resultant libraries would have a genome coverage as a shotgun library constructed from DNA fragments physically sheared. The partially digested DNA was selected on an agarose gel, and the DNA fragments that best reflected the average size of cotton genes were

selected and cloned in the *Eco*RV site of the pGEM5 vector. The regular DNA libraries were constructed by partially digested the DNA with one 4-bp restriction enzyme, *Mbo*I, size-selected on an agarose gel and cloned in the *Bam*HI site of the pUC18 vector. The titers, percentages of clones containing inserts and insert sizes of the libraries were determined by plating on agar selective medium and insert analysis of random clones on agarose gels.

The libraries were blotted onto the Hybond-N+ nylon membrane and screened by hybridization, as described above, using the same 15 selected cotton NBS-LRR-encoding genes as a probe that represent the entire gene family. The actual number of positive clones were counted and used to calculate the number of gene members in the NBS-LRR-encoding gene family in the genomes of the genotypes (Figure 4). From this experiment we expected to get a second estimation of copy number of the genes in the three accessions, which were used to further verify the results of the membrane array analysis and provide additional references for each lineage in estimation of the number of gene members in the family.

2.3 Data analysis

2.3.1 *Variation of number of genes in the NBS-LRR-encoding gene family among species and within a species*

Since the membrane array experiment was replicated for six times, ANOVA and linear regression could be used to determine the variation of number of gene members in the gene family among the *Gossypium* species and within a species. The actual number

of genes of the family or its $\log_{10}$-transformed number of genes in each accession were analyzed using ANOVA in the statistical program SPSS.

2.3.2 *Impact of genome size change on the number of genes in the NBS-LRR-encoding gene family*

Because the genome sizes of most of the *Gossypium* species are available and vary significantly (Hendrix and Stewart 2005; see Figure 1) at and between the diploid and polyploid levels, the relationships between the variation of number of genes in the family identified among the *Gossypium* species and the genome size variation could be determined. Therefore, we tested whether there is any correlation between the genome size and the number of genes in the NBS-LRR-encoding gene family at the diploid level, and among all the *Gossypium* species studied using a linear regression model in the SPSS statistical program.

2.3.3 *Impact of polyploidization on the number of genes in the NBS-LRR-encoding gene family*

As indicated above, the *Gossypium* genus consists of diploid and tetraploid species (for review, see Zhang et al. 2008). The tetraploid species, including *G. hirsutum* $(AD)_1$, *G. barbadense* $(AD)_2$, *G. tomentosum* $(AD)_3$, *G. mustelinum* $(AD)_4$ and *G. darwinii* $(AD)_5$, all contain A and D genomes. Previous studies showed that they were originated from the A-genome species and D-genome species sometimes 1 – 2 million years ago (for review, see Zhang et al. 2008). These allowed determining the

relationships between the variation in number of genes in the family and polyploidization. Although it is most accepted that the A genome of the polyploid species was contributed by *G. herbaceum* (A1) and/or *G. arboreum* (A2), while the D genome of the polyploid species by *G. raimondii* (D5), strong arguments exist (e.g., Rong et al. 2007). Therefore, we conducted T-test between the number of genes in the five polyploid species (AD) and the artificial sums of gene number of the average of two A-genome species and all thirteen D-genome species (artificial A+D) to infer the impact of polyploidization on the genetic variation of number of genes in the NBS-LRR-encoding gene family. Since the rates of the gene family size evolution may have varied significantly among individual species during evolution, the mean sum of the A- and D-genome species would provide a reasonable representation in the evolutionary rate of the gene family size in the species and thus could be used to make comparison in the rate of the gene family size evolution in the polyploid species. The results would provide knowledge about whether the number of NBS-LRR-encoding genes have been gained or lost after polyploid cottons originated. Again, the analysis was performed by t-test using the SPSS statistical programs.

2.3.4 *Impact of domestication/breeding on the number of genes in the NBS-LRR-encoding gene family*

The *Gossypium* species consists of both cultivated and wild species, two, *G. herbaceum* (A1) and *G. arboreum* (A2), at the diploid level, and two, *G. hirsutm* and *G. barbadense*, at the polyploid level. This provided us a possibility of estimating the

impact of domestication/breeding on the number of genes in the NBS-LRR-encoding gene family. Moreover, considering the impact of genome size variation on the number of genes in the family, we selected *G. tomentosum* [$(AD)_3$], *G. mustelinum* [$(AD)_4$] and *G. darwinii* [$(AD)_5$] that have similar genome sizes as the wild polyploid reference species and the E- and G-genome species that have similar genome sizes as the diploid wild reference species for the analysis. The differences in number of genes between the cultivated and wild species were subjected to t-test using the SPSS statistical programs.

2.3.5 *Impact of natural selection on the number of genes in the NBS-LRR-encoding gene family*

As shown in Figure 2, the *Gossypium* species were originated in three continents, America, Africa-Asia and Australia; therefore, it is expected that they are significantly different in growing climates and environments. These environmental differences may have resulted in different pressures of natural selection on species evolution. Therefore, we grouped the *Gossypium* species into three groups according to their geographical origin, American group including D- and AD-genome species, Africa-Asian group including A-, B-, E- and F-genome species, and Australian group including C-, G- and K-genome species, and subjected to statistical analysis in the variation of numbers of genes in the family among the species native to the three geographical regions. Moreover, considering the ploidy level of the polyploid species and that the A-genome species are cultivated, the AD- and A-genome species were excluded. Since there is a significant correlation between the number of genes in the family and genome size, the

number of genes in the family in each accession or cultivar were first modified according to their correlation formulas in diploid species and then subjected to ANOVA using the SPSS statistical programs.

CHAPTER III

RESULTS

3.1 Estimation of the number of genes in the NBS-LRR-encoding gene family in individual accession or cultivar of each *Gossypium* species

To estimate the number of genes in the NBS-LRR-encoding gene family in the genome of each accession of the *Gossypium* species (Table 1), the membrane array and library screening methods were used. The results are presented in following tables.

3.1.1 *Membrane array method*

For the membrane array experiment, we conducted six replicates for each DNA sample. Figure 6 shows an example of the membrane array hybridization with the 15 NBS-LRR-encoding genes representing the gene family. The hybridization intensity of each accession DNA spot was quantified; and the hybridization noise background was removed using the negative controls of the printing buffer and non-homologous DNA. Then, we compared the net hybridization intensity of each accession with those of the 15 NBS-LRR-encoding gene positive controls of known copy number to estimate the copy number of the gene family in each genome. Table 2 shows the six-replicate mean number of genes in the gene family in the genome of each accession studied. Among species with the same genome (such as D), the numbers of genes in the NBS-LRR-encoding gene family varied from 88 in *G. schwendimanii* (D11) to 1,710 in *G. trilobum* (D8), with a difference of nearly 20 fold at the diploid level. At the polyploidy

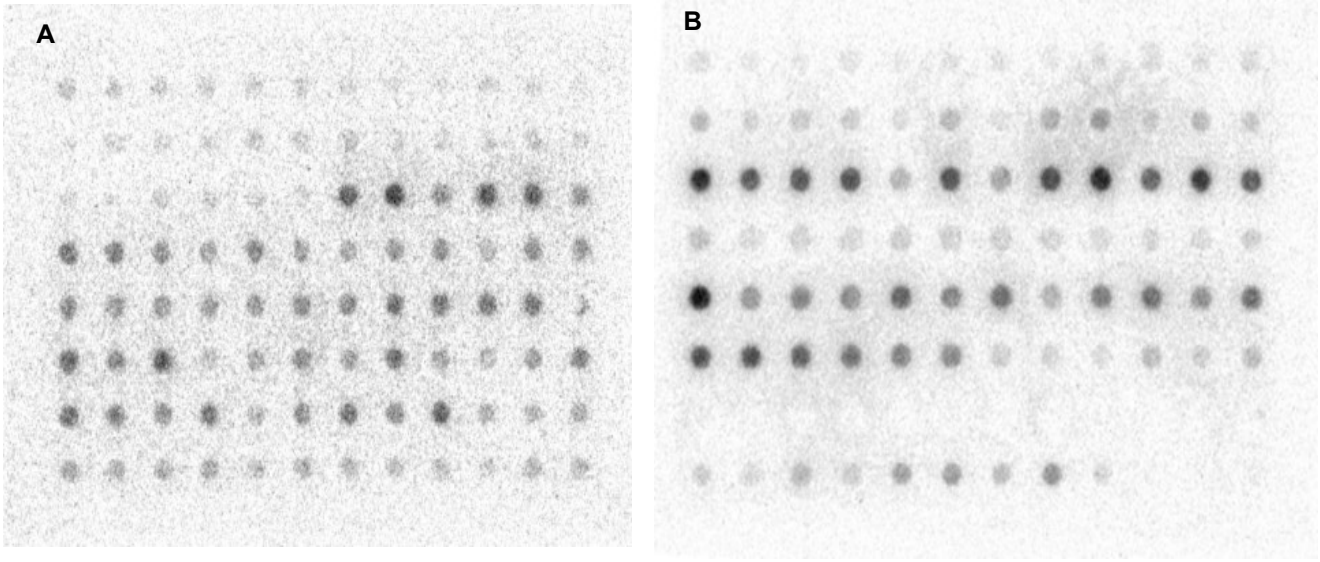

FIGURE 6.—Example of the membrane arrays of Gossypium species nuclear DNA hybridized with 15 NBS-LRR-encoding genes representing the entire gene family (see Figure 3). The 15 NBS-LRR-encoding genes of known copy No. were used as the positive controls and references to estimate the number of genes in the gene family in each accession genome. The printing buffer and non-homologous fish DNA were used as the negative controls to remove the noise background.

TABLE 2

Numbers of gene members of NBS-LRR-encoding gene family in the different accessions or cultivars of *Gossypium* species estimated by membrane array

Species	Genome	Accession /cultivar	Acc. or cultivar mean [a] (No. /C)	Species mean (No. /C)	Genome mean (No. /C)
G. sturtianum	C1	C1-4	170.04		
		C1-7	271.52	220.78	
G. nandewarense	C1-n	C1-n-5	330.77		
		C1-n-6	747.96	539.37	380.07
G. costulatum	K	C5-3	406.77		
		C5-4	319.14	362.96	
G. nobile	K	NWA35	397.82	397.82	
G. pulchellum	K	C8-1	693.48	693.48	
G. marchantii	K	NWA-6	383.87	383.87	459.53
G. australe	G	C3-1	428.97		
		C3-4	515.90	472.44	
G. nelsonii	G	C9-1	516.07		
		C9-2	353.15	434.61	
G. bickii	G1	G1-1	336.95		
		G1-3	388.57	362.76	423.27
G. thurberi	D1	D1-1	281.48		
		D1-7	765.52	523.50	
G. trilobum	D8	D8-7	1167.19		
		D8-8	1999.34		
		D8-9	1963.52	1710.02	
G. davidsonii	D_{3d}	D_{3d}-1	110.99		
		D_{3d}-2	70.20	90.60	
G. klotzchianum	D_{3-k}	D_{3-k}-57	174.81		
		D_{3-k}-58	132.44		
		D_{3-k}-59	62.55	123.27	
G. armourianum	D_{2-1}	D_{2-1}-7	187.42		
		D_{2-1}-9	452.14	319.78	
G. harknessii	D_{2-2}	D_{2-2}-4	816.26	816.26	
G. turneri	D10	D10-1	1017.46		
		D10-2	802.19	909.83	

TABLE 2

(Continued)

Species	Genome	Accession /cultivar	Acc. or cultivar mean [a] (No. /C)	Species mean (No. /C)	Genome mean (No. /C)
G. aridum	D4	D4-5	185.65	185.65	
G. lobatum	D7	D7-4	159.49		
		0208082.07	112.66	136.08	
G. laxum	D9	D9-3	210.24		
		0208021.08	169.22	189.73	
G. schwendimanii	D11	D11-1	88.39	88.39	
G. gossypioides	D6	D6-6	269.81		
		0208082.05	269.91	269.86	
G. raimondii	D5	D5-3	412.10		
		D5-6	550.13		
		D5-8	326.94	429.72	445.59
G. herbaceum	A1	A1-108	425.83		
		A1-111	1361.61		
		A1-120	1184.23		
		A1-127	786.31		
		A1-128	449.68		
		A1-129	268.56		
		A1-153	511.19		
		A1-154	1370.49		
		A1-172	1179.37		
		A1-180	1465.77	900.30	
G. arboreum	A2	0208083.10	862.06		
		A2-142	1147.00		
		A2-47	881.51		
		A2-84	1720.13	1152.68	1026.49
G. anomalum	B1	B1-1	279.41		
		B1-7	255.02	267.22	
G. capitis-viridis	B3	B3-1	430.68	430.68	348.95
G. longicakyx	F1	F1-1	224.70		
		F1-4	155.75	190.23	190.23
G. stocksii	E1	E1-3	386.28		

TABLE 2

(Continued)

Species	Genome	Accession /cultivar	Acc. or cultivar mean [a] (No. /C)	Species mean (No. /C)	Genome mean (No. /C)
G. stocksii	E1	E1-4	393.31	389.80	
G. areysianum	E3	E3-1	328.16	328.16	
G. incanum	E4	0208081.07	314.28		
		E4-4	247.17	280.73	332.89
G. hirsutum	(AD)1	TM1	1713.07		
		Wild Mexico	2160.79		
		Jack Jones			
		Clevewilt 6	885.55		
		Auburn 56	1493.21		
		Stoneville 213	1479.36		
		Coker 201	758.24		
		Coker 310	1419.06		
		Deltapine 16	1182.92		
		Deltapine 61	1037.18	1347.71	
G. barbadense	(AD)2	Pima S6	723.89		
		3-79	1005.91		
		(AD)2-81	573.70		
		(AD)2-372	659.08		
		K101	588.56	710.23	
G. tomentosum	(AD)3	(AD)3-15	549.74		
		(AD)3-16	620.50		
		(AD)3-17	718.36		
		(AD)3-25	741.97		
		0208081.05	762.63		
		(AD)3-26	628.8		
		(AD)3-1	933.96		
		(AD)3-3	793.65		
		(AD)3-4	942.22		
		(AD)3-5	974.52		
		(AD)3-7	934.01		
		(AD)3-11	1040.68	803.42	

TABLE 2

(Continued)

Species	Genome	Accession /cultivar	Acc. or cultivar mean [a] (No. /C)	Species mean (No. /C)	Genome mean (No. /C)
G. mustelinum	(AD)4	0208082.04	1359.47		
		(AD)4-9	1027.31		
		(AD)4-7	1672.38	1353.05	
G. darwinii	(AD)5	(AD)5-3	562.42		
		(AD)5-7	739.11	650.77	973.04

"*" indicates the mean of six replicates.

level, the variation among species all having AD genomes ranged from 650 in *G. darwinii* [(AD)5] to 1347 in *G. hirsutum* [(AD)1], with a difference of 2 fold. The variation of the number of genes in the family was much larger among diploid species than that among polyploidy species. This may be due to the fact that the polyploid species originated much later than the diploid species in the course of evolution. Alternatively, the process of the cotton polyploidization has promoted the variation of number of genes in the family. Within a species, significant variation was also observed in the number of genes in the family, even though only 3 - 12 accessions or cultivars were analyzed for each species. The within species variation ranged up to more than 5 fold within the species, *G. herbaceum* (A1).

3.1.2 *Library screening*

The sensitivity, accuracy, and reproducibility of membrane array method have been previously examined extensively (see the Materials and Methods). In this study, we further verified the copy number data obtained from the membrane array assay using another method, i.e., the library screening method (Table 3). We constructed 1 or 2 partially-digested libraries with three or four 4-bp restriction enzymes from the DNA of the three selected species representing the three major lineages of the *Gossypium* species, *G. herbaceum* (A1-120), *G. raimondii* (D5-8) and *G. hirsutum* (TM-1, AD1). This would provide a reasonable representation of the libraries for the genome. The insert sizes of the libraries were determined by analyzing 50 – 100 random clones from each library on agarose gels. The result showed that the libraries had average insert

TABLE 3

Number of genes in the NBS-LRR-encoding gene family estimated by library screening using the 15 NBS-LRR-encoding genes representing the entire gene family

Description	Library				
	G. herbaceum (A1-120)		*G. raimondii* (D5-8)	*G. hirsutum* (TM-1)	
	pGEM5	pUC18	pGEM5	pGEM5	pUC18
No. of clones screened	3,600	2,760	6,210	3,771	6,405
No. of clones with inserts	32,40 (90.00%)	1,756 (63.63%)	4,916 (79.17%)	3,186 (84.00%)	5,124 (80.00%)
Average insert size (bp)	4,748	2,894	5,105	6,840	4,067
Genome coverage (Mb)	15.38	5.08	25.10	21.80	20.84
Genome coverage (%)	0.932	0.299	2.852	0.899	0.859
No. of positive clones	14	2	20	11	16
Total:					
Genome coverage (Mb)	20.47		25.10	42.63	
Genome coverage (%)	1.231		2.852	1.758	
No. of positive clones	16		20	27	
No. of the genes in the genome	1,326.66		701.24	1,535.74	

sizes ranging from 2,894 – 6,840 bp. These insert sizes seemed to represent the gene sizes of the NBS-LRR-encoding gene family. The numbers of clones, 4,916 to 8,310, were randomly selected from the libraries of each genotype, blotted onto nylon membrane and screened with 15 NBS-LRR-encoding genes representing the entire family. A total of 16 – 27 positive clones were obtained for each genotype. Therefore, 1,326, 701, and 1,535 were estimated to be contained in the genomes of *G. herbaceum* A1-120, *G. raimondii* D5-8 and *G. hirsutum* TM-1, respectively. While the numbers of genes are being subjected to further verification by additional library screening, the gene numbers, 1,326, 701, and 1,535, of *G. herbaceum* A1-120, *G. raimondii* D5-8 and *G. hirsutum* TM-1 are very close to those of the accessions estimated by the membrane array method. This result has further confirmed the gene number results obtained from the membrane array analysis.

3.2 Variation in number of genes in the NBS-LRR-encoding gene family among species and within a species

To further confirm the variation of the number of genes in the gene family among species and within a species observed (Table 2), they were subjected to ANOVA among all species, among diploid species, among polyploid species, and among accessions within a species. Since the number of genes in the species were away from a normal distribution, we first transformed the number of genes in each accession to $\log_{10}$ *100 for ANOVA. The variation among all species, diploid species and polyploid species was all significant ($P \leq 0.001$) (Table 4). Within a species, the variation ($P \leq$

TABLE 4

Variation of Log_{10}-transferred number of genes in the NBS-LRR-encoding gene family among *Gossypium* species.

Species	No. of Species	d.f.	F value	*P*
All *Gossypium* species	35	569	18.045	≤ 0.0001 ***
Diploid species	30	371	9.162	≤ 0.0001 ***
Polyploid species	5	197	18.600	≤ 0.0001 ***

[a] "***" indicates that the numbers of genes in the family among the species studied are different at a significance level of $P \leq 0.001$.

0.05 or $P \leq 0.001$) was observed among different cultivars of each of the four cultivated species, *G. hirsutum* (AD_1), *G. barbadense* (AD_2), *G. herbaceum* (A_1), *G. arboretum* (A_2), although no significant variation was observed among the accessions of other species analyzed (Table 5).

3.3 Impact of genome size change on the number of genes in the NBS-LRR-encoding gene family

What evolutionary force directed the variation in the number of genes in the NBS-LRR-encoding gene family among species and within a species? To answer that question, we analyzed the membrane array data by linear regression for genome size variation and number of genes in the NBS-LRR-encoding gene family at the diploid and polyploid level (Figure 7), and at diploid level only (Figure 8). At the diploid and polyploid level, the correlation coefficient value was 0.442 ($P \leq 0.001$). At the diploid species level only, the correlation coefficient value was 0.299 ($P \leq 0.05$). The results suggest that the variation in the number of genes in the gene family is significantly and positively correlated with genome size, i.e., the larger genomes tend to have more NBS-LRR genes.

3.4 Impact of polyploidization on the number of genes in the NBS-LRR-encoding gene family

Since it has been accepted that polyplodization has played an important role in the variation of number of genes in a species, we also estimated the roles of

TABLE 5
Variation of Log_{10}- transferred number of genes in the NBS-LRR encoding gene family among different accessions or cultivars of a species

Species	Genome	No. of lines	d.f.	F value [a]	*P*
G. hirsutum	AD_1	9	53	6.114	≤ 0.0001 ***
G. barbadense	AD_2	5	29	3.018	0.037 *
G. herbaceum	A_1	10	59	19.395	≤ 0.0001 ***
G. arboreceum	A_2	4	23	9.326	≤ 0.0001 ***
G. tomentosum	AD_3	12	77	1.404	0.192
G. mustelinum	AD_4	3	17	2.312	0.133
G. darwinii	AD_5	2	17	2.038	0.173
G. trilobum	D_8	3	16	0.026	0.974
G. klotzchianum	$D_{3\text{-}k}$	3	14	0.168	0.847
G. armourianum	$D_{2\text{-}1}$	2	10	0.245	0.632
G. gossypioides	D_6	2	11	0.429	0.527
G. raimondii	D_5	3	17	0.302	0.744

[a] "*" and "***" indicates that the numbers of genes in the family among the accessions or cultivars studied within a species are different at a significance level of $P \leq 0.05$ and $P \leq 0.001$, respectively.

$$y = 13.51x + 220.66 \quad (r = 0.442, p \leq 0.0001)$$

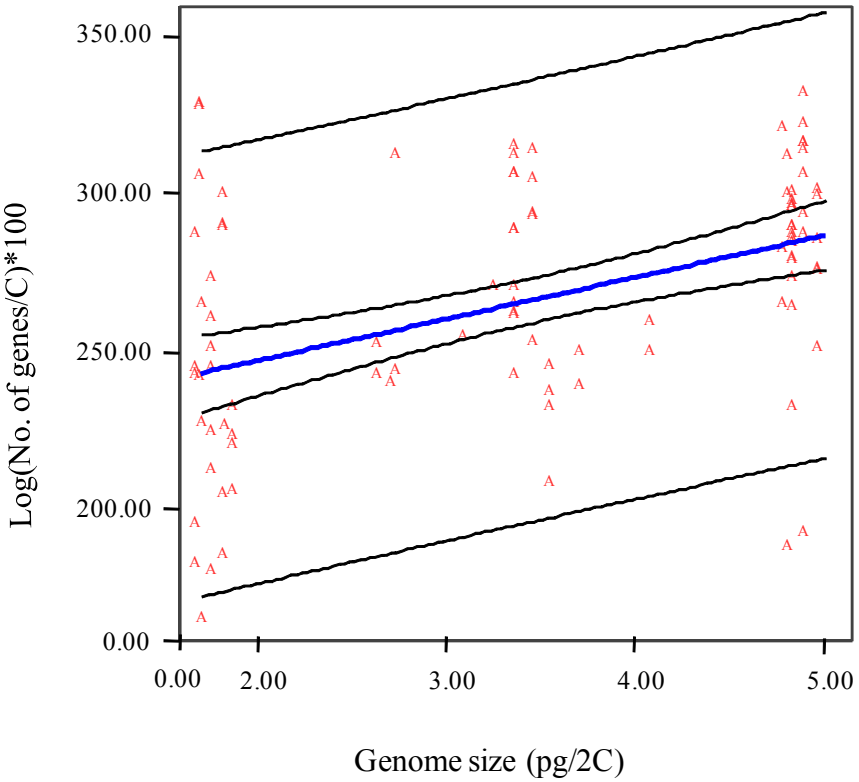

FIGURE 7.—Linear regression between genome size and number of genes in the NBS-LRR gene family inferred from all species.

$$y = 14.08x + 219.36 \quad (r = 0.299, p = 0.02)$$

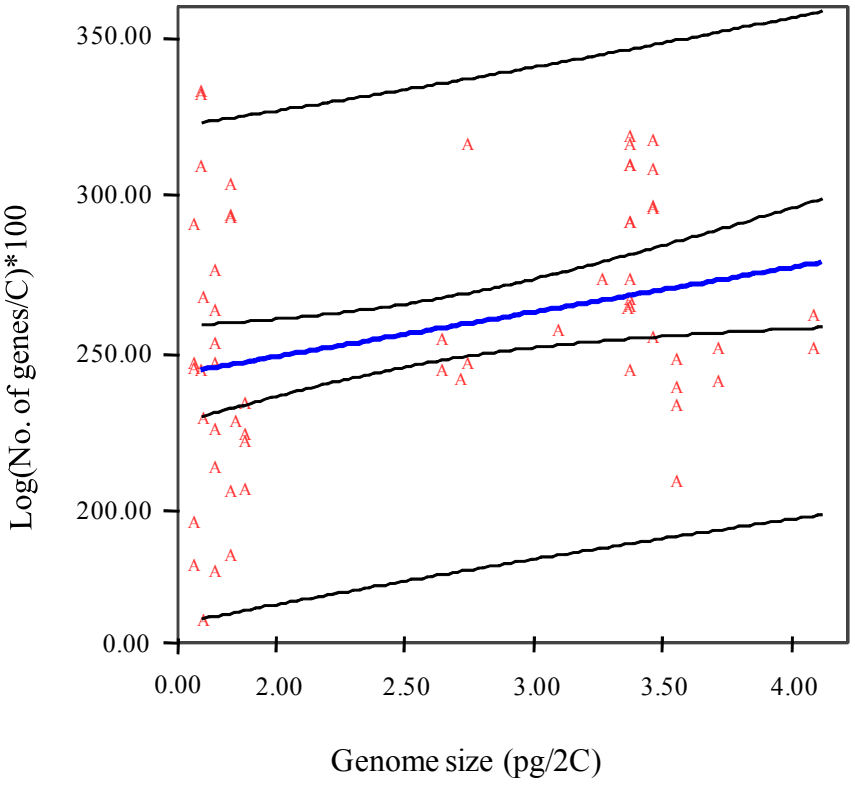

FIGURE 8.—Linear regression between genome size and number of genes in the NBS-LRR gene family inferred from diploid species only.

polyploidization in the variation of number of genes in the gene family among the species. The AD-genome polyploid species were proposed to originate from crosses between a A-genome diploid species and a D-genome diploid species, followed by chromosome doubling, 1 - 2 million years ago (Zhang et al., 2008) We compared by t-test the number of genes in the family in the five natural AD-genome polyploid species [$(AD)_1$, $(AD)_2$, $(AD)_3$, $(AD)_4$ and $(AD)_5$] and the artificial sum of the average number of genes of combinations A-genome D-genome diploid species (A+D-genome artificial polyploid species) (Table 6). The results showed that the natural AD-genome polyploid species had almost 500 fewer NBS-LRR-encoding genes than the A+D-genome artificial polyploid species, suggesting that polyploidization either slowed down the increase of the number of genes in the family or led to a significant loss of the genes during their post-polyploidization. To further confirm this hypothesis, we examined the difference in genome size between natural and artificial polyploids by t-test as well. The result showed that the natural AD-genome species also have a significant smaller genome ($P \leq 0.001$) than the A+D-genome artificial species. This finding further supported our hypothesis that a significant number of genes or other elements have been lost, or the evolution of some gene or other sequence families have been slow down significantly after polyploid cottons originated.

TABLE 6

Influence of ploypoloidization on the number of NBS-LRR genes and genome size.

Genome	Mean ± s.d.	Mean difference [a]
A: Number of genes in the family		
(AD)	973.04 ± 348.74	-499.05*
(A+D)	1472.08 ± 464.67	
B: Genome size (pg/2C)		
(AD)	4.90 ± 0.07	-0.37***
(A+D)	5.27 ± 0.07	

[a] "*" and "***" indicates that the numbers of genes in the family or genome size between the natural polyploid (AD) and artificial (A+D) polyploid species are different at a significance level of $P \leq 0.05$ and $P \leq 0.001$, respectively.

3.5 Impact of domestication/breeding on the number of genes in the NBS-LRR-encoding gene family

Domestication and breeding represents a major force in crop plant evolution. To test the roles of plant domestication and breeding in the variation in the number of genes in the family, we compared the numbers of NBS-LRR-encoding genes between the cultivated cotton species and the wild species having similar genome sizes at both diploid and polyploid levels by t-test (Table 7). At the polyploid level, we compared the number of NBS-LRR genes in the cultivated species, *G. hirsutum* and G. *barbadense*, with that of the wild species, *G. tomentosum*, *G. darwinii* and *G. mustelinum*. The results revealed that cultivated polyploid cottons have 390 more NBS-LRR genes than the wild polyploid species ($P \leq 0.05$). At the diploid level, we compared the number of the NBS-LRR genes in the cultivated A-genome species, *G. herbaceum* and *G. arboreum*, with that in the E- and G-genome species. The t-test also showed that the cultivated diploid species have more ($P \leq 0.001$) NBS-LRR genes than the wild diploid species. The results from both diploid and polyploid species indicated that cultivated species have more NBS-LRR genes than the wild species with similar genome sizes, suggesting that man's activities has played an important role in the variation and evolution of number of genes in the NBS-LRR-encoding gene family. These results also suggest that plant breeding has not only selected for favorable alleles, genes, and their combinations, but also changed the number of genes in a gene family.

TABLE 7
Influence of domestication and breeding on the number of NBS-LRR genes and genome size

Genome	Mean ± s.d.	Mean difference [a]
A: Cultivated vs. wild species: at polyploid level		
(AD_1, AD_2)	1101.23 ± 597.73	249.41**
(AD_3, AD_4, AD_5)	851.82 ± 438.57	
B: Cultivated vs. wild species: at diploid level		
(A_1, A_2)	972.41 ± 523.24	599.60***
(E, G)	372.81 ± 210.65	

[a] "**" and "***" indicate that the numbers of genes in the family between the cultivated and wild species are different at a significance level of $P \leq 0.01$ and $P \leq 0.001$, respectively.

3.6 Impact of natural selection on the number of genes in the NBS-LRR-encoding gene family

In addition to plant domestication and breeding, natural selection may have played an important role in the variation and evolution of a gene family. To test this hypothesis, we comparatively analyzed the numbers of NBS-LRR-encoding genes in the wild diploid species native to North America (D-genome species), Africa-Asia (B-, E-, and F-genome species), and Australia (C-, G-, and K-genome species). We first modified the number of genes in each accession by using the correlation formulas between genome size and number of genes because the genome sizes of the species vary significantly, and then conducted ANOVA and T-test with the modified number of genes in the family (Table 8). First, ANOVA showed that the numbers of NBS-LRR-encoding genes were different ($P \leq 0.001$) among the wild diploid species native to the three continents. Further analysis showed that the number of NBS-LRR-encoding genes were also different ($P \leq 0.01$ or 0.001) between the wild diploid species native to each pair of the three continents. These results suggest that natural selection has indeed played an important role not only in gene mutation but also in the variation and evolution of number of genes in the NBS-LRR-encoding gene family.

TABLE 8
Multiple comparisons of natural selection on the Log_{10}- transferred number of NBS-LRR genes in the genome of a wild diploid species

Continents	Mean[a] ± s.d.	Mean difference
A. Between the species native to America and Africa-Asia		
D-genome species	212.52 ± 56.78	27.91 ***
B-, E-, F-genome species	240.42 ± 22.88	
B. Between the species native to America and Australia		
D-genome species	212.52 ± 56.78	41.12 ***
C-, G-, K-genome species	253.64 ± 32.60	
C. Between the species native to Africa-Asia and Australia		
B-, E-, F-genome species	240.42 ± 22.88	13.22 **
C-, G-, K-genome species	253.64 ± 32.60	

[a] The mean was modified according to the correlation formulas between the genome size and number of genes shown in Figure 8.
[b] "**" and"***" indicates that the numbers of genes in the family between the species native to the two continents are different at a significance level of $P \leq 0.01$ and $P \leq 0.001$, respectively.

CHAPTER IV

DISCUSSION AND CONCLUSION

Large-scale genome sequencing and analysis have revealed that most, if not all, of genes exist in forms of families in the genomes of human, plants, and animals (Gibson and Muse 2004; Demuth et al. 2006). However, little is known about their variation and evolution in terms of family size or number of genes in a family in the course of speciation and evolution. This study represents the first report of the number of genes in a multigene family in a large number of plant accessions, cultivars, and species, and its variation and evolution in the course of genome evolution (Table 2). While several methods, including membrane array, microarray, shotgun sequencing, and qrtPCR, have been used to assay the copy number of genes or other elements in a genome, this study has further confirmed that the membrane array method is a reliable, economical and efficient one for assay of the variation of number of genes in a gene family in a genome in a large-scale manner. This not only supports the conclusions of Diaz et al. (2007) who used the method to detect individual gene deletion and dosage, and our previous studies in which we used the method to determine the copy number of genes in gene families with multiple plants of a cultivar (unpublished), but also is confirmed by the shotgun-like and regular DNA library screening method.

It is expected that the number of genes in a genome varies significantly among diverged species such as between Arabidopsis (26,000 genes) and rice (40,000 genes). This study, for the first time, shows that the number of genes in a gene family varies (*P*

≤ 0.001), not only among related species having the same genome (Table 4), but also within a species, i.e., among different cultivars or accessions (Table 5) even though only a few species with 3 – 12 cultivars or accessions were analyzed in the study (Table 3). While research using a large number of cultivars or populations of each species remains to further explore the degree of the intra-specific variation, the finding obtained in this study is surprising sufficiently to raise many significant questions on the current knowledge and research in genetics, evolution, biology, and breeding. For instance, does the variation in number of genes in a gene family affect the morphology, biology, and complexity of an organism? What kind of role does such variation play in plant speciation, adaptation, and evolution? Does the intra-specific variation or the variation between cultivars in number of genes in a gene family affect the current plant and animal breeding practices of selecting for favorable alleles and/or their combination?

Variation in number of genes in a gene family could be attributed to many natural and artificial forces. The results of this study provide several lines of evidence that the size variation of the NBS-LRR-encoding gene family in the *Gossypium* species is affected significantly by genome size variation (Figures 7 and 8), polyploidization (Table 6), domestication/breeding (Table 7), and natural selection (Table 8). The roles of genome size variation, polyploidization, and natural selection in the number of genes in the NBS-LRR-encoding gene family were predictable, but that of domestication/breeding is surprising since it is known that breeding is to select for and/ or pyramid favorable genes and/or their combinations. Since most of traits that are

important to agriculture are quantitative and controlled by multiple genes, the accumulation of agronomic genes in a cultivar may have resulted from selection for such quantitative traits. The positive correlation ($P \leq 0.05$) between genome size and number of genes in the family (Figures 7 and 8) suggest that species with larger genomes tend to have more NBS-LRR-encoding genes. This result agrees with previous findings that *Arabidopsis* with a genome size of 145 Mb/1C has many fewer NBS-LRR-encoding genes (150 NBS-LRR-encoding genes, Meyers et al. 2003) than rice (480 NBS-LRR genes, Zhou et al. 2004), which has a genome size of 400 Mb/1C. It is apparent that polyploidization, perhaps post-polyploidization as well, either led to the loss of these genes or slowed the process of gene number increase after polyploidization occurs. This conclusion is supported by the observation that the polyploid *Gossypium* species have fewer ($P \leq 0.001$) NBS-LRR-encoding genes than their putative diploid ancestors (Table 6). This trend can be observed in the genomes of other genera where polyploid species have smaller genomes ($P \leq 0.001$) than their diploid ancestors. The highly significant difference in number of genes in the NBS-LRR-encoding gene family among the species native to different geographical regions suggests that natural selection plays an important role in the variation in number of genes in the NBS-LRR-encoding gene family. The gene members that are favorable for fitness at the time are selected and accumulated in the genomes, but those that are not favorable for fitness at the time are lost in natural selection. It is surprising that the number of genes in the family is so different ($P \leq 0.001$) between the cultivated cottons and wild species with similar genome sizes at both diploid and polyploid levels, and

that the cultivated cottons have significantly more NBS-LRR-encoding genes than the wild species. Because the NBS-LRR-encoding gene family contributes 80% of the known genes conferring resistance to different pathogens, including bacteria, fungi, viruses, and nematodes (Takken and Joosten 2000), it is likely that the wild *Gossypium* species have more NBS-LRR-encoding genes than the cultivated cottons. The fact that the number of genes in the family in the cultivated cottons is larger ($P \leq 0.001$) than that in the wild species indicates that plant breeding likely allows accumulation of NBS-LRR-encoding genes that potentially provide resistance to pathogens. Therefore, plant breeders, in fact, have selected not only for favorable alleles and favorable allele combinations, as expected, but also the number of genes.

Since this is the first report in this field, further studies remain. These include, but not limited to, the universality of the findings in plants and animals, the universality of the findings in different gene families, genetics of the gene family size variation, relationship between the gene family size variation and variation in organism biology, morphology and complexity, gene family size variation and breeding, etc. Nevertheless, the finding resulted from this study will shed light on many fundamental questions in biology, diversity and complexity of plants and animals.

REFERENCES

Chung, Y.-J., J. Jonkers , H. Kitson , H. Fiegler , S. Humphray , C. Scott , S. Hunt , Y. Yu , I. Nishijima , A. Velds , H. Holstege , N. Carter , A. Bradley , 2004 A whole-genome mouse bac microarray with 1-mb resolution for analysis of DNA copy number changes by array comparative genomic hybridization. Genome Res. **14**: 188-196.

Demuth, J. P., De Bie T., Stajich, J. E., Cristianini, N., Hahn, M. W. 2006 The evolution of mammalian gene families. PLoS ONE Issue **1**:e85.

Diaz, M.G.Q., M. Ryba, H. Leung, R. Nelson, and J. E. Leach , 2007 Detection of deletion mutants in rice via overgo hybridization onto membrane spotted arrays. Plant Mol. Biol. Rep. **25**:17-26.

Ferreira, I.D., do Rosário, V.E., and Cravo, P.V.L. 2006 Real-time quantitative PCR with SYBR Green I detection for estimating copy numbers of nine drug resistance candidate genes in Plasmodium falciparum. Malaria J. 5:1

Gibson, G. and S. V. Muse, 2004 *A Primer of Genome Science*, second edition. Sinauer Associates, Sunderland MA.

Hawkins, J.S., H. Kim, , J. D. Nason, R.A. Wing, and J. F. Wendel, 2006 Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res. **16**:1252-1261.

He, L., C. Du,, L. Covaleda, Z. Xu,, A. F. Robinson, J.Z. Yu, R. J. Kohel, H-B. Zhang, 2004 Cloning, characterization, and evolution of the NBS-LRR-encoding

resistance gene analogue family in polyploid cotton (*Gossypium hirsutum* L.) MPMI **17**: 1234–1241.

Hendrix, B. and J. McD. Stewart, 2005 Estimation of the nuclear DNA content of Gossypium species, Annals of Botany **95**:789–797.

Irwin, D. M., and A. C. Wilson, 1990 Concerted evolution of ruminant stomach lysozymes: characterization of lysozyme cDNA clones from sheep and deer. J. Biol. Chem. **265**:4944-4952.

Meyers, B.C., A. Kozik, A. Griego, H. Kuang, and R.W. Michelmore, 2003 Genome-wide analysis of NBS-LRR–encoding genes in Arabidopsis. Plant Cell **15**:809-834.

Michelmore, R.W., and B. C. Meyers, 1998 Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. **8**:1113–1130.

Nei, M., X. Gu, and T. Sitnikova, 1997 Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. U.S.A. **94**:7799-7806.

Ota T, M. Nei 1994 Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol Biol Evol **11**: 469-482.

Smith, G. P., 1973 Unequal crossover and the evolution of multi-gene families. Cold Spring Harbor Symp. Quant. Biol. **38**:507-513.

Takken, F. L.W. and M. H. A. J. Joosten, 2000 Plant resistance genes: their structure, function and evolution. European Journal of Plant Pathology **106**: 699–713.

Yi, C. X., J. Zhang, K. M. Chan, X. K. Liu, and Y. Hong, 2008 Quantitative real-time PCR assay to detect transgene copy number in cotton (*Gossypium hirsutum*). Analytical Biochemistry **375**:150-152.

Zhang H-B, Y. Li , B. Wang , P. Chee, 2008 Recent advances in cotton genomics. International Journal of Plant Genomics, Vol. 2008, Article ID 742304.

Zhou, T., Y. Wang, J.-Q. Chen, H. Araki, Z. Jing, K. Jiang, J. Shen, D. Tian, 2004 Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol Gen Genomics 271: 402–415.

Zimmer, E. A., S. L. Martin, S. M. Beverley, Y. W. Kan, and A. C. Wilson, 1980 Rapid duplication and loss of genes coding for the α chains of hemoglobin. Proc. Natl. Acad. Sci. U.S.A. **77**: 2158-2162.

VITA

Name: Yen-Hsuan Wu

Address: Heep Center 609, 370 Olsen Blvd. College Station, Texas 77843-2474

Email Address: zxcvbnmokm@tamu.edu

Education: B.S, Agronomy, National Taiwan University, 2005
M.S, Plant Breeding, Texas A&M University, 2009