PHYSICAL ORGANIZATION OF THE GENOME OF *GOSSYPIUM RAIMONDII* AND AN IN-DEPTH STUDY OF *LIGON LINTLESS-2* GENE REGION IN TETRAPLOID COTTON SPECIES

by

LIFENG LIN

(Under the direction of Andrew H. Paterson)

ABSTRACT

The cotton (*Gossypium*) genus contains four domesticated species that are among the most important crop species for modern society. In my thesis projects, I studied the cotton species from two different levels. In the first part, I approached the cotton D genome through a team effort in whole genome physical mapping and comparative genomic analysis. This provided insights into cotton genome composition and proved helpful to the cotton research community in efforts such as gene cloning and whole genome sequence assembly. In the second part, I focused on a specific region near the top of Chr.18 of tetraploid cotton, fine-mapping the *Ligon lintless-2* (*Li2*) gene. With a large mapping population, I have determined the closest marker to be ~0.1 cM or ~ 500 kb away from the gene. A physical map contig spanning >500kb near the gene region is also identified. Sequence analysis of BACs from the contig identified homologous regions in the genomes of other species, as well as to begin to explore the likely gene content of the *Li2* region.

INDEX WORDS: physical map, genetic map, microsynteny, cotton, seed trichome, *Ligon lintless-2*, *Li2*

PHYSICAL ORGANIZATION OF THE GENOME OF *GOSSYPIUM RAMONDII* AND AN IN-DEPTH STUDY OF LIGON LINTLESS-2 GENE REGION IN TETRAPLOID COTTON SPECIES

by

LIFENG LIN

B.S., Fudan University, China, 2003

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2010

PHYSICAL ORGANIZATION OF THE GENOME OF *GOSSYPIUM RAMONDII* AND AN IN-DEPTH STUDY OF LIGON LINTLESS-2 GENE REGION IN TETRAPLOID COTTON SPECIES

by

LIFENG LIN

Major Professor: Andrew H. Paterson

Committee: John M Burke
Robert Ivarie
O. Lloyd May
Zhenghua Ye

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2010

DEDICATIONS

To my parents: my father Guilin Lin and my mother Ruying Shi.

献给我的父母，林桂林先生和石如英女士

&

To vivi

ACKNOWLEDGEMENTS

These projects would never be possible without the help from past and present members of the Paterson Lab. I am especially thankful for Dr. Paterson, who has patiently guided me through the years, tolerating my mistakes and giving me room to better myself. Dr. John E Bowers and Dr. Junkang Rong have taught me lab techniques and analysis method with uttermost patience. Dr. Haibao Tang has not only provided me with many useful ideas through our discussions, but also, perhaps more importantly, has been a truthful friend that carried me through some of the most difficult periods during this endeavor. I would also like to thank members of my advisory committee for their continuous support.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This thesis involves two research projects that I have undertaken during my doctoral studies. Chapter 1 is an overview of the structure of the thesis; Chapter 2 is a review of our current understandings of the physical composition of cotton genomes; Chapter 3 describes my first project, the construction of a whole genome physical map of D genome cotton; Chapter 4 and 5 focus on the second project: progress towards cloning a gene involved in cotton fiber development.

The literature review chapter (Chapter 2) describes the background and the foundation on which my two projects are based. It introduces our knowledge of the cotton genomes including the evolution history of the diploid and tetraploid cotton, composition of repetitive elements, genome size variation and current genetic and physical mapping efforts.

Chapter 3 describes the construction of a BAC-based physical map for a D genome cotton species. We used both agarosed-based fingerprinting and high information content fingerprinting (HICF) in this process, and integrated thousands of molecular markers through BAC hybridization. Our result shows that cotton genome is composed of two qualitatively different components. The contigs were integrated onto a consensus genetic map, and anchored on two of the sequenced genomes. The map will be helpful in studies such as gene cloning and the assembly of the whole genome sequence.

Chapter 4 describes the effort toward the identification of the *Ligon lintless*-2 gene in tetraploid cotton. Fine mapping and chromosome walking was carried out in a region close to one end of chromosome 18. Physical map contigs that correspond to this region were identified through probe hybridization and re-evaluated, from which we selected several BACs for shotgun sequencing. Several of the predicted genes on these BACs showed annotated functions that are likely to be related to cotton fiber development, providing us with a list of candidate genes for further validation.

The BACs sequenced in the gene cloning project (Chapter 4) provided us with material to study sequence evolution in this region. In Chapter 5, we undertook comparative genomic analysis using these sequences and the homologous grape regions and discovered new evidence supporting that cotton is an ancient polyploid. Whole genome level dot-plot analysis also showed evidence of an ancient genome duplication event in cotton after its divergence with grape.

Chapter 6 summarizes the major findings and conclusions of Chapters 3 through 5, and gave perspectives for future research.

Among these Chapters, Chapter 2 was published as a book chapter in *Genetics and Genomics of Cotton*, edited by Andrew H Paterson, published by Springer Press. Chapter 3 has been submitted to BMC *Genomics*. Chapter 5 will also be submitted to BMC *Genomics*. Chapters submitted or to be submitted to journals were formatted in manuscript style, with contents organized into Abstract, Introduction, Materials and Methods, Results, and Discussion. Chapter 4 is organized as a regular chapter.

CHAPTER 2

PHYSICAL COMPOSITION AND ORGANIZATION OF THE GOSSYPIUM GENOMES[1]

[1] Lin, L and Paterson, A.H. Physical composition and organization of the *Gossypium* genomes. In Genetics and Genomics of Cotton. Page 141-155 Paterson, A.H., Eds.; Springer Press, 2009

Abstract

The 8 different diploid *Gossypium* genomes vary about three-fold in genome size, ranging from less than 900 Mb to over 2,000 Mb. DNA renaturation kinetic analyses more than 30 years ago suggested that much of this variation was attributable to the repetitive DNA, and subsequent cloning and sequencing studies have revealed specific DNA elements and families that contribute to this variation. The relationship between physical quantity of DNA and genetic distance (recombination fraction) in a region shows striking variation along individual *Gossypium* chromosomes, but an appreciable degree of correspondence across subgenomes and species due largely to conserved locations of centromeres. A substantial and growing collection of bacterial artificial chromosome (BAC) libraries for *Gossypium* species and genotypes provides a platform for studies of local organization of specific genomic regions, and for global physical characterization (which is in progress for several genomes). Of particular importance in planning for the sequencing of members of the *Gossypium* genus is the nearly two-fold difference in size between the A and D diploid genome types that have contributed to tetraploid cotton, and the finding that repetitive DNA has been transmitted between these two genomes, especially from A to D, in tetraploid cottons. Additional information currently being assembled about the diversity among different members of the major repetitive element families and the degree of inter-genomic exchange following polyploidization will be important to devising cost-effective sequencing strategies.

2.1. Overview

From a common ancestor thought to have existed about 5-10 million years ago, the eight different diploid genome types in the *Gossypium* genus have evolved striking differences in physical composition and nearly three-fold variation in genome size. The

size and evolutionary relationships among the cotton species are shown in Figure 2.1. Among the diploid genomes, the K genome is the largest, with an estimated genome size even larger than that of the tetraploid (WENDEL *et al.* 2002b). The D genome is the smallest, at about 880 Mb. All other cotton diploid genomes have genome sizes between 1300 Mb and 2600 Mb.

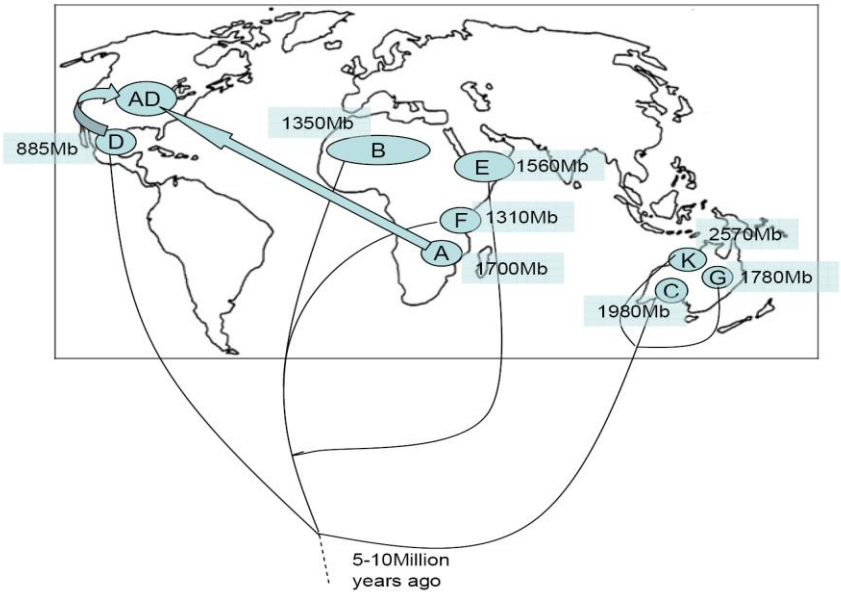


Figure 2.1 The genome size and evolutionary relationship among different cotton species.
Modified from http://www.eeob.iastate.edu/faculty/WendelJ/images/map2.jpg.

A variety of approaches have been employed dissecting of the molecular basis of this variation. Cotton was an early subject of DNA reassociation kinetics studies, which yielded a general picture of cotton genome organization and comparative evolution of genome size that continues to be applicable today. However, newer methods have permitted us to dissect the 'kinetic components' of cotton DNA into individual DNA element families with different genomic distributions and evolutionary strategies; to identify particular chromosomal regions in which there are striking deviations from the

'average' physical/genetic distance relationship; and to clone and characterize selected chromosomal segments. Repetitive DNA is the largest component of eukaryotic genomes and is a key consideration in whole genome sequencing (PATERSON 2006). Therefore, current and ongoing research in this area is important to designing cost-effective strategies by which to capture the unique sequence information that distinguishes the respective cotton genomes from one another and from those of other organisms.

2.2. Characterization of Cotton Genome Composition.

2.2.1 Comparison of cotton genomes by DNA reassociation kinetics.

In early efforts to analyze the DNA composition of the cotton genomes, quantitative measurements of the DNA content of the A, D and AD genomes were obtained using DNA reassociation kinetics, or 'C_0t analysis' (GEEVER *et al.* 1989; KADIR 1976; WALBOT and DURE 1976). In this procedure, genomic DNA is sheared into fragments and denatured, and then allowed to reassociate under controlled conditions with continuous monitoring of the portion of DNA that has renatured. DNA elements that are present in many (thousands of) copies in a genome renature rapidly, while elements present in few copies such as many genes renature slowly. In recent years this procedure has been used in conjunction with cloning to selectively clone and characterize DNA element families with differing abundance in a genome (PETERSON *et al.* 2002a; PETERSON *et al.* 2002b).

An early reassociation kinetic analysis of tetraploid cotton (*G. hirsutum*) provided our first glimpse into cotton genome composition (WALBOT and DURE 1976). Highly repetitive DNA elements, with an average C_0t value of less than 0.1, comprised

about 8% of the genome, and moderately repetitive elements with a C_0t value of 5.42 comprised about 27% of the genome. The remaining 60% of the tetraploid genome showed high C_0t values consistent with low copy number (excluding the small portion of DNA that is invariably damaged during such experiments).

C_0t analysis of the A1 *(G. herbaceum)* and D5 *(G. raimondii)* genomes (GEEVER *et al.* 1989) revealed substantial differences in the composition and organization of these, the putative ancestors of the tetraploid cottons. Specifically, the "zero time" (C_0t around 10^{-3}, extremely repetitive or self-annealing), moderately repetitive and single copy fragments comprise 7%, 54% and 39% of the A1 genome, respectively; and 7%, 30% and 63% of the D5 genome (GEEVER *et al.* 1989), indicating that the D genome is substantially less repetitive than the A genome. The tetraploid genome (*G. hirsutum*) was re-evaluated, with 6%, 46% and 48% of the respective components, differing from Walbot and Dure's estimate due to somewhat different circumscription of the three components. Notably, the tetraploid values (GEEVER *et al.* 1989) are intermediate between those of the constituent A and D genome diploids, albeit somewhat closer to the A genome values. This is consistent with the fact that roughly two-thirds of the tetraploid DNA is A-genome derived, in that the A genome contains roughly twice as much DNA as the D.

The sequence similarity between the A1 and D5 genome was estimated by reannealing of mixtures of DNA from the two species, comparing interspecific hybridization results with intraspecific hybridization to estimate the similarity between genomes. Reciprocal experiments showed 76.4% -- 78.7% re-naturation, which we now know to be explicable by very high similarity of low-copy sequence, with appreciable divergence of many repetitive DNA families in the two genomes (see below).

2.2.2 Cloning and Characterization of Cotton Repetitive Element Families

DNA cloning permitted individual *Gossypium* repetitive elements to be isolated and studied. In a detailed characterization of repetitive elements in tetraploid cotton, a genomic library was screened by hybridization to labeled total genomic DNA, identifying 313 putatively repetitive clones that showed particularly strong hybridization signal. The clones were cross-hybridized to one another, and grouped into 103 families that differed in genome organization, methylation pattern, abundance, and DNA variation (ZHAO *et al.* 1995). High abundance families were estimated by slot blot analysis to range from 15,000 to 100,000 copies, while moderate-abundance families ranged from 4,000 to 10,000 copies, and low abundance families ranged from 100-4,000 copies. Using this estimation, 25 elements that were highly abundant and another 8 representative moderately abundant elements made up 24.5% of the haploid genome; the remaining 46 moderately abundant elements make up another 7.2%. The 24 low abundance elements make up less than 0.5% of the haploid genome (ZHAO *et al.* 1995). So in this estimation, the repeat families comprise 29-35% of the haploid genome of *G. hirsutum* (ZHAO *et al.* 1995), which roughly agrees with estimates from C_0t analysis (WALBOT and DURE 1976). Based on patterns of hybridization to genomic Southern blots, most (83/103) of the repetitive element families are interspersed or partially interspersed, with the remaining 20 being tandem or partially tandem. Based on analysis of genomic digests with isoschizomers, most interspersed repetitive elements are methylated, and most tandem repeats are not methylated (ZHAO *et al.* 1995).

More recently, Hawkins *et al.* categorized repetitive sequences of *Gossypium* (Figure 2.2) by BLASTing sequences from whole genome shotgun libraries against NCBI databases (HAWKINS *et al.* 2006). Four genomes were randomly-sampled for sequences

that resembled known repetitive element families: the A1 (*G. herbaceum*), D5 (*G. raimondii*) and K (*G. exiguum*) genomes, and an outgroup: *Gossypioides kirkii*. The repetitive sequences identified were further characterized into different groups of transposable elements and tandem repeats.

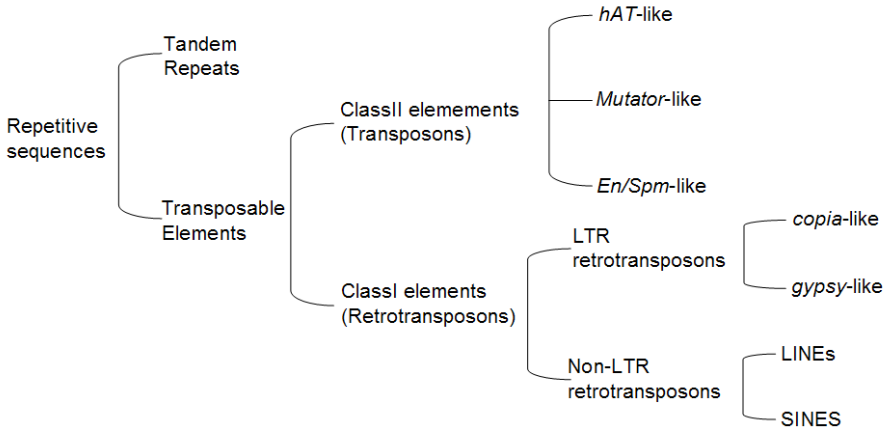


Figure 2.2 The categorization of repetitive sequences in cotton.

Tandem repeats were identified using *Tandem Repeat Finder* (BENSON 1999). 5SrDNA was identified in all four genomes tested. The copy numbers are estimated for the D (7675±3826) and A (5073±3379) genomes. No significant differences were shown between the genomes, however the precision of the estimated copy numbers was relatively low (note large standard deviations). In earlier study, the copy number of 5SrDNA was found to vary several-fold even among species of the same diploid genome (CRONN *et al.* 1996). Another previously published *Gossypium* repeat: pXP1-80 (ZHAO *et al.* 1998) was also identified in all four genomes, with copy numbers: *G. kirkii*:12,263±6098; *G. raimondii*:6573±3956; *G. herbaceum*:10,101 ±5391) and *G. exiguum* (23,795±8528) (HAWKINS *et al.* 2006). Some unknown types of tandem repeats were also found in low copy numbers (HAWKINS *et al.* 2006).

En/Spm-like, *Mutator*-like, and *hAT*-like are the three major superfamilies of Class II (DNA) transposons identified in the cotton libraries. No evidence of MITEs, TRIMs, LARDs, or Helitrons was found in the libraries evaluated, noting however that these libraries are a relatively small sampling of the genome. *En/Spm*-like sequences make up less than 1% of the genomes of all three cotton species and the outgroup, and so does the *hAT*-like sequences. *Mutator*-like sequences were identified, but the copy numbers were not estimatable due to lack of a confident length estimate of the element. All together, Class II transposons make up <2% of the whole genomes.

Class I transposons are much more abundant, making up about 45-60% of each of the genomes tested, indicating that these elements have amplified roughly proportionally to the size of the genome. *Copia*-like element numbers are proportional to genome sizes, except that the D genome has a higher than expected number. LINE-like elements are similar in number in the D genome and the outgroup, but have significantly higher copy numbers in the A and K genomes. SINE-like elements were not identified in these libraries. *Gypsy*-like families have the closest relationship between copy number and relative genome size, and are considered a major component of the size differences between different cotton genomes (HAWKINS *et al.* 2006).

2.2.3 Repetitive sequence evolution in tetraploid *Gossypium*

The diploid origins of repetitive elements in the tetraploid cotton species can be deduced by comparative analysis with the diploid ancestors. A total of 83 noncross-hybridizing clones from *G. barbadense* containing dispersed nuclear repetitive DNA were radioactively labeled and hybridized to quantitative slot blots of genomic DNA from a series of cotton genotypes representing the respective subgenomes (ZHAO *et al.*

1998). Hybridization intensities of repetitive elements are largely consistent with our present understanding of *Gossypium* phylogeny (WENDEL and ALBERT 1992). With the exception of one D genome species (*G. gossypioides*), all the A genome "specific" elements are largely confined to closely related Old World B, E and F genomes, showing only low levels of signal in the Australian C and G genomes. The few D genome "specific" elements are confined to New World cottons, showing little signals in the Old World A, B, E and F genomes. Only 4 of the 83 repetitive fragments tested were D genome-enriched or D genome-specific. Most dispersed repeat families in tetraploid cotton are derived from the physically larger A-genome diploids (ZHAO *et al.* 1998).

The finding of otherwise A genome-specific repetitive elements in the D genome species *G. gossypioides* is particularly interesting because it indicates cross-genome transfer of DNA elements. The signals from A genome specific repeat probes hybridized on *G. gossypioides* are, on average, only~36% of the level of A genome diploids, but this is 600% higher than the levels in other D genome cotton species. *G. gossypioides* is sister to *G. raimondii,* long suspected to be the closest extant relative to the tetraploid D-genome progenitor. One could envision that *G. gossypioides* may have been the tetraploid progenitor, or may have been an additional lineage spawned by the illegitimate A-D hybridization that led to polyploid formation.

The discovery of otherwise A genome-specific repetitive elements in *G. gossypioides* also suggested the possibility that repetitive elements may spread outside of their original genome following polyploid formation. Such spread has been demonstrated by fluorescence in situ hybridization (FISH): many previously A genome specific elements have spread to the D subgenome chromosomes of tetraploid cotton (ZHAO *et al.* 1998). The extent of spread between subgenomes varies among different

families of dispersed elements: some families remain confined to the A subgenome chromosomes of tetraploid cotton (pXP137), and others (pXP224) confined to the D subgenome.

Tandemly repetitive DNA element families show evidence of concerted evolution in tetraploid cotton, with fixation of different diploid alleles in different lineages. For example, rDNA ITS sequences from 10 A, D, and AD genome species and an outgroup C genome species were tested for phylogenetic relationships. Bidirectional homogenization of tandem repeats has occurred within the AD tetraploid genome after polyploidization. One clade of tetraploid species had all rDNA homogenized to the A genome type, and the other clade had most rDNA homogenized to the D genome type (WENDEL *et al.* 1995a; WENDEL *et al.* 1995b).

2.3 Cotton Genome Size Evolution

Their wide range of genome sizes, well understood phylogeny, and relatively short history of divergence, makes the cotton genomes well suited to research into genome size evolution. Studies of genome size variation in cotton include two general approaches: the comparison of corresponding regions between different genomes, and global comparisons between genomes. The former approach examines closely how intron size variation, differences in the size and number of insertions and/or deletions (indels), and illegitimate recombination affect genome size. The latter compares globally the types and numbers of transposable elements between genomes.

2.3.1 Causes of Genome Size Variation among Cotton Genomes

Increased genome size can be caused by polyploidization, transposable element (TE) amplification, increase in pseudogene number and/or intron size, and

incorporation of organellar genome fragments into the nucleus. Polyploidization and TE amplification usually lead to large scale changes in genome size, while other mechanisms have smaller effects. Compared to the relatively established routes of genome expansion, genome size shrinkage is less well understood (BENNETZEN and KELLOGG 1997). However, several possible mechanisms for reduction of genome size have been suggested, including the loss of whole chromosomes, unequal intrastrand recombination, and illegitimate recombination. The loss of whole chromosomes has not yet been observed, but evidence for intrastrand recombination and illegitimate recombination has already been found in other plant genomes (BENNETZEN 2002).

2.3.2 One group of Class I transposable elements is largely responsible for genome size variation among different diploid cotton species

As in many plant species, genome size expansion in diploid cotton is mostly due to Class I transposable elements, i.e., retrotransposons.

Reassociation kinetics analysis showed little difference in the low copy sequences of different diploid genomes, but the complexity and copy numbers of repetitive elements were roughly proportional to genome size (GEEVER *et al.* 1989). Further, among different classes of repetitive sequences, copy numbers of tandem repeats are similar among different species. Class I transposable elements constitute 45-60% (HAWKINS *et al.* 2006) of the cotton genome, and their "copy-and-paste" mechanism results in a net increase of genome size. In three different cotton genomes and an out group (A, D and K, and *Gossypioides kirkii*), the copy number of Class I elements varied 4.4-fold ranging from 45,515 ± 9241 in the outgroup, to 197,294 ± 18,935 in the K genome species. The majority of repetitive elements found in these 4 genomes are LTR

retrotransposons. Class II (DNA) transposable elements, using "cut-and-paste" movement make up less than 2% of the cotton genome.

One specific group of gypsy-like retrotransposons (*Gossypium* retrotransposable gypsy-like elements group 3, i.e., Gorge3) has has similar copy numbers in D genome cotton and the out group *Gossyploides kirkii* (genome size 588 Mb), but significantly higher copy numbers in the larger A and K genomes (HAWKINS *et al.* 2006). From a purely quantitative standpoint, the propagation of Gorge3 family members is responsible for much variation in genome size among different *Gossypium* genomes.

2.3.3 Other Mechanisms of Genome Size Variation in Cotton

In addition to the effects of transposable elements, intron size differences, small indel number differences and illegitimate recombination have been examined for their possible contributions to genome size variation in cotton.

Contiguous sequence from BACs containing the cellulose synthase gene *CesA1* was compared between the two sub-genomes (At and Dt) of tetraploid cotton (*G. hirsutum*) (GROVER *et al.* 2004). The overall gapped aligned length is 123.8 kb. The *CesA1* region appeared to be within a "gene island". A total of 14 genes were detected, all present in collinear order in each of the two genomes, and totaling about 29.2 kb in size. Only two transposable elements were found and shared between the two homeologous genomes (GROVER *et al.* 2004), indicating relatively ancient origins preceding the A-D divergence of 5-10 million years ago (SENCHINA *et al.* 2003). The high level of conservation of microsynteny in the *CesA1* region might be due to its euchromatic property. Comparative genomic research in other taxa (BOWERS *et al.* 2005) has suggested that genome rearrangements may be somewhat deleterious, and more likely

to happen in heterochromatic regions. It is very likely that *Gossypium* genomes, although more recently diverged from one another, may show a similar pattern, with conserved gene content and order in euchromatin and rearrangement and size variation in heterochromatin. Integration of plastid DNA into the nucleus may contribute to cotton genome expansion. A plastid gene, *ycf2*, inserted in the At genome, accounting for 5.6% of the At genome-specific sequence. On the other hand, intron sizes showed little difference between At and Dt genomes (a mere gain of 3 bp in At) (GROVER *et al.* 2004). Other studies concurred that there exists little intron size variation among *Gossypium* species irrespective of genome size (WENDEL *et al.* 2002a).

Small indel numbers were also evaluated in the *CesA1* BACs. Overall, small indels accounted for 14% and 18% of the total length in the At and Dt subgenome, respectively, but do not contribute significantly to the overall size difference in the region.

Among the indels discovered, 38% were flanked by short direct repeats of 2-15 bp associated with illegitimate recombination (DEVOS *et al.* 2002; MA *et al.* 2004). These putative illegitimate recombinations were not equally distributed between At and Dt genomes, with Dt genome have nearly twice as many as the At genome (36 vs 19); but at the same time, they cover a similar amount of sequence. This suggests that illegitimate recombination is very likely a common mechanism of sequence evolution in cotton, and may also play a role in the evolution of cotton genome size.

2.4 Variation in the Genetic/Physical Distance Relationship

Genetic distances are measured by recombination rates between markers, but recombination events do not happen uniformly across the genome. Enormous variations in recombination frequencies exist even among different regions of the same

chromosomes. This brings about variation between recombination-based genetic distances and nucleotide-based physical distances.

Genetic/physical distance variation can often be inferred based upon marker density information from genetic maps. Chromosomal regions that are densely populated with DNA markers are often characteristic of heterochromatic regions in which recombination in rare, and therefore have a low genetic/physical distance ratio. In euchromatic regions, while there is generally more low-copy DNA (including genes) than in heterochromatic regions, this difference is outweighed by a much higher frequency of recombination, leading to an overall increase in the genetic/physical distance ratio.

To explore variations in genetic marker density, detailed cotton genetic maps composed of 2584 loci on the AD tetraploid map and 763 on the D genome map (RONG *et al.* 2004) were used. Each linkage group was partitioned into intervals of 10 cM in length. A total of 65 intervals comprising 49 clusters were statistically marker rich. These intervals occurred in an average of 1-3 clusters on each chromosome, except for tetraploid chromosomes 1 and 25 and D-genome linkage groups D3, D6, D7, D8 and D10 with no marker-rich intervals.

On most chromosomes, at least one significant concentration of loci occurs, possibly corresponding to the centromeric regions. Virtually all marker-rich regions corresponded between the D and Dt genomes, and most also corresponded with the At genome, suggesting that these may be the locations of many of the cotton centromeres. In several cases, the breakpoints of structural rearrangements between the A and D subgenome locate squarely in these regions (RONG *et al.* 2004), consistent with the widespread observation that chromosomal inversion breakpoints often lay at or near

centromeres. A total of three marker-rich regions are unique to Dt and 9 are unique to At, generally consistent with the much larger quantity of repetitive DNA in the A genome (ZHAO *et al.* 1998).

A total of 17 intervals comprising 12 clusters were marker poor, all on the tetraploid genomes (RONG *et al.* 2004). Marker-poor regions showed little correspondence, and in the At genome occurred only at the false-positive level, but did seem to be real in the Dt genome. These clues await more information about cotton genome organization to unravel their significance, if any.

2.5 BAC-Based Physical Mapping Projects Underway

Bacterial artificial chromosome (BAC) libraries, containing genomic DNA clones that are typically 100 kb or more in length and maintained at high fidelity by virtue of low copy-number plasmids, have proven to be valuable for study of genome organization, genome-wide physical mapping and sequencing, and isolation of key features surrounding a gene (such as promoter regions). Extensive BAC resources for global physical characterization of cotton genomes are available (Table 2.1). A high priority has been their use in development of scaffolds of genetically and physically-anchored sequence-tagged sites that can provide a foundation for eventual assembly of whole-genome sequences. Anchoring of these resources to DNA marker maps that have been employed in a host of genetic, evolutionary and functional studies over the past two decades will link the eventual cotton sequences to a rich history of prior research.

Table 2.1 A summary of BAC resources known to be publicly available, and their locations

***Species*/ genotype**	**Enzyme**	**Insert size (kb)**	**Genome coverage**	**Source***
G. hirsutum				
Acala Maxxa	H*ind*III	137	8.3	CUGI
TM-1	B*am*HI	130	4.4	TAMU
TM-1	H*ind*III	150	5.2	ARS
TM-1	E*co*RI	175	6.0	TAMU
Auburn 623	B*am*HI	140	2.7	TAMU
Tamcot HQ95	H*ind*III	93	2.3	TAMU
O-613-2R	H*ind*III	130	5.7	NAU
G. barbadense				
Pima S6	H*ind*III	100	5.0	PGML
Pima 90	BamHI/H*ind*III	130	6.5	Agr Univ Hebei, China
G. raimondii				
unnamed acc.	H*ind*III	97	10.0	PGML
unnamed acc.	E*co*RI		in validation	PGML
G. arboreum				
AKA8401	M*bo*I	115	6.0	PGML
AKA8401	H*ind*III	144	9.0	PGML
G. longicalyx				
F1-1	H*ind*III	125	4.4	PGML
F1-1	E*co*RI		in validati	PGML
Gossypioides kirkii				
unnamed acc.	H*ind*III	132	8.4	PGML
unnamed acc.	E*co*RI		in validation	PGML

*ARS: http://algodon.tamu.edu/cropgerm.htm
CUGI: http://www.genome.clemson.edu/
NAU: cotton@njau.edu.cn
PGML: http://www.plantgenome.uga.edu/catalog/
TAMU: http://hbz7.tamu.edu

A total of 10 genome-equivalent coverage of *G. raimondii* BACs has been fingerprinted at the Plant Genome Mapping Laboratory (Univ. Georgia) using standard procedures (MARRA *et al.* 1997). To anchor the fingerprints genetically onto an integrated physical map, virtually all genetically mapped probes have been applied to the fingerprinted BACs using the overlapping oligonucleotides (overgo) method (CAI *et al.* 1998). Manual editing and revision of the physical map is in progress, incorporating genetic marker hybridization data with BAC fingerprint data, and assembly into contigs using FingerPrint Contigs (FPC) (SODERLUND *et al.* 2000; SODERLUND *et al.* 1997). The assembly will be publicly available via a WebFPC site. Additional coverage of *Eco*RI BACs for the same genotype has recently been generated, in validation.

A BAC library of *G. hirsutum* acc. 'TM-1' has been used for whole genome physical mapping by capillary based technology (XU *et al.* 2004), through collaborative research with the Kohel/Yu (USDA-ARS) and the Zhang laboratories (TAMU). Nearly ~100,000 clones (~5x coverage) have been fingerprinted on capillary sequencers. Preliminary contig assembly from the fingerprints showed that at least 20% looked to contain clones originating from homoeologous subgenomes and/or duplicated loci. To help resolve the duplicate fragments, a new TM-1 BAC library with a much larger average insert size (~175 kb) is being constructed.

Two libraries of *G. arboreum* acc. AKA8401, totaling about 15 genome-equivalent coverage, are being genetically anchored by hybridization to genetically mapped DNA probes. These data will be incorporated into the existing 'BACMan resource' at the Plant Genome Mapping Laboratory web site (www.plantgenome.uga.edu), which already includes similar anchoring data for BAC libraries for *G. hirsutum* 'Acala Maxxa', *G. barbadense* 'Pima S6', and *G. raimondii*.

A BAC library from a male-sterile fertility restorer line 0-613-2R (*G. hirsutum* L.) has been used for identification of *Rf1* gene in a 100 kb region (YIN *et al.* 2006). FISH of landed BACs recently completed the assignment of linkage groups to identified chromosomes (WANG *et al.* 2006c).

2.6 Perspectives

2.6.1 Implications of Physical Organization of the *Gossypium* Genomes for Whole-Genome Sequencing

Efficient strategies for capturing the sequence diversity represented within the *Gossypium* genus will be greatly influenced by the large differences in genome size and organization that differentiate species and genome types within the genus. The 3-fold variation in diploid genome size appears to have been generated in about 5-10 million years since the diploid clades diverged from a common ancestor (SENCHINA *et al.* 2003). Much of this size variation arises from dispersed repetitive DNA (ZHAO *et al.* 1998), which appears to be largely LTR retrotransposon-like elements (HAWKINS *et al.* 2006). There have been particularly large expansions of repetitive DNA content in the A/B/E/F and C/G/K clades in the 5-10 million years since their divergence; thus many repetitive element families in these clades may include large numbers of relatively recently-derived members – this condition would be especially problematic for whole-genome shotgun sequencing approaches, which require individual sequencing reads to be distinguishable (even if only by a single nucleotide) from all other sequences in the genome. By contrast, the D genome clade appears to have few such recently amplified repetitive DNA families, and is expected to be more amenable to whole-genome shotgun approaches, that permit rapid production and assembly of a sequence with a minimum

of background information (although favored by, and fully incorporating, any such information that exists, such as genetic and physical maps). That there exists a high degree of colinearity and synteny among the A, D, and tetraploid genomes (BRUBAKER *et al.* 1999; DESAI *et al.* 2006; REINISCH *et al.* 1994; RONG *et al.* 2004) suggests that complete sequencing of a D-genome by an economical whole-genome shotgun approach, together with reduced-representation sequencing of representatives of additional branches of the *Gossypium* family tree by a combination of EST sequencing, Cot-based, and methylation-based methods, might be a cost-effective means to capture quickly much of the genomic diversity among the diploid cottons.

DNA content of the allopolyploids is approximately the sum of those of the A and D-genome progenitors. However, recent polyploidy introduces new dimensions into the evolution of these genomes. The tetraploid clades combine the properties of the A and D genome diploids with modification by intergenomic concerted evolution, already clearly documented for the repetitive DNA fraction (CRONN *et al.* 1996; WENDEL *et al.* 1995a; WENDEL *et al.* 1995b; ZHAO *et al.* 1998). The possibility of intergenomic exchange of low-copy DNA remains somewhat unclear, with tenuous evidence for it from genetic mapping (REINISCH *et al.* 1994), and against it from localized comparisons of small numbers of corresponding sequences (CRONN *et al.* 1999), but growing data from other taxa strongly suggest that it may be an important dimension of polyploid evolution (CHAPMAN *et al.* 2006; GAO and INNAN 2004; HUGHES and HUGHES 1993; MOORE and PURUGGANAN 2003). Recent data from computational analysis of the rice genome suggests concerted evolution of even low-copy sequences that are diverged by a few million years (WANG *et al.* Accepted), roughly the degree of divergence among the cotton diploids. In the tetraploid cotton genome(s), the possibility of intergenomic concerted

evolution both among repetitive and low-copy DNA families may strengthen the case for a BAC-based rather than a whole-genome shotgun approach.

2.6.2 Future Directions

Despite much progress (detailed above), there still exist numerous gaps in infrastructure and information needed to clarify our knowledge of cotton genome structure. First and foremost, the *Gossypium* community lacks a high-quality reasonably complete genome sequence to use as a reference, the nearest one phylogenetically being that of *Arabidopsis* (RONG *et al.* 2007) and of some value but also suffering numerous limitations. A recent investment by the US Department of Energy Joint Genome Institute 'Community Sequencing Program' will provide about 0.5 genome-equivalent coverage of *G. raimondii*, sufficient to clarify whether this smallest and least repetitive of *Gossypium* genomes is amenable to whole-genome shotgun sequencing, guided by its genetic (Rong et al 2004) and physical (see above) maps.

Second, we need not only to sequence one diploid progenitor, but also a *Gossypium* tetraploid as well. A host of data show that the polyploid formation and associated ~1-2 million year period of adaptation to the polyploid state have been of both fundamental and practical importance in *Gossypium* evolution and improvement. Issues raised above regarding the degree of homogeneity of repetitive fractions, and the degree of intergenomic concerted evolution of low-copy DNA that has taken place, need to be clarified in order to formulate an effective strategy for this undertaking. The *Gossypium* community is acutely aware of these needs, and actively working to bring them to fruition.

Third, a host of interesting and potentially important genetic variation exists within members of the *Gossypium* genus that are difficult to access by sexual crosses. Further progress is needed to complete BAC resources for the various genome types (and preferably for multiple diverse representatives within each type), and to use multiple complementary approaches detailed above to extend *Gossypium* sequence information to these additional taxa.

CHAPTER 3

A DRAFT PHYSICAL MAP OF A D-GENOME COTTON SPECIES (G. RAIMONDII)[2]

[2] Lin, L, Pierce, GJ and Bowers, JE, *et al*. Submitted to *BMC Genomics*, 03/11/2010

Abstract

Cultivated tetraploid cottons, *Gossypium hirsutum* L. and *G. barbadense* L., share a common ancestor formed by a merger of the A and D genomes about 1-2 million years ago. Here we report a whole-genome physical map of *G. raimondii*, the putative D genome ancestral species of tetraploid cottons, integrating genetically-anchored overgo hybridization probes, agarose-based fingerprints, and 'high information content fingerprinting' (HICF). A total of 13,662 BAC-end sequences and 2,828 overgo probes were used in genetically anchoring 1585 contigs to a consensus map inferred from genetic maps of the respective diploid cotton genomes and tetraploid subgenomes. Several lines of evidence suggest that the *G. raimondii* genome is comprised of two qualitatively different components, one that is gene-rich and recombinogenic with gene repertoire and order similar to those in members of other angiosperm families (*Vitis, Arabidopsis*), and another that is repeat-rich and recombinationally-recalcitrant with relatively few genes and highly rearranged gene order. A total of 370 and 438 contigs, respectively, could also be aligned to *Arabidopsis thaliana* (*At*) and *Vitis vinifera* (*Vv*) whole-genome sequences. While *Vitis* may be more informative about cotton genome organization, translational genomics from *Arabidopsis* offers singular benefits in identifying the functions of cotton genes. The integrated genetic-physical map is of value as a component of assembling and validating a planned reference sequence. The alignment of GR contigs on *At* and *Vv* genomes shows promise for utilizing translational genomic approaches in understanding this important genome and its resident genes.

3.1 Introduction

The *Gossypium* (cotton) genus, composed of 50 species among which four provide the major raw material for one of the world's largest industries (textiles), has a large impact on our economy and everyday life. Diploid cottons are classified into 8 genome types, denoted A-G and K, based on chromosome pairing relationships (WENDEL and ALBERT 1992). All diploid cotton species are believed to have shared a common ancestor about 5-10 million years ago (WENDEL and ALBERT 1992). The cotton genome types diverged into genome groups that vary in haploid genome size from 2500 Mb in the K genome, to less than 900 Mb in the D genome (HAWKINS *et al.* 2006; HENDRIX and STEWART 2005), while retaining common chromosome number (n=13) and largely-collinear gene order (BRUBAKER *et al.* 1999; DESAI *et al.* 2006; REINISCH *et al.* 1994; RONG *et al.* 2004). The tetraploid cotton genome is thought to have formed by an allopolyploidy event about 1-2 million years ago, involving species similar to the modern New World D genome species *G. raimondii* (GR) (WENDEL 1989) or *G. gossypioides* (GG) (WENDEL *et al.* 1995a) and the Old World A genome species *G. herbaceum* (GH).

There exist at least a dozen published genetic maps for various *Gossypium* crosses, most involving members of the superior-fiber-quality *G. barbadense* species crossed with high-yielding *G. hirsutum*. These maps collectively include at least 5,000 public DNA markers (~3,300 RFLP (Restriction Fragment Length Polymorphism), 700 AFLP (Amplified Fragment Length Polymorphism), 1,000 SSR (Simple Sequence Repeats), and 100 SNP (Single Nucleotide Polymorphism)). Many thousands of additional SSRs have been described , but only a subset of these have been mapped (GUO *et al.* 2007; LACAPE *et al.* 2003; RONG *et al.* 2004; XIAO *et al.* 2009; YU *et al.* 2007). The most detailed sequence tagged site (STS)-based map, and a source of probes for

many of the other maps, are reference genetic maps for diploid (D) and tetraploid (AtDt[3]) *Gossypium* genomes that include, respectively, 2584 loci at 1.72 cM (~600 kb) intervals based on 2007 probes (AtDt); and 1014 loci at 1.42 cM (~600 kb) intervals detected by 809 probes (D) (RONG *et al.* 2004; RONG *et al.* 2005a). A high degree of collinearity among the respective genome types permitted inference of the gene order of a hypothetical common ancestor of the At, Dt, and D genomes for 3016 loci identified by 2337 probes, spanning 2324.7 cM (RONG *et al.* 2005b). Additional maps that are particularly marker-rich and/or have been widely used as reference maps for QTL studies have been developed from three additional interspecific crosses (GUO *et al.* 2007; LACAPE *et al.* 2007; YU *et al.* 2007). Other important resources include aneuploid substitution stocks that were derived from tetraploid genotypes TM-1 (*G. hirsutum*) x 3-79 (*G. barbadense*) (ENDRIZZI and RAMSAY 1979) and TM-1 x *G. tomentosum* (SAHA *et al.* 2006). Together, monosomics and telosomics have been used to assign 20 of the 26 cotton linkage groups to chromosomes, and the remaining six linkage groups were assigned to chromosomes by translocation and fluorescence *in situ* hybridization mapping. (WANG *et al.* 2006d)

Cotton genetic maps have been employed in identification of diagnostic DNA markers for a wide range of traits related to fiber yield and quality (ABDURAKHMONOV *et al.* 2007; ABDURAKHMONOV *et al.* 2008; ASIF *et al.* 2008; CHEE *et al.* 2005; DRAYE *et al.* 2005; GUO *et al.* 2003; GUO *et al.* 2008; HE *et al.* 2005; HE *et al.* 2007; HE *et al.* 2008; JIANG *et al.* 1998; KOHEL *et al.* 2001; MEI *et al.* 2004; MIR *et al.* 2008; PATERSON *et al.*

---

[3] Dt refers to the D-subgenome found in tetraploid cottons (to distinguish it from the genome of D-diploid cottons). Likewise, At refers to the A-subgenome of tetraploid cottons.

2003; QIN *et al.* 2008; REN *et al.* 2002; SAHA *et al.* 2008; SHEN *et al.* 2007; SHEN *et al.* 2005; SHEN *et al.* 2006b; ULLOA *et al.* 2005; WAN *et al.* 2007; WANG *et al.* 2006a; WANG *et al.* 2007a; WU *et al.* 2007; ZHANG *et al.* 2003; ZHAO *et al.* 2008); drought tolerance (SARANGA *et al.* 2004; SARANGA *et al.* 2001; ZHAO *et al.* 2008); and resistance to diseases (BOLEK *et al.* 2005; NIU *et al.* 2008; RUNGIS *et al.* 2002; WANG *et al.* 2008; WRIGHT *et al.* 1998; YANG *et al.* 2008), and pests (NIU *et al.* 2007; SHEN *et al.* 2006a; WANG *et al.* 2006b; WANG and ROBERTS 2006; YNTURI *et al.* 2006). Interest in hybrid cottons in some countries has drawn attention to a nuclear restorer of cytoplasmic male sterility (FENG *et al.* 2005; GUO *et al.* 1998; LAN *et al.* 1999; WANG *et al.* 2007b; ZHANG and STEWART 2004). Morphological features such as the pubescence that is characteristic of *G. hirsutum* (ALI *et al.* 2009b; DESAI *et al.* 2008; LACAPE and NGUYEN 2005; WRIGHT *et al.* 1999), leaf morphology (HAO *et al.* 2008; JIANG *et al.* 2000; SONG *et al.* 2005; WAGHMARE *et al.* 2005) and color (ALI *et al.* 2009a), and unique features such as nectarilessness (MEI *et al.* 2004; SAJID UR *et al.* 2008; WAGHMARE *et al.* 2005) have also received attention. The value of cotton seed has led to interest in mapping variation in seed physical characteristics and nutritional value (SONG and ZHANG 2007). Meta-analysis of multiple QTL mapping experiments by alignment to a common reference map has begun to reveal the genomic organization of trait variation (RONG *et al.* 2007). Although members of the D genome clade do not make spinnable fiber, genetic mapping has shown that the majority of fiber QTLs mapped in tetraploid cotton fall on D genome (*G. raimondii*-derived) chromosomes, suggesting that the D genome has been crucial to the evolution of the higher fiber quality and yield of cultivated tetraploid cottons (RONG *et al.* 2007).

Toward the long-term goal of characterizing the spectrum of diversity among the 8 *Gossypium* genome types and three polyploid clades, the worldwide cotton community has prioritized the D-genome species *Gossypium raimondii* for complete sequencing (CHEN *et al.* 2007; PATERSON 2007). *Gossypium raimondii* is a diploid with a ~880 Mb genome (HENDRIX and STEWART 2005), the smallest genome in the *Gossypium* genus at ~60% of the size of the diploid A genome and 40% of the tetraploids. It is largely inbred, and a largely-homozygous genotype has been used in both a reference genetic map (RONG *et al.* 2004) and for a BAC library (herein). DNA renaturation kinetics shows that 30-32% of the *G. raimondii* genome contains repetitive DNA, with a kinetic complexity of 1.6×10^6 bp and an average iteration frequency of ~120 copies per haploid genome (GEEVER *et al.* 1989). This has been subdivided into a highly-repetitive component of about 5% of the genome, composed of elements in 10,000 or more copies; and a middle-repetitive component accounting for 27% of the genome (WALBOT and DURE 1976). A random sampling of 0.04% of the tetraploid cotton genome, enough to sample repetitive element families that occur in 2500 or more copies, revealed only 4 D-genome-derived elements ranging in estimated copy number up to about 15,000, versus dozens of A-genome-derived repeats at much higher copy numbers (ZHAO *et al.* 1998). Pilot sequencing studies (X. Wang, D. Rokhsar, A.H. Paterson, unpubl.) show that most D-genome repetitive DNA families are sufficiently heterogeneous to be compatible with a whole-genome shotgun approach.

Genetically anchored physical maps of large eukaryotic genomes have proven useful both for their intrinsic merit and as an adjunct to genome sequencing. In species where no whole-genome sequence is yet available, a physical map is a useful tool in a wide range of activities including comparative genomics and gene cloning. Physical

mapping also provides a method of genome assembly independent of a sequence, and is useful in contributing to and/or validating whole-genome shotgun sequences (PATERSON *et al.* 2009). For BAC-based sequencing of a genome, a physical map is a prerequisite. Recent study of chromosomes 12 and 26 of upland cotton (*Gossypium hirsutum*) (XU *et al.* 2008) suggests that physical mapping of polyploid cotton may be complicated by homoeologous genome fragments.

As an important step toward its genome-wide characterization, we describe here a genetically anchored, BAC-based physical map for *G. raimondii*. By incorporating thousands of DNA markers, the physical map is tightly integrated with the rich history of cotton molecular genetics research described above, and expedites a host of studies of *Gossypium* biology and evolution. Moreover, comparison of the physical map to the sequences of *Arabidopsis thaliana* and *Vitis vinifera* shows promise for utilizing translational genomic approaches in better understanding the structure, function, and evolution of this important genome and its resident genes.

3.2 Materials and Methods

3.2.1 BAC library construction

The *Gossypium raimondii* (GR) BAC library was constructed according to Peterson *et al.* (PETERSON *et al.* 2000). The library consists of 92,160 individually-archived clones and is available through the Plant Genome Mapping Laboratory (http://www.plantgenome.uga.edu). To estimate mean insert size and false positive percentage, two clones were selected from each of the library's 240 384-well plates, and minipreps of these clones were digested with *Not*I and analyzed by pulsed-field gel electrophoresis. Of the 480 digested clones, 448 produced interpretable banding

patterns; the remaining 32 were not visible on the gels suggesting that the DNA was lost in the miniprep procedure. Three of the 448 clones appear to be false positives.

3.2.2 Probe design and hybridization

A total of 2828 sequence-tagged site probes were hybridized to the GR library: 357 were overgos designed from *Arabidopsis* genic sequences (prefixed AOG); 1751 were designed from genetically mapped cotton markers (prefixed COV for cotton overgos, or CM/COAU/PAR for PCR based probes); and 252 from cotton EST sequence reads (prefixed COV). The rest were designed and probed from cotton genes of interest related to multiple projects. Overgo probes (CAI *et al.* 1998) were designed and hybridized to the libraries as described (BOWERS *et al.* 2005). Briefly, source sequences were aligned to all known plant sequences to using BLAST to find conserved domains, and compared to known plant repeats to screen out possible repetitive sequences. The selected sequences were then chopped into 40 bp segments and screened for GC content of between 40% and 60%.

Probes were labeled using 32P and applied to macroarrays of 18,432 BACs per membrane in a multiplex of 576 probes, using pools of 24 probes per bottle, by rows, columns and diagonals of a 24x24 array of probes. Films were manually scored, scores digitized using text-recognition software (ABBYY FINEREADER), and data deconvoluted and stored in the MS Access database system "BACMan".

3.2.3 Fingerprinting

Agarose based fingerprinting methods were adapted from Marra *et al.* (MARRA *et al.* 1997). Plasmids were extracted in batches of 96-well plates and digested using *Hin*dIII. Fragments were separated on a 121-lane 1% agarose TAE (Tris-Acetate-EDTA

buffer) gel, with a size standard every 5 lanes. Band migration distances and molecular weights were digitized using IMAGE (SULSTON *et al.* 1989), before importing into FPC (Fingerprinted Contigs) (SODERLUND *et al.* 2000; SODERLUND *et al.* 1997).

High information-content fingerprinting (HICF) was adapted from published methods (LUO *et al.* 2003). Plasmids were digested with *EcoRI, BamHI, XbaI, XhoI* and *HhaI*. The ends of restriction fragments were differentially labeled using fluorochrome tagged ddNTPs after the first four enzyme cuts, and the last enzyme further reduced fragment size and produced a blunt end. Fingerprints were generated using an ABI3730xl sequencer and size files generated by GeneMapper v4.0 after processing the chromatograms.

3.2.4 Physical map assembly

Agarose-based fingerprints were assembled first by FPC using a cut-off value of 1e-10 and a tolerance value of 8. CpM (contigs plus markers) tables were used to integrate the marker hybridization results: the cut-off value was relaxed to 1e-8, 1e-7 and 1e-6 when two BACs shared one, two and three markers respectively.

After the preliminary assembly, two BACs from each end of the largest 4608 agarose FPC contigs were subjected to HICF. These fingerprints were assembled separately in FPC using a cut-off value of 1e-50 and a tolerance of 3. Overgo hybridization information was not used in HICF assembly. Results from HICF were formatted into a marker file, and fed into the final, integrated assembly in the same manner as probe hybridization results. In this assembly, cutoff was set to 1e-12 and tolerance was set to 7. CpM tables were used in integrating the data. Cut-off values were

relaxed to 1e-10, 1e-9 and 1e-8 when two BACs shared one, two and three markers (or HICF contig) respectively.

In each of the three iterations of assembly, the final stringency settings (tolerance and cut-off) were determined by comparing results of different cut-off and tolerance value combinations. For HICF, tolerance values of 2 through 5 and cut-off value of 1e-20 through 1e-50 were tested; for agarose fingerprints, tolerance values of 6 through 9 were and cut-off value of 1e-10 through 1e-12 were tested. Possible cross-well contaminations were identified and rendered as singletons using the built-in function under "search commands" in FPC v 9.3.

3.2.5 Finalizing the assembly

End-to-end auto-merges were done recursively by lowering the cut-off value one step at a time, from 1e-12 through 1e-6. Singletons were also merged into the assembly recursively using the Keyset-to-FPC function in the FPC program. The CB (Consensus Band) maps for each contig with 2 or more Q clones were recalculated using a higher stringency cutoff value. Q-contigs were thus split up by FPC into smaller contigs and singletons. This was done recursively by raising the cutoff value by 1 level at a time until each one of the splitted contigs contains no more than 1 Q clone. A compressed file containing all data (both agarose-based fingerprints and HICF) is available at http://www.plantgenome.uga.edu/pgml_image_data/.

3.2.6 BAC-end sequencing

Two BACs from each end of the largest 2016 contigs were end-sequenced by the Arizona Genome Institute using methods as described (AMMIRAJU *et al.* 2006).

3.2.7 Anchoring contigs onto genetic maps

To achieve a maximum number of anchor points, a 13-linkage-group consensus map of cotton, constructed by integration of At, Dt, and D genome genetic maps (RONG *et al.* 2005a) was used to anchor contigs. Probes that hit only one BAC in a contig were considered possible hybridization artifacts and were not used; probes that hit 30 or more BACs in the GR library were considered repetitive and were also excluded. 482 BACs with 8 or more different probes hybridized to them were excluded as possible contamination artifacts produced in hybridization. Contigs were aligned to the consensus map using the remaining anchor markers.

On average, we had less than one hybridization marker per contig, and the vast majority of contigs had less than three anchor probes. Thus, instead of requiring the contig to have two or more anchor markers from proximal regions on the genetic map to call an anchor, we listed all the contigs anchored by one or more genetic markers alongside the marker's location(s) on the genetic map.

3.2.8 Aligning contigs to whole-genome sequences

BAC-end sequences (BES) and source sequences of overgo probes were used to BLAST against *Arabidopsis thaliana* and *Vitis vinifera* genomes, using a penalty score of -2 (instead of -3 as the default value) and an e-value of 1e-5 in BLASTn. The penalty score was changed to fit the sequence divergence among genomes surveyed, so that longer hits with lower similarity (66.7%) can be retained. *Arabidopsis* and *Vitis* genome sequences were downloaded from TAIR (ftp://ftp.*Arabidopsis*.org/home/tair/Sequences/whole_chromosomes/) and Genoscope

(http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/assembly/goldenpath/unmasked/), respectively.

Sequences with 10 or more BLAST hits in either genome were considered repetitive and excluded from later analysis. Probe hybridization results used the same filters described for anchoring to genetic maps. BAC contigs were then linked to the genomic sequences through the BLAST data in a MS Access database query. The query results were processed by a Python script aligning the contigs to a genomic region of *At* or *Vv* when two or more sequences from the same contig hit a genomic region less than 200 kb (against *At*) or 1 Mb (against *Vv*) apart.

3.3 Results

3.3.1 BAC library

The *Gossypium raimondii* BAC library used in physical mapping consists of 92,160 clones. Pulsed-field gel electrophoresis-based examination of 448 *Not*I digested clones indicates a mean insert size of 100 kb. Of note, there was little variation in insert size among clones (standard error of mean = 0.76). Three of the 448 interpretable *Not*I-digested clones (i.e., 0.67%) appear to be false positives. Likewise, three of the 4032 BAC end sequences generated from the library exhibit homology to chloroplast DNA (0.07%) indicating that the methods employed in constructing the library (PETERSON *et al.* 2000) were successful in keeping chloroplast contamination low. Collectively, the library affords 10X coverage of the *G. raimondii* genome.

3.3.2 Agarose-based fingerprints

The entire 92,160 GR BAC library was fingerprinted using slight modification of established methods (MARRA *et al.* 1997). Preliminary assembly formed 9,290 contigs and 26,716 singletons at a tolerance value of 8 and cutoff value of 1e-10. The average agarose-based fingerprint band number of individual BACs was 17.4. Band number distribution across the library is shown in Figure2.1A. A total of 3266 BACs failed to produce usable fingerprints.

3.3.3 HICF fingerprints

Two terminal BACs from each end of the largest 4608 agarose contigs (four BACs per contig, totaling 18,432 BACs) from the preliminary assembly were fingerprinted using HICF. The average HICF band number per BAC was initially 203.6. HICF batches with extremely high or low band numbers (approximately top or bottom 5%) were re-fingerprinted. The average band number dropped to 178. These 18,432 BACs formed 3508 contigs and 2570 singletons. The final band number distribution is shown in Figure2.1B.

3.3.4 Overgo hybridizations

Thousands of probes were applied to the GR library using a multiplex hybridization scheme (see Methods). A total of 2828 probes from *Arabidopsis* genes, cotton ESTs, and genetic markers showed hybridization signal attributable to one or more BACs by this approach. On average, each probe hit 17.3 BACs. A total of 46 probes hit more than 100 BACs and are considered highly repetitive. To minimize false associations, probes with >50 hits were not used in the contig assembly process, and

probes with >30 hits were not used in the contig anchoring process (detailed later). Thus, 2658 probes (with <50 hits) were integrated into the assembly using the CpM table in FPC: stringency (cutoff value) was relaxed by 2, 3, or 4 denary (ten-fold) intervals when 1, 2 and 3+ common markers were found between two BACs.

3.3.5 Integrated assembly

Since agarose-based fingerprinting and HICF use different sets of restriction enzymes, a different band-calling scheme with different error rates and band size tolerances, data from these two different methods cannot be merged directly. Further, while we targeted HICF to contig-terminal BACs, it would be imprudent to declare a join in the agarose assembly whenever HICF suggests a merge of contig-terminal BACs, overlooking potential false joins in HICF. To circumvent this, if two agarose contig-terminal BACs were suggested to be joined by HICF, we lowered the cutoff value for joining agarose contig-terminal BACs by two denary intervals, e. g., when the overall cutoff was set to 1e-12, we would accept an overlap at the cutoff at 1e-10 if the two BACs were found in the same HICF contig. The agarose assembly was thus reassembled, only forming a merged contig if it was supported by both data types (see Methods), and integrating 2658 hybridization markers based on 2828 overgos.

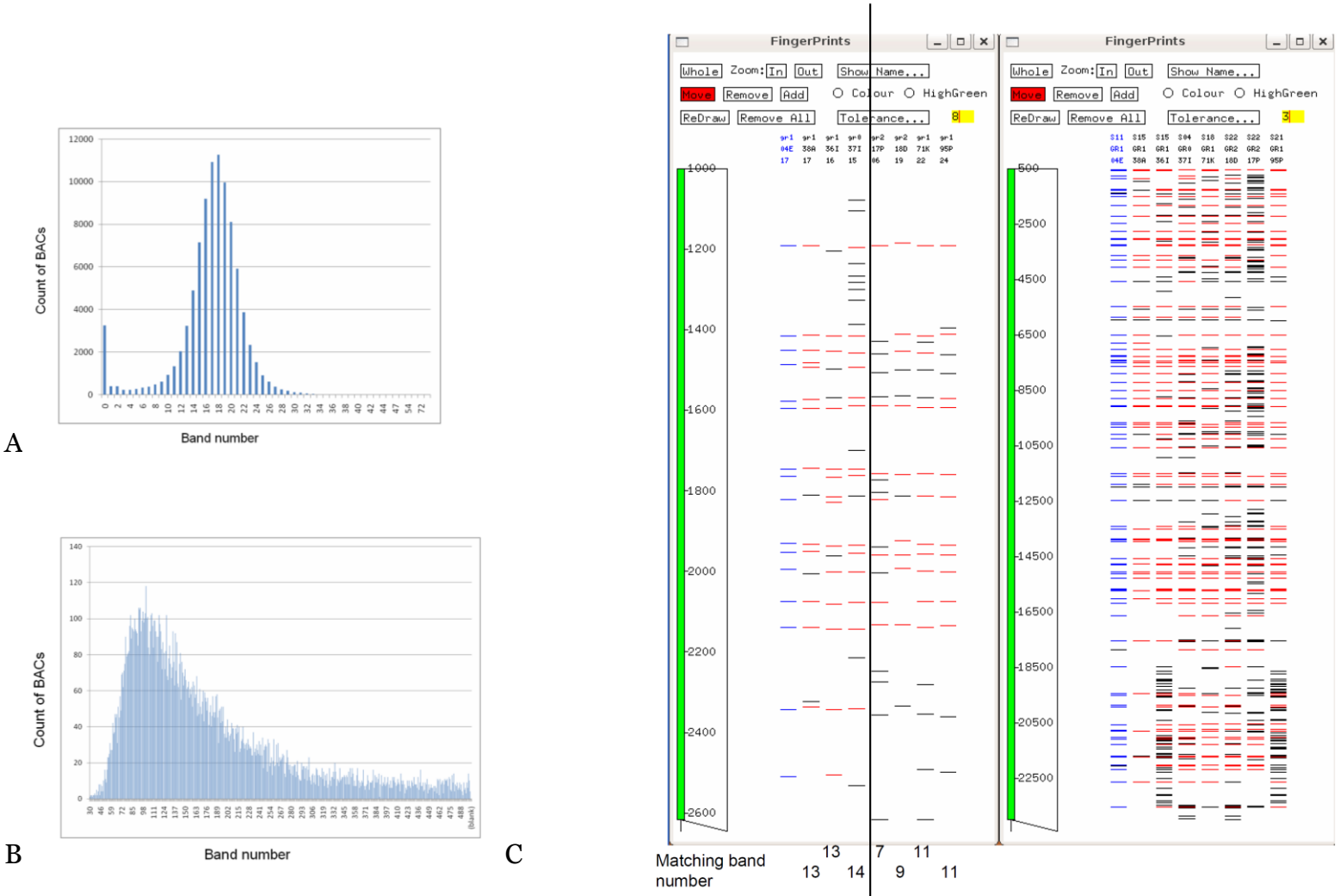


Figure 3.1. Band number comparison between agarosed-based and HICF fingerprints.
A. Band number distribution of agarose-based fingerprints; B. Band number distribution of HICF fingerprints. C. An example of two agarose FPC contig joined in HICF. Red bands are matching bands to the highlighted (in blue) BAC. Count of matching bands to the BAC are listed below each lane. The four BACs on the right were not assembled into the same contig.

Collectively, the agarose fingerprints, targeted HICF fingerprints, and overgo hybridization data joined a total of 67,343 BACs into 4208 contigs, leaving 21,551 singletons. Based on the average insert size estimate of 100 kb, and an estimated genome size of 880 Mb (HENDRIX and STEWART 2005), the 67,343 BACs in contigs provide ~7.7x coverage of the GR genome. The majority of contigs (61.5%) contain between 3 and 25 BACs. The distribution of BAC numbers per contig is shown in Figure 3.2.

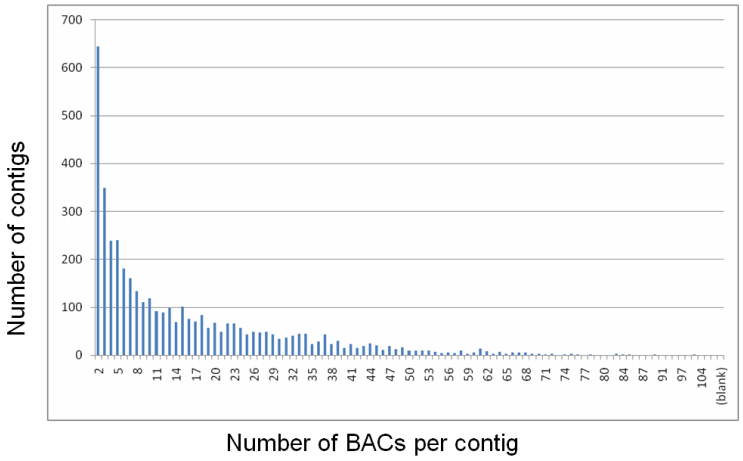


Figure 3.2 Distribution of contig sizes measured in number of BACs per contig.

Singletons differed in several ways from BACs in contigs. The average agarose-based fingerprint band number was 13.4 for singletons, versus 17.9 for BACs in contigs. A total of 9476 (44% of) singletons contained less than 12 bands. This could reflect either shorter length of singleton BACs, or the presence of tandem repeats that produce fingerprint bands that comigrate, reducing the scoreable band number and perhaps contributing to failure of some BACs to form contigs (see more discussion of band numbers below). A total of 1904 overgo probes hit singleton BACs, among which 364 overgos were repetitive and 1540 were low copy (having <30 hits total). Compared to

the probes that hit BACs in contigs (376 repetitive and 2129 low copy), singletons show some enrichment in repetitive DNA content. A total of 585 singletons were identified as possible cross-well contaminations.

3.3.6 Anchoring contigs to the cotton consensus map

After filtering out 381 (of 2828) repetitive overgo probes that hit more than 30 BACs in the GR library, and 357 BACs (out of 34,713 BACs with at least one marker hit) with more than 8 markers hybridized as suspected hybridization artifacts, the remaining probes and BACs produced 40,152 BAC-probe pairs. A total of 7772 of these were produced by BACs that were not in contigs (singletons); 5946 of the markers on contigs were "weak anchors" produced by a single BAC-probe pair for the contig. Weak anchors were not used in aligning the contigs onto the genetic map. The remaining 26,434 BAC-probe pairs derive from 1920 probes, and were distributed in 2154 contigs.

A 'consensus' cotton genetic map built from the At, Dt and D genome genetic maps contains 13 homologous groups made up of 3016 loci based on 2337 unique sequence tags (RONG *et al.* 2005a). Among these, 2109 have probes designed (961 RFLP probes and 1744 overgos, 596 have both, most of the remainder could not be sequenced). After filtering out probes with >30 hits in the library, 1468 loci on the consensus map have anchored 1586 contigs. (Table 3.1, Figure 3.3, Figure 3.4, Appendix 1). On average, each marker anchored 2.42 contigs.

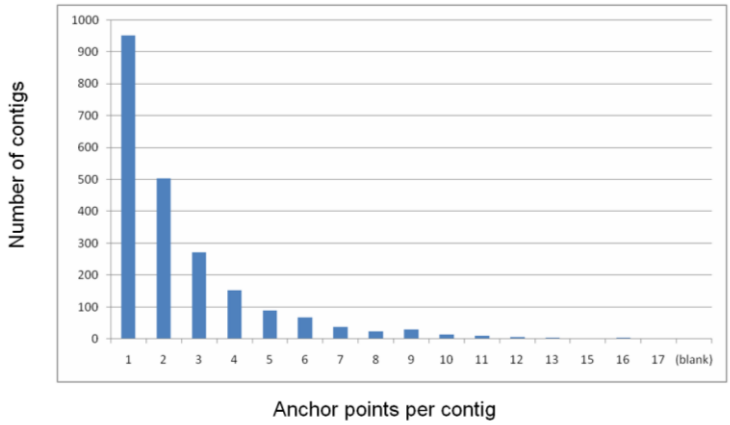


Figure 3.3 Distribution of usable anchor probes per contig after removal of contaminant and repetitive anchors.

BAC-probe relationship are produced through hybridization. BACs with 8 or more probe hits were excluded; probes hit 40+ BACs were excluded; probes that hit only one BAC in a contig were excluded. The remaining BAC-probe information was used as "anchors". The x-axis denotes the anchors per contig.

3.3.7 Aligning contigs to *Arabidopsis thaliana* and *Vitis vinifera* whole-genome sequences

A total of 8064 BACs selected from the ends of the largest 2016 contigs from the preliminary assembly were used for paired-end sequencing. The resulting 13,662 high-quality sequences, along with the 1920 low copy probes (after filtering described above), were used in comparing the GR contigs to *Arabidopsis thaliana* (*At*) and *Vitis vinifera* (*Vv*) chromosomes.

BAC end-sequences (BES) and the source sequences of the hybridization probes were aligned to the *At* and *Vv* whole-genome sequences using BLASTn. A total of 2607 sequences (1370 BES and 1237 overgo source sequences) had between 1 and 9 BLAST hits in the *At* genome, and 2968 sequences (1557 BES and 1411 overgo source sequences) have between 1 and 9 hits in the *Vv* genome. (Sequences with >10 hits were excluded as repetitive.)

Table 3.1 Distribution of anchored contigs on consensus chromosomes**.**

Homologous Group	Number of loci	Contig anchoring markers	Anchored contigs	Average # of contigs per marker
1	245	68	149	2.19
2	194	48	101	2.10
3	149	40	80	2.00
4	208	47	145	3.09
5	246	54	121	2.24
6	235	65	163	2.51
7	290	55	121	2.20
8	247	55	119	2.16
9	382	91	251	2.76
10	164	50	141	2.82
11	227	57	132	2.32
12	187	50	126	2.52
13	242	53	125	2.36
Grand Total	3016 (2234 unique)	733 (715 unique)	1774 (978 unique)	2.42

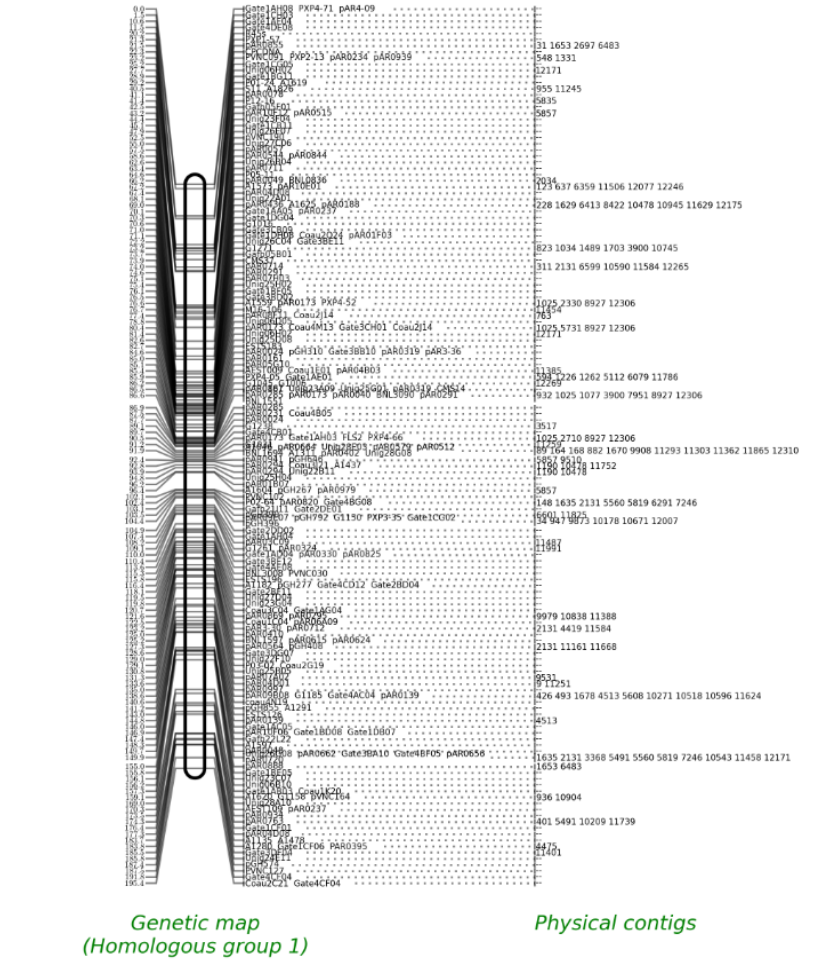


Figure 3.4 Homologous Group 1 of the integrated genetic-physical map. The genetic map is drawn using data from Rong et al. (RONG *et al.* 2005a)**.**

A total of 370 contigs were aligned to *Arabidopsis* chromosomes, 438 to *Vitis* chromosomes, and 242 to both (Table 3.2, Figure 3.5). All 566 that aligned contained 64 CB units (consensus band units, the number of total non-overlapping bands in a contig) per contig on average, about 50% larger than the overall average contig size (42 CB units). Based on an estimated size of 4097 bp per band (average of all band sizes from all BACs fingerprinted), these contigs cover a minimum of 13% (contigs anchored on *Vv*) and 11% (contigs anchored on *At*) of the GR genome, noting that band numbers somewhat underestimate contig sizes because both very large and very small bands are excluded from bandcalling. A second estimate of coverage of the target genomes by aligned contigs was obtained by adding up the distances between anchor marker BLAST matches and excluding overlaps. This suggests that 27.7% of the *Arabidopsis* genome and 22.8% of the *Vitis* genome is covered by aligned GR contigs. Some contigs have significant association with two or more positions on a target genome. The distributions of contigs along *At* and *Vv* chromosomes are shown in Figure 3.5. Contigs are more likely to be anchored to two or more locations in *At* than *Vv* (159 or 43% of contigs anchor to multiple *At* locations versus 111 or 25.4% of contigs anchored to *Vv*), consistent with the fact that the *Arabidopsis* lineage has experienced two more whole-genome duplication (WGD) events than grape (TANG *et al.* 2008a).

The GR contigs anchored on *Vv* are not evenly distributed across the chromosomes, but rather are clustered in several regions/chromosome arms that tend to have higher than average gene densities. Gene density distribution across the *Vitis* genome was extracted by counting the number of genes in 200 kb bins along the chromosomes. Gene density is largely uniform across the *Arabidopsis* chromosomes

except for the centromeric regions; while in the *Vitis* genome, we observed greater heterogeneity of gene density. The regions on which we were able to anchor GR contigs (Figure 3.5) had an average of 20 genes per 200 kb window, versus an average of 14.8 for the remainder of the genome. Among the 30% of *Vv* 'windows' with highest gene density, 37.9% were covered by GR contigs; versus 22.8% of the genome as a whole.

3.3.8 Nature of repetitive probes

A total of 46 probes are classified as highly-repetitive with >100 BAC hits and came from several sources: 28 were derived from cotton EST sequences (COV), 3 from low-copy genes in *Arabidopsis* (AOG), and 15 from cotton RFLP probes used in genetic mapping. Six of the highly repetitive cotton overgo sequences were found to be located within known repetitive elements using Repbase (http://www.girinst.org/repbase/). The overgo with the most hits (COV1526, which hits 1593 BACs) is in a helitron. The remaining five were from two *hAT*-like DNA transposons, one EnSpm element, one ERV/ERV2 element and one *Gypsy* element. Four of the 15 highly repetitive PCR-based probe sequences contain repetitive elements. The three *Arabidopsis* genes from which highly repetitive overgos were designed (At5g10360, At2g30740 and AtGRF2) showed no known repetitive elements in their sequences, which might indicate cotton lineage-specific gene multiplications. Given that Repbase does not include a comprehensive set of cotton repetitive sequences (due to lack of a complete *Gossypium* genome), it is likely that the remaining highly repetitive overgos that did not match repetitive sequences from Repbase may reveal cotton elements not previously known to be repetitive.

Table 3.2 Number of anchored contigs on each chromosomes of *At* and *Vv* genomes

Vv chr	number of contigs anchored
1	36
2	21
3	19
4	21
5	28
6	53
7	41
8	96
9	14
10	8
11	21
12	18
13	59
14	45
15	17
16	11
17	17
18	57
19	18
Total	600

At chr	number of contigs anchored
1	132
2	126
3	168
4	72
5	152
Total	650

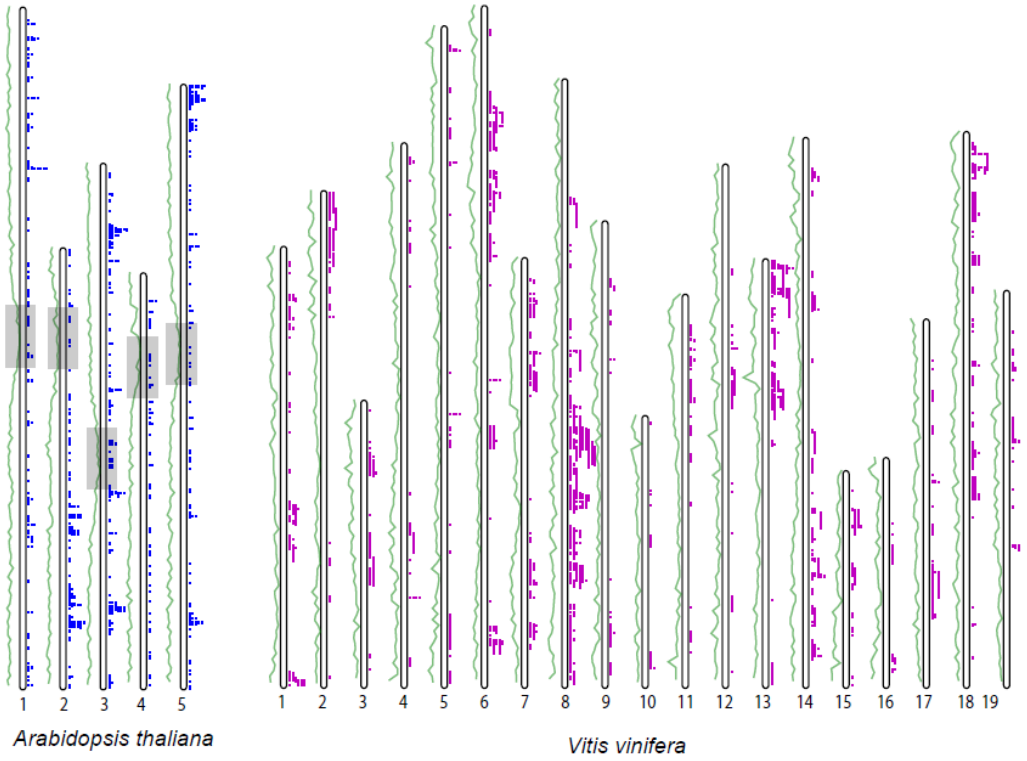


Figure 3.5 The alignment of GR contigs onto *Arabidopsis* chromosomes and *Vitis* chromosomes.

Blue and red bars next to the chromosomes show the GR contigs. The length of the bar represents the physical distance between anchoring markers on the target genome. Contigs are anchored only when two or more BES or marker sequences hit within a certain interval. In *Arabidopsis*, the interval is set to 200 kb, and in *Vitis*, 1 Mb**.**

3.3.9 Low-copy and repetitive DNA loci were concentrated in different regions of the genome

A total of 3060 contigs contain BACs to which one or more probes hybridized. Probes were classified as low copy (<30 hits total), moderately repetitive (31-97 hits), or highly repetitive (>100 hits). Accordingly, contigs were tentatively classified as repetitive or low-copy based on the ratio of repetitive probes versus low-copy probes hybridized to each contig. A total of 761 contigs contain only repetitive probes, and 1262 contigs contains mostly (>60%) low copy probes. Because a large number of the probes are designed from cotton EST sequences or *Arabidopsis* genes, contigs with relatively more hybridization anchors from low copy probes and relatively fewer from repetitive probes are likely to be gene rich. The 1262 low-copy probe enriched contigs contain 1786 of the 2300 non-repetitive probes. The majority of the low-copy probe enriched contigs (901 out of 1262, or 71.4%) are anchored to the cotton consensus map (Figure S1). By comparison, only 37.7 % (1586 out of 4208) of contigs overall could be anchored to the consensus map.

Repetitive contigs are slightly shorter than contigs enriched in low copy probes (average 38.32 CB units versus 44.35 CB units). This could be caused by co-migrating fragments produced by the repetitive sequences that reduce the total number of bands.

3.3.10 Low-copy probe enriched contigs appear to be largely euchromatic

Among the 438 contigs that showed microsynteny to *Vv* chromosomes, 218 are enriched in low-copy probes and only 14 are repetitive probe-enriched. Similarly, among the 370 contigs that showed microsynteny to *At* chromosomes, 166 were enriched in low

copy probes, and only 17 are repeat-enriched. This is consistent with our findings in other taxa that microsynteny tends to be preserved in gene-rich euchromatic regions but not in repeat-rich heterochromatic regions (BOWERS *et al.* 2005). We assume that the 761 repeat-enriched contigs are likely to be largely from heterochromatic regions of the genome and the 1262 low-copy sequence-enriched contigs are likely to be from euchromatic regions of the genome. The 1262 low-copy contigs can be estimated to cover 26% of the genome based on the estimated genome size of 880 Mb and average band size of 4097 bp. Based on the 68% of the genome estimated to be low-copy by renaturation kinetics(GEEVER *et al.* 1989), these contigs may cover about 38.2% of the low-copy DNA. Contigs aligned to *Vv* and *At* genomes contains 1150 (50%) and 954 (41.5%) of all non-repetitive probes. The low copy probes that were unable to align were partly due to the limitation of BLAST in searching across distant related species and the variation in gene density in *Vv* genome.

3.3.11 Consequences of ancient duplications in the *Arabidopsis thaliana* genome

To illustrate the alignment of GR contigs on the *At* and *Vv* genomes, ctg500 provides an example. The contig is anchored to a single *Vv* chromosomal location at about 14.7 Mb on chr8, and to four different locations on the *At* genome, at 15 Mb on chr2, 2.7 Mb on chr3, 20 Mb on chr3 and 0.1 Mb on chr5 respectively (Table 3.3, Figure 3.6A). These four *At* regions were previously shown to be paralogous segments created by two rounds of whole-genome duplication (BOWERS *et al.* 2003). The chromosomal region in *Vitis* has also been identified using MCScan (TANG *et al.* 2008b), to have conserved collinearity with the four *At* regions (Table 3.3, Figure 3.6B). Ctg500 is

Table 3.3 Anchored regions of contig500 on *At* and *Vv* chromosomes

GR ctg	*Vv* chr	Starting anchor (bp)	end anchor (bp)	*At* chr	Starting anchor (bp)	End anchor (bp)
ctg500	chr8	14578390	14980935	Chr2	15885694	15917946
				Chr3	2780924	2787267
				Chr3	20018330	20046188
				Chr5	54439	94007

A

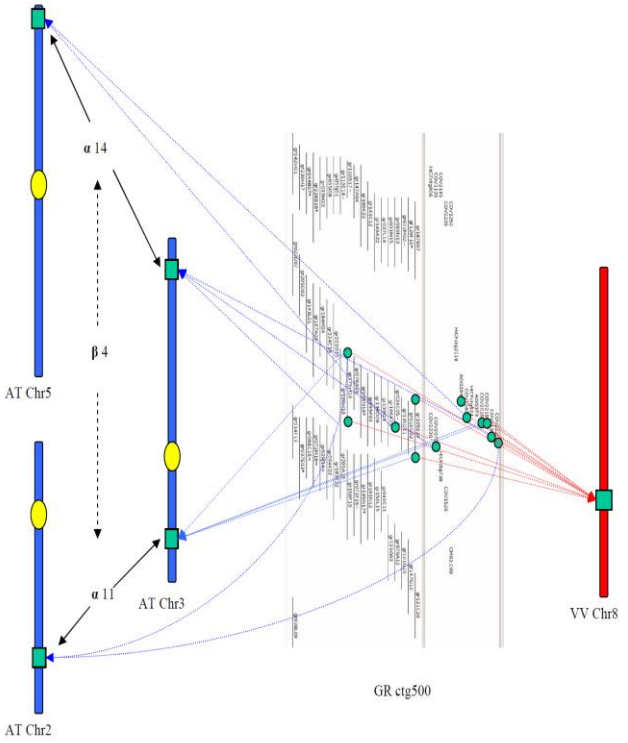
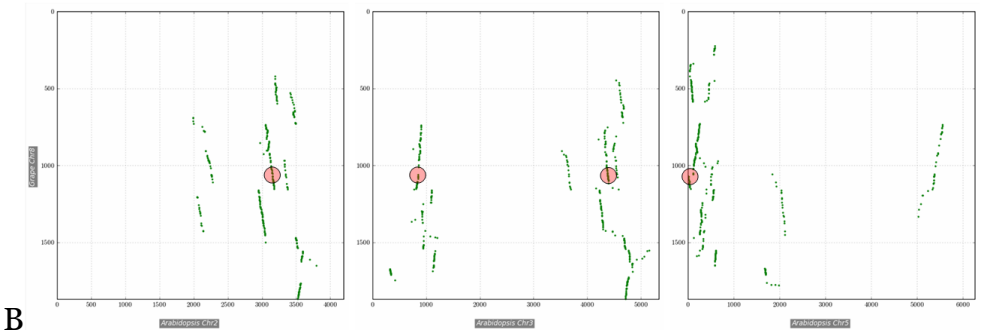


Figure 3.6 A sample contig (ctg500) showing homology to *Arabidopsis* and grape genome sequences.
A. The contig is mapped to four regions in *Arabidopsis*, which are identified as being paralogues produced by the alpha and beta duplications after the cotton-*Arabidopsis* divergence. The contigs is only anchored to a single *Vitis* chromosomal location. B. dot plot generated by MCscan on Plant Genome Duplication Database, showing conserved syntenic blocks between *Vitis* chr.8 and *Arabidopsis* chromosomes. The region corresponding to GR ctg 500 is marked by red circles.

B



anchored on cotton consensus homologous group 2, at around 67 cM. Based on cotton DNA markers, this region has shown evidence of homology to *Arabidopsis* alpha11 and alpha14 groups (RONG *et al.* 2005a).

3.3.12 The *G. raimondii* chloroplast

By aligning to the chloroplast DNA sequence of upland cotton (*Gossypium hirsutum*) using BLAST, BAC-end sequences and probes likely to be of chloroplast origin were identified. Ctg11556 is identified as a chloroplast contig. The contig contains 20 BACs, 10 of which are "buried" in FPC, meaning they have nearly identical band patterns as other BACs in the contig, indicating very high similarity among these BACs. COV1960, an overgo probe designed from the sequence of the chloroplast *psaJ* gene, hits 17 of the 20 BACs in the contig. Three BACs from the contig have end sequences, all of which correspond to the published *G. hirsutum* chloroplast sequence. (Figure 3.7). Based on low-coverage genomic sequencing with some targeted finishing, a D-genome chloroplast sequence has been assembled and is being described (M. Rahman, A. H. Paterson, in prep.)

3.3.13 GO analysis of BES and shotgun sequences

The 13,662 BES (BAC-end sequences) were analyzed using Blast2Go to obtain a distribution of functional gene groups. A total of 9042 did not have significant hits using BLASTx against NCBI nr database, 3234 of the sequences are annotated, and 963 were mapped, but not annotated. No significant differences were observed between the GO distribution of BES and random shotgun sequences except that more genes involved in

localization processes were represented in the random shotgun sequences. (Table 3.4, Figure 3.8).

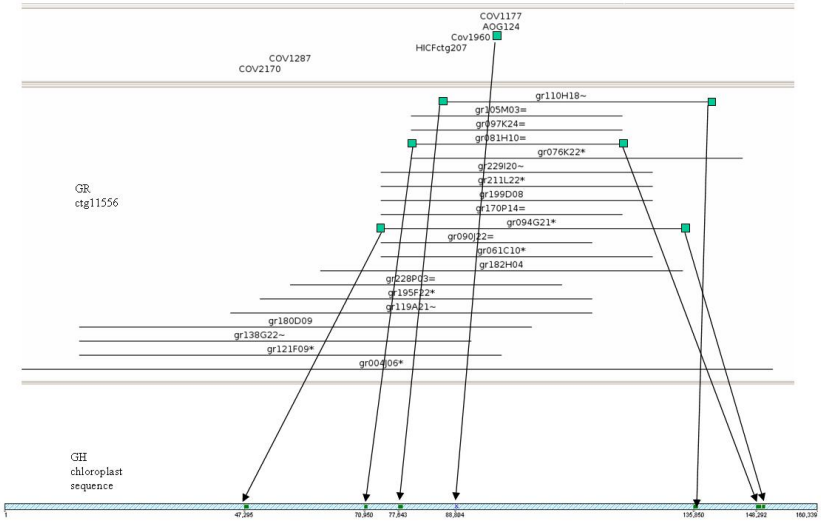
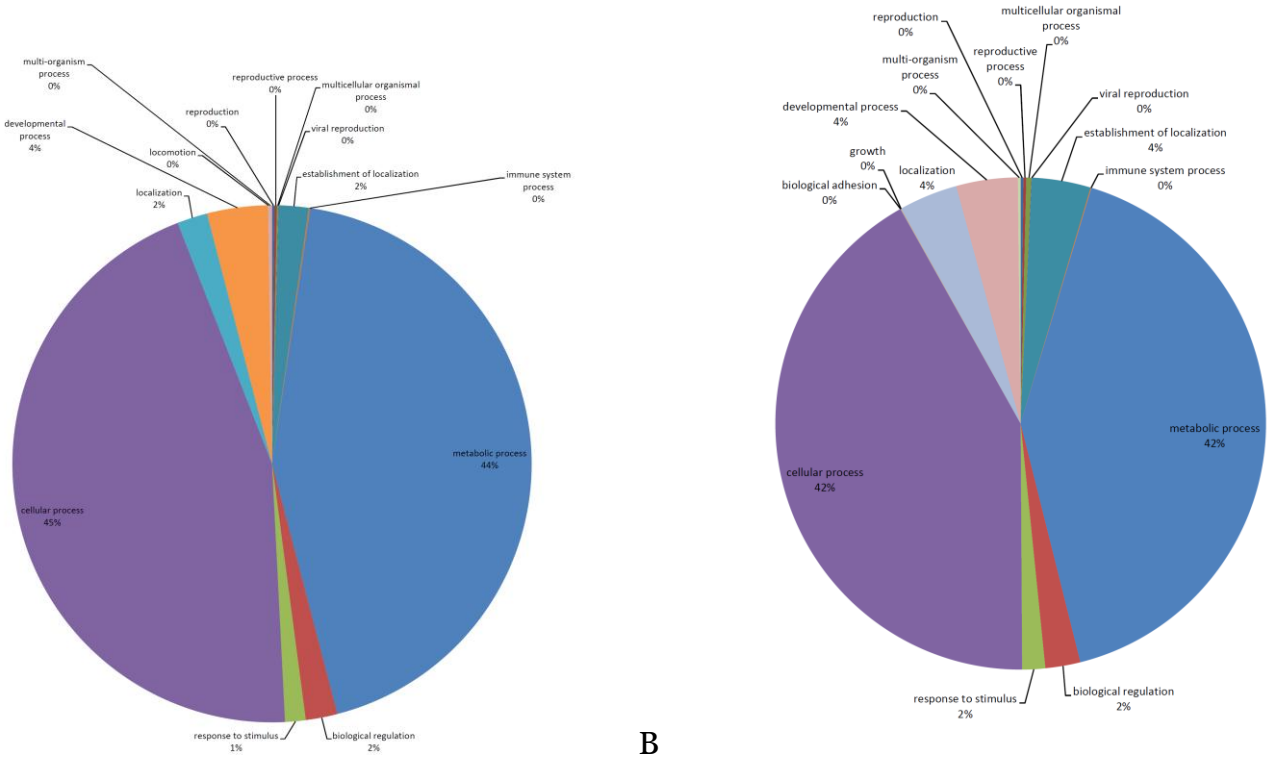


Figure 3.7 The GR chloroplast contig.
Contig11556 is identified as a chloroplast contig, with BAC-end sequences and an overgo probe aligned to the GH chloroplast sequence.

3.4 Discussion

The first whole-genome physical map of a cotton species has provided new tools and information, and foreshadows a picture of cotton genome organization prior to the completion of the D-genome sequence currently in progress. The genetically anchored contigs are potentially helpful in efforts such as gene cloning and local sequence analysis, by providing region-specific BAC resources for marker development and chromosome walking. On a genomic level, comparative analysis between cotton, *Arabidopsis,* and *Vitis* genomes illustrates the potential for translational genomics across these species, and several regions with an unusually high degree of conserved collinearity may be interesting for further research.



A B

Figure 3.8 GO analysis of BAC end sequences
(A) and random sampling of whole genome shotgun sequences (B)

Table 3.4 GO classification results generated from 13662 BAC-end sequences and 13661 random shotgun sequences, using Blast2Go at an ontology level of 2.

	BAC-end Sequences		random shotgun	
reproduction	7	0.15%	7	0.18%
reproductive process	7	0.15%	7	0.18%
multicellular organismal process	4	0.09%	13	0.34%
viral reproduction	1	0.02%	1	0.03%
establishment of localization	87	1.86%	148	3.87%
immune system process	4	0.09%	2	0.05%
metabolic process	2034	43.59%	1585	41.44%
biological regulation	92	1.97%	88	2.30%
response to stimulus	60	1.29%	58	1.52%
cellular process	2094	44.88%	1604	41.93%
locomotion	1	0.02%	-	-
biological adhesion	-	-	1	0.03%
growth	-	-	1	0.03%
localization	88	1.89%	148	3.87%
developmental process	177	3.79%	155	4.05%
multi-organism process	10	0.21%	7	0.18%
Total	4666		3825	

Several lines of evidence herein suggest that the *G. raimondii* genome is comprised of two qualitatively different components, specifically one that is gene-rich and recombinogenic with gene identities and order that is still recognizably similar to those in members of other angiosperm families (*Vitis, Arabidopsis*), and another that is repeat-rich and recombinationally-recalcitrant with relatively few genes that are highly rearranged relative to their homologs in other taxa. This general picture of cotton genome organization is similar to that which emerged from comparison of the two monocot genomes, rice and sorghum (BOWERS *et al.* 2005; PATERSON *et al.* 2009).

Curiously, we were able to anchor more contigs on the *Vitis* genome despite the closer relationship of cotton to *Arabidopsis*. This difference is attributable in part to differences in anchoring parameters (see Methods), but also reflects the relatively slow

evolution of *Vitis* (TANG *et al.* 2008b), and highlights the value of the *Vitis* genome as a botanical model for cross-taxon comparative genomic studies.

The present genome assembly remains somewhat fragmented and may be further improved as more information and new technology emerges. Adding more genetically anchored STS to the BACs, as well as mapping more BAC-derived sequences will permit anchoring of more contigs to their corresponding chromosomal locations.

3.4.1 Average band number is crucial in agarose based fingerprinting

The use of both agarose-based and HICF methods in this physical map assembly gave us the opportunity to directly compare these two methods which have been widely used in genome projects. Using only the agarose-based fingerprints, we obtained a large number of small contigs. This is mainly due to low average band number. FPC uses the Sulston score (SULSTON *et al.* 1989) as a cutoff criterion to call overlaps, $S=\sum_{j}^{n} C_{n,j} P^{j}(1-P)^{n-j}$. This is the probability of finding j matching bands in two BACs with n bands each. When S=cutoff=1e-12, the minimum matching band number required to call an overlap between two clones having 17 bands each (the average in the present study) is 12. One can predict expected contig numbers using the Lander-Waterman formula (LANDER and WATERMAN 1988), $E(contig\#)=Ne^{-LN/G\cdot(1-T/L)}$, where G is the genome length (genome size/average band size), L is the average band number; N is the number of BACs fingerprinted and T is the number of bands needed to call an overlap. If the gel length is 5000 bands, the genome size is 880 Mb, and the average band size is 4096 bp (for a 6-cutter), with a tolerance value of 7 and cutoff of 1e-12, the expected contig number in the

assembly would be over 9000 if the average band number per clone is 17. In other words, our agarose-based assembly yielded the expected result.

The expected contig number drops rapidly with increased average band number. From the Lander-Waterman formula, if the average band number is increased to 20, the expected contig number would be about 5000. With an average band number of 30, one would expect only about 400 contigs. This should be an underestimation because we are not considering physical gaps and under-represented parts of the genome in the BAC libraries, but nevertheless, shows how critical band numbers are to an agarose-based fingerprinting project. BACs with fewer than 8 bands offer too little information to form statistically-supported contigs, even with identical band patterns.

Our success with using HICF in a targeted manner to improve the physical map stems from much higher band numbers. HICF merged contig-end BAC pairs had average agarose band numbers that are not significantly different from the overall band number (18.02 vs. 18.15 in all BACs in contigs). The reason why they failed to join is due to the high percentage of matching bands needed to call an overlap. FPC was unable to call an overlap even if 11 bands were matching (Figure 3.1C).

3.4.2 Cross-contamination and chimeric clones in HICF

HICF has gained favor in recent physical mapping projects due to its high throughput and the large amount of data it provides per BAC. However, certain pitfalls still need to be considered. Unlike agarose-based fingerprinting, because of the way the size files are generated in HICF, it is difficult to go back and quality check the band calling for each clone and eliminate non-specific bands. This may help explain the

rightward-skew in the histogram for average band number across all BACs (Figure 3.1B). These non-specific bands would be potential causes of false joins (NELSON *et al.* 2005).

Cross-contamination seems to be a more severe problem in HICF than in agarose-based FPC assembly. In our first HICF assemblies, we encountered a very large contig containing as many as ~50% of all BACs, depending on the assembly stringency. This "dust ball" can be taken care of by excluding clones with suspiciously high band numbers (possible chimeras) and also by a newly implemented function in FPC to identify potential cross-contaminations. After the exclusion of 1166 BACs from the assembly as potential contaminants, the size of the contigs returned to normal, with no contigs containing >24 BACs.

3.4.3 Further improvements of the genetic-physical map

While contigs covering ~40% of the genome have been genetically anchored, a higher density of genetic markers may permit anchoring of many more contigs. Some genetically mapped probes hit only singleton BACs and were not incorporated into the physical map in the interest of minimizing false positives. Nearly 1000 probes that hybridized to GR BACs are from sequences that have not yet been genetically mapped, so they are not useful in linking the genetic and physical maps. Designing new overgo probes from mapped sequence-tagged sites can be done recursively as more densely populated genetic maps become available. Conversely, new SSR markers can be developed from BES and put onto the genetic map, which would help anchor more contigs and help confirm the position of those already anchored.

3.4.4 Probes targeted at specific regions of interest

Marker density on the physical map reflects efforts to enrich specific genomic regions containing genes of interest for DNA markers. Most prominent are probes aimed at the *Li1* (Ligon lintless-1) and *Li2* (Ligon lintless-2) genes of cotton. About 300 overgo probes were designed from genetic markers and EST reads that showed relationship to the regions of these genes. This enrichment created "hotspots" where more GR contigs could be aligned to both *Arabidopsis* and *Vitis* (Figure 3.5). In the *At* genome, there is an excess of anchored GR contigs near one end of chromosome 2, the upper and lower parts of chromosome 3, and the telomeric region of chromosome 5. These four regions were identified in earlier studies (BOWERS *et al.* 2003) to have been produced by two rounds of whole-genome duplication, all belonging to the consensus group ß4. Likewise, the regions near the top of *Vv* chromosome 13 and bottom of chromosome 8 anchor a higher than average number of GR contigs.

A closer look at these "hotspots" revealed that the majority of the contigs anchored here contain probes from the *Li1* and *Li2* regions. There are 114 contigs anchored in the *At* regions described above, 94 or 82.5% of which contain *Li1* and/or *Li2* probes. In 87 out of these 94 cases, the *Li* probes provided one or more anchor points in microsynteny detection. In grape, a total of 134 contigs fell into the most densely anchored regions on grape chromosome 6, 8 and 13; 111 or 82.8% of these contigs contain *Li1* or *Li2* probes of which 92 provided one or more anchor points in microsynteny detection. Compared to the whole-genome average of 23% (970) contigs that contains *Li* probes, these regions shows a significant enrichment in *Li* contigs and the ability to align to the *At* and *Vv* genomes.

This illustrates the potential use of the contig assembly in cross genome comparisons, and that the power to detect synteny and align contigs across genomes can be greatly increased by targeted enrichment of specific regions for hybridization probes.

3.4.5 The grape genome as a model

Aligning physical map contigs with sequenced genomes has proven informative in several ways (BOWERS *et al.* 2005; SNELLING *et al.* 2007). Comparative mapping data and BES alignments to the human genome helped in assigning bovine physical map contigs to their respective chromosomes (SNELLING *et al.* 2007). The pattern of sorghum physical map contigs along rice chromosomes has given empirical evidence that gene rearrangement is generally deleterious (BOWERS *et al.* 2005). Cross-species synteny information has also enabled us to make better use of the sequenced genome data on other genomes.

For cotton, *Arabidopsis* is the currently most closely-related genome for which a sequence is published. The rapid evolution of, and two additional WGD (Whole Genome Duplication) events in, the *Arabidopsis* lineage may reduce our ability to align these respective genomes. The *Vitis* genome, on the other hand, evolves relatively slowly (TANG *et al.* 2008b) and has experienced no WGD events apart from the hexaploidy (γ) event that is likely to be shared by all dicots (JAILLON *et al.* 2007; TANG *et al.* 2008a). The grape genome might prove to be more useful than that of *Arabidopsis* in comparative genomics across distantly related species.

One disadvantage of using the grape genome as a model for cotton lies in its relative low gene density compared to *Arabidopsis*. Unlike sorghum and rice, where the euchromatic regions have a similar gene density (BOWERS *et al.* 2005; PATERSON *et al.*

2009), gene density is at least twice as high in *Arabidopsis* as in *Vitis*. Gene density across the currently assembled grape pseudomolecules fluctuates from about 20 to 25 genes per 200 kb in higher gene density regions to 10 to 15 genes per 200 kb in lower gene density regions. Similar analysis showed that gene density is uniformly 50 to 60 genes per 200 kb across the *Arabidopsis* genome, except for the centromeric regions and a few low density points with 30 to 40 genes per 200 kb. This lower gene density in *Vitis* reduces our ability to anchor cotton contigs, and look for synteny using contig information. Here, we were able to anchor cotton contigs onto most of the gene dense regions of the *Vitis* genome, but large parts of the low-gene-density chromosomal regions are not covered.

3.4.6 Using the genetic-physical map in gene cloning

Map-based cloning has always been a long and tedious process. The genetic-physical map provides a shortcut by which contigs spanning a target gene region can be readily identified through flanking markers. Markers immediately upstream and downstream of a target gene can be used to identify neighboring anchored contigs, and sequencing of BACs within the contig(s) could provide candidate genes for further study. In efforts to characterize a gene involved in cotton fiber development, we were able to identify a contig that anchors to a genetic region of interest using this method, and design new genetic markers very close to the gene (unpublished data).

The value of the physical map for positional cloning would be further enhanced by anchoring more contigs onto the genetic maps efficiently and accurately. We have provided a framework on which more than 1500 contigs have been aligned. In genomic regions of high priority to specific research efforts (positional cloning, etc.), many

unanchored contigs might be tentatively merged into the anchored contigs, given a lower stringency or higher tolerance for questionable clones, then seeking additional corroborative data such as additional BAC ends, hybridization anchors, or targeted genetic mapping of hybridizing elements. For regions where no contigs have been anchored yet, a simple probing of the library using flanking genetic markers should be able to help build a local genetic, physical map. Contigs upstream and downstream of a target contig can be identified by manually searching for similar contigs at a lower cutoff, and rebuilding the contigs for the region of interest.

Microsynteny information permits one to utilize new ways of developing genetic markers targeted to a region of interest (FELTUS *et al.* 2006) that may be of high value in translating functional information from botanical models to cotton. The contigs aligned to the *At* and *Vv* genomes cover about ¼ of these respective genomes, primarily in regions that are likely to be gene-rich. Earlier research has identified some *Arabidopsis* genes with well defined roles in trichome and root hair development that approximately correspond to the locations of cotton fiber QTLs. Some of these genes are in regions which showed conserved organization with the GR physical map contigs. e.g., an α-tubulin gene (*TUA6*) is found in a region spanned by contig1653 and contig3177; the *TTG2* gene, which is involved in trichome pattern formation (ISHIDA *et al.* 2007), is in a region spanned by contig937; the *ACT2* gene, which involves in trichome morphogenesis(NISHIMURA *et al.* 2003), is in a region spanned by contig908; the *GL2* gene is spanned by contig601. These anchorings may provide a good starting point to search for candidate genes and QTLs with similar functions in cotton fiber development, and help elucidate the similarities and differences in trichome formation in different tissues.

CHAPTER 4

PROGRESS TOWARD CLONING THE LIGON LINTLESS-2 (*LI2*) GENE INVOLVED IN COTTON FIBER DEVELOPMENT

4.1 Introduction

The cotton fiber is one of the most important natural resources in our everyday life. The four species of cultivated cotton generate hundreds of billions of dollars in Gross Domestic Product (GDP) worldwide, through the textile industry. Cotton fiber is also thought to be the longest single cell in the plant kingdom (KIM and TRIPLETT 2001), making it a unique system for studying trichome development and cell elongation, as well as secondary cell wall synthesis and cellulose metabolism.

There are two different forms of cotton fibers (ovular trichomes): lint or fuzz fibers (LANG 1938). Lint fibers are normally several centimeters long, and initiate at anthesis; fuzz fibers, which are much shorter (a few millimeters), initiate a week later. Several mutants that show either no lint fiber but normal fuzz fiber growth, or no fuzz but normal lint growth have been identified and mapped (RONG *et al.* 2005b), suggesting that lint and fuzz fiber development are at least partly controlled by different sets of genes. However, double mutants of two fuzz-less loci (with intact lint fiber) lack both fuzz and lint fibers (TURLEY and KLOTH 2008), indicating that the two processes overlap to some degree.

To elucidate the processes involved in cotton fiber initiation and development, it would be valuable to characterize and clone major genes in the associated biochemical pathways. Despite many efforts made in mapping of discrete mutants and QTLs

influencing fiber properties, to date, no genes determining genetic variation of these properties have been identified. A widely adopted method for plant gene cloning consists of three major steps: a. fine-scale genetic mapping of the targeted region; b. characterization of a set of contiguous clones or sequences that contains the gene of interest; and c. identification of the causative loci by complementation tests. Such 'positional cloning' in cotton has been difficult due to the lack of genetic (step a) and physical (step a and b) resources. However, recent years have seen a boost in cotton genome research. Databases of cotton genetic markers, multiple genetic maps and ESTs were set up (BLENDA *et al.* 2006); a BAC based D genome physical map was built (Chapter 2 of this Thesis); comparative genomic analysis between cotton and *Arabidopsis* (RONG *et al.* 2005a), and gene expression analysis of many different tissues have been performed (ALABADY *et al.* 2008; CHAUDHARY *et al.* 2009; RAPP *et al.* 2009; TALIERCIO and BOYKIN 2007; UDALL *et al.* 2007; WU *et al.* 2005). These datasets are gradually increasing knowledge of the cotton genome and its genes, and may help in expediting the chromosome walking process.

The *Li2* mutant (KOHEL *et al.* 1992; NARBUTH and KOHEL 1990) was discovered in a cotton breeding nursery. It had no lint fiber on the seed coat, while the fuzz fiber was intact. The lintless phenotype of *Li2* mutants resembles that of the previously discovered *Ligon lintless-1* (*Li1*) mutant (GRIFFEE and LIGON 1929), hence the name. However, *Li2* mutants have normal vegetative growth while *Li1* has deformed leaves and stems. *Li2* seeds are also significantly lighter than *Li1* seeds(NARBUTH and KOHEL 1990). The mutant phenotype is completely dominant, and is controlled by a single gene (RONG *et al.* 2005b). While *Li1* maps to the suspected centromeric region of Chr 22, *Li2* maps to the top of chromosome 18 (RONG *et al.* 2005b), in a region thought to have a

favorable genetic/physical distance ratio. According to earlier publications (RONG *et al.* 2005b), 9 markers within a 15cm region of *Li2* were available. Only one marker, A1552 was thought to map to the distal side of the gene.

We have taken three interconnected approaches toward identifying the *Li2* gene: (1) fine mapping was carried out using published SSR markers as well as RFLP –derived PCR markers; (2) the D genome physical map was used to identify BACs and contigs that anchor to the *Li2* region, and; (3) synteny information among cotton, *Arabidopsis* and *Vitis* was used to facilitate marker development and gene prediction. This approach yielded a detailed genetic map of the immediate vicinity of the *Li2* gene, as well as a BAC contig that covers most of this region.

4.2 The identification of BAC contigs anchoring to the *Li2* region

The position of the *Li2* gene (tip of the chromosome) has made it very hard to find existing markers that flank the telomeric side of the gene. Starting with a preliminary mapping of all published SSR markers from different maps in this region, we decided to screen for BACs/contigs that contain genomic sequences from this region. By sequencing these BACs, we hoped to be able to derive new markers that would allow us to move closer to the gene.

4.2.1 BAC library screening

Overgo probes were designed from genetic markers mapping closely to the gene. By applying these probes to cotton BAC libraries, we identified BACs that contain genome fragments from the *Li2* region.

From previous genetic mapping (RONG *et al.* 2005b), RFLP markers A1552 and COAU2K07 were identified as closest to the *Li2* gene. From cotton EST sequences

identified by *Arabidopsis* synteny to the *Li2* region (detailed in later sections), we identified another two markers that were closely linked to the gene: *Li2*-01 and *Li2*-02. These four markers hit 124 BACs (Table 4.1) across 5 different libraries, namely GAD (standing for *Gossypium* AD genome)(*G. barbadense*), GAHIN (standing for *Gossypium* A genome HindIII digest) (*G. arboreum*) , GAMBO (standing for *Gossypium* A genome MboI digestion) (*G. arboreum*), GR (*G. raimondii*) (Chapter 2 of this thesis) and Acala MAXXA (*G. hirsutum*) (TOMKINS *et al.* 2001).

Table 4.1 Number of BACs hit by probes derived from the *Li2* region**.**

	GAD	GAHIN	GAMBO	GR	MAXXA	Total hit BACs
A1552	5	0	0	9	7	21
COAU2K07	3	10	5	1	5	24
Li2-02	22	5	5	9	0	41
Li2-01	8	7	3	20	0	38
Total hit BACs	38	22	13	39	12	124

4.2.2 Identification of a physical map contig of the *Li2* region

We made use of the GR physical map (as described in Chapter 2) to identify contigs that anchor to the *Li2* region, as a starting point for chromosome walking. This has several advantages: first, with more anchor points, we are more confident in distinguishing real *Li2* BACs from false positive hits or homoeologs. Secondly, the contigs span a much larger physical distance and allow us to walk faster toward the gene.

GR BACs and their corresponding physical map contigs are identified by hybridization using probes designed from closely linked genetic markers. With four probes from genetic markers closely mapped to the *Li2* gene showing correspondence, one contig is selected for further analysis.

In order to further confirm the identity of the contig, we selected the BAC with most *Li2* probe hits from the identified contig (*Li2* BAC01 here after) for shotgun sequencing (sequence analysis detailed in later sections). The sequencing reads assembled into two separate contigs. The two contigs are oriented using BAC vector as a bridge. From the BAC sequence, we designed 15 pairs of CISP (Conserved Intron-Scanning Polymorphism (FELTUS *et al.* 2006)) primers. Of these primers, 1 failed to amplify, and 11 produced amplicons that were not polymorphic. From the 3 pairs of polymorphic primers, we picked the one with the clearest bands and mapped it with a population of 154 plants. The marker (*Li2*-36) was found to co-segregate with the *Li2* phenotype in our mapping population. This validates the anchoring of this contig at the tip of chromosome 18.

4.2.3 Rebuilding and extension of the GR contig that anchors to the *Li2* region

The *Li2* contig we have identified is from a whole genome physical map assembly (Chapter 2 of this Thesis), which involved multiple rounds of auto-merging and splitting of all contigs of the whole genome. Although this method is robust in building a large set of reasonably accurate contigs on a whole-genome scale, there could be assembly errors and missing clones in any one specific contig. In order to further validate the assembly of the *Li2* contig, and to attempt to extend the existing contig, we manually built a contig starting from the confirmed *Li2* BAC01 and extending in both directions. Each BAC in the identified contig was used to search for matching BACs among the 92,160 GR BACs that were fingerprinted (Chapter 2 of this Thesis). Contigs that contain a number of BACs that overlap with the *Li2* contig were forced to merge in FPC. This was done

recursively until no significant overlapping contigs could be identified (Figure 4.1). The merged contig was then evaluated in FPC by recalculation of to CB maps at cutoff value of 1e-6. “Q-clones” (questionable clones) were excluded from the final merging (Figure 4.1).

Probe hybridization data was also taken into consideration in the manual merging of contigs, but mainly as a secondary confirmation to add confidence in the merges. From the sequenced BAC, three genetically mapped DNA markers were found. These markers, along with previously identified markers through hybridization, suggested possible overlaps with other BACs/contigs from hybridization results.

4.2.4 Sequencing of additional BACs and orientation of the *Li2* BAC contig

From the extended BAC assembly, two BACs, one upstream (*Li2* BAC02) and one downstream (*Li2* BAC03) of *Li2* BAC01 were shotgun sequenced. In both cases, the size of the largest sequence contigs assembled is approximately the estimated insert size of BACs in the GR library (Chapter 2 of this Thesis). The unassembled contigs contain mostly low quality reads and are very short (<1 kb). The three BAC sequences were checked in Sequencher for overlaps, and *Li2* BAC03 is found to overlap with *Li2* BAC01 contig1.

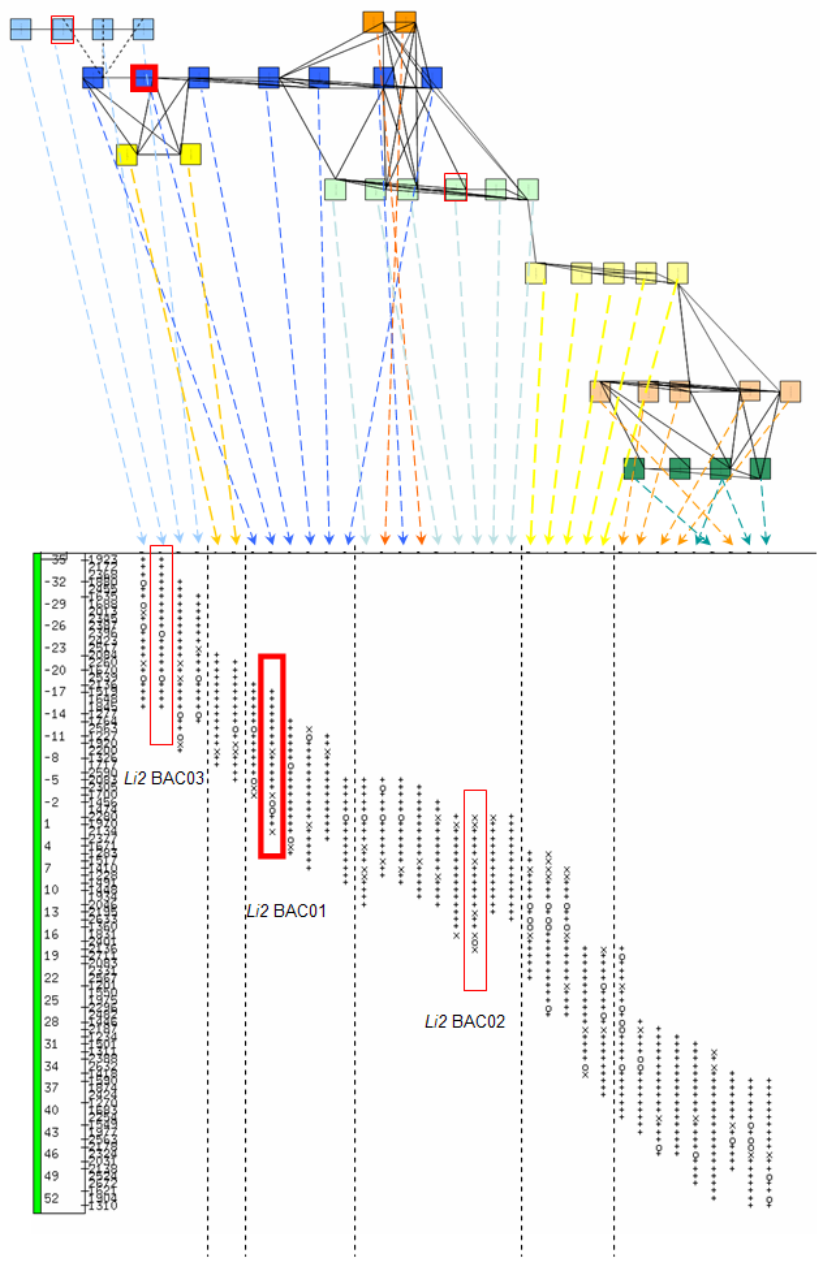


Figure 4.1 Extension of the *Li2* contig.

In the upper panel, each square represents a GR BAC; different colors indicate a different contig in the preliminary assembly. BACs that show significant overlap relative to the threshold of 1e-6 are connected by thin lines. The lower panel shows the CB (Consensus Bands) map view from FPC, indicating the order of the BACs within the new contig. The sequenced BAC is marked out by a red rectangle.

4.2.5 The development and fine mapping of new markers

From these BACs, new SSR and CISP markers were developed. SSR markers were designed from BAC sequences using CID (http://www.shrimp.ufscar.br/cid/index.php) (FREITAS *et al.* 2008). CISP (Conserved Intron Scanning Polymorphisms) markers were developed as described (FELTUS *et al.* 2006): Cotton EST assembly sets were used with BLASTn to identify gene structure from BAC sequences; primers are then designed to specifically amplify putative intron sequences. A Perl script was used to automate the process.

Twenty-eight pairs of SSR primers and 19 pairs of CISP primers were generated for *Li2* BAC02; 12 pairs of CISP primers were designed for *Li2* BAC03. Out of these primers, only three are polymorphic: 2 dominant SSR markers and one co-dominant CISP markers were mapped to the *Li2* region, validating the contig assembly. (Figure 4.2)

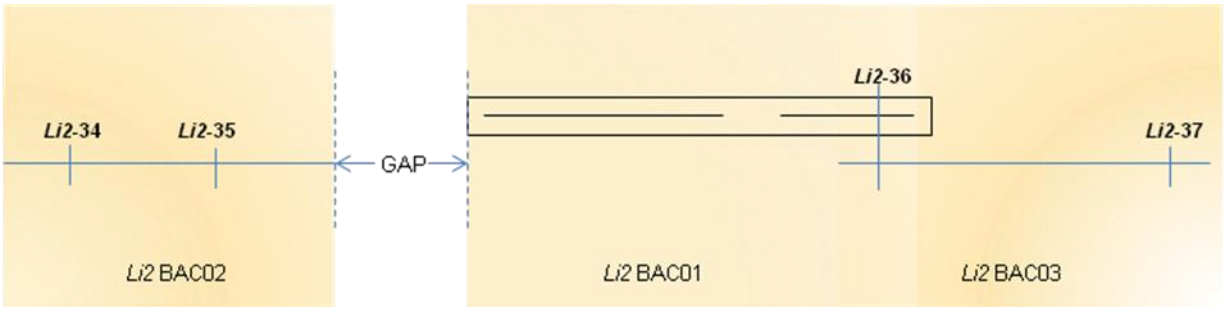


Figure 4.2 The relative position of the three sequenced BACs and the position of new markers developed and candidate genes**.**

4.3 The fine mapping of the *Li2* region

4.3.1 Preliminary results and plant materials

In a previous study, the *Li2* gene was genetically mapped along with 6 other fiber mutants (RONG *et al.* 2005b), using markers from a reference map (RONG *et al.* 2004). In a population of 154 F2 plants, *Li2* was placed near the tip of chromosome 18, with one dominant marker mapping to the distal side of the gene. The resolution of this mapping was restricted by the size of the mapping population and the dominant nature of several of the markers used. In order to get a higher resolution map of the region, and confirm the ordering of markers with more confidence, we needed to select/develop more markers that map to the top of Chr. 18, as well as generate a larger mapping population.

Figure 4.3 Phenotype of the seeds containing the *Li2* mutant allele, and homozygous WT allele.

The *Ligon-lintless 2* (*Li2*) mutant strain is in a *G. hirsutum* background (RONG *et al.* 2005b). *G. barbadense* cultivar Pima S-7 was used as a common parent to cross with the mutant strain. All F1 progeny displayed the mutant phenotype.

In order to develop a larger mapping population, F2 seeds were collected from selfed F1 plants. A total of 980 F2 individuals were planted in batches of 96 plants, in Ray Leach "Cone-tainers"™ (Stuewe & Sons, Inc.) in the green house at the University of Georgia, Athens. The mutant and wild type phenotype of the F2 plants are shown in Figure 4.3.

4.3.2 A list of candidate markers to the *Li2* region

To identify more markers targeted at the *Li2* region, we used the following approaches: 1. SSR markers were selected from published genetic maps; 2. Markers were developed from cotton-*Arabidopsis* synteny relationships; 3. Markers were developed from selected BAC sequences.

Developing markers from *Arabidopsis* synteny is a relatively new approach. Homologous regions to the cotton *Li2* region in the *Arabidopsis* genome were identified in an earlier study (RONG *et al.* 2005a). *Arabidopsis* genomic sequences from the homologous regions were used to BLAST against cotton EST databases, identifying candidate ESTs that are tentatively related to the *Li2* region. PCR primers were designed from the identified cotton ESTs and screened for polymorphisms.

In the third approach, BACs were identified from cotton genomic libraries. As described earlier, a D genome (*Gossypium raimondii*) BAC contig is anchored to the *Li2* region. SSR and CISP markers were developed from the BAC sequences, and four markers have shown polymorphism that map to the *Li2* region.

Altogether, 37 markers from these different sources were tried using the previous mapping population (Table 4.2). Some were monomorphic in our population. Among

the ones that are polymorphic in our population, we were able to map 16 markers to the *Li2* region.

4.3.3 Selection of markers for fine mapping

To integrate markers from different genetic maps, the selected markers (Table 4.2) were genotyped using 135 plants from the previous mapping population, and genetic distance was recalculated. The genotyping data were analyzed in JoinMap3.0. The rebuilt linkage is shown in Figure 4.4. With the integration of more co-dominant markers, the position of the markers on the previous map, especially dominant markers changed appreciably. In particular, all markers mapped to the centromeric side of the gene.

In order to gain a higher resolution map of the *Li2* region, 8 markers were selected according to these preliminary screening results. Three markers from published maps with the clearest band patterns were selected: *Li2*-10, *Li2*-26 and *Li2*-19. These markers, along with the 4 new markers developed from BAC sequences from this region (detailed in the next section), were fine mapped using a combined population of ~700 F2 individuals. Possible recombinants were rechecked a second time from DNA extraction to rule out possible scoring and experimental errors. The final order of these markers is shown in Figure 4.5.

Table 4.2 A summary of markers used in fine mapping of the *Li2* region

Marker name	Polymorphic in our population	mapped to *Li2* region
Li2-01	CAPS	Yes
Li2-02	Yes	Yes
Li2-03	No	-
Li2-04	No	-
Li2-05	No	-
Li2-06	No	-
Li2-07	Yes	Yes
Li2-08	Yes	unstable
Li2-09	Yes	unstable
Li2-10	Yes	Yes
Li2-11	No	-
Li2-12	Yes	Yes
Li2-13	No	-
Li2-14	No	-
Li2-15	No	-
Li2-16	No	-
Li2-17	No	-
Li2-18	No	-
Li2-19	Yes	Yes
Li2-20	No	-
Li2-21	No	-
Li2-22	No	-
Li2-23	No	-
Li2-24	No	-
Li2-25	Yes	Yes
Li2-26	Yes	Yes
Li2-27	Yes	Yes
Li2-28	Yes	Yes
Li2-29	No	-
Li2-30	Yes	Yes
Li2-31	No	-
Li2-32	No	-
Li2-33	No	-
Li2-34	Yes	Yes
Li2-35	Yes	Yes
Li2-36	Yes	Yes
Li2-37	Yes	Yes

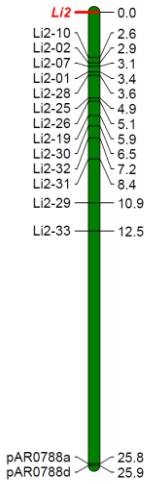


Figure 4.4 The reconstructed linkage map of the *Li2* region combining markers from different genetic maps.

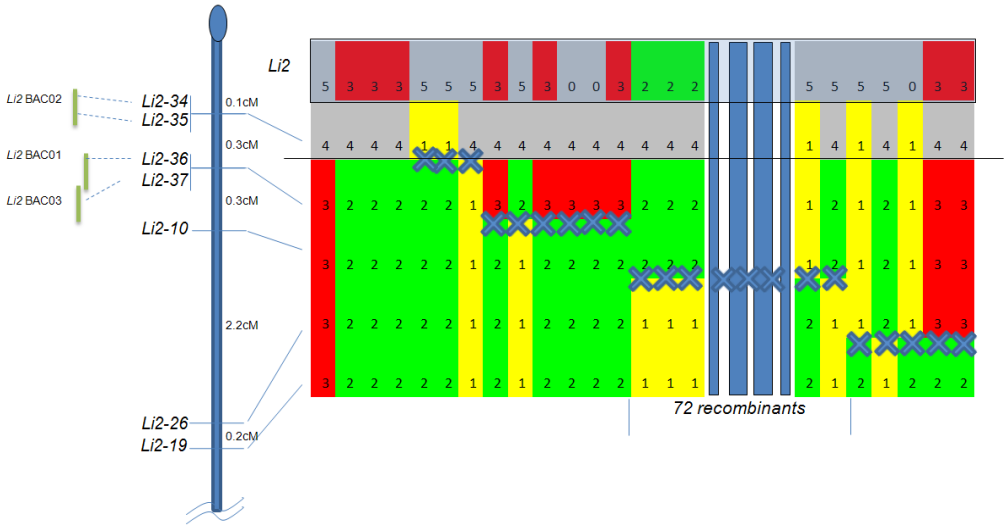


Figure 4.5 Placements of BAC-sequence-derived markers in the *Li2* region.
The number in the genotype table indicate: 1, homozygous for *G. hirsutum* allele; 2, heterozygous; 3, homozygous for *G. barbadense* allele; 4, either heterozygous or homozygous for *G. barbadense* allele and 5, either heterozygous or homozygous for *G. hirsutum* allele and 0 missing data**.** Cross-overs are indicated by blue crosses.
*The ordering between *Li2* phenotype and the markers *Li2*-34 and *Li2*-35 cannot be confidently determined. MapMaker has calculated identical log-likelihood values for different orders among the three loci. The log-likelihood value for placing *Li2* before and after *Li2*-36 and *Li2*-37 differs by 8.37, large enough for a confident placement of the gene at the distal side of the markers *Li2*-36 and *Li2*-37.

Among these new markers developed from BAC sequences, *Li2*-36 and *Li2*-37 were co-dominant markers, while *Li2*-34 and *Li2*-35 were dominant markers. Unfortunately, we were not able to identify co-dominant markers from *Li2* BAC02.

The four markers from the sequenced BAC were found to be distal to all existing markers, confirming that the BAC contig we have identified has advanced progress toward the gene. *Li2* BAC02 is closer to the gene than the other two BACs, orienting the contig. Due to the fact that markers from *Li2* BAC02 are both dominant, we are not able to resolve the relative position between the *Li2* gene and the BAC. To do this would require further study of F3 phenotypes and the development of new co-dominant markers, to distinguish homozygote from heterozygotes for both the phenotype and the markers.

4.3.4 Segregation distortion

The *Li2* mutant parent bears seeds with no lint fiber, while fuzz fiber development seemed normal. The F1 plants closely resemble the mutant parent, indicating that the Li2 mutant phenotype is completely dominant, thus the ratio of lintless individuals to individuals with lint fiber in a F2 population produced from the selfing of F1 plants is expected to be 3:1. However, in the F2 plants whose phenotypes are available so far, 192 showed a lintless phenotype, and 105 were wild type phenotype (the phenotyping is still ongoing), or 1.83:1. The neighboring markers also showed patterns of segregation distortion favoring the wild type parent (Table 4.3).

Segregation distortion can be caused by many different factors including experimental techniques, residual heterozygosity in parental lines, the existence of a segregation distortion locus (SDL), etc. What we have observed here is a distortion

biased towards the *G. barbadense* alleles. It could be that this part of the GB genome contains a favorable allele compared to the GH counterpart. It is also possible that the mutation in the *Li2* gene not only resulted in the lintless phenotype, but also affects the viability of the seeds containing the homozygous mutant allele.

Table 4.3 Segregation distortion of three representative markers tested in the *Li2* region.

	Li2-36			*Li2*-19			*Li2*-26		
Genotype	observed	expected	Chi test	observed	expected	Chi test	observed	Expected	Chi test
GH/GH	130	175	0.0002	140	165.75	0.030	129	175	0.0003
GH/GB	366	350		335	331.5		378	350	
GB/GB	204	175		188	165.75		193	175	
Total	700			663			700		

4.4 Gene identification from BAC sequences and next steps

4.4.1 Possible roles of *Li2*

It is relatively rare for a mutant phenotype to be dominant over the wildtype phenotype. In the case of *Li2*, the F1 and the heterozygous plants are lintless, indicating that the mutant allele is dominant over the wild type allele. One possible explanation is that since all other cotton species besides the A genome lineage and the tetraploid lineage (which contains the A subgenome) are lintless, the production of lint fiber is by itself a "mutant" phenotype, thus making our *Li2* plants revertants. Intuitively, this might be caused by a transposon insertion in the A genome lineage that caused the production of lint fiber, with transposon excision restoring the lintless phenotype in the *Li2* mutant plants (Paterson, unpublished discussions). Another possibility is that *Li2* is

a suppressor (e.g., miRNA gene) that is specific to the A genome lineage, and the *Li2* mutation disables the suppressor, releasing the gene that restores the lintless phenotype.

4.4.2 Next steps

We have identified and fine mapped several SSR markers that are very closely linked to the *Li2* locus; established an F2 population consisting of ~700 plants for future fine mapping of new markers; and identified a physical map contig that anchored to the region. However, several pieces are still missing before we can confidently locate the gene. First, to place the gene in an interval, we would need to identify marker(s) from the telomeric side of the chromosome. Unfortunately, all markers published so far appear to be on the centromeric side of the gene. The first round of chromosome walking has determined the orientation of the contig. Further sequencing of BACs and development of new markers may be needed to construct a contig that spans the gene, and provide an upper bound of candidate genes. To increase the current resolution of the mapping, F3 phenotyping will be done for a subset of individuals with dominant GH phenotypes to distinguish homozygotes and heterozygotes.

CHAPTER 5

NEW EVIDENCE OF ANCIENT GENOME DUPLICATION EVENTS IN DIPLOID COTTON GENOMES[4]

[4] Lin, L, Tang, H., *et al*. To be submitted to *BMC Genomics*

Abstract

It has been suggested by earlier genome mapping studies that diploid cotton might be an ancient polyploid. Here, we used the *Vitis* genome as an out group and used two different approaches to further explore evidence regarding ancient whole genome duplication (WGD) in the diploid cotton lineage. All-against-all gene dotplots showed several cases where one grape chromosomal segment is collinear with two separate regions on the cotton consensus map. A local level comparative analysis using cotton BAC sequences and their homologous regions in sequenced eudicot genomes also showed that an appreciable number of homologous genes have been lost in the cotton lineage, resembling the pattern of diploidization after WGD. Gene densities in corresponding regions from cotton, grape, *Arabidopsis* and papaya genomes are similar, despite their huge genome size difference and different number of WGDs each genome has experienced, which supports the notion that genome expansions are usually caused by transposon insertions that happen in heterochromatic regions.

5.1 Introduction

Whole genome duplication (WGD) events have been more frequent in the lineages of flowering plant species than in most other taxa. With more plant genomes being sequenced and released, and the emergence of new tools for genome comparisons, our understanding of the history of genome duplication and its importance in angiosperm evolution is becoming clearer. An ancient genome triplication event is very likely to have been shared by all eudicots (JAILLON *et al.* 2007; TANG *et al.* 2008a), and different lineages have undergone additional rounds of WGD (BOWERS *et al.* 2003; TANG *et al.* 2008a). Indeed, all eudicot genome sequences released so far except *Vitis* and *Carica* have lineage specific genome duplication events.

WGD profoundly affects the genomic landscape of modern plants (SEMON and WOLFE 2007). Synthetic polyploid plants experience abrupt CpG methylation changes after genome doubling (LUKENS *et al.* 2006). Interchromosomal rearrangements increase after WGD in teleost fish (POSTLETHWAIT *et al.* 2005). Duplicated genes created by WGD behave differently from single gene duplications, showing a longer life span before one copy is deleted (LYNCH and CONERY 2000). In the study of the cotton physical map (Lin *et al.* unpublished), we were able to align more cotton contigs to the *Vitis* genome despite its much longer divergence time from cotton than *Arabidopsis*. One explanation of this is that *Arabidopsis* has experienced two more rounds of WGD than *Vitis*, indicating that WGD and subsequent genome changes have a more profound effect on the preservation of synteny than millions of years of genome evolution (Lin *et al.* unpublished). Multiple rounds of WGD and associated diploidization (gene loss) complicated comparative genomic analysis. Elucidation of the histories of WGD in angiosperms helps to mitigate this complication.

The fact that cotton has a base number of 13 and several related genera have many species with n=6 has long hinted that there might be at least one round of WGD in the cotton lineage. The history of duplication of the cotton lineage is not yet well understood. It appears likely that "diploid" (2n=26) cotton might have experienced at least one round of whole genome duplication event since the triplication shared by all eudicots (MURAVENKO *et al.* 1998; RONG *et al.* 2004; RONG *et al.* 2005a), based on the classic cytogenetic analysis, Ks distributions of all duplicated gene pairs from the cotton unigene set, and possible homoeologous relationships among multiple chromosomal segments within the cotton genome (RONG *et al.* 2005a). However, as stated by the authors, the homology detected is correlated to marker density, which might indicate

some false inferences; and intercentromeric gene movement may also cause false positives (Rong *et al.* 2005a). Thus, although ancient lineage specific WGD in cotton has been strongly indicated, definitive proof is still lacking.

In sequenced genomes, a common way to search for evidence of ancient WGD is "all-against-all" dotplots. In this method, ancient homologous genes are identified using BLAST, with duplicated segments reflected by consecutive strings of homologous genes preserved in a linear pattern parallel to the diagonal or anti-diagonal (the latter indicating segmental inversion). Compared to the Ks distribution plot, this not only provides evidence of ancient duplication events, but also the current location of the duplicated segment pairs, facilitating downstream studies such as reconstruction of the hypothetical ancient genome landscape. However, this method is less efficient in genomes that lack complete sequences or other abundant information about their genes and relative positions.

Lacking whole genome data, local gene loss patterns can also be indicative of the history of WGD (Ku *et al.* 2000). After genome duplication, one homologous gene is widely thought to be freed from selective pressure, and may adapt new functions (neofunctionalization), share the original gene function with its paralogue (subfunctionalization) or become pseudogenized or lost. Indeed, the vast majority of duplicated gene copies are lost. Thus, if a eudicot genome (such as *Gossypium*) has experienced WGD with associated gene loss after its divergence from *Vitis*, one would predict that many genes would no longer be found in their corresponding ancestral locations in the two genomes (Figure 5.1).

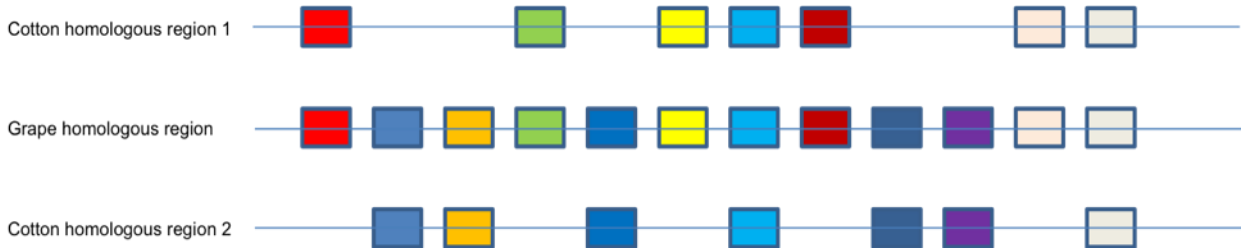


Figure 5.1 A model of stratification of cotton genome after whole genome duplication.

To further our understanding of its evolutionary history, we studied the *Gossypium* genome using two different methods: a whole genome level dotplot analysis, and a local level comparative study of a specific region of cotton-grape synteny using two sequenced *Gossypium* BACs covering ~184 kb. Both the whole genome dotplots and local level sequence comparisons provide new evidence of *Gossypium* lineage specific genome duplication after the Vitales-Malvales split. Comparison of homologous sequences between the two species also provides insight into mechanisms of genome size variation.

5.2 Materials and Methods

5.2.1 Genetic map and genome sequences

Cotton genetic map and marker sequence data were retrieved from a previously published map (RONG *et al.* 2004). Gene peptide sequences and position information for grape, papaya and *Arabidopsis* were all downloaded from the Plant Genome Duplication Database (PGDD: http://chibba.agtec.uga.edu/duplication/).

5.2.2 *Gossypium-Vitis* whole genome dot plot

Cotton marker sequences were blasted against *Vitis* genes using BLASTx, with an e value cut-off of 1e-10. The top 5 best hits were retained in the BLAST results. The dot-

plot was generated using a Python script (provided by Haibao Tang). ColinearScan (WANG *et al.* 2006e) was used to detect collinear blocks. The maximum gap allowed within a syntenic block on a grape chromosome was set to 1 Mb, and the maximum genetic distance allowed on the consensus map was set to 10cM.

5.2.3 BAC sequencing

Each BAC DNA sample was sheared using a Hydroshear (Genemachine) to ensure random fragmentation. The ends of the BACs were repaired using End-it DNA End Repair Kit (Epicenter, Madison, Wisconsin, USA). Fragment sizes around 4-5 kb were selected on a 1% low melting agarose gel, eluting the DNA with appropriate size from the gel using Qiagen QIAEX II (Qiagen, Valencia, California, USA) gel extraction system. DNA fragments were then ligated into the PCR-Blunt II-TOPO vector and transformed into DH10B *E.coli* host cells using an electroporator. The transformed cells were spread onto Q-plates and picked by a Q-bot into 96-well plates. Sequencing was performed on an ABI 3730-XL Sequence Analyzer using BigDye Terminator v3.1 Cycle Sequencing Kit. Chromotographs were assembled using PhredPhrap. Quality of sequence assemblies were checked using Sequencher V.4.1.4.

5.2.4 Gene and repetitive elements identification from BAC sequences

Genes were identified from BAC sequences using FGENESH (http://linux1.softberry.com/berry.phtml). In cotton, the species parameter was set to "Dicot plants"; for grape the parameter was set to "*Vitis vinifera*". Repetitive elements were identified using RepBase repeat masking service (http://www.girinst.org/), with species set to *Arabidopsis thaliana*.

5.2.5 Collinearity searches

The grape peptide sequences were used to BLAST against the BAC sequences using tBLASTn, with a cutoff value of 1e-20. The BLAST results were manually checked for collinearity. For *Arabidopsis-Vitis* genomes synteny, multiple collinearity search and alignment was performed using MCscan (TANG *et al.* 2008b).

5.3 Results

5.3.1 Limitations of *Gossypium-Gossypium* whole genome dotplot analysis

Detecting ancient WGD often requires relatively complete information, i.e., sequences and arrangement of most genes in a genome, in order for the signals to be discernible after extensive gene loss, single gene duplications and translocations. For cotton, which is not yet sequenced and has only ~2000 (~10% of) genes genetically mapped, there are simply too few homologous gene pairs available so far to distinguish paleopolyploidy from background noise (RONG *et al.* 2004).

The problem of too few data points to infer paleopolyploidy by intra-genomic comparison can be partially mitigated by using a consensus genetic map (i.e., integrated genetic map with merged chromosomes consisting of markers from both homeoelogous chromosomes resulting from in recent polyploidy) and intergenomic comparison to an outgroup genome. This approach has two advantages: 1.by using a consensus genetic map, we approximately doubled the number of cotton data points available; 2. by using an outgroup genome, we might be able to detect "ghost duplication" (SIMILLION *et al.* 2002) segments that are not detectable in self-plots due to the loss of one homolog.

5.3.2 *Gossypium-Vitis* whole genome dot plot

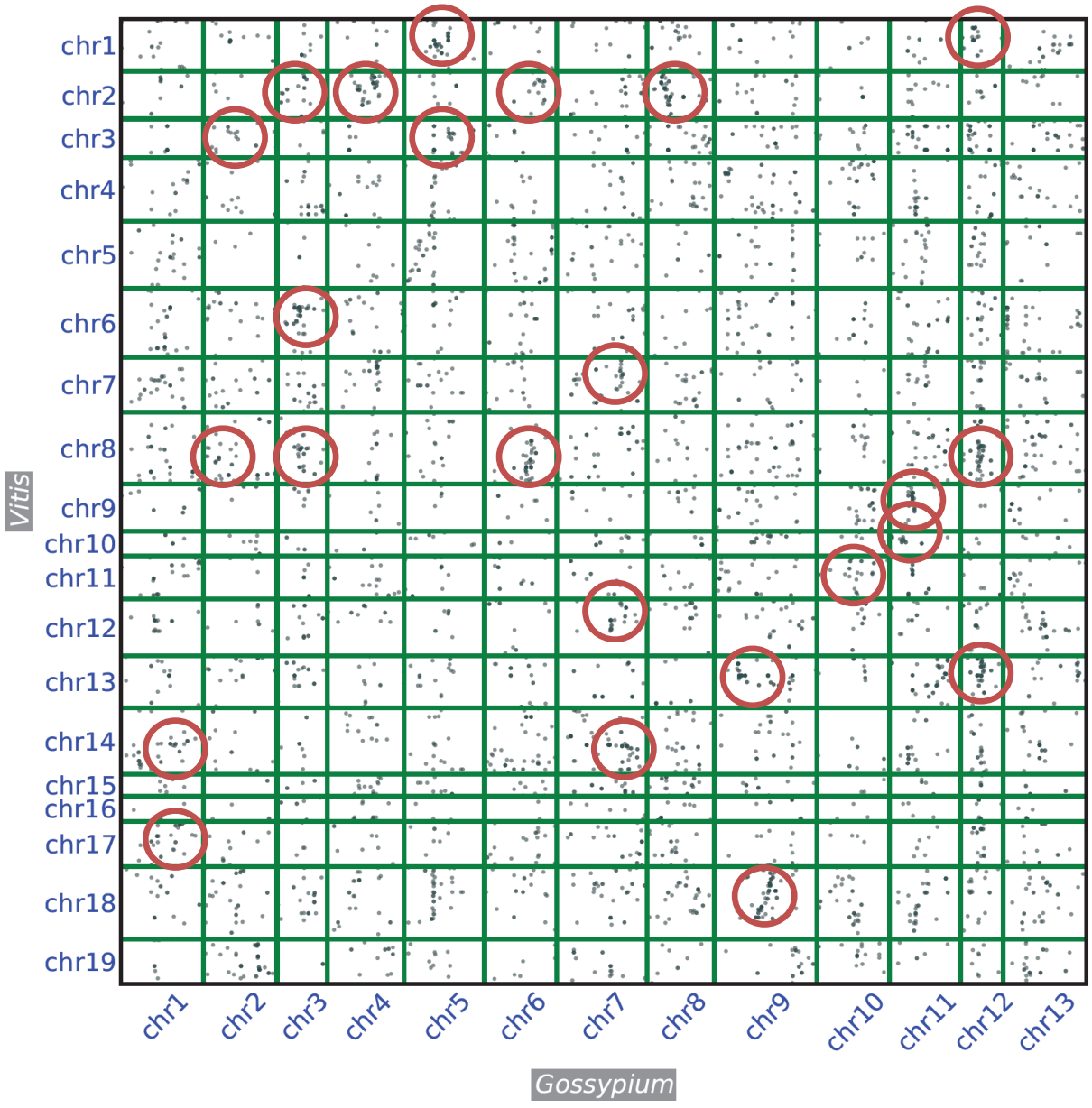
All markers from the cotton consensus map were plotted against all *Vitis* genes. By this threshold, we were able to detect 24 syntenic blocks (Figure 4.2A)

We found four cases in which one *Vitis* chromosomal segment shows collinearity with segments from different cotton chromosomes. e.g., an inverted collinear segment was detected between *Vitis* chromosome 2 and cotton consensus chromosomes 3, 4, 6 and 8 respectively; *Vitis* chromosome 8 showed collinearity with cotton consensus chromosomes 2, 3, 6 and 12 respectively; syntenic blocks were detected between *Vitis* chromosome 13 and cotton consensus chromosomes 9 and 12; and between *Vitis* chromosome 14 and cotton consensus chromosomes 1 and 7. This shows that duplicated segments are widely distributed in the cotton diploid genomes. (Figure 4.2A).

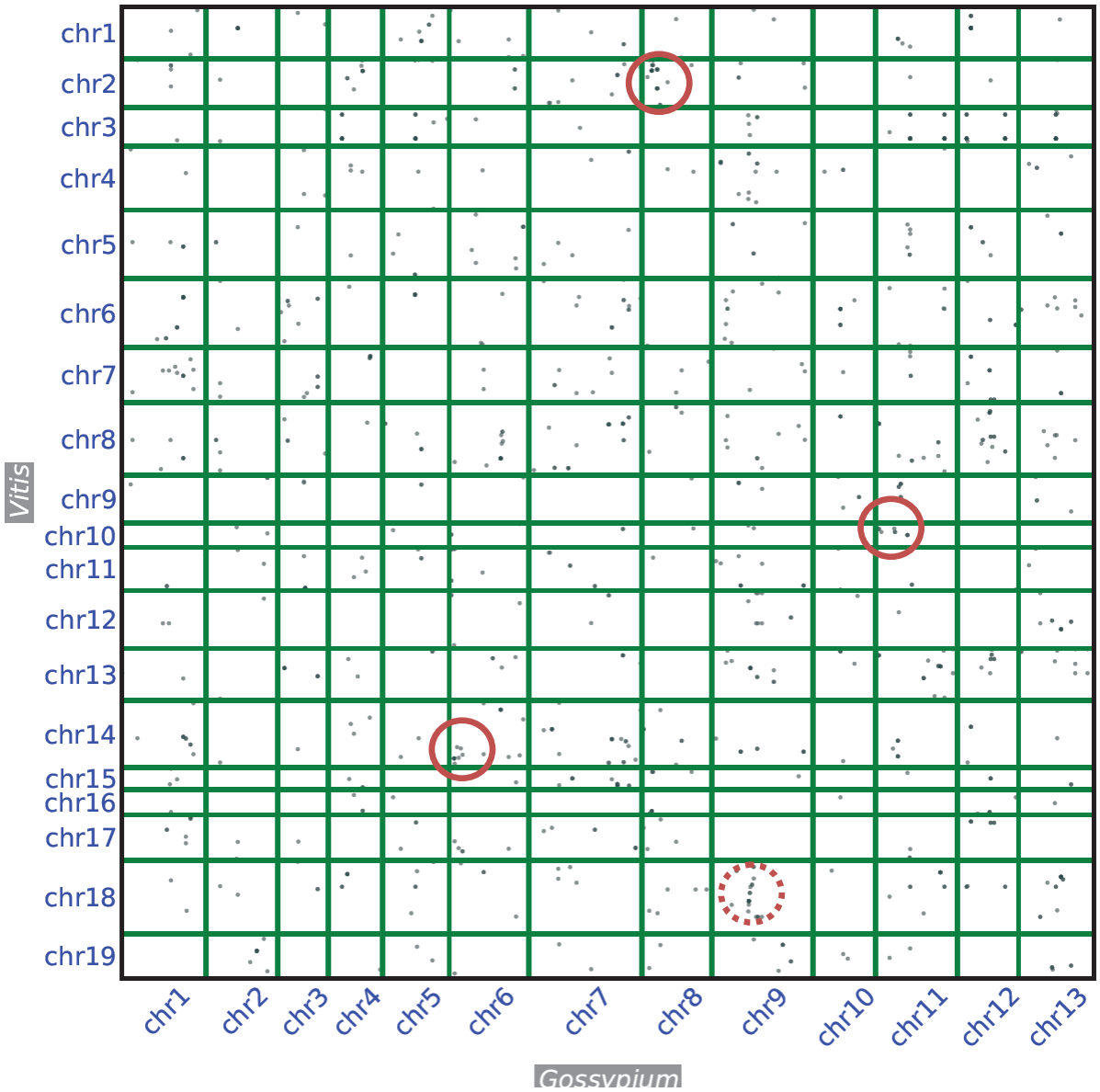
The power of detecting collinearity in whole genome dotplots depends on the quantity of gene position data available. In addition to study of the consensus map, we also tried to detect collinearity using its individual components, the AD tetraploid reference map and the D genome genetic map (RONG *et al.* 2004) separately. With the same parameters used in the consensus map study, we were only able to detect four syntenic blocks between the AD tetraploid map and the grape genome, and no discernible synteny between the D genome map and the grape genome. By lowering the stringency in ColinearScan and allowing the maximum distance between two hit points on the grape genome to be 1.5 Mb instead of 1 Mb, we were able to detect 17 blocks using the AD map (Figure 4.2C), but still only 3 blocks using the D genome map (Figure 4.2B). Although there are cases where homoeologous tetraploid cotton chromosomes were found to be syntenic to the same grape chromosome region, this analysis provided little information about ancient polyploidy and genome rearrangements. There are many

places where syntenic blocks detected in the plot using the consensus map were missed in the two plots using the individual reference maps due to lack of data points, as indicated by dashed circles in Figure 5.2.

A



B


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
Vitis
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
Gossypium

C

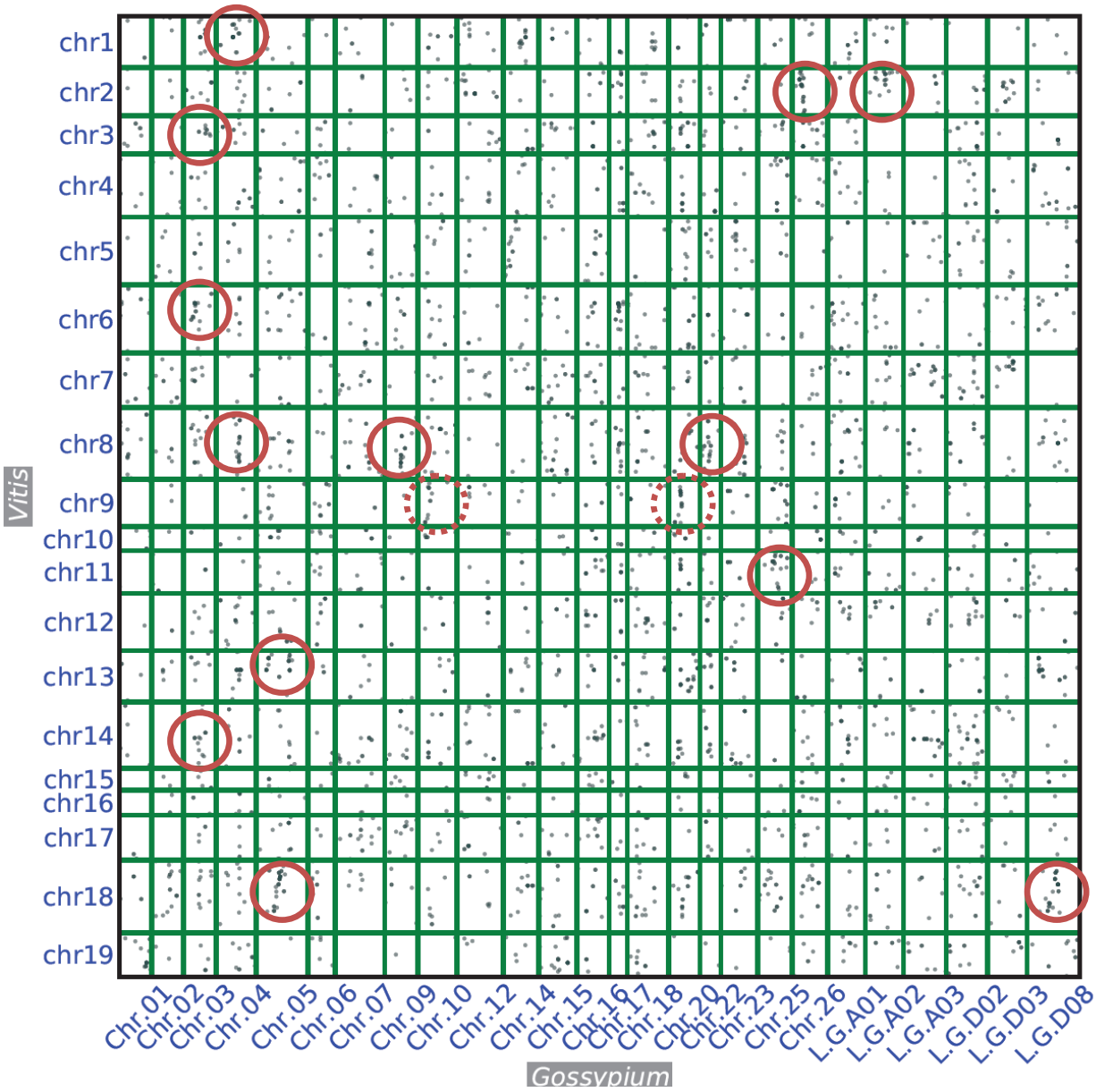


Figure 5.2 Whole genome dotplots between different cotton genetic maps and vitis whole genome peptide sequences.
A. Dotplot generated using cotton consensus map; **B**. Dotplot generated using D genome genetic map; **C**. Dotplot generated using *At* tetraploid genome map.

The consensus map has limitations also. Some places show a high density of hits, but lack a collinear relationship. Even in places where we could discern significant

collinear relationships, considerable fluctuation is evident around the predicted linear order. This is likely due at least in part to the process of consensus map construction. The consensus cotton map was assembled by combining the At, Dt and D genome genetic map relative to common "anchor" markers. "Unique" markers were interpolated between the common anchor markers based on the relative recombinational distance from the nearest anchor marker. This approach is potentially erroneous in inferring marker orders on a local scale, both because the maps are relatively low resolution (ca. 1 cM) and because the genetic/physical distance ratio can fluctuate widely.

5.3.3 Cotton BAC sequencing and microsyteny detection

Three BACs from the D genome physical map (Lin *et al*. unpublished) were selected for shotgun sequencing. The BACs selected were GR174O23, GR109E22 and GR163B08, in the order arranged by FPC. Two sequence contigs were assembled for GR109E22 with the size of 30,903 bp (GR109E22contig1) and 78,650 bp (GR109E22contig2) respectively. The two contigs were ordered and oriented using the vector sequence as a bridge. The assembled length is 97,267 bp for GR174O23 and 134,012 bp for GR163B08. Analysis in Sequencher revealed GR174O23 to overlap with GR109E22contig1, with a merged sequence 104,965 bp long. No overlaps among other BAC sequence fragments were found.

Vitis genes 1597 to 1637 on chromosome 6 were found to be collinear with GR174O23 and part of GR109E22; a region from genes 801-829 on chromosome 6 was found to be syntenic to the rest of GR109E22. We were not able to detect syntenic relationships using GR163B08.

For easy interpretation, we divided the collinear relationships found between cotton BACs and the *Vitis* genome into two regions. Region 1 contains the consensus sequence combining GR174O23, GR109E22 contig1 and part of GR109E22 contig2 that is immediately downstream of contig 1 across the sequencing gap in the BAC. This region contains 10 collinear genes that aligned to 21.8 Mb to 22.3 Mb on *Vitis* chromosome 6. Region 2 contains the remaining portion of GR109E22 contig2, which corresponds to 7.5 Mb to 7.8 Mb on *Vitis* chromosome 6, with 9 genes in collinear order. (Table 5.1, Figure 5.3).

Region 1 and region 2 are contiguous on the cotton genome, but are located on separate arms of chromosome 6 in *Vitis* (Figure 5.3). The syntenic regions of the *Arabidopsis* genome are ordered the same way as in the cotton genome, indicating that the rearrangement happened prior to the cotton-*Arabidopsis* divergence.

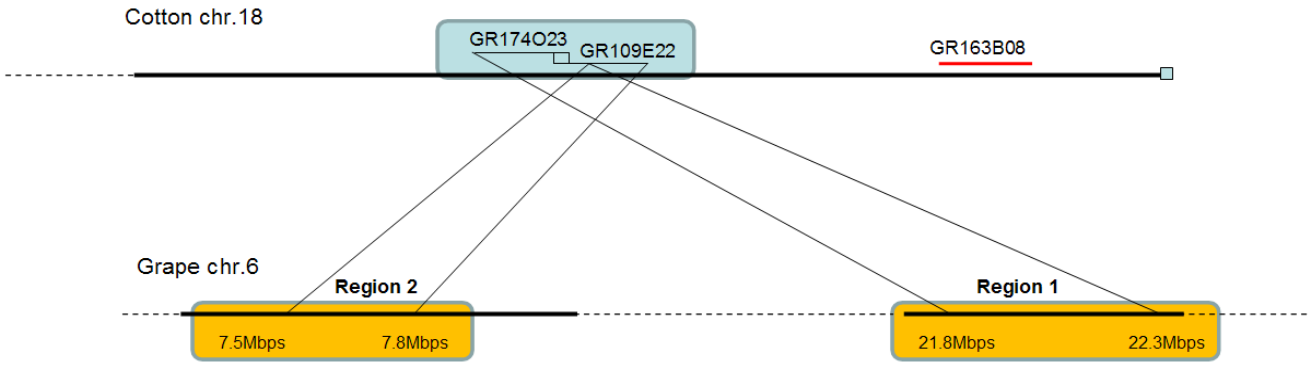


Figure 5.3 Positions of the homologous fragments of cotton sequenced BACs on grape chromosome6**.**

Table 5.1 Grape homologous region to cotton sequenced BACs

	Vitis gene number	GR BAC number	Hit position on GR BACs (Apprx. kb)
Region1	Vv6g1597	GR174O23_GR109E22contig1	22
	Vv6g1599	GR174O23_GR109E22contig1	64
	Vv6g1600	GR174O23_GR109E22contig1	79
	Vv6g1602	GR174O23_GR109E22contig1	84
	Vv6g1615	GR174O23_GR109E22contig1	89
	Vv6g1617	GR174O23_GR109E22contig1	97
	Vv6g1624	GR174O23_GR109E22contig1	105
	Vv6g1625	GR109E22Contig2	2
	Vv6g1627	GR109E22Contig2	6
	Vv6g1637	GR109E22Contig2	16
Region2	Vv6g0801	GR109E22Contig2	24
	Vv6g0802	GR109E22Contig2	29
	Vv6g0806	GR109E22Contig2	34
	Vv6g0814	GR109E22Contig2	43
	Vv6g0817	GR109E22Contig2	49
	Vv6g0819	GR109E22Contig2	49
	Vv6g0823	GR109E22Contig2	54
	Vv6g0826	GR109E22Contig2	58
	Vv6g0829	GR109E22Contig2	77

5.3.4 The non-syntenic BAC is enriched for repetitive DNA

GR163B08 is distal to GR109E22 in the same BAC contig (Figure 5.3). We were able to identify 19 genes from this BAC through FGENESH, however, no collinearity can be detected with the grape, papaya or *Arabidopsis* genomes.

The content of this BAC differs markedly from those of the other two BACs sequenced. Homology searches in Genbank showed that 8 (out of 19) predicted genes on this BAC are retrotransposon related, and the remaining 11 showed either no significant homology to known proteins, or homology to unknown proteins. Eight of these genes found no homologs in *Vitis* or *Arabidopsis* with a cutoff e-value of 1e-5 in BLASTp and the other 11 showed similar homologies to multiple genes, indicating a repetitive nature. A total of 11% of the BAC sequence is made up of transposable elements, but unlike the

other two BACS, these are almost exclusively (97%) LTR-retrotransposons. The number of tandem repeats found in this BAC is 3 to 8 times higher than in other two BACs.

GR163B08 is closer to the end of the chromosome than the other sequenced BACs (Lin *et al.* unpublished) and may be in or near a transitional region from gene rich euchromatin to the subtelomeric region. Common features of subtelomeric regions include the enrichment of tandem repeats and large transposable element insertions (KUO *et al.* 2006), consistent with the sequence composition of GR163B08.

5.3.5 Gene loss in cotton resembles the pattern of diploidization after WGD

To investigate gene loss in the cotton lineage after its split from *Vitis*, a putative ancestral gene order is needed. From the genes conserved in collinear arrangements in all four *Arabidopsis* homologous regions, we were able to distinguish genes in putative ancestral locations from putative lineage specific single gene insertions in both genomes. Genes found in collinear blocks across genomes were inferred to be in putative ancestral locations; other genes are likely to be lineage specific gene insertions. Figure 5.4 shows an example using Region 2.

In Region 1, 24 genes were in putative ancestral locations on the *Vitis* chromosome, of which 10 are still preserved in *Gossypium*; in Region 2 (Figure 5.4), 9 genes are preserved in *Gossypium* out of 17 in putative ancestral locations in *Vitis*. In both cases, roughly half the *Vitis* genes in ancestral locations are still identifiable in *Gossypium*, consistent with appreciable gene loss after one whole genome doubling event in the *Gossypium* lineage after the split from *Vitis*.

We compared the extent of gene loss in the *Gossypium* regions with the corresponding regions in the papaya genome (which has experienced no WGDs after its

divergence with *Vitis*) and the *Arabidopsis* genome (which has experienced two WGDs and diploidization after its divergence from grape). Gene number in the papaya genomic regions is similar to *Vitis*, with approximately twice the number of genes found in collinear positions in *Gossypium* (Table 5.2). In the *Arabidopsis* regions, the preserved gene number is significantly lower than that of *Gossypium* (Table 5.2), closer to ¼ of the genes in putative ancestral locations. This may suggest that only one round of WGD has happened in the *Gossypium* lineage. However, we would not yet exclude the possibility of two rounds of WGD in *Gossypium* due to the following concerns: first, in inferring the ancestral gene repertoire, we would inevitably miss genes that are lost either in *Vitis* or in all *Arabidopsis* homologous regions, or both. So the real ancestral gene number may be larger than what we infer, and thus the apparent 2:1 ratio of ancestral gene number to cotton preserved gene number may actually be not significantly different from 4:1 (indicative of two rounds of WGD). Secondly, although on average, *Arabidopsis* homologous regions have fewer duplicated genes preserved, the number of duplicated genes preserved in *Gossypium* is not significantly larger than what is found in the best preserved *Arabidopsis* homologous region (Table 5.2, bold numbers). With more BACs sequenced, we would be more confident of inferring the number of WGDs in the cotton lineage.

Table 5.2 Number of ancestral genes preserved in cotton and *Arabidopsis* in the sequence BACs

	Region 1	Region 2
Size of the region in *Vitis*	476 kb	290 kb
number of *Vitis* genes in homology	24	17
Size of the region in GR	116 kb	53 kb
number of GR genes in homology	10	9
Size of the region in *Carica*	~250 kb	328 kb
Number of *Carica* genes in homology	~20	18
Size of the region in AT	22-70 kb	23-40 kb
number of AT genes in homology	6,4,**9**,5	5,**8**,3,6

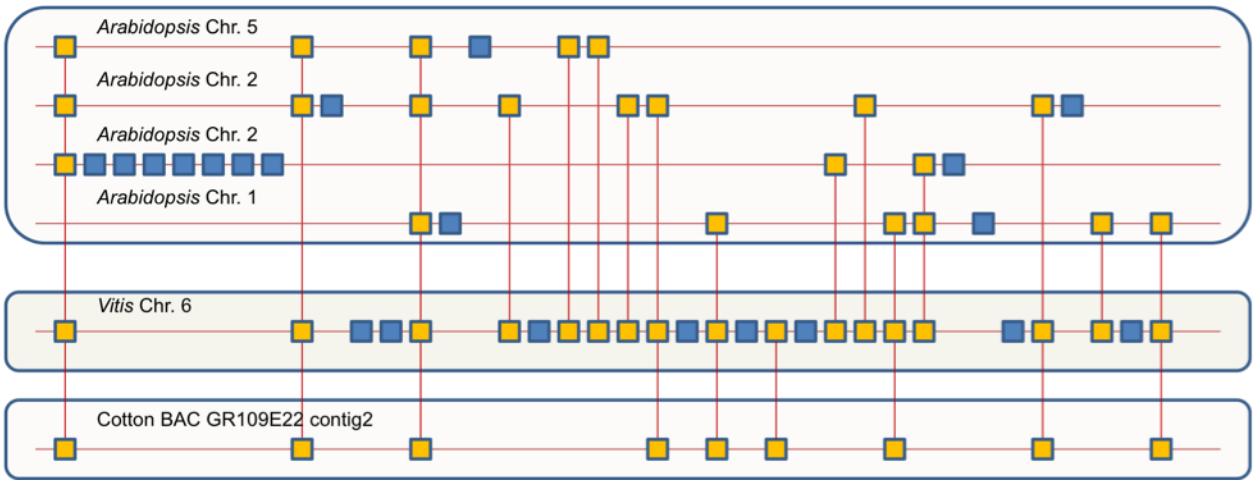


Figure 5.4 Pattern of cotton homologous gene loss in Region2.

Genes that showed collinearity across genomes are represented by orange squares; genes not preserved in collinear arrangement (putative lineage specific insertions) are represented by blue squares. Out of the 17 genes at putative ancestral locations in grape, only 9 are still identifiable in cotton.

There are still many genes in the collinear regions of these genomes that do not fit into the putative ancestral gene positions. These are likely to be lineage specific gene insertions. In particular, in *Arabidopsis* Region 2 (Figure 5.4), seven consecutive genes find no homology in *Vitis*, *Carica* or *Gossypium* in this region, but are found in a collinear block on grape chromosome 13 and a separate papaya scaffold of the current assembly, indicating translocation of a large fragment to this region in the *Arabidopsis* lineage.

5.3.6 The *Vitis* homologous region spans a larger physical distance than the corresponding regions of cotton and *Arabidopsis*

Although the *Vitis* genome is only about 55% of the size of the cotton D genome (HENDRIX and STEWART 2005; JAILLON *et al.* 2007), the syntenic region on *Vitis* is much

larger in size than the corresponding cotton regions in both cases. Region 1 covers a *Vitis* genomic region of ~476 kb, and a *Gossypium* region of 116 kb; region 2 covers a *Vitis* region of 290 kb and a *Gossypium* region of 52.7 kb. In both cases, the *Vitis* region is 5-10 times as large as the corresponding *Gossypium* region. *Arabidopsis* syntenic regions had physical sizes similar to the cotton regions (approximately 53 kb and 43 kb for Region 1 and 2 respectively).

5.3.7 Causes of size differences between the *Vitis* and *Gossypium* homologous regions

The size difference between corresponding regions of cotton and grape could be caused by either extensive expansions in the grape genome or condensation in the cotton genome, or very likely, both.

Transposons

We analyzed the distribution of transposable elements in these different regions (Figure 5.5 -Figure 5.6) using RepBase (http://www.girinst.org/) and default parameters. TEs comprise a larger proportion of the sequence in the *Vitis* homologous regions, at 25% and 17% of Region 1 and 2, as compared to 13% and 7% in *Gossypium*. Both DNA transposons and retroelements comprise a larger portion of the grape sequences than the cotton sequences. The difference in quantity of transposons explains 30% and 18% of the size differences between the compared regions in the two genomes (Figure 5.5 A).

Gene loss

Even excluding transposable elements, there is still a 3x to 4x difference in the size of the corresponding sequence (Figure 5.5). Therefore, we counted the number of

genes in these regions in both species, and found that this variation in physical length of syntenic regions is approximately proportional to the number of genes identified; indicating that diploidization (gene loss) in the cotton lineage also played an important role in the size difference.

In order to compare gene number from *Gossypium* and *Vitis* regions, we used FGENESH (http://linux1.softberry.com/berry.phtml) predictions. In *Gossypium*, because no species-specific gene model profile is available, the "Organism" parameter was set to "Dicot plants"; for grape the parameter was set to "*Vitis vinifera*". In grape, we have identified 68 and 44 genes in regions 1 and 2 respectively, which indicates a gene density of 7 kb and 6.59 kb per gene respectively. The corresponding cotton regions have a gene density of 5.80 kb and 5.27 kb per gene. Collectively, the sequences that encode genes comprised 141 kb and 131 kb in grape Region 1 and 2, and 50 kb and 38 kb in cotton. This explains 25% and 40% of the size difference in the compared regions (Figure 5.5B).

By plotting the positions of genes and TEs on these regions from the two genomes (Figure 5.6), one can see that many of the "extra" gene sequences in the grape regions are in ancestral positions, suggesting that they may have been lost in this particular region of *Gossypium* during diploidization. One would predict that the missing genes may be found in paralogous regions of the *Gossypium* genome (once we have its sequence).

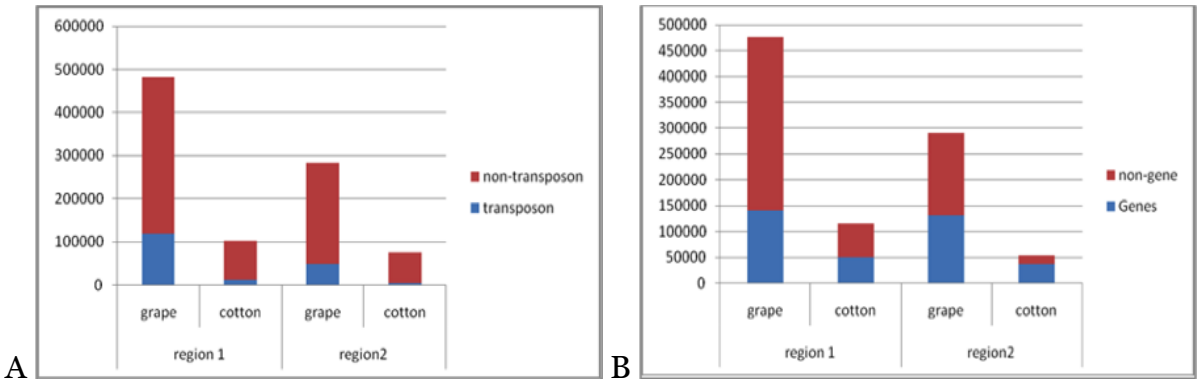


Figure 5.5 The proportion of transposable elements (A) and genes (B) in the homologous regions compared.

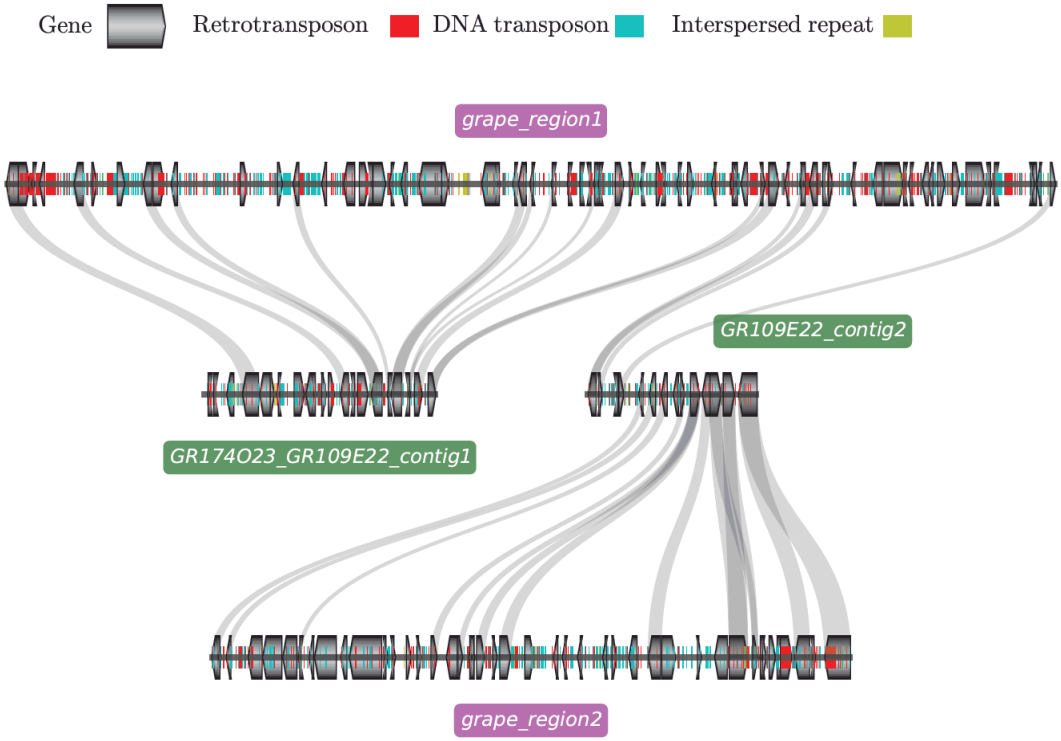


Figure 5.6 Distribution of gene and transposable elements in the cotton and grape regions compared.
Lines connecting different regions indicates syntenic genes.

In partial summary, a huge size difference was observed between the cotton and grape syntenic regions compared. The grape regions contain a larger portion of

transposable elements than the cotton regions, indicating that grape lineage specific TE insertions could play a role in the expansion of these regions; nonetheless, loss of ancestral gene sequences in the cotton lineage after WGD seemed to account for a larger portion of the difference of physical length in the compared regions.

5.4 Discussion

Earlier genome mapping studies suggested that diploid cotton might be an ancient polyploid. In this study, we used two different approaches to search for evidence of ancient whole genome duplication in the diploid cotton lineage. Whole genome dotplot analysis using all mapped genes in cotton against all genes in the sequenced *Vitis* genome, was constrained by the limited number of informative cotton genes. Nonetheless we still observed several cases in which one grape chromosome segment corresponded to two segments in cotton, which strongly suggests at least one round of WGD in the diploid cotton lineage. Detailed dissection of one of the collinear regions has revealed genome stratification in cotton that fits the expected behavior after WGD events. These findings, along with earlier published findings using different methods, strongly support the hypothesis that cotton is an ancient polyploid.

Despite the much smaller genome size of grape compared to cotton, the homologous regions in grape that we have analyzed are far larger than the cotton regions. Although transposon insertions do play a role in the size differences, we have observed that diploidization in the cotton genome explains a larger portion of the difference in segment of the genome. The deleted genes in the cotton regions in this study are likely to be preserved in paleo-duplicated fragments elsewhere in the cotton genome. Therefore, although gene loss has caused the cotton regions in comparison to be shorter than the grape regions, given the similar gene densities, it is likely that the

overall genome size is not affected much by gene deletion. The similar gene density of this region observed in grape, cotton and *Arabidopsis* echoes the finding that the cotton genome is composed of two distinctive components where genes are densely packed in euchromatic regions, and the other being heterochromatic regions that explain the majority of genome size differences between cotton, grape and *Arabidopsis*.

5.4.1 New evidence supporting a history of WGD in cotton

Earlier research suggested that the cotton lineage experienced at least one WGD (DESAI *et al.* 2006; RONG *et al.* 2004), largely based on intragenomic comparisons of genetic marker positions and use of the current gene/marker order to deduce the ancestral gene order. Two new lines of evidence further support the hypothesis that cotton has indeed experienced at least one round of WGD subsequent to the triplication affecting most if not all dicots.

Compared to earlier analyses using CrimeStat2 and FISH in the detection of ancient duplicated segments, dotplots of mapped genes reveal fewer segments. However, as the dot plot analysis requires the genes to be ordered in a collinear manner, segment detected this way are more likely to be ancient paralogues than false positives.

The grape (*Vitis vinifera*) genome is an excellent reference for efforts to determine numbers of WGDs in eudicots. The grape genome has experienced no WGD since the ancient hexaploidy event shared by most if not all eudicots. A slow evolutionary rate in grape appears to have helped in preserving ancestral gene order(JAILLON *et al.* 2007). The grape genome is advantageous in many ways: first of all, it may better reflect the gene order before WGD than any other eudicot lineage; it also is a good phylogenetic outgroup for comparative analysis of many eudicot species. These

attributes are very helpful in elucidating the ancient duplication history of a new genome.

Our study here also includes a local level comparative analysis, started with a putative ancestral gene order from a different species (*Vitis*) predating the duplication event, and using a topdown approach (TANG *et al.* 2008a). The local analysis, based on BAC sequences, has the advantage of having more conserved genes in collinear order than intragenomic studies using self-comparisons between cotton homologs alone. The preserved gene number in collinearity in all cotton regions studied is less than 50%, but appreciably higher than that of any one *Arabidopsis* segment, which experienced two rounds of WGD after its divergence from grape. However, across the four *Arabidopsis* segments (Figure 4.4), a total of 17 genes are preserved in collinear locations in at least one segment, versus only 9 in the single cotton segment. This is consistent with the fact that we do not have the sequence of the paleo-duplicated cotton region yet.

While we hypothesize that cotton has experienced only a single WGD in this time period, we cannot yet rule out the possibility of two rounds of WGD in cotton. To further elucidate the question, more cotton BACs that are homologous to these grape regions need to be sequenced to compare the gene loss and preservation patterns in all the cotton homologous fragments to this region. The finishing of the whole genome sequencing of a D genome cotton (*G. raimondii*) in the near future will provide us with a relatively complete list of cotton genes and their arrangements, and much clearer picture of the history of genome duplications in cotton species.

5.4.2 Effect of genome duplication on genome size

The effect of WGD on genome size is complicated. Genomes with history of WGD vary greatly in genome size. e.g., sorghum and rice genome share a similar history of WGD, while the sorghum genome is 72% larger (740 Mb vs 430 Mb). The *Arabidopsis* genome, with a history of one genome triplication and two genome duplications, has one of the smallest genomes in higher plants, while the maize genome, with at least three rounds of WGD, has one of the largest. There are no obvious correlations between the number of WGDs a genome has experienced, and the size of its genome.

Gene deletion is common after whole genome duplication events. The diploidization process maintains a relatively stable gene number and gene space before and after genome duplications. From our results, the gene density of homologous regions between genomes with and without WGD is similar, suggesting that sizes of gene-rich regions are not affected much by genome duplication. This seemed to support the idea that the expansion of genome size is not much affected by genome duplication, but rather mostly caused by transposon accumulations in heterochromatic regions. Comparative study between rice and sorghum, in which sizes of gene space are thought to be nearly identical, has shown that heterochromatin alone can account for huge genome size differences (BOWERS *et al.* 2005; PATERSON *et al.* 2009). In the regions of our study, however, fewer transposon insertions were detected in the cotton sequences. This might be because the cotton BACs selected came from a gene-rich region. Transposon insertions tend to accumulate in centromeres and heterochromatic regions (BENNETZEN *et al.* 2005; BOWERS *et al.* 2005). In euchromatic regions, during the diploidization process after a WGD, duplicated genes in one paralogous region might be removed along with neighboring sequences, and cause the cotton homologous region to

be even shorter. It is conceivable from our observation here that cotton genes are densely packed in euchromatin, with gene poor heterochromatic regions making up a large portion of cotton chromosomes.

Many studies of genome size evolution focus on the effects of transposable elements, particularly the insertion and deletion patterns of LTR-retrotransposons (BENNETZEN 2002; BENNETZEN *et al.* 2005). The rapid expansion of one or a few cotton transposon families may have contributed to variations in genome size of Gossypium species (HAWKINS *et al.* 2006; ZHAO *et al.* 1995; ZHAO *et al.* 1998). A burst of transposon activity has been described in synthesized polyploids, and retrotransposons alone can account for genome size doubling in some species even without WGD (PIEGU *et al.* 2006). Our findings here show that the size of gene-rich regions do not vary much regardless of the number of WGDs a genome has experienced, and also supports the idea that activation of transposable elements, rather than the genome doubling caused by WGD, is the major cause of the huge size difference among plant species.

CHAPTER 6

CONCLUSIONS

My research has focused on whole genome physical mapping of D genome cotton, and investigation of a chromosomal region in which a gene that is likely to be involved in fiber initiation is located. Both projects are thought to be among the first such efforts in cotton research.

Draw backs in physical map assembly due to the low average band number in agarose fingerprinting were compensated by refingerprinting of a subset of the library using HICF, to produce an assembly of 4208 contigs. The alignment of these contigs on the consensus map is useful for researchers interested in understanding a particular region in the cotton genome. This potential has been illustrated by our usage of the contig in dissecting the *Li2* region. The map we used to anchor the contigs is a consensus map, integrating reference maps of At, Dt and D genomes, encompassing over 3000 loci. Although these loci all came from the reference map (RONG *et al.* 2004), recent approaches to integrate different genetic maps and resources such as CMAP databases (RONG *et al.* 2007) has made it easier for features from other mapping efforts to find their corresponding region on our consensus map, and further, their corresponding contigs that anchor to the region.

In facilitating and validating physical map assemblies, a whole genome sequence of a close relative has proven to be very beneficial. In animals and monocot plant species, large regions, that often encompass whole chromosome arms retain well preserved gene order across tens of millions of years of evolution. In eudicots, however, syntenic

relationships between segments from different species are very noisy, with extensive genome rearrangements specific to every species. Thus, we were not able to use the *Arabidopsis* or the *Vitis* genome or any recently published dicot whole genome sequences to validate the cotton physical map assembly to the degree that the rice genome was useful in validating the assembly of sorghum genome contigs. Nevertheless, we were still able to detect microsynteny between cotton contigs and the genomic sequences of *Arabidopsis* and *Vitis*. These alignments of cotton fragments on sequenced genomes provide a foundation to utilize translational genomics in the characterization of important cotton genes and the improvement of cotton crop species.

An important observation derived from our analysis of probe hybridization data is that the cotton genome is likely to be composed of two qualitatively different components, as recently suggested for several other angiosperm genomes (sorghum, rice, soybean). One of these is gene rich and high in recombination frequency, and the other is repeat rich and recalcitrant to recombination. Our sequence comparison in euchromatic regions between cotton, grape, *Arabidopsis* and papaya has shown similar gene densities, which also indicate that, the variation of size among these genomes happened elsewhere. This has several implications in cotton research. First, the gene density we have observed in cotton is close to that of *Arabidopsis*, which suggests that cotton might have a small gene space. Second, variation in recombination rate between euchromatic and heterochromatic regions will be reflected in a huge difference in genetic/physical distance ratios. In genetic mapping and positional cloning of a specific gene, it is important to determine the characteristics of the gene region before determining the best mapping strategy to use.

Sequence analysis has confirmed that *Li2* gene is in a gene-rich region. The fine mapping of the *Li2* gene provides a foundation toward later gene cloning. With newly designed markers and a large F2 population, we were able to find DNA markers very close to the gene. Perhaps the most important result in the gene cloning project is the identification of a BAC contig that maps closely to the *Li2* gene. With multiple pieces of evidence validating the anchoring, the physical map contig gave us a solid foundation from which chromosome walking can be carried out. Three BACs were sequenced and their positions ordered by genetic mapping. Designing of overgo probes from the closest BAC sequences will help us identify BACs that reaches further in the direction of the gene. Fingerprint data would be a good way to validate candidate clones for a next round of chromosome walking.

The eventual identification of the *Li2* gene will be a big step toward understanding cotton fiber development. The phenotype of the *Li2* mutant has indicated the gene is likely to function at the initiation or elongation stage, rather than the secondary cell wall synthesis stage, of cotton fiber development. An analysis of genes on the BACs closely mapped to *Li2* showed several genes that have functions potentially related to fiber initiation/elongation. Their relationship with the *Li2* gene is to be determined through further mapping.

The recently published *Vitis vinifera* genome is a preferred reference genome in cross genome comparative analysis in dicots, based on the absence of genome duplication after the eudicot divergence, and its slow evolutionary rate. It allows us to infer homology with higher confidence, because of its suspected close resemblance to ancestral gene order in eudicots. The *Vitis* genome could be used as a bridge when trying to detect homology between distantly related eudicot genomes, or between

genomes that have gone through extensive genome rearrangements after their divergence. We have used the *Vitis* genome in the validation of *Li2* contig assembly, as well as the validation of *Arabidopsis* homology detection. The pattern of loss of homologous genes in cotton closely resembles the pattern of diploidization after a whole genome duplication event, thus adding new evidence to the notion that diploid cottons are ancient polyploids.

REFERENCES

ABDURAKHMONOV, I. Y., Z. T. BURIEV, S. SAHA, A. E. PEPPER, J. A. MUSAEV *et al.*, 2007 Microsatellite markers associated with lint percentage trait in cotton, Gossypium hirsutum. Euphytica **156:** 141-156.

ABDURAKHMONOV, I. Y., R. J. KOHEL, J. Z. YU, A. E. PEPPER, A. A. ABDULLAEV *et al.*, 2008 Molecular diversity and association mapping of fiber quality traits in exotic G. hirsutum L. germplasm. Genomics **92:** 478-487.

ALABADY, M. S., E. YOUN and T. A. WILKINS, 2008 Double feature selection and cluster analyses in mining of microarray data from cotton. BMC Genomics **9:** -.

ALI, I., M. ASHRAF, MEHBOOB-UR-RAHMAN, Y. ZAFAR, M. ASIF *et al.*, 2009a Development of Genetic Linkage Map of Leaf Red Colour in Cotton (Gossypium Hirsutum) Using DNA Markers. Pakistan Journal of Botany **41:** 1127-1136.

ALI, I., A. KAUSAR, MEHBOOB-UR-RAHMAN, Y. ZAFAR, M. ASIF *et al.*, 2009b Development of Genetic Linkage Map of Leaf Hairiness in Gossypium Hirsutum (Cotton) Using Molecular Markers. Pakistan Journal of Botany **41:** 1627-1635.

AMMIRAJU, J. S. S., M. Z. LUO, J. L. GOICOECHEA, W. M. WANG, D. KUDRNA *et al.*, 2006 The Oryza bacterial artificial chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. Genome Research **16:** 140-147.

ASIF, M., J. I. MIRZA and Y. ZAFAR, 2008 Genetic analysis for fiber quality traits of some cotton genotypes. Pakistan Journal of Botany **40:** 1209-1215.

BENNETZEN, J. L., 2002 Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115:** 29-36.

BENNETZEN, J. L., and E. A. KELLOGG, 1997 Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell **9:** 1509-1514.

BENNETZEN, J. L., J. X. MA and K. DEVOS, 2005 Mechanisms of recent genome size variation in flowering plants. Annals of Botany **95:** 127-132.

BENSON, G., 1999 Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27:** 573-580.

BLENDA, A., J. SCHEFFLER, B. SCHEFFLER, M. PALMER, J. M. LACAPE *et al.*, 2006 CMD: a Cotton Microsatellite Database resource for Gossypium genomics. BMC Genomics **7:** -.

BOLEK, Y., K. M. EL-ZIK, A. E. PEPPER, A. A. BELL, C. W. MAGILL *et al.*, 2005 Mapping of verticillium wilt resistance genes in cotton. Plant Science **168:** 1581-1590.

BOWERS, J. E., M. A. ARIAS, R. ASHER, J. A. AVISE, R. T. BALL *et al.*, 2005 Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. Proc Natl Acad Sci U S A **102:** 13206-13211.

BOWERS, J. E., B. A. CHAPMAN, J. RONG and A. H. PATERSON, 2003 Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433-438.

BRUBAKER, C. L., A. H. PATERSON and J. F. WENDEL, 1999 Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome **42:** 184-203.

CAI, W. W., J. RENEKER, C. W. CHOW, M. VAISHNAV and A. BRADLEY, 1998 An anchored framework BAC map of mouse chromosome 11 assembled using multiplex oligonucleotide hybridization. Genomics **54:** 387-397.

CHAPMAN, B. A., J. E. BOWERS, F. A. FELTUS and A. H. PATERSON, 2006 Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. Proc Natl Acad Sci U S A **103:** 2730-2735.

CHAUDHARY, B., L. FLAGEL, R. M. STUPAR, J. A. UDALL, N. VERMA *et al.*, 2009 Reciprocal Silencing, Transcriptional Bias and Functional Divergence of Homeologs in Polyploid Cotton (Gossypium). Genetics **182:** 503-517.

CHEE, P., X. DRAYE, C. X. JIANG, L. DECANINI, T. A. DELMONTE *et al.*, 2005 Molecular dissection of interspecific variation between Gossypium hirsutum and Gossypium barbadense (cotton) by a backcross-self approach: I. Fiber elongation. Theor Appl Genet **111:** 757-763.

CHEN, Z. J., B. E. SCHEFFLER and E. DENNIS, 2007 Toward Sequencing cotton (Gossypium) Genomes. Plant Physiology **145:** 1303-1310.

CRONN, R. C., R. L. SMALL and J. F. WENDEL, 1999 Duplicated genes evolve independently after polyploid formation in cotton. Proc Natl Acad Sci U S A **96:** 14406-14411.

CRONN, R. C., X. ZHAO, A. H. PATERSON and J. F. WENDEL, 1996 Polymorphism and concerted evolution in a tandemly repeated gene family: 5S ribosomal DNA in diploid and allopolyploid cottons. J Mol Evol **42:** 685-705.

DESAI, A., P. W. CHEE, O. L. MAY and A. H. PATERSON, 2008 Correspondence of trichome mutations in diploid and tetraploid cottons. Journal of Heredity **99:** 182-186.

DESAI, A., P. W. CHEE, J. RONG, O. L. MAY and A. H. PATERSON, 2006 Chromosome structural changes in diploid and tetraploid A genomes of Gossypium. Genome **49:** 336-345.

DEVOS, K. M., J. K. BROWN and J. L. BENNETZEN, 2002 Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res **12:** 1075-1079.

DRAYE, X., P. CHEE, C. X. JIANG, L. DECANINI, T. A. DELMONTE *et al.*, 2005 Molecular dissection of interspecific variation between Gossypium hirsutum and G-barbadense (cotton) by a backcross-self approach: II. Fiber fineness. Theoretical and Applied Genetics **111:** 764-771.

ENDRIZZI, J., and G. RAMSAY, 1979 Monosomes and telosomes for 18 of the 26 chromosomes of Gossypium hirsutum Canadian Journal of Genetics **21:** 531-536.

FELTUS, F. A., H. P. SINGH, H. C. LOHITHASWA, S. R. SCHULZE, T. D. SILVA *et al.*, 2006 A comparative genomics strategy for targeted discovery of single-nucleotide polymorphisms and conserved-noncoding sequences in orphan crops. Plant Physiol **140:** 1183-1191.

FENG, C. D., J. M. D. STEWART and J. F. ZHANG, 2005 STS markers linked to the Rf(1) fertility restorer gene of cotton. Theoretical and Applied Genetics **110:** 237-243.

FREITAS, P. D., D. S. MARTINS and P. M. GALETTI, 2008 CID: a rapid and efficient bioinformatic tool for the detection of SSRs from genomic libraries. Molecular Ecology Resources **8:** 107-108.

GAO, L. Z., and H. INNAN, 2004 Very low gene duplication rate in the yeast genome. Science **306:** 1367-1370.

GEEVER, R. F., F. R. H. KATTERMAN and J. E. ENDRIZZI, 1989 DNA hybridization analyses of a *Gossypium* allotetraploid and two closely related diploid species. Theor Appl Genet **77:** 553-559.

GRIFFEE, F., and L. LIGON, 1929 Occurrence of "lintless" cotton plants and inheritance of character "lintless" *Journal of the American Society of Agronomy* **21:** 711-717.

GROVER, C. E., H. KIM, R. A. WING, A. H. PATERSON and J. F. WENDEL, 2004 Incongruent patterns of local and global genome size evolution in cotton. Genome Res **14:** 1474-1482.

GUO, W. Z., C. P. CAI, C. B. WANG, Z. G. HAN, X. L. SONG *et al.*, 2007 A microsatellite-based, gene-rich linkage map reveals genome structure, function and evolution in gossypium. Genetics **176:** 527-541.

GUO, W. Z., T. Z. ZHANG, J. J. PAN and R. J. KOHEL, 1998 Identification of RAPD marker linked with fertility-restoring gene of cytoplasmic male sterile lines in upland cotton. Chinese Science Bulletin **43:** 52-54.

GUO, W. Z., T. Z. ZHANG, X. L. SHEN, J. Z. YU and R. J. KOHEL, 2003 Development of SCAR marker linked to a major QTL for high fiber strength and its usage in molecular-marker assisted selection in upland cotton. Crop Science **43:** 2252-2256.

GUO, Y. F., J. C. MCCARTY, J. N. JENKINS and S. SAHA, 2008 QTLs for node of first fruiting branch in a cross of an upland cotton, Gossypium hirsutum L., cultivar with primitive accession Texas 701. Euphytica **163:** 113-122.

HAO, J. J., S. X. YU, Z. D. DONG, S. L. FAN, Q. X. MA *et al.*, 2008 Quantitative inheritance of leaf morphological traits in upland cotton. Journal of Agricultural Science **146:** 561-569.

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, D. H., Z. X. LIN, X. L. ZHANG, Y. C. NIE, X. P. GUO *et al.*, 2005 Mapping QTLs of traits contributing to yield and analysis of genetic effects in tetraploid cotton. Euphytica **144:** 141-149.

HE, D. H., Z. X. LIN, X. L. ZHANG, Y. C. NIE, X. P. GUO *et al.*, 2007 QTL mapping for economic traits based on a dense genetic map of cotton with PCR-based markers using the interspecific cross of Gossypium hirsutum x Gossypium barbadense. Euphytica **153:** 181-197.

HE, D. H., Z. X. LIN, X. L. ZHANG, Y. X. ZHANG, W. LI *et al.*, 2008 Dissection of genetic variance of fibre quality in advanced generations from an interspecific cross of Gossypium hirsutum and G-barbadense. Plant Breeding **127:** 286-294.

HENDRIX, B., and J. M. STEWART, 2005 Estimation of the nuclear DNA content of gossypium species. Ann Bot (Lond) **95:** 789-797.

HUGHES, M. K., and A. L. HUGHES, 1993 Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. Mol Biol Evol **10:** 1360-1369.

ISHIDA, T., S. HATTORI, K. OKADA and T. WADA, 2007 Role of TTG2 in genetic network of epidermal cell differentiation in Arabidopsis. Plant and Cell Physiology **48:** S82-S82.

JAILLON, O., J. M. AURY, B. NOEL, A. POLICRITI, C. CLEPET *et al.*, 2007 The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463-467.

JIANG, C., R. J. WRIGHT, S. S. WOO, T. A. DELMONTE and A. H. PATERSON, 2000 QTL analysis of leaf morphology in tetraploid *Gossypium* (cotton). Theor. Appl. Genet. **100:** 409-418.

JIANG, C. X., R. J. WRIGHT, K. M. EL-ZIK and A. H. PATERSON, 1998 Polyploid formation created unique avenues for response to selection in Gossypium (cotton). Proceedings of the National Academy of Sciences of the United States of America **95:** 4419-4424.

KADIR, Z., 1976 DNA evolution in the genus gossypium. Chromosoma **56:** 85.

KIM, H. J., and B. A. TRIPLETT, 2001 Cotton fiber growth in planta and in vitro. Models for plant cell elongation and cell wall biogenesis. Plant Physiol **127:** 1361-1366.

KOHEL, R. J., E. V. NARBUTH and C. R. BENEDICT, 1992 Fiber Development of Ligon Lintless-2 Mutant of Cotton. Crop Sci **32:** 733-735.

KOHEL, R. J., J. YU, Y. H. PARK and G. R. LAZO, 2001 Molecular mapping and characterization of traits controlling fiber quality in cotton. Euphytica **121:** 163-172.

KU, H. M., T. VISION, J. P. LIU and S. D. TANKSLEY, 2000 Comparing sequenced segments of the tomato and Arabidopsis genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. Proceedings of the National Academy of Sciences of the United States of America **97:** 9121-9126.

KUO, H. F., K. M. OLSEN and E. J. RICHARDS, 2006 Natural variation in a subtelomeric region of arabidopsis: Implications for the genomic dynamics of a chromosome end. Genetics **173:** 401-417.

LACAPE, J. M., D. DESSAUW, M. RAJAB, J. L. NOYER and B. HAU, 2007 Microsatellite diversity in tetraploid Gossypium germplasm: assembling a highly informative genotyping set of cotton SSRs. Molecular Breeding **19:** 45-58.

LACAPE, J. M., and T. B. NGUYEN, 2005 Mapping quantitative trait loci associated with leaf and stem pubescence in cotton. Journal of Heredity **96:** 441-444.

LACAPE, J. M., T. B. NGUYEN, S. THIBIVILLIERS, B. BOJINOV, B. COURTOIS *et al.*, 2003 A combined RFLP-SSR-AFLP map of tetraploid cotton based on a Gossypium hirsutum x Gossypium barbadense backcross population. Genome **46:** 612-626.

LAN, T.-H., C. COOK and A. PATERSON, 1999 Identification of a RAPD marker linked to a male-fertility restoration gene in cotton (*Gossypium hirsutum* L.). J. Agr. Genomics **4**.

LANDER, E. S., and M. S. WATERMAN, 1988 Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics **2:** 231-239.

LANG, A. G., 1938 The origin of lint and fuzz hairs of cotton. Journal of Agricultural Engineering Research **56:** 507-521.

LUKENS, L. N., J. C. PIRES, E. LEON, R. VOGELZANG, L. OSLACH *et al.*, 2006 Patterns of sequence loss and cytosine methylation within a population of newly resynthesized Brassica napus allopolyploids. Plant Physiology **140:** 336-348.

LUO, M. C., C. THOMAS, F. M. YOU, J. HSIAO, S. OUYANG *et al.*, 2003 High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics **82:** 378-389.

LYNCH, M., and J. S. CONERY, 2000 The evolutionary fate and consequences of duplicate genes. Science **290:** 1151-1155.

MA, J., K. M. DEVOS and J. L. BENNETZEN, 2004 Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res **14:** 860-869.

MARRA, M. A., T. A. KUCABA, N. L. DIETRICH, E. D. GREEN, B. BROWNSTEIN *et al.*, 1997 High throughput fingerprint analysis of large-insert clones. Genome Res **7:** 1072-1084.

MEI, M., N. H. SYED, W. GAO, P. M. THAXTON, C. W. SMITH *et al.*, 2004 Genetic mapping and QTL analysis of fiber-related traits in cotton (Gossypium). Theoretical and Applied Genetics **108:** 280-291.

MIR, R. R., S. RUSTGI, S. SHARMA, R. SINGH, A. GOYAL *et al.*, 2008 A preliminary genetic analysis of fibre traits and the use of new genomic SSRs for genetic diversity in jute. Euphytica **161:** 413-427.

MOORE, R. C., and M. D. PURUGGANAN, 2003 The early stages of duplicate gene evolution. Proc Natl Acad Sci U S A **100:** 15682-15687.

MURAVENKO, O. V., A. R. FEDOTOV, E. O. PUNINA, L. I. FEDOROVA, V. G. GRIF *et al.*, 1998 Comparison of chromosome BrdU-Hoechst-Giemsa banding patterns of the A(1) and (AD)(2) genomes of cotton. Genome **41:** 616-625.

NARBUTH, E. V., and R. J. KOHEL, 1990 Inheritance and Linkage Analysis of a New Fiber Mutant in Cotton. Journal of Heredity **81:** 131-133.

NELSON, W. M., A. K. BHARTI, E. BUTLER, F. WEI, G. FUKS *et al.*, 2005 Whole-genome validation of high-information-content fingerprinting. Plant Physiol **139:** 27-38.

NISHIMURA, T., E. YOKOTA, T. WADA, T. SHIMMEN and K. OKADA, 2003 A semi-dominant mutation in the ACT2 gene affects the root hair development in Arabidopsis. Plant and Cell Physiology **44:** S205-S205.

NIU, C., D. J. HINCHLIFFE, R. G. CANTRELL, C. L. WANG, P. A. ROBERTS *et al.*, 2007 Identification of molecular markers associated with root-knot nematode resistance in upland cotton. Crop Science **47:** 951-960.

NIU, C., H. E. LISTER, B. NGUYEN, T. A. WHEELER and R. J. WRIGHT, 2008 Resistance to Thielaviopsis basicola in the cultivated A genome cotton. Theoretical and Applied Genetics **117:** 1313-1323.

PATERSON, A. H., 2006 Leafing through the genomes of our major crop plants: strategies for capturing unique information. Nat Rev Genet **7:** 174-184.

PATERSON, A. H., 2007 Sequencing the cotton genomes, pp. in *World Cotton Research Conference*. International Cotton Advisory Committee, Lubbock TX.

PATERSON, A. H., J. E. BOWERS, R. BRUGGMANN, I. DUBCHAK, J. GRIMWOOD *et al.*, 2009 The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551-556.

PATERSON, A. H., Y. SARANGA, M. MENZ, C. X. JIANG and R. J. WRIGHT, 2003 QTL analysis of genotype x environment interactions affecting cotton fiber quality. Theoretical and Applied Genetics **106:** 384-396.

PETERSON, D. G., S. R. SCHULZE, E. B. SCIARA, S. A. LEE, J. E. BOWERS *et al.*, 2002a Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res **12:** 795-807.

PETERSON, D. G., J. P. TOMKINS, D. A. FRISCH, W. R. A. and A. H. PATERSON, 2000 Construction of Plant Bacterial Artificial Chromosome (BAC) Libraries: An Illustrated Guide. **5**.

PETERSON, D. G., S. R. WESSLER and A. H. PATERSON, 2002b Efficient capture of unique sequences from eukaryotic genomes. Trends Genet **18:** 547-550.

PIEGU, B., R. GUYOT, N. PICAULT, A. ROULIN, A. SANIYAL *et al.*, 2006 Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Research **16:** 1262-1269.

POSTLETHWAIT, J. H., Y. L. YAN, A. AMORES, B. CRESKO, A. SINGER *et al.*, 2005 Consequences of genome duplication for the evolution of developmental mechanisms in teleost fish. Integrative and Comparative Biology **45:** 1058-1058.

QIN, H. D., W. Z. GUO, Y. M. ZHANG and T. Z. ZHANG, 2008 QTL mapping of yield and fiber traits based on a four-way cross population in Gossypium hirsutum L. Theoretical and Applied Genetics **117:** 883-894.

RAPP, R. A., J. A. UDALL and J. F. WENDEL, 2009 Genomic expression dominance in allopolyploids. Bmc Biology **7:** -.

REINISCH, A. J., J. M. DONG, C. L. BRUBAKER, D. M. STELLY, J. F. WENDEL *et al.*, 1994 A detailed RFLP map of cotton, Gossypium hirsutum x Gossypium barbadense: chromosome organization and evolution in a disomic polyploid genome. Genetics **138:** 829-847.

REN, L. H., W. Z. GUO and T. Z. ZHANG, 2002 Identification of quantitative trait loci (QTLs) affecting yield and fiber properties in chromosome 16 in cotton using substitution line. Acta Botanica Sinica **44:** 815-820.

RONG, J., C. ABBEY, J. E. BOWERS, C. L. BRUBAKER, C. CHANG *et al.*, 2004 A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium). Genetics **166:** 389-417.

RONG, J., J. E. BOWERS, S. R. SCHULZE, V. N. WAGHMARE, C. J. ROGERS *et al.*, 2005a Comparative genomics of Gossypium and Arabidopsis: unraveling the consequences of both ancient and recent polyploidy. Genome Res **15:** 1198-1210.

RONG, J., F. A. FELTUS, V. N. WAGHMARE, G. J. PIERCE, P. W. CHEE *et al.*, 2007 Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. Genetics **176:** 2577-2588.

RONG, J., G. J. PIERCE, V. N. WAGHMARE, C. J. ROGERS, A. DESAI *et al.*, 2005b Genetic mapping and comparative analysis of seven mutants related to seed fiber development in cotton. Theor Appl Genet **111:** 1137-1146.

RUNGIS, D., D. LLEWELLYN, E. S. DENNIS and B. R. LYON, 2002 Investigation of the chromosomal location of the bacterial blight resistance gene present in an Australian cotton (Gossypium hirsutum L.) cultivar. Australian Journal of Agricultural Research **53:** 551-560.

SAHA, S., J. N. JENKINS, J. WU, J. C. MCCARTY and D. M. STELLY, 2008 Genetic analysis of agronomic and fibre traits using four interspecific chromosome substitution lines in cotton. Plant Breeding **127:** 612-618.

SAHA, S., D. RASKA and D. M. STELLY, 2006 Upland (Gossypium hirsutum L.) x Hawaiian cotton (G. tomentosum Nutt. ex Seem.) F1 hybrid hypoaneuploid chromosome substitution series. Journal of Cotton Science **10:** 263-272.

SAJID UR, R., T. A. MALIK, M. ASHRAF and M. AHSAN, 2008 Identification of DNA marker for nectariless trait in cotton using random amplified polymorphic DNA. Pakistan Journal of Botany **40:** 1711-1719.

SARANGA, Y., C. X. JIANG, R. J. WRIGHT, D. YAKIR and A. H. PATERSON, 2004 Genetic dissection of cotton physiological responses to arid conditions and their inter-relationships with productivity. Plant Cell and Environment **27:** 263-277.

SARANGA, Y., M. MENZ, C. X. JIANG, R. J. WRIGHT, D. YAKIR *et al.*, 2001 Genomic dissection of genotype x environment interactions conferring adaptation of cotton to arid conditions. Genome Research **11:** 1988-1995.

SEMON, M., and K. H. WOLFE, 2007 Consequences of genome duplication. Current Opinion in Genetics & Development **17:** 505-512.

SENCHINA, D. S., I. ALVAREZ, R. C. CRONN, B. LIU, J. RONG *et al.*, 2003 Rate variation among nuclear genes and the age of polyploidy in Gossypium. Mol Biol Evol **20:** 633-643.

SHEN, X. L., W. Z. GUO, Q. X. LU, X. F. ZHU, Y. L. YUAN *et al.*, 2007 Genetic mapping of quantitative trait loci for fiber quality and yield trait by RIL approach in Upland cotton. Euphytica **155:** 371-380.

SHEN, X. L., W. Z. GUO, X. F. ZHU, Y. L. YUAN, J. Z. YU *et al.*, 2005 Molecular mapping of QTLs for fiber qualities in three diverse lines in Upland cotton using SSR markers. Molecular Breeding **15:** 169-181.

SHEN, X. L., G. VAN BECELAERE, P. KUMAR, R. F. DAVIS, O. L. MAY *et al.*, 2006a QTL mapping for resistance to root-knot nematodes in the M-120 RNR Upland cotton line (Gossypium hirsutum L.) of the Auburn 623 RNR source. Theoretical and Applied Genetics **113:** 1539-1549.

SHEN, X. L., T. Z. ZHANG, W. Z. GUO, X. F. ZHU and X. Y. ZHANG, 2006b Mapping fiber and yield QTLs with main, epistatic, and QTL X environment interaction effects in recombinant inbred lines of upland cotton. Crop Science **46:** 61-66.

SIMILLION, C., K. VANDEPOELE, M. C. E. VAN MONTAGU, M. ZABEAU and Y. VAN DE PEER, 2002 The hidden duplication past of Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America **99:** 13627-13632.

SNELLING, W. M., R. CHIU, J. E. SCHEIN, M. HOBBS, C. A. ABBEY *et al.*, 2007 A physical map of the bovine genome. Genome Biol **8:** R165.

SODERLUND, C., S. HUMPHRAY, A. DUNHAM and L. FRENCH, 2000 Contigs built with fingerprints, markers, and FPC V4.7. Genome Res **10:** 1772-1787.

SODERLUND, C., I. LONGDEN and R. MOTT, 1997 FPC: a system for building contigs from restriction fingerprinted clones. Comput Appl Biosci **13:** 523-535.

SONG, X. L., W. Z. GUO, Z. G. HAN and T. Z. ZHANG, 2005 Quantitative trait loci mapping of leaf morphological traits and chlorophyll content in cultivated tetraploid cotton. Journal of Integrative Plant Biology **47:** 1382-1390.

SONG, X. L., and T. Z. ZHANG, 2007 Identification of quantitative trait loci controlling seed physical and nutrient traits in cotton. Seed Science Research **17:** 243-251.

SULSTON, J., F. MALLETT, R. DURBIN and T. HORSNELL, 1989 Image analysis of restriction enzyme fingerprint autoradiograms. Comput Appl Biosci **5:** 101-106.

TALIERCIO, E. W., and D. BOYKIN, 2007 Analysis of gene expression in cotton fiber initials. Bmc Plant Biology **7:** -.

TANG, H., J. E. BOWERS, X. WANG, R. MING, M. ALAM *et al.*, 2008a Synteny and collinearity in plant genomes. Science **320:** 486-488.

TANG, H. B., X. Y. WANG, J. E. BOWERS, R. MING, M. ALAM *et al.*, 2008b Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Research **18:** 1944-1954.

TOMKINS, J. P., D. G. PETERSON, T. J. YANG, D. MAIN, T. A. WILKINS *et al.*, 2001 Development of genomic resources for cotton (Gossypium hirsutum L.): BAC library construction, preliminary STC analysis, and identification of clones associated with fiber development. Molecular Breeding **8:** 255-261.

TURLEY, R. B., and R. H. KLOTH, 2008 The inheritance model for the fiberless trait in upland cotton (Gossypium hirsutum L.) line SL1-7-1: variation on a theme. Euphytica **164:** 123-132.

UDALL, J. A., L. E. FLAGEL, F. CHEUNG, A. W. WOODWARD, R. HOVAV *et al.*, 2007 Spotted cotton oligonucleotide microarrays for gene expression analysis. BMC Genomics **8:** -.

ULLOA, M., S. SAHA, J. N. JENKINS, W. R. MEREDITH, J. C. MCCARTY *et al.*, 2005 Chromosomal assignment of RFLP linkage groups harboring important QTLs on an intraspecific cotton (Gossypium hirsutum L.) joinmap. Journal of Heredity **96:** 132-144.

WAGHMARE, V. N., J. RONG, C. J. ROGERS, G. J. PIERCE, J. F. WENDEL *et al.*, 2005 Genetic mapping of a cross between Gossypium hirsutum (cotton) and the Hawaiian endemic, Gossypium tomentosum. Theor Appl Genet **111:** 665-676.

WALBOT, V., and L. S. DURE, 3RD, 1976 The developmental biochemistry of cotton seed embryogenesis and germination. VII. Characterization of the cotton genome. Biochim Biophys Acta **101:** 503-536.

WAN, Q., Z. S. ZHANG, M. HU, L. CHEN, D. J. LIU *et al.*, 2007 T-1 locus in cotton is the candidate gene affecting lint percentage, fiber quality and spiny bollworm (Earias spp.) resistance. Euphytica **158:** 241-247.

WANG, B. H., W. Z. GUO, X. F. ZHU, Y. T. WU, N. T. HUANG *et al.*, 2006a QTL mapping of fiber quality in an elite hybrid derived-RIL population of upland cotton. Euphytica **152:** 367-378.

WANG, B. H., Y. T. WU, W. Z. GUO, X. F. ZHU, N. T. HUANG *et al.*, 2007a QTL analysis and epistasis effects dissection of fiber qualities in an elite cotton hybrid grown in second generation. Crop Science **47:** 1384-1392.

WANG, C., M. ULLOA and P. A. ROBERTS, 2006b Identification and mapping of microsatellite markers linked to a root-knot nematode resistance gene (rkn1) in Acala NemX cotton (Gossypium hirsutum L.). Theoretical and Applied Genetics **112:** 770-777.

WANG, C. L., and P. A. ROBERTS, 2006 Development of AFLP and derived CAPS markers for root-knot nematode resistance in cotton. Euphytica **152:** 185-196.

WANG, F., J. M. STEWART and J. ZHANG, 2007b Molecular markers linked to the Rf(2) fertility restorer gene in cotton. Genome **50:** 818-824.

WANG, H. M., Z. X. LIN, X. L. ZHANG, W. CHEN, X. P. GUO *et al.*, 2008 Mapping and quantitative trait loci analysis of verticillium wilt resistance genes in cotton. Journal of Integrative Plant Biology **50:** 174-182.

WANG, K., X. SONG, Z. HAN, W. GUO, J. Z. YU *et al.*, 2006c Complete assignment of the chromosomes of Gossypium hirsutum L. by translocation and fluorescence in situ hybridization mapping. Theor Appl Genet **113:** 73-80.

WANG, K., X. L. SONG, Z. G. HAN, W. Z. GUO, J. Z. YU *et al.*, 2006d Complete assignment of the chromosomes of Gossypium hirsutum L. by translocation and fluorescence in situ hybridization mapping. Theoretical and Applied Genetics **113:** 73-80.

WANG, X., H. TANG, J. E. BOWERS, F. A. FELTUS and A. H. PATERSON, Accepted Extensive concerted evolution of rice paralogs and the road to regaining independence. Genetics.

WANG, X. Y., X. L. SHI, Z. LI, Q. H. ZHU, L. KONG *et al.*, 2006e Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. Bmc Bioinformatics **7:** -.

WENDEL, J. F., 1989 New World tetraploid cottons contain Old World cytoplasm. Proc Natl Acad Sci U S A **86:** 4132-4136.

WENDEL, J. F., and V. A. ALBERT, 1992 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 **17:** 115-143.

WENDEL, J. F., R. C. CRONN, I. ALVAREZ, B. LIU, R. L. SMALL *et al.*, 2002a Intron size and genome size in plants. Mol Biol Evol **19:** 2346-2352.

WENDEL, J. F., R. C. CRONN, J. S. JOHNSTON and H. J. PRICE, 2002b Feast and famine in plant genomes. Genetica **115:** 37-47.

WENDEL, J. F., A. SCHNABEL and T. SEELANAN, 1995a Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc Natl Acad Sci U S A **92:** 280-284.

WENDEL, J. F., A. SCHNABEL and T. SEELANAN, 1995b An unusual ribosomal DNA sequence from Gossypium gossypioides reveals ancient, cryptic, intergenomic introgression. Mol Phylogenet Evol **4:** 298-313.

WRIGHT, R., P. THAXTON, K. EL-ZIK and A. H. PATERSON, 1998 D-subgenome bias of Xcm resistance genes in tetraploid *Gossypium* (Cotton) suggests that polyploid formation has created novel avenues for evolution. Genetics **149:** 1987-1996.

WRIGHT, R. J., P. M. THAXTON, K. H. EL-ZIK and A. H. PATERSON, 1999 Molecular mapping of genes affecting pubescence of cotton. Journal of Heredity **90:** 215-219.

WU, J. X., J. N. JENKINS, J. C. MCCARTY, M. ZHONG and M. SWINDLE, 2007 AFLP marker associations with agronomic and fiber traits in cotton. Euphytica **153:** 153-163.

WU, Y. G., S. ROZENFELD, A. DEFFERRARD, K. RUGGIERO, J. A. UDALL *et al.*, 2005 Cycloheximide treatment of cotton ovules alters the abundance of specific classes of mRNAs and generates novel ESTs for microarray expression profiling. Molecular Genetics and Genomics **274:** 477-493.

XIAO, J., K. WU, D. FANG, D. M. STELLY, J. YU *et al.*, 2009 New SSR Markers for Use in Cotton (*Gossypium* spp.*)* Improvement. The Journal of Cotton Science **13:** 75-157.

XU, Z., R. J. KOHEL, G. SONG, J. CHO, J. YU *et al.*, 2008 An integrated genetic and physical map of homoeologous chromosomes 12 and 26 in Upland cotton (G. hirsutum L.). BMC Genomics **9:** 108.

XU, Z., S. SUN, L. COVALEDA, K. DING, A. ZHANG *et al.*, 2004 Genome physical mapping with large-insert bacterial clones by fingerprint analysis: methodologies, source clone genome coverage, and contig map quality. Genomics **84:** 941-951.

YANG, C., W. Z. GUO, G. Y. LI, F. GAO, S. S. LIN *et al.*, 2008 QTLs mapping for Verticillium wilt resistance at seedling and maturity stages in Gossypium barbadense L. Plant Science **174:** 290-298.

YIN, J., W. GUO, L. YANG, L. LIU and T. ZHANG, 2006 Physical mapping of the Rf1 fertility-restoring gene to a 100 kb region in cotton. Theor Appl Genet **112:** 1318-1325.

YNTURI, P., J. N. JENKINS, J. C. MCCARTY, O. A. GUTIERREZ and S. SAHA, 2006 Association of root-knot nematode resistance genes with simple sequence repeat markers on two chromosomes in cotton. Crop Science **46:** 2670-2674.

YU, J. W., S. X. YU, C. R. LU, W. WANG, S. L. FAN *et al.*, 2007 High-density linkage map of cultivated allotetraploid cotton based on SSR, TRAP, SRAP and AFLP markers. Journal of Integrative Plant Biology **49:** 716-724.

ZHANG, J. F., and J. M. STEWART, 2004 Identification of molecular markers linked to the fertility restorer genes for CMS-D8 in cotton. Crop Science **44:** 1209-1217.

ZHANG, T. Z., Y. L. YUAN, J. YU, W. Z. GUO and R. J. KOHEL, 2003 Molecular tagging of a major QTL for fiber strength in Upland cotton and its marker-assisted selection. Theoretical and Applied Genetics **106:** 262-268.

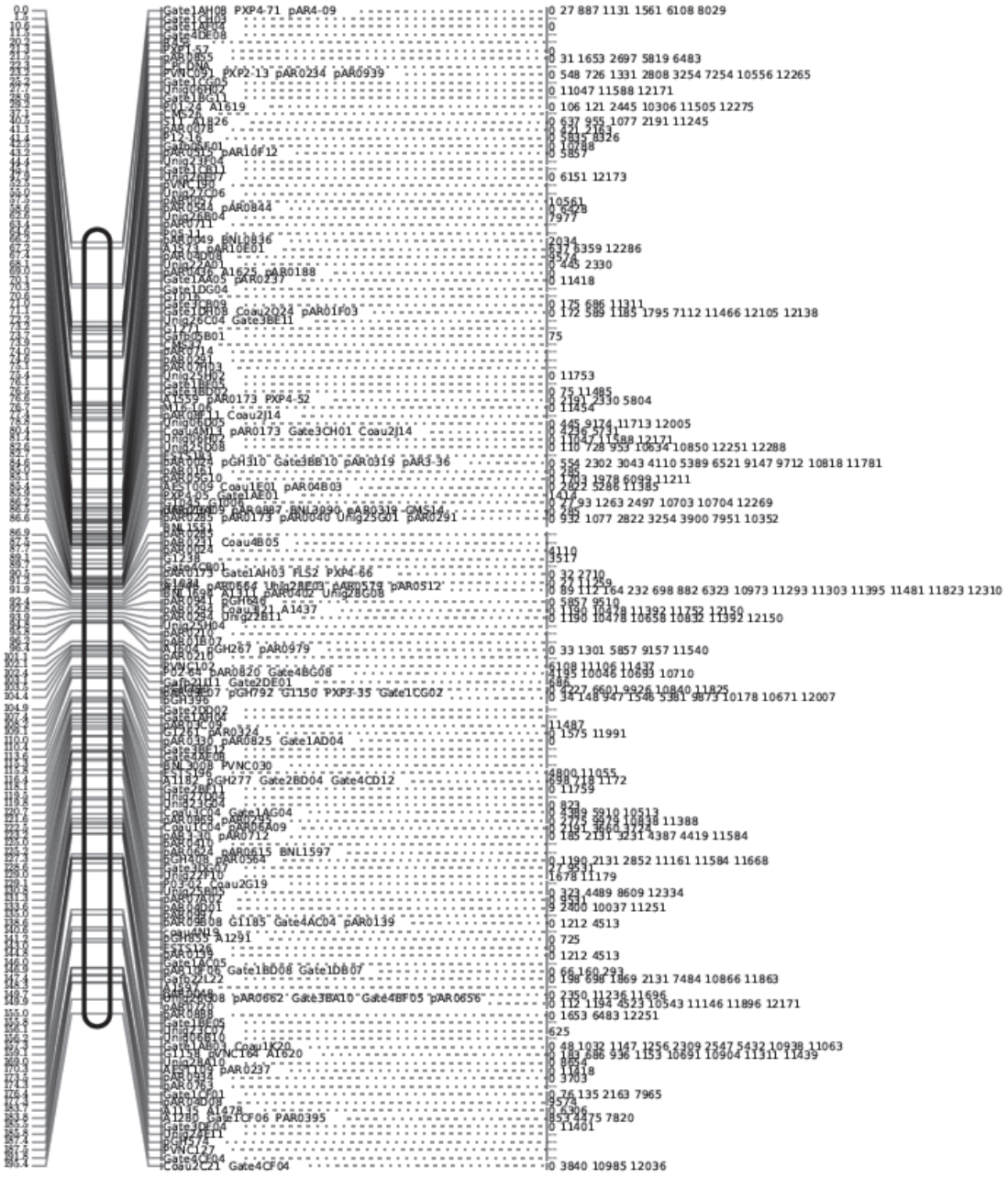
ZHAO, X., R. A. WING and A. H. PATERSON, 1995 Cloning and characterization of the majority of repetitive DNA in cotton (Gossypium L.). Genome **38:** 1177-1188.

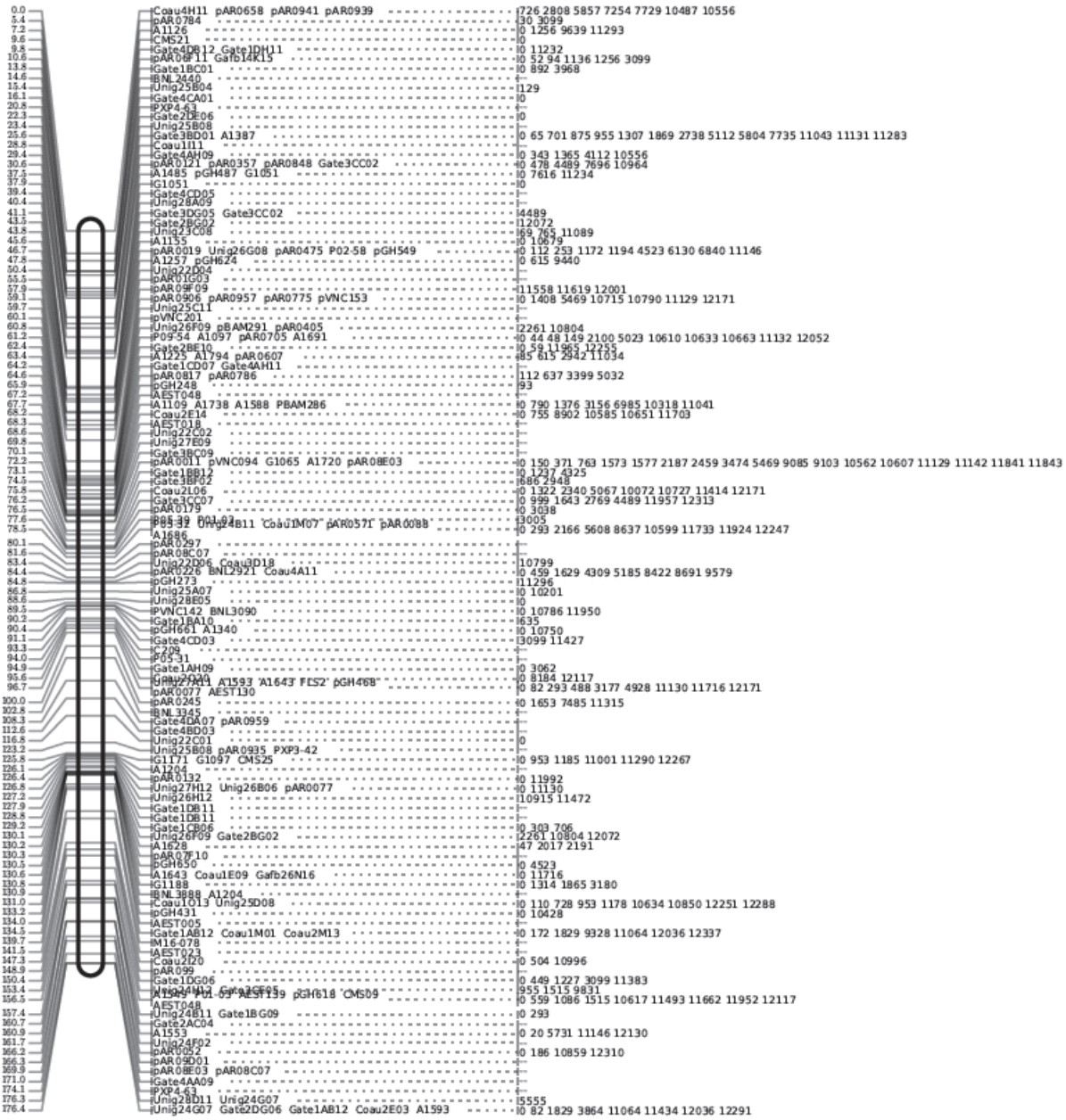
ZHAO, X. P., Y. SI, R. E. HANSON, C. F. CRANE, H. J. PRICE *et al.*, 1998 Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res **8:** 479-492.

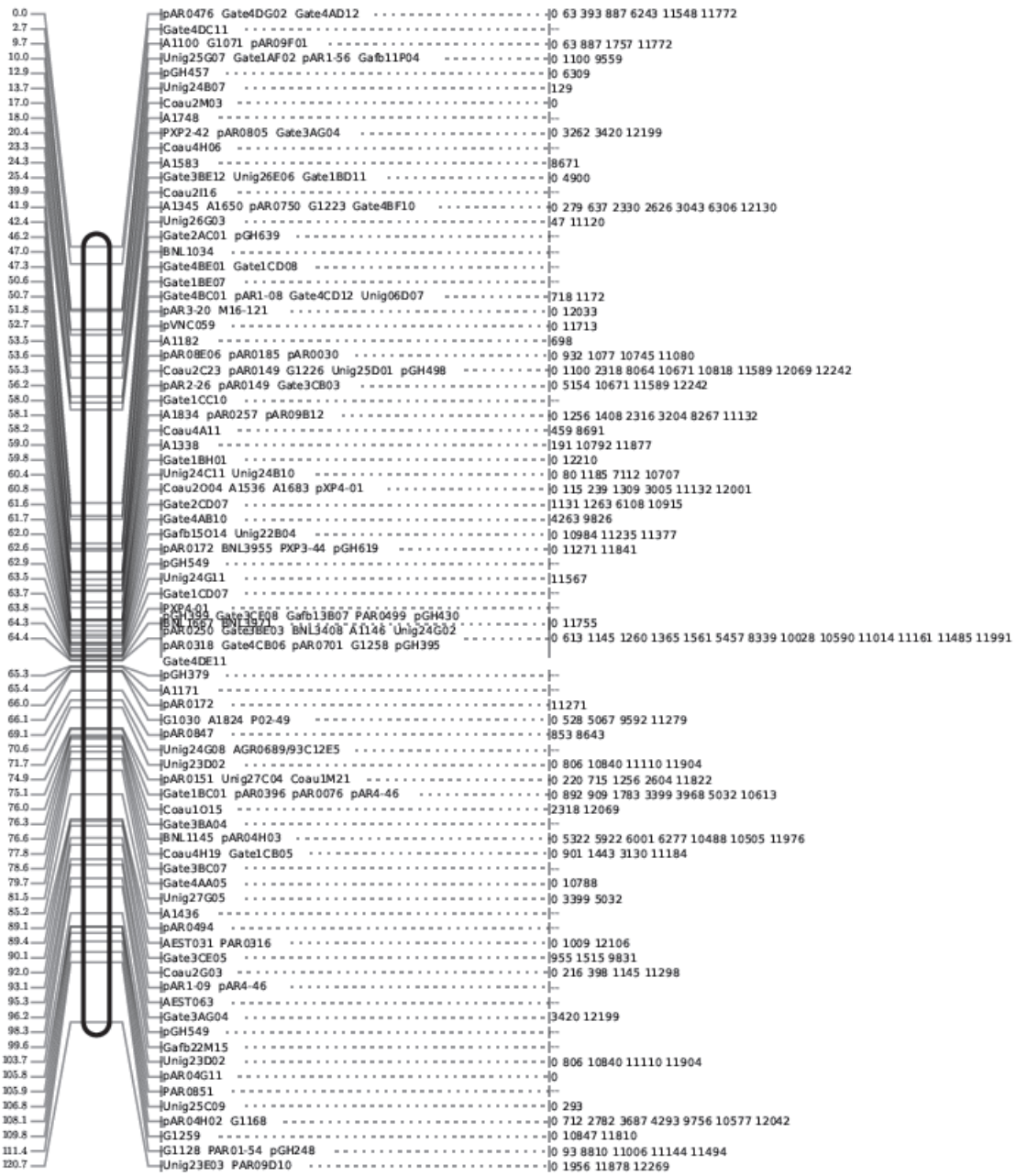
ZHAO, X. Q., J. L. XU, M. ZHAO, R. LAFITTE, L. H. ZHU *et al.*, 2008 QTLs affecting morph-physiological traits related to drought tolerance detected in overlapping introgression lines of rice (Oryza sativa L.). Plant Science **174:** 618-625.

APPENDICES

APPENDIX 1 THE ANCHORING OF COTTON D GENOME CONTIGS ON THE CONSENSUS MAP

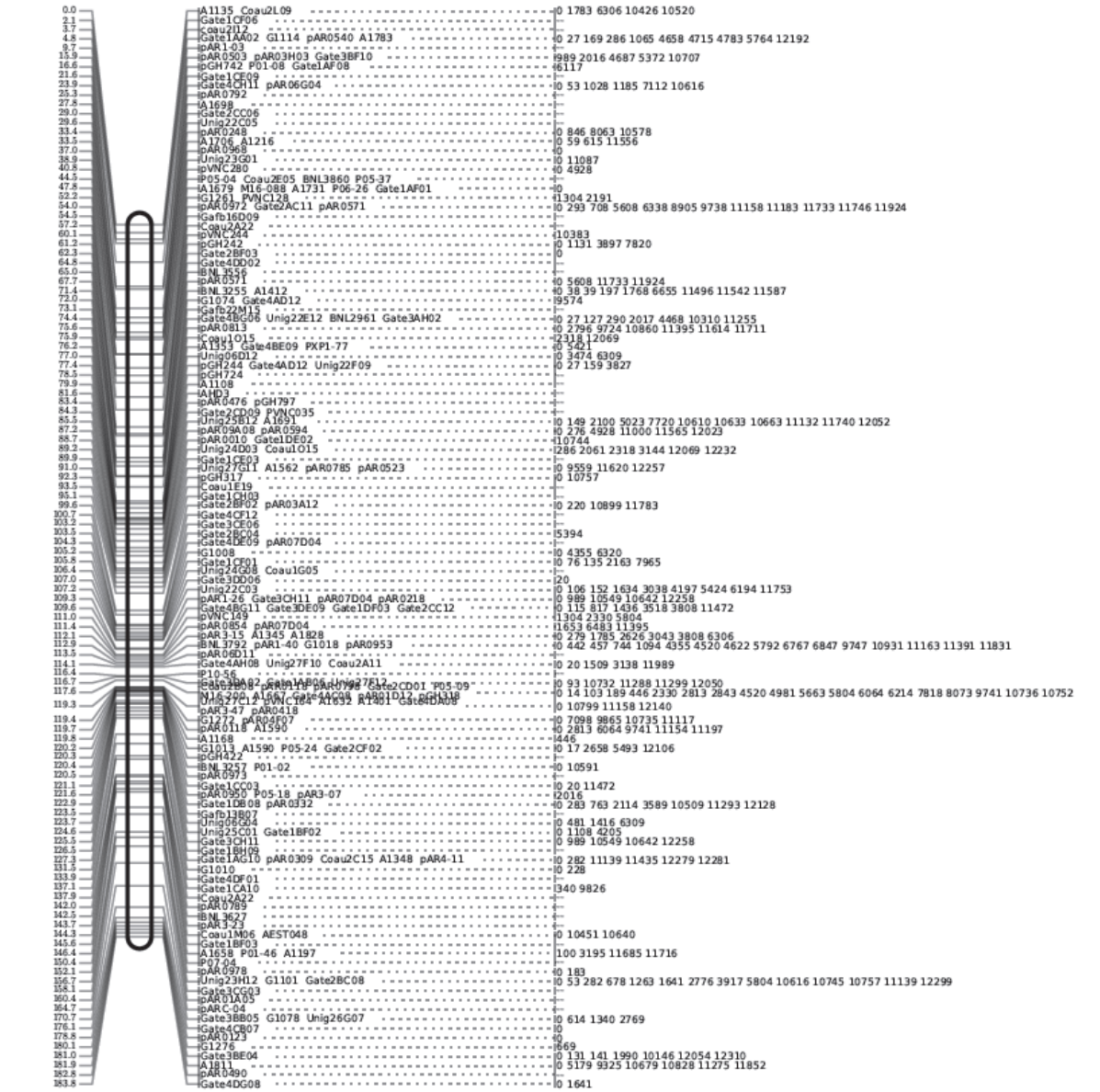



Genetic map
(Homologous group 2)
Physical contigs



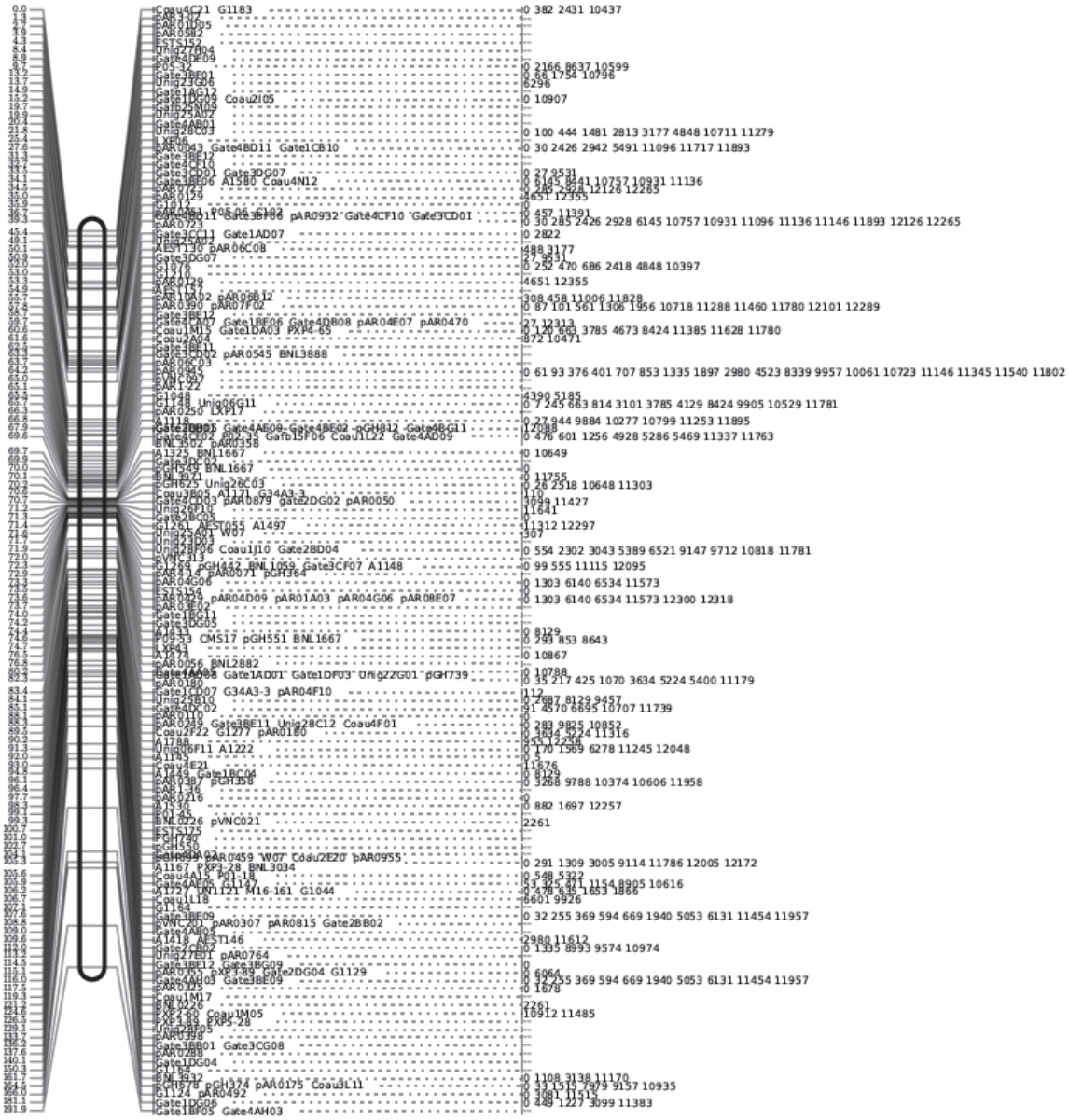
Genetic map (Homologous group 3)

Physical contigs



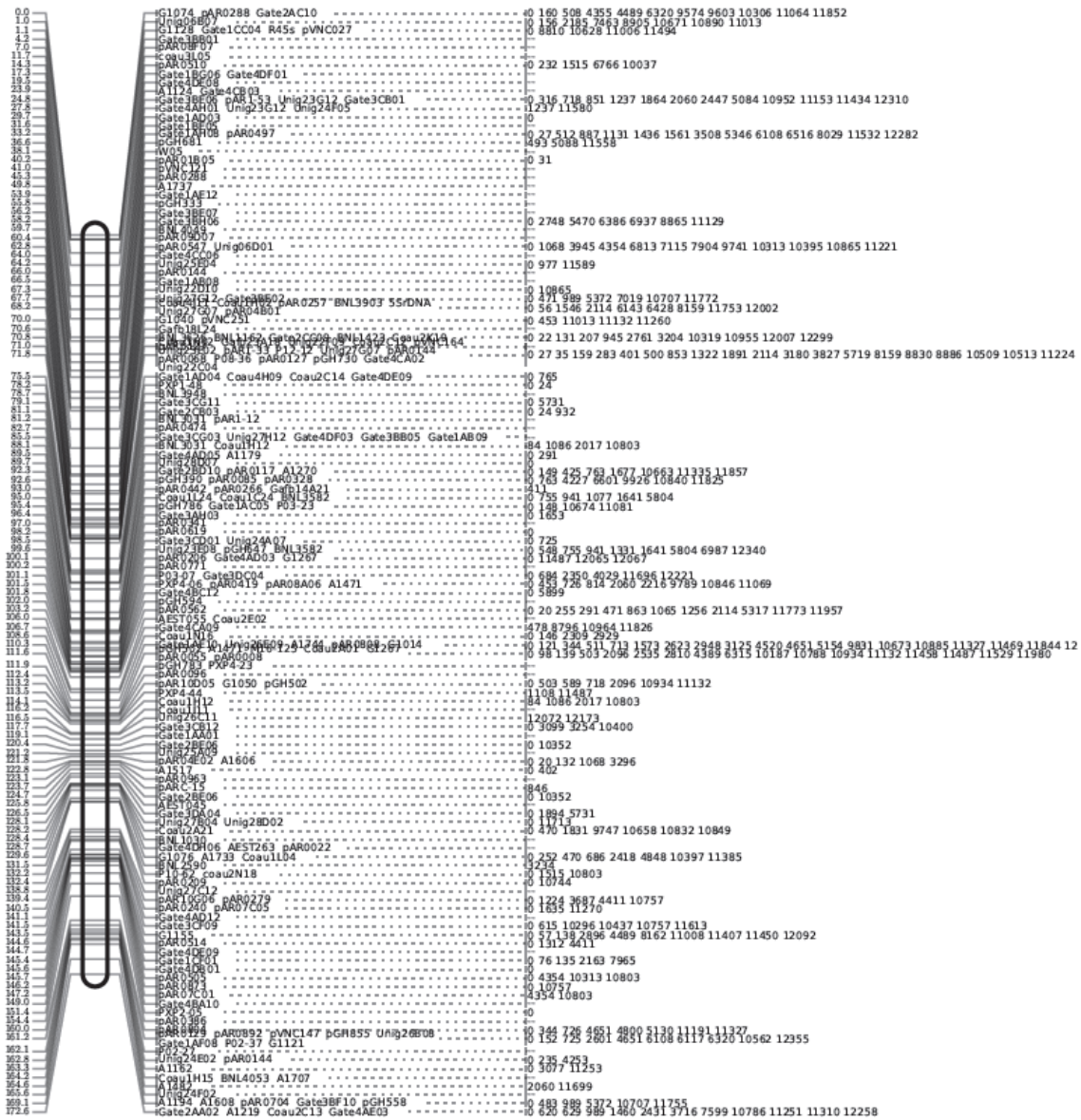
Genetic map (Homologous group 4)

Physical contigs



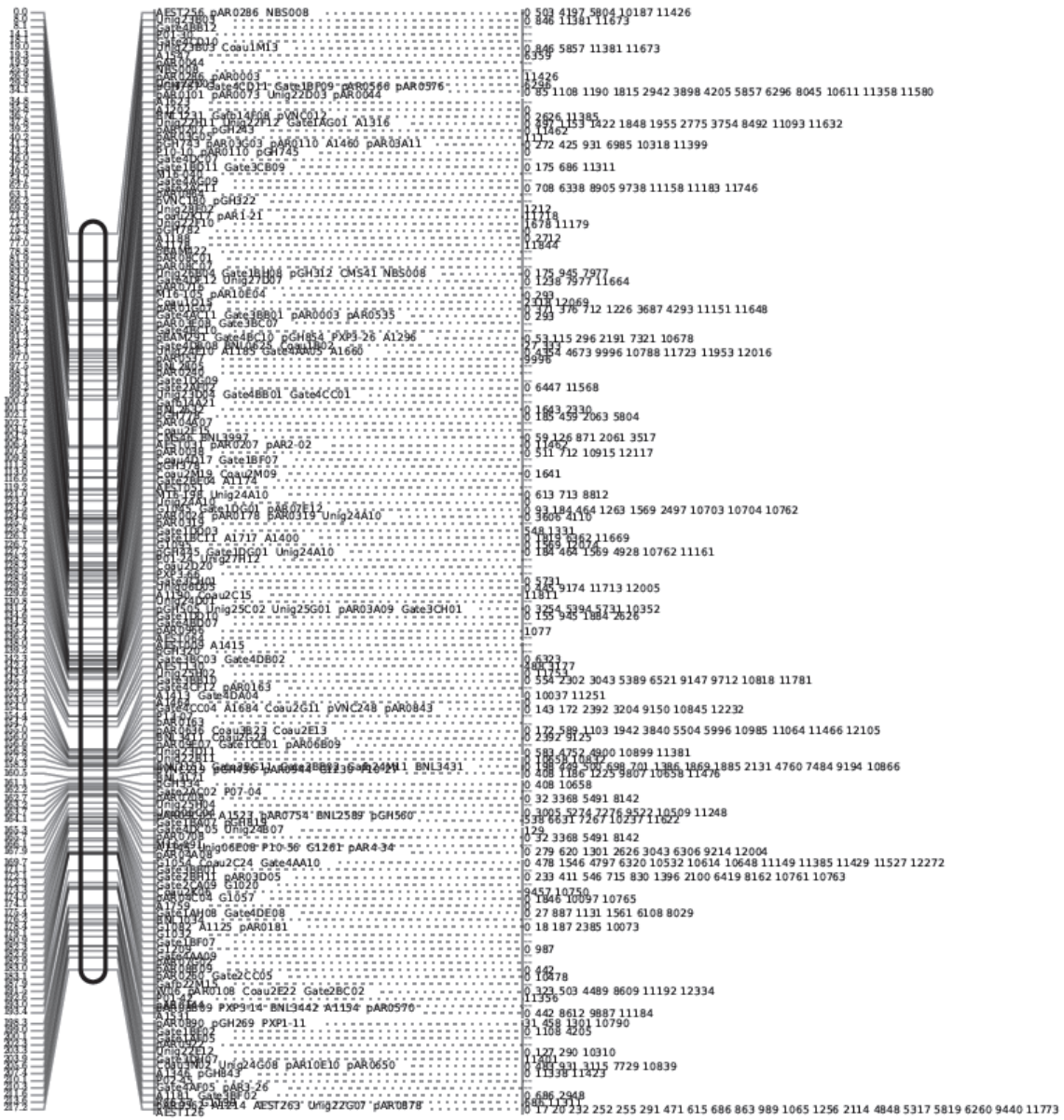
Genetic map (Homologous group 5)

Physical contigs



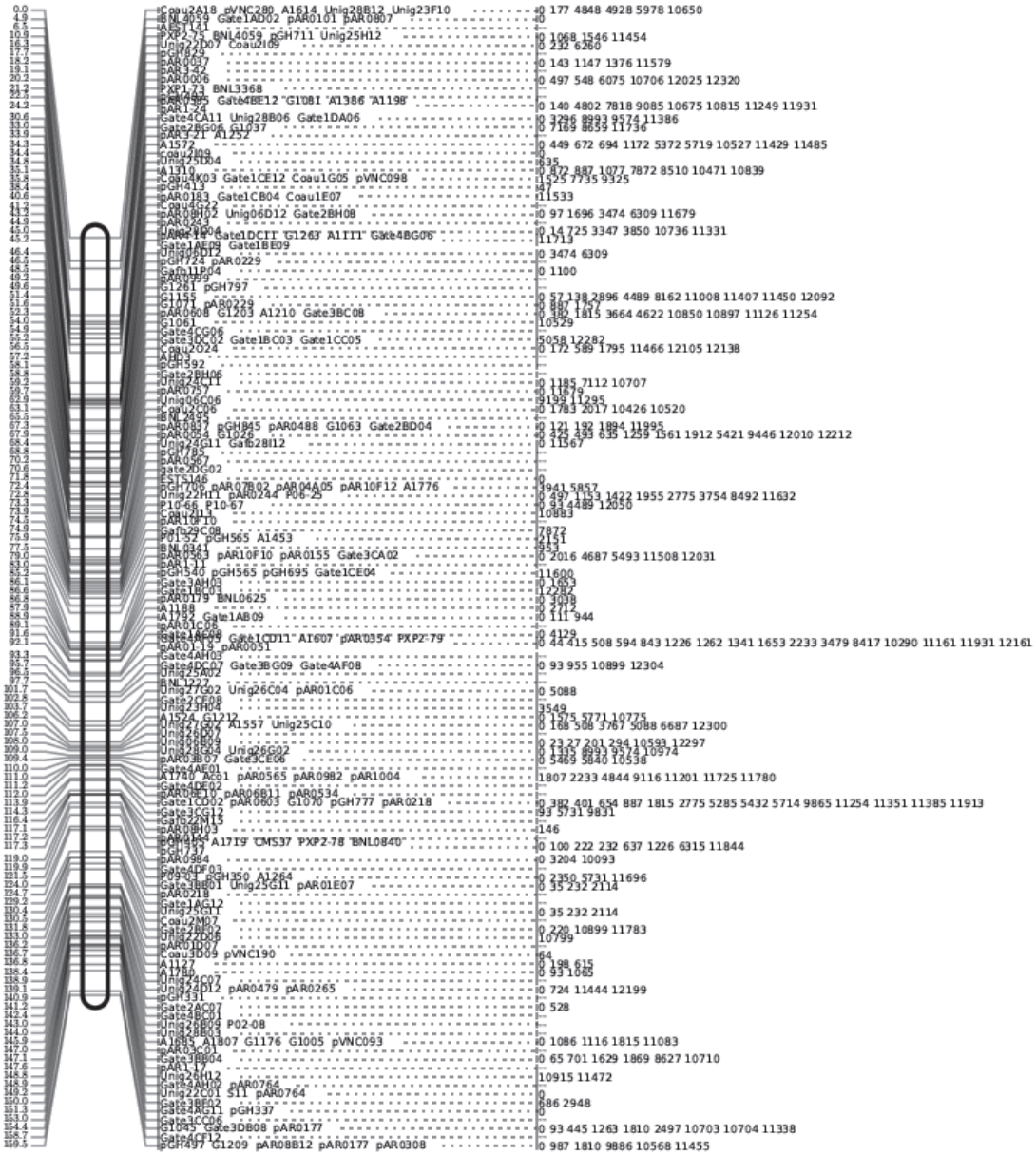
Genetic map
(Homologous group 6)

Physical contigs



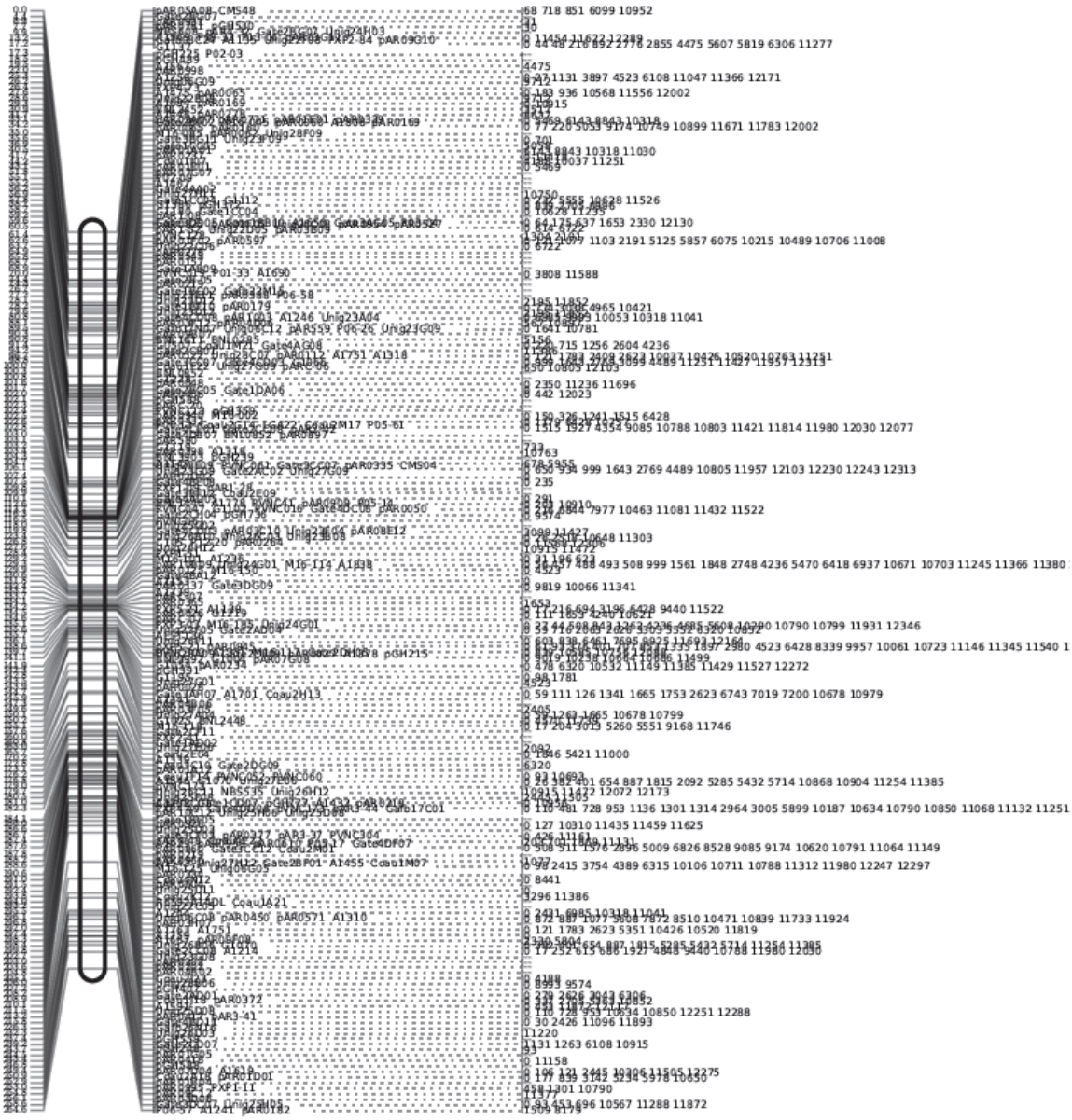
Genetic map (Homologous group 7)

Physical contigs



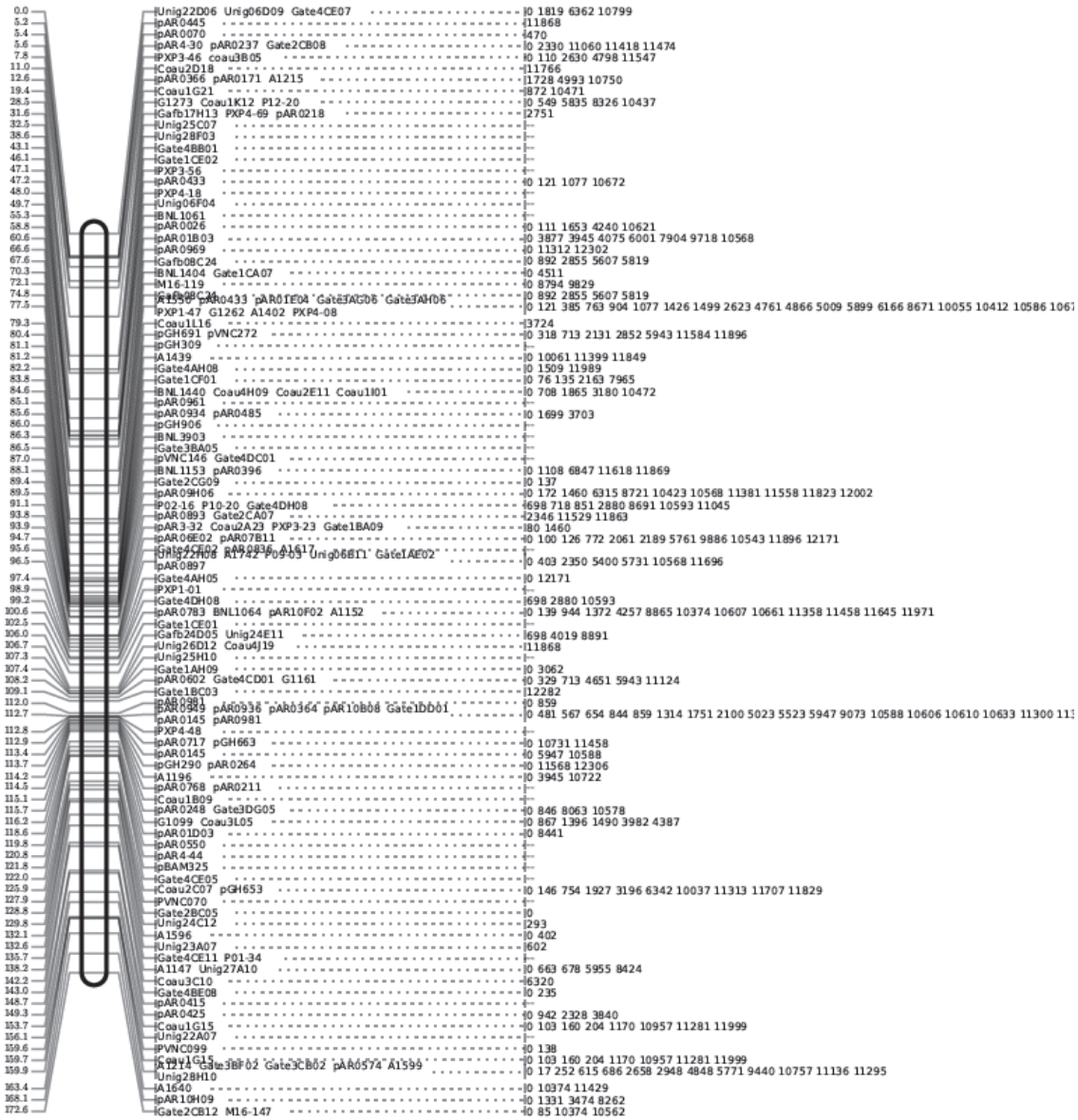
Genetic map (Homologous group 8)

Physical contigs



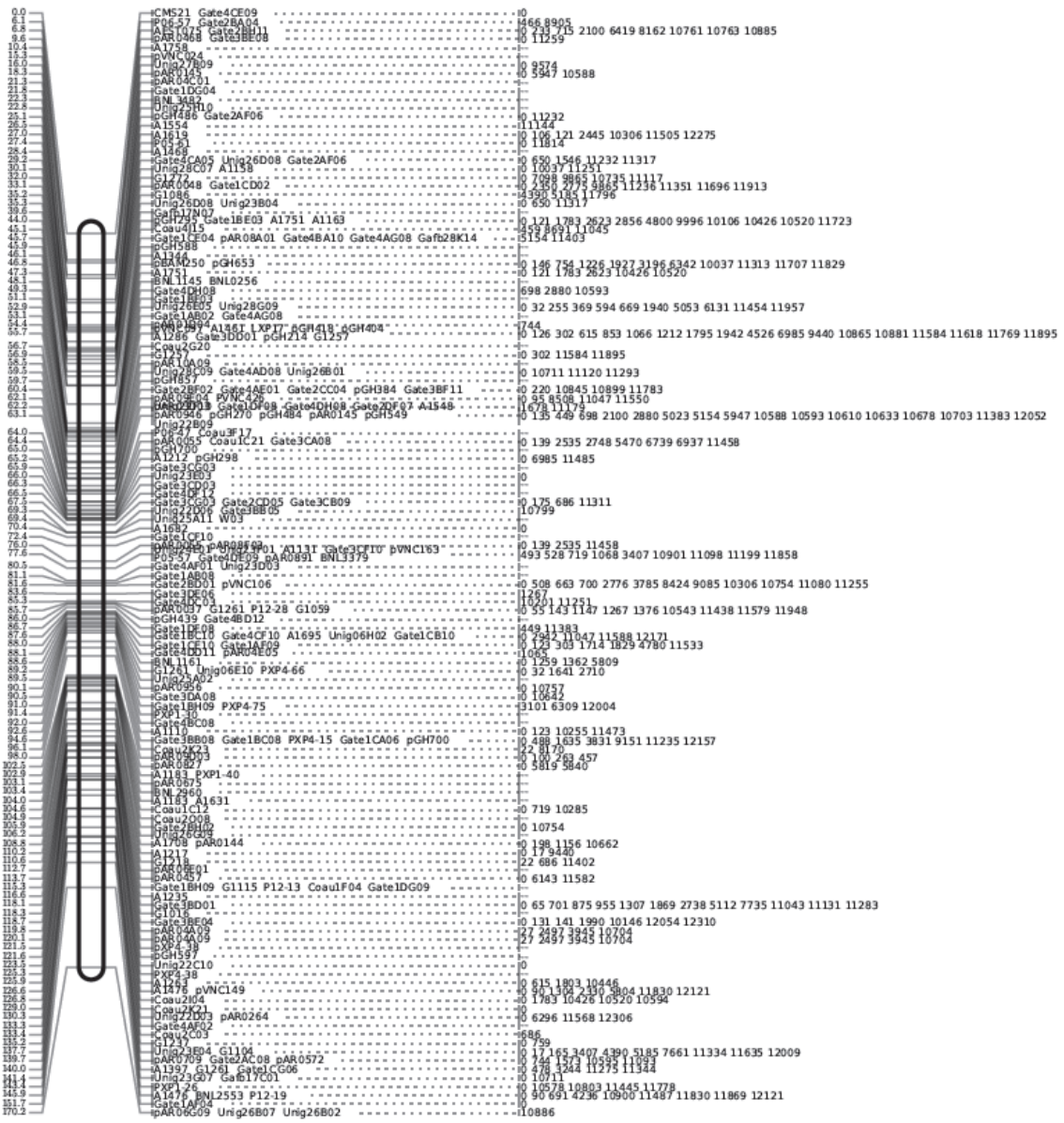
Genetic map
(Homologous group 9)

Physical contigs



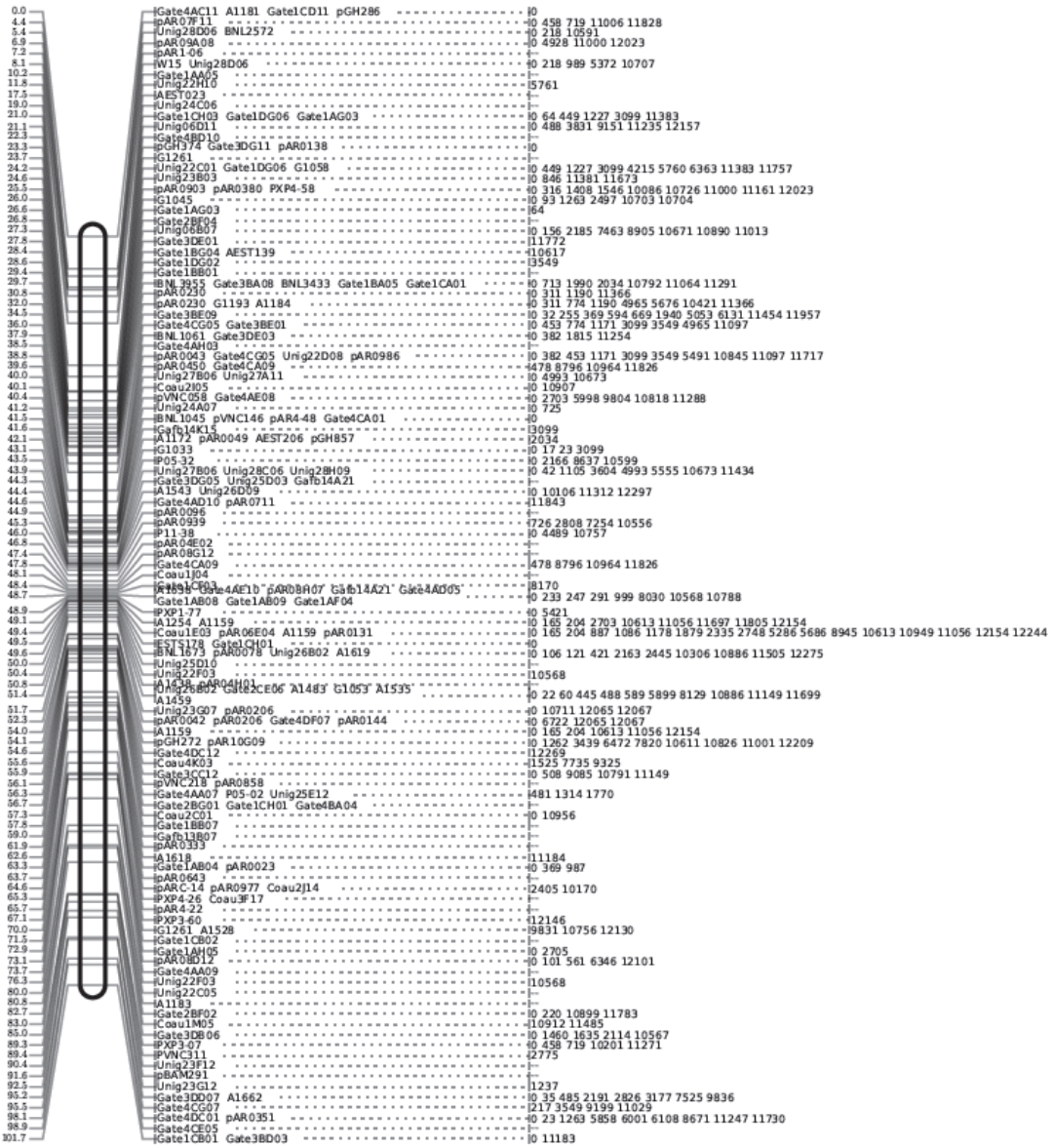
Genetic map (Homologous group 10)

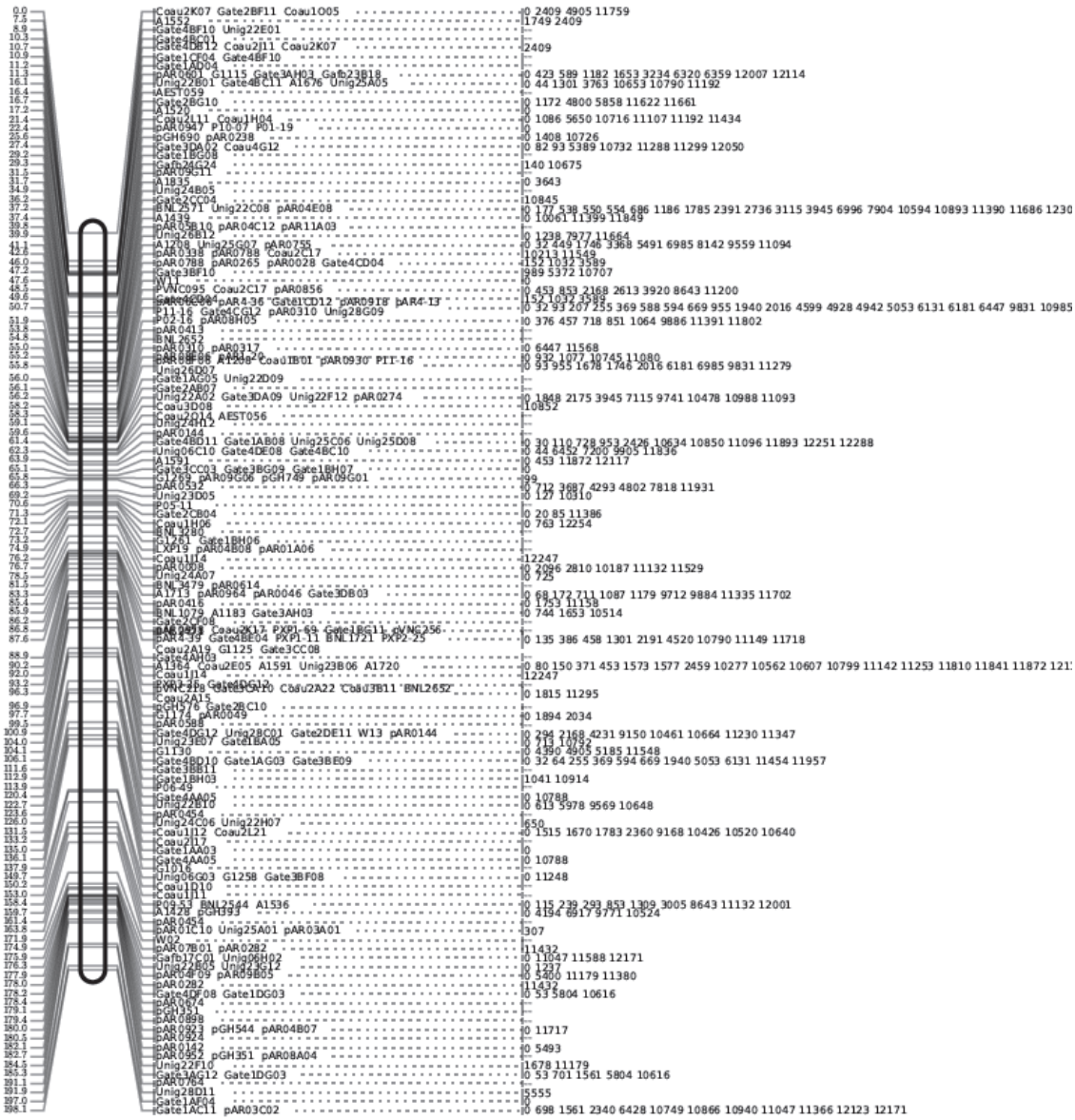
Physical contigs



Genetic map (Homologous group 11)

Physical contigs


Genetic map
(Homologous group 12)
Physical contigs



Genetic map (Homologous group 13)

Physical contigs