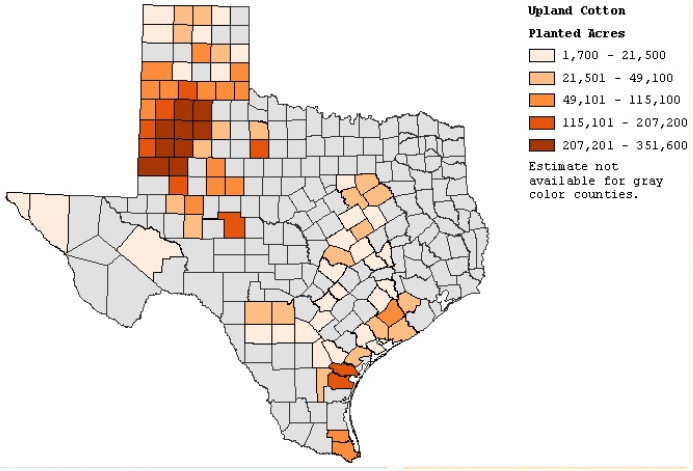
Toward Development of Robust Integrated Physical and Genetic Maps for Individual Chromosomes of Upland Cotton (*Gossypium hirsutum* L.) for Accurately Sequencing Its Genome



YANG ZHANG, MEIPING ZHANG, Yun-Hua Liu, C. Wayne Smith, Steve Hague, David M. Stelly, Hong-Bin Zhang*

Department of Soil and Crop Sciences, Texas A&M University, College Station, Texas

Outline

1. Introduction
2. Strategy and Methodology
3. Results
4. Conclusions
5. Future Plan

Introduction

Key words:
Polyploid genomes, physical mapping, cottons

1 Introduction: Polyploid Genomes

- Polyploidy is a significant evolutionary process in higher organisms. The genomes of most angiosperms are thought to have incurred one or more polyploidization events during evolution.
- Studies have demonstrated that genome doubling has also been significant in the evolutionary history of all vertebrates and in many other eukaryotes.
- It is estimated that about 70% of the extant angiosperms are polyploids, including many world-leading field, forage, horticultural and environmental crops such as cotton, wheat, potatoes, canola, sugarcane, oats, peanut, tobacco, rose, alfalfa, coffee and banana.

1 Introduction: Polyploid Genomes

- Genomics research of polyploid is generally behind that of diploid species due to their polyploidy nature that could significantly complicate genome research, especially genome physical mapping with large-insert bacterial artificial chromosome (BAC) and/or transformation-competent binary BAC (BIBAC) clones.

1 Introduction: Physical Mapping

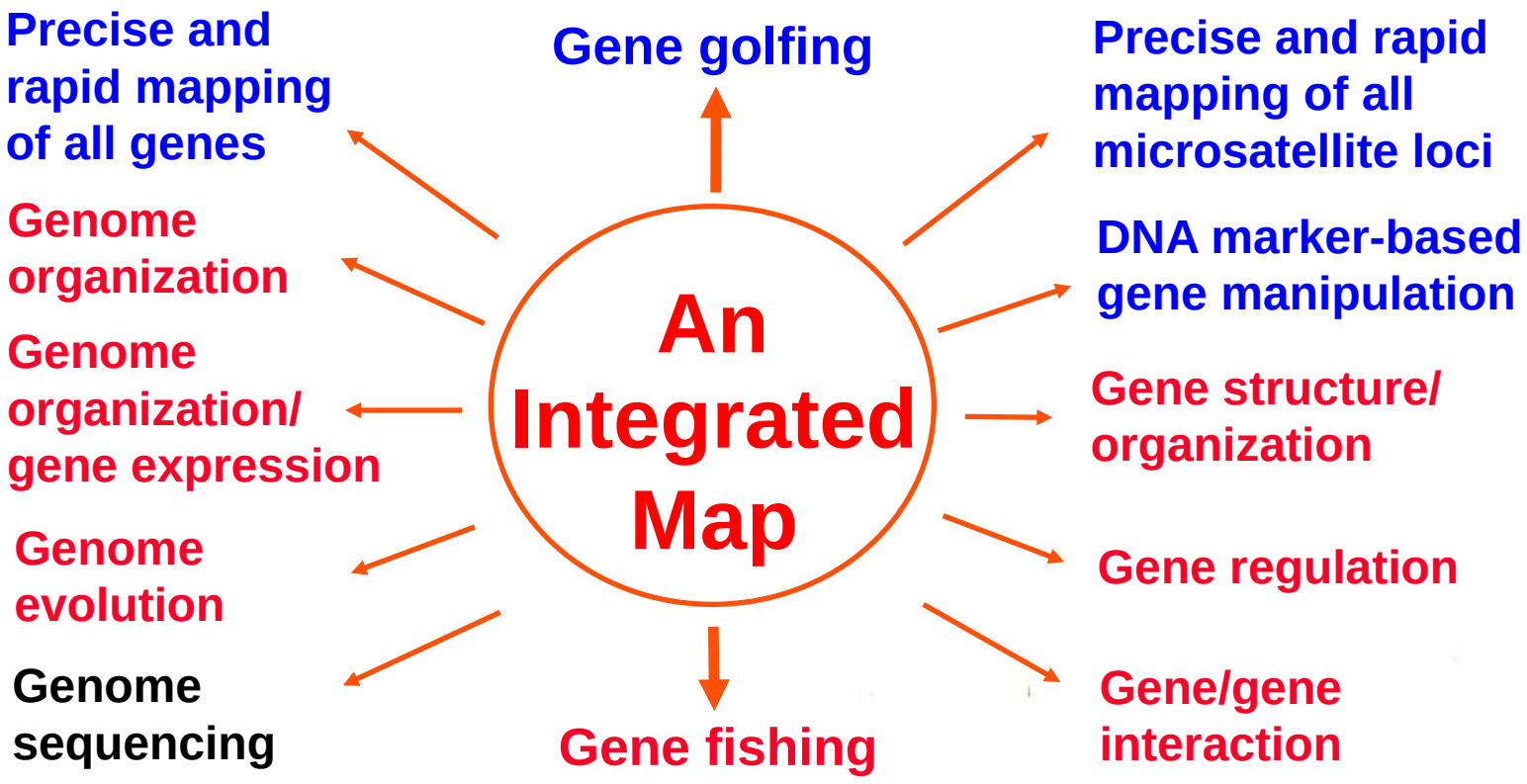
- Whole genome physical maps have been demonstrated to be the centerpiece essential for many areas of advanced studies, such as gene and QTL fine mapping and cloning, functional genomics, comparative genomics, and genome sequencing and assembly.
- Physical maps constructed in a number of plant species: *Arabidopsis*, rice, soybean, maize, chickpea, apple, poplar …

1 Introduction: Physical Mapping

- However, no physical map has been developed and no genome sequenced to date for a polyploid species that contains two or more homoeologous genomes.

Introduction: Physical Mapping

An integrated physical and genetic map provides a platform and "freeway" for high-throughput target marker development, gene mapping, gene isolation, and many advanced studies of functional and comparative genomics



1 Introduction: Why *Gossypium hirsutum* L.

- Upland cotton is an allotetraploid, consisting of A- and D-subgenomes, and has a genome size of 2,400 Mb/1C.
- It was originated around 1 - 2 million years ago via allopolyploidization between a diploid species containing an A genome and a diploid species containing a D genome, whereas the A- and D-subgenomes are homoeologous, their diploid progenitors having split from a common ancestor some 5 - 7 million years ago.

1 Introduction: Why *Gossypium hirsutum* L.

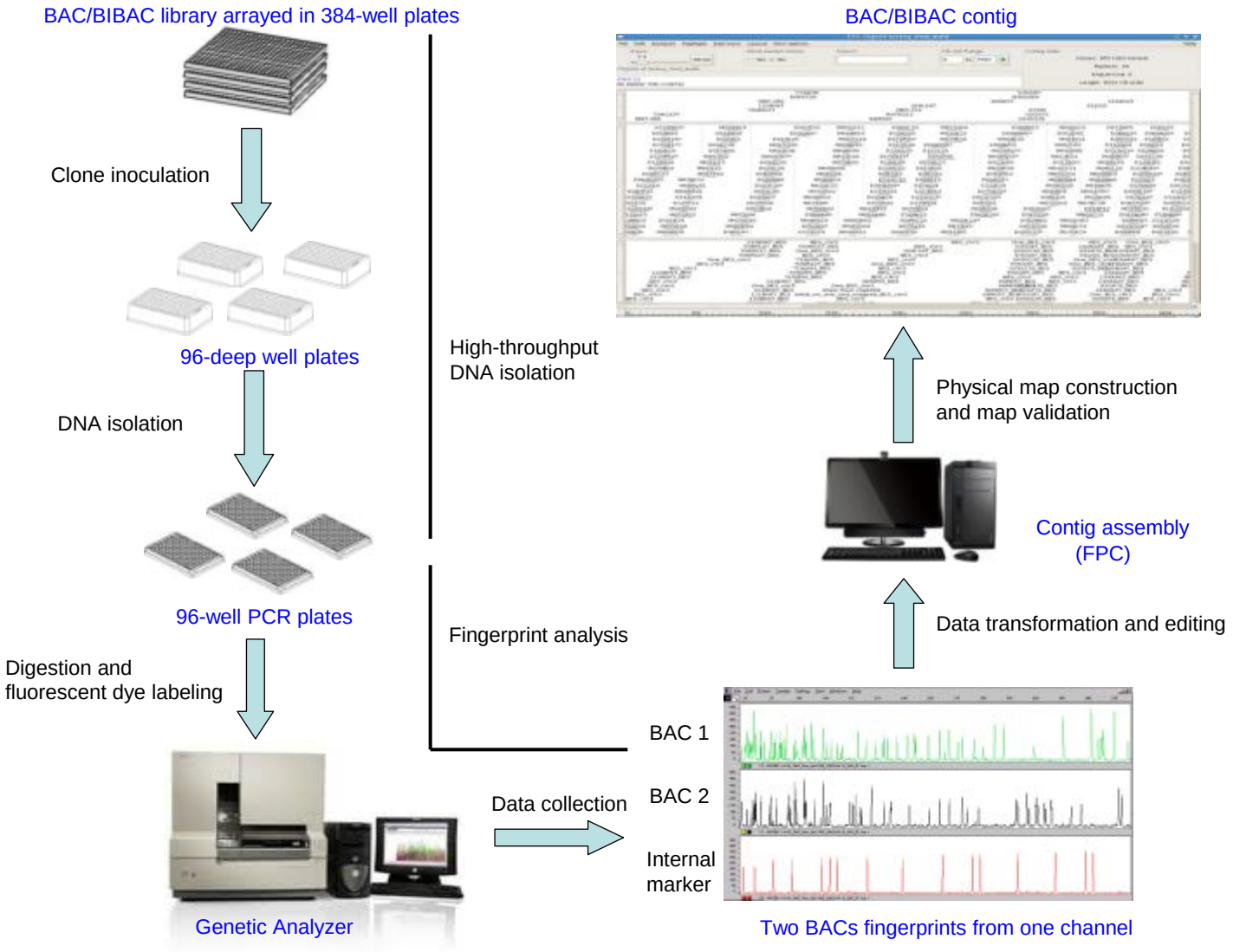
- Cottons are a world-leading fiber and an important oilseed crop. Upland cotton provides over 90% of the world's cotton fibers and oilseeds.
- Cotton has long been used as a model species for studies of plant polyploidization, speciation and evolution.
- Cotton fibers are a model system for studies of cellulose biosynthesis that is crucial to bioenergy production and plant cell wall biogenesis that makes the largest portion of biomass on the earth.

1 Introduction: What we've done in this study

- We addressed the concerns of genome physical mapping of polyploids with BIBACs using Upland cotton, *Gossypium hirsutum* L.
- We developed a whole-genome BIBAC physical map for cultivated tetraploid Upland cotton, which is essential for many advanced studies of cotton genomes.
- We sorted the BIBAC contigs according to their A- or D-subgenome origin for accurately sequencing its genome.
- We sequenced over 10,000 BIBAC ends from the physical map using the Sanger sequencing method, with ~250 kb/STS.
- We identified 15,277 MTP clones from the physical map.
- We compared the subgenomes of upland cotton against to the genome of *G. raimondii* (Cotton D V1.0).

Strategy and Methodology

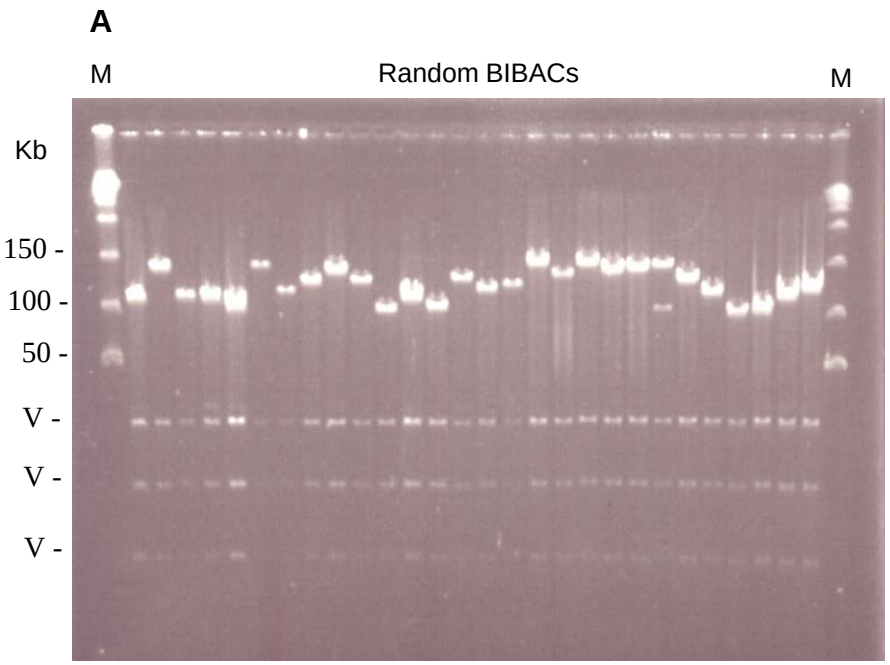
Strategy and Methods: Workflow Chart

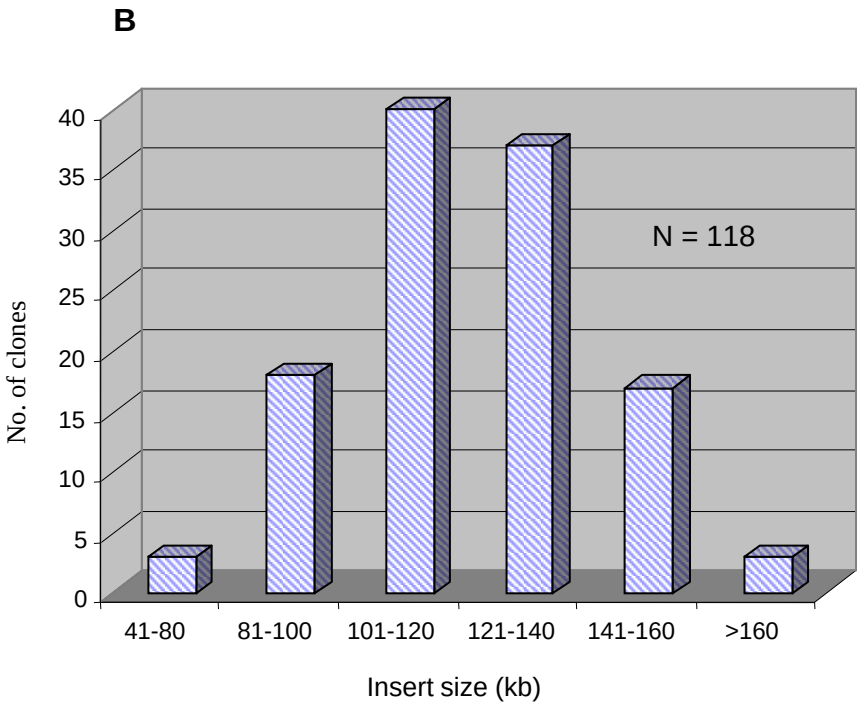


2 Strategy and Methods: Source BIBAC library

1. **Upland Cotton**: *G. hirsutum* cv. Texas Marker-1 (TM-1).
2. **BIBAC Vector:** pCLD04541 *Bam*H I site.
3. **BIBAC Library:** 76,800 clones, with an average insert size of 135 kb and providing a 4.3-fold coverage of the haploid Upland cotton genome.

Strategy and Methods: Source BIBAC library



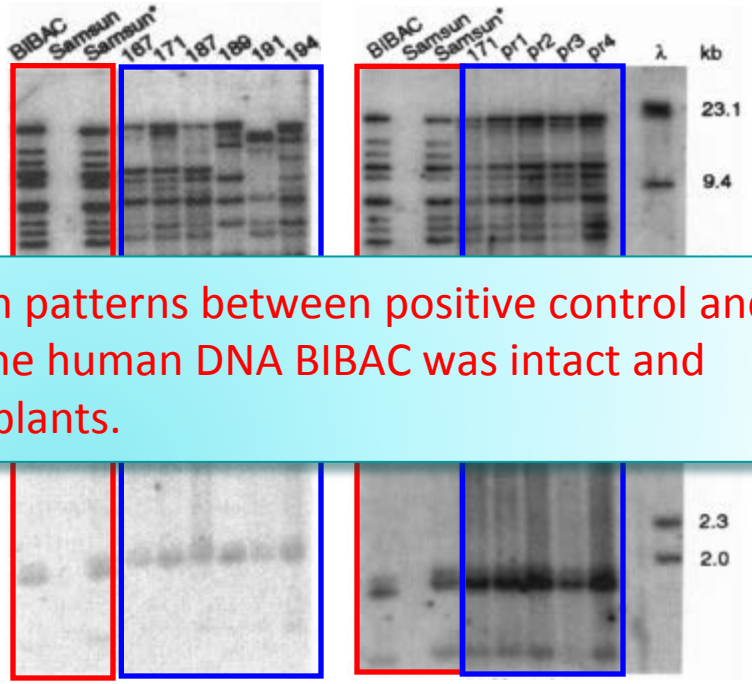


Strategy and Methods: Why BIBACs?

- ✓ **BIBACs are capable of stably cloning and maintaining foreign DNA fragments of over 300 kb in bacteria as BACs, containing 10 – 50 genes**
- ✓ **BIBACs can be directly transformed into a variety of plant species using the traditionally widely used methods: *Agrobacterium* or particle bombardment, without need of any modifications**
- ✓ **A cluster of genes involved in a pathway or a biological process can be transformed into different plants through BIBACs**
- ✓ **A target gene with its regulatory elements and related neighboring genes can be transformed into plants through BIBACs, thus minimizing the gene silencing problems that are often encountered in the traditional gene transformation method**

Strategy and Methods: Why BIBACs?

150-kb BIBAC



The similarity of the hybridization patterns between positive control and transgenic plants suggests that the human DNA BIBAC was intact and stable in the tobacco transgenic plants.

Tobacco transgenic plants
(150-kb human DNA BIBAC)

Probe: BIBAC 150-kb insert

Hamilton et al. 1996, PNAS USA 93: 9975-9979

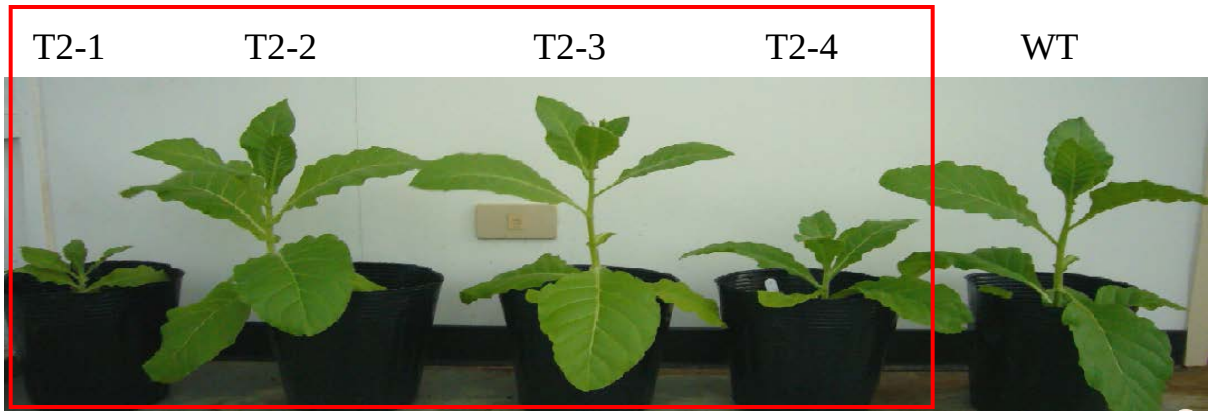
Strategy and Methods: Why BIBACs?

Arabidopsis

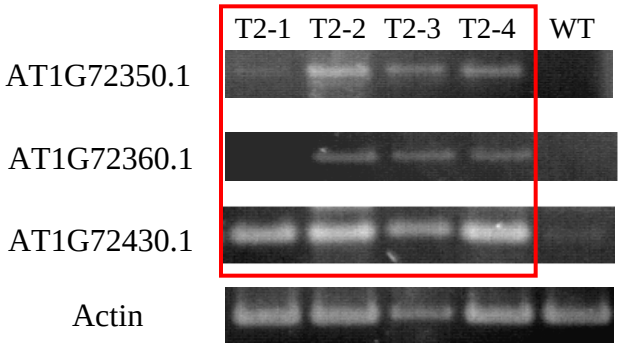
83-kb BIBAC (20 genes)

Tobacco

A: Transgenic plants

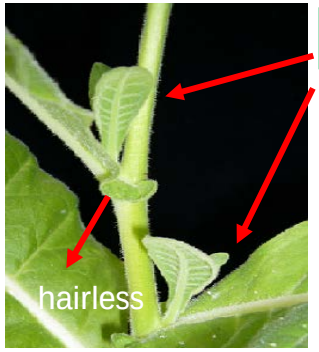


B: Gene expression in T2 plants



C: transgenic plant phenotype

Wild-type plant

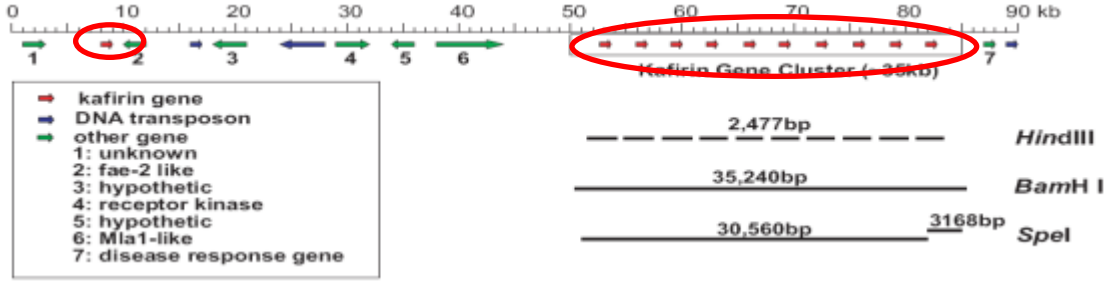


Transgenic plant

Chang et al. 2010 Genome 54:437-447

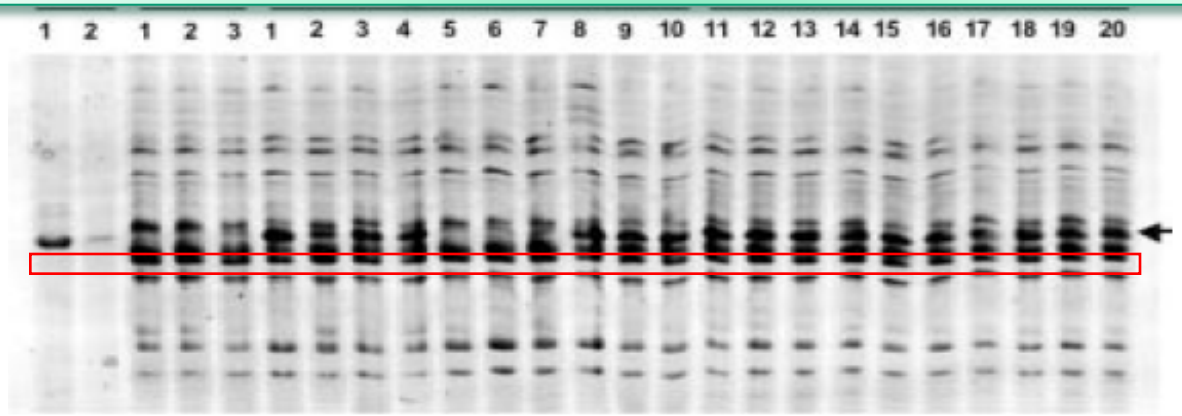
Strategy and Methods: Why BIBACs?

Sorghum SB40L16 (18 genes)



SB40L16 BAC

The genes contained in a large-insert BIBAC or BAC were actively expressed and gave new or significantly varied phenotypes, suggesting that BIBACs could be used as a tool for large-scale genetic engineering and molecular breeding.



Song et al. 2004, NAR 32:e189

Strategy and Methods: BIBAC contig assembly, BES incorporation and MTP identification

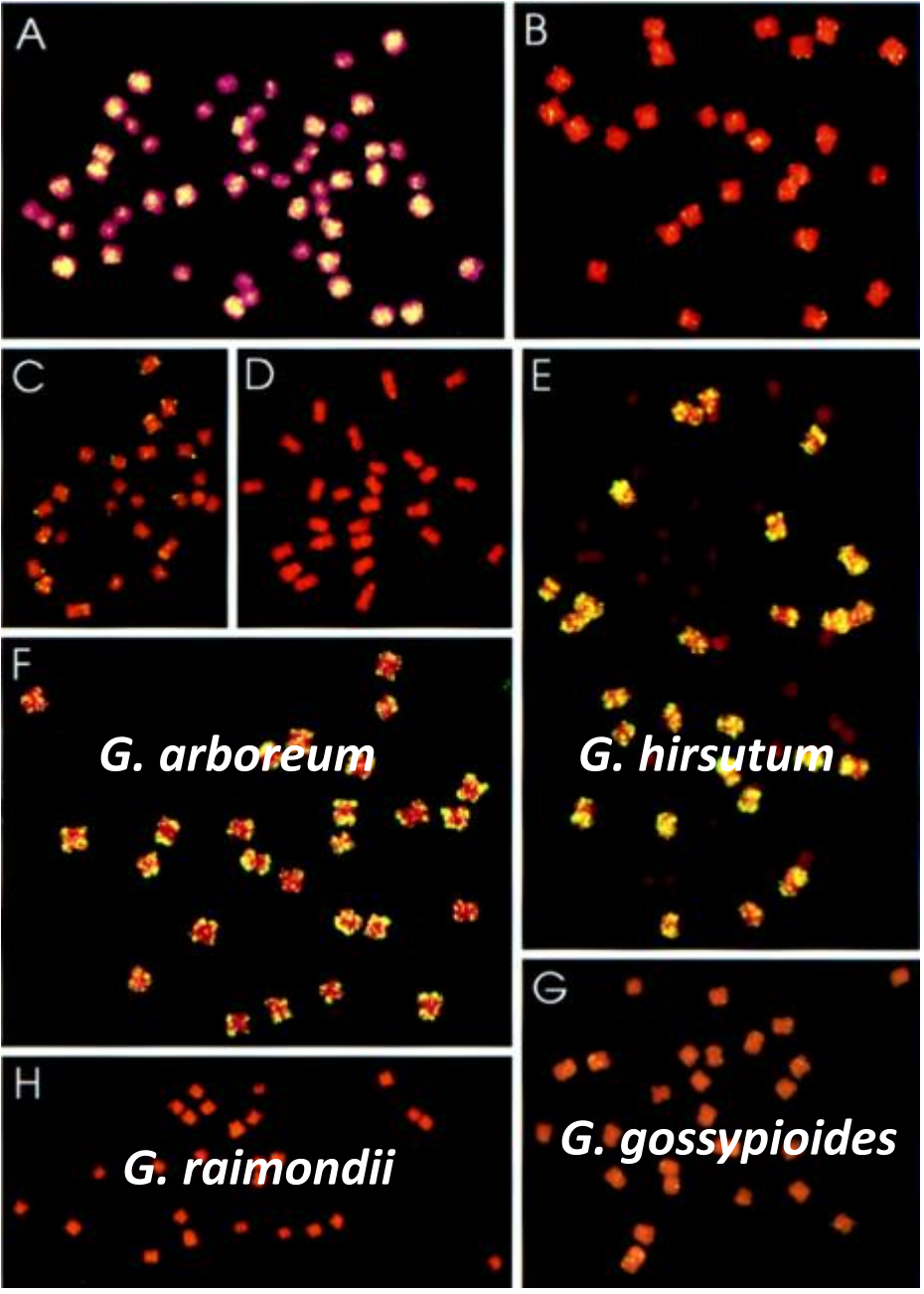
- FingerPrinted Contig (FPC) V9.3 was used to assemble the physical map contigs from the BIBAC fingerprints, with tolerance = 4 and cutoff = 1e-05.
- The BESs were generated from the physical map BIBACs using the Sanger sequencing method.
- The MTP picking function of FPC V9.3 was used to identify the MTP clones in the cotton contig physical map.

2 Strategy and Methods: Contig verification, Gene-containing contig identification

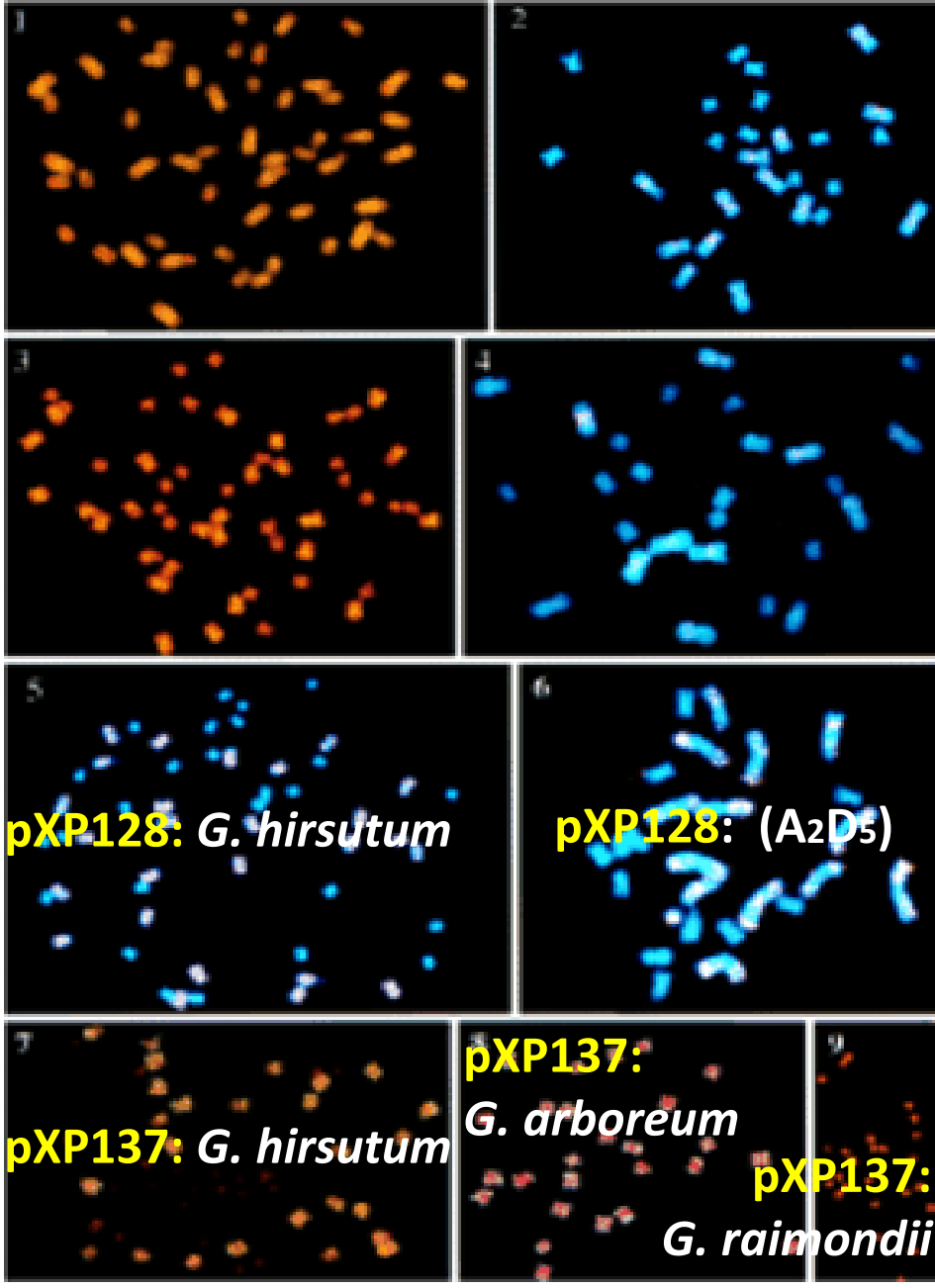
- High-density clone filters were prepared from the Upland cotton physical map BIBAC library.
- The high-density clone filters were hybridized with 13 gene-specific overgo probes to further verify the map contigs and to identify the BIBAC contigs containing the loci of genes significant to fiber development, cellulose biosynthesis, seed fatty acid metabolism, cell wall biogenesis and cotton host-nematode interaction.

Strategy and Methods: Subgenome-specific contig sorting

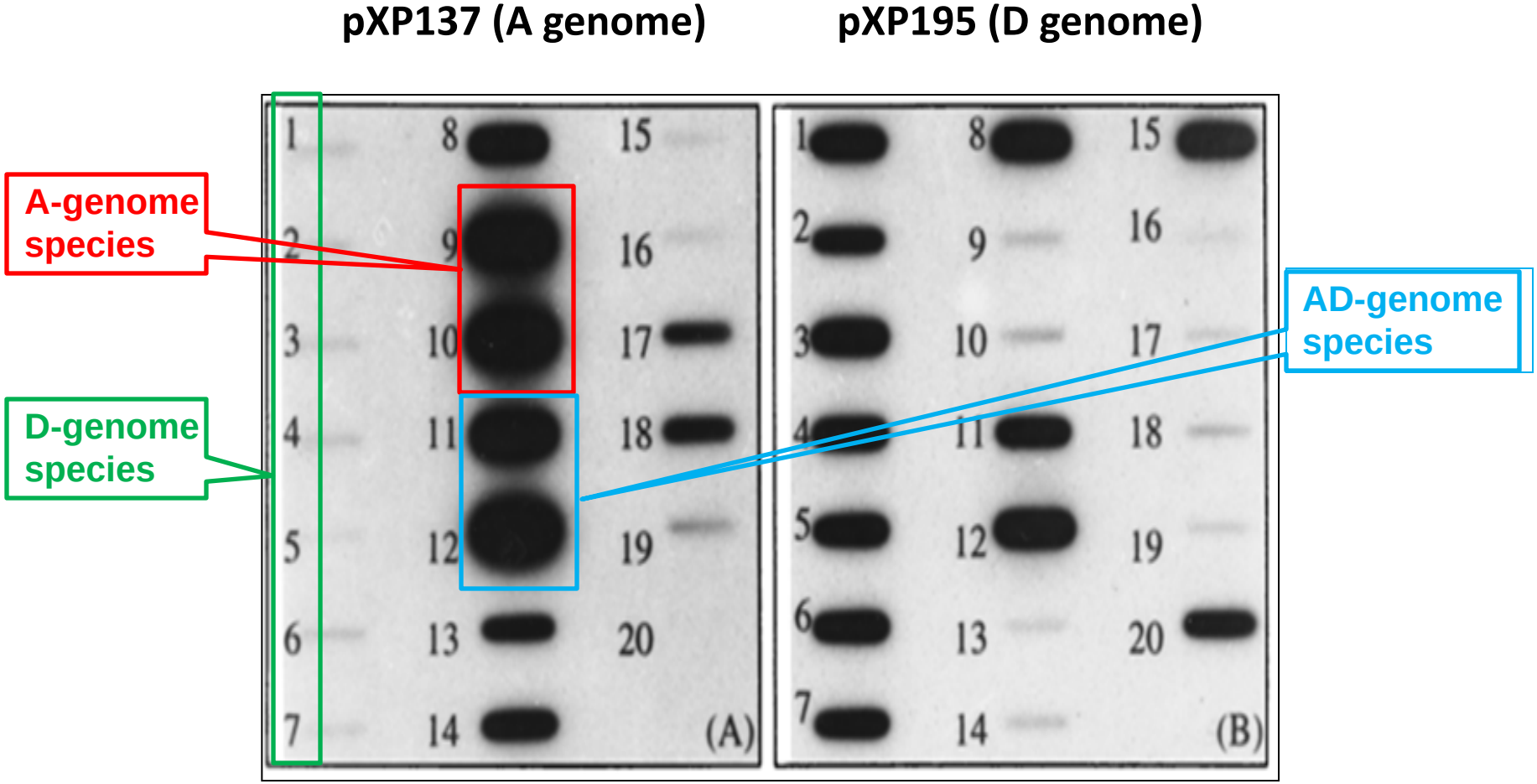
- The library filters were also hybridized with the probes made from three A-and D-subgenome-specific, interspersed repetitive elements, pXP128, pXP137 and pXP195 (Hansen et al. 1998; Zhao et al. 1998), to determine the subgenome origin of the physical map contigs and to test the feasibility of genome physical mapping of polyploids from BACs and/or BIBACs by fingerprint analysis.
- Of the three repetitive elements, pXP128 and pXP137 were A-subgenome-specific whereas pXP195 was D-subgenome-specific.



pXP137 (Zhao et al. 1998)



pXP128, pXP137 (Hansen et al. 1998)



Zhao et al. 1998

Results

Table 1. Summary of the allotetraploid Upland cotton genome physical map.

Clones fingerprinted	76,800 (4.3 x)
Clones used in the physical map construction	73,983 (4.2 x)
Singletons	9,963
Clones contained in the contigs	64,020
Contigs assembled	3,450
Contigs containing >100 clones	39
Contigs containing 50 - 99 clones	198
Contigs containing 25 – 49 clones	509
Contigs containing 10 – 24 clones	1,018
Contigs containing 4 – 9 clones	1,273
Contigs containing 3 clones	413
Average contig size	650 kb
N50 contig size	863 kb
Largest contig size	6,380 kb
Average clone number per contig	19
Average band number per clone	38.3
Physical length contribution per clone	35 kb
BESs contained in the contigs	9,711
Consensus bands of the contigs	636,530
Total physical length of the physical map	2,244 Mb
Clones in the physical map MTP	15,277
Total physical length of the MTP clones	1,955 Mb

Evaluation of physical map contigs and identification of the contigs containing genes of interest

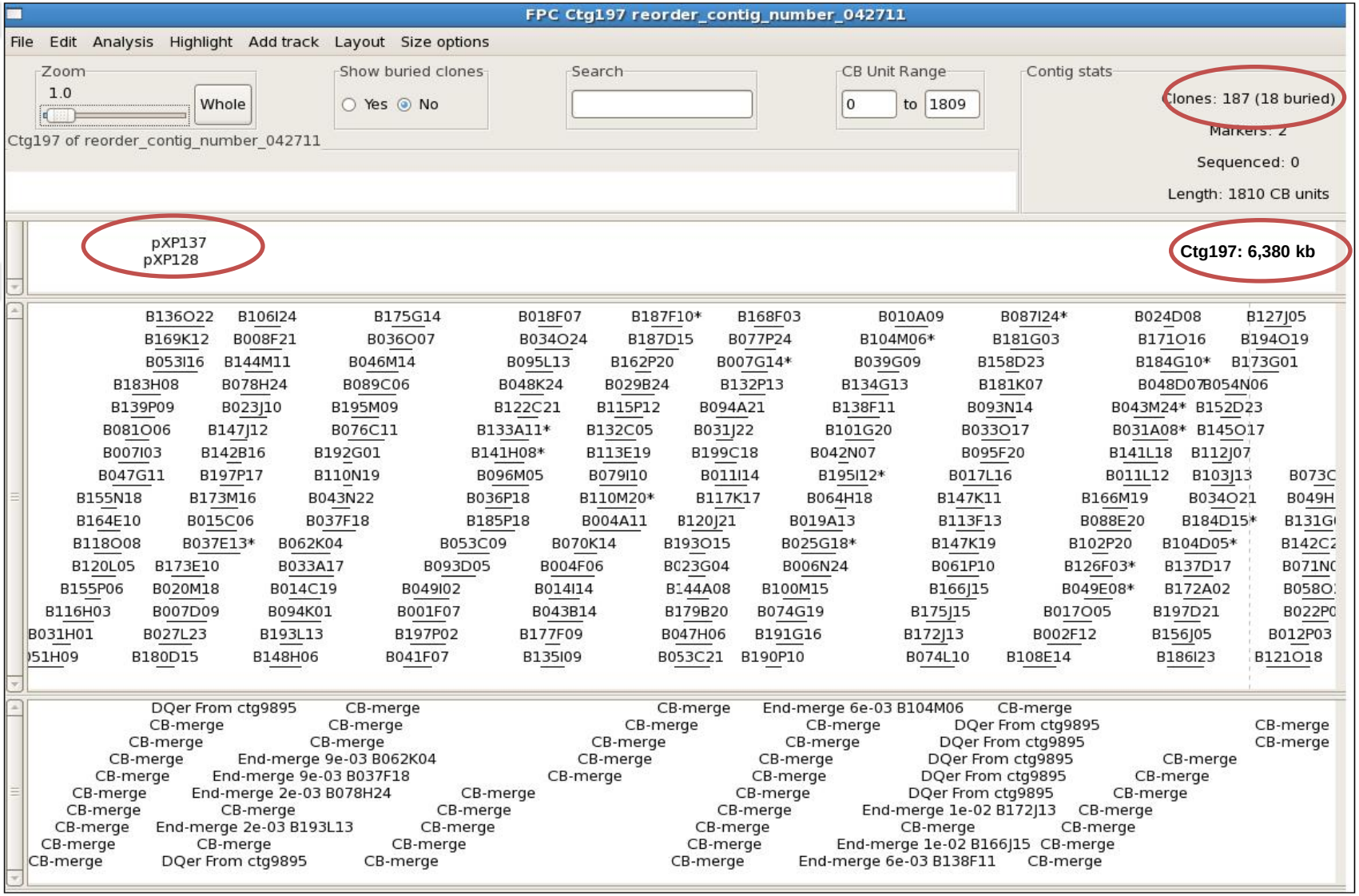
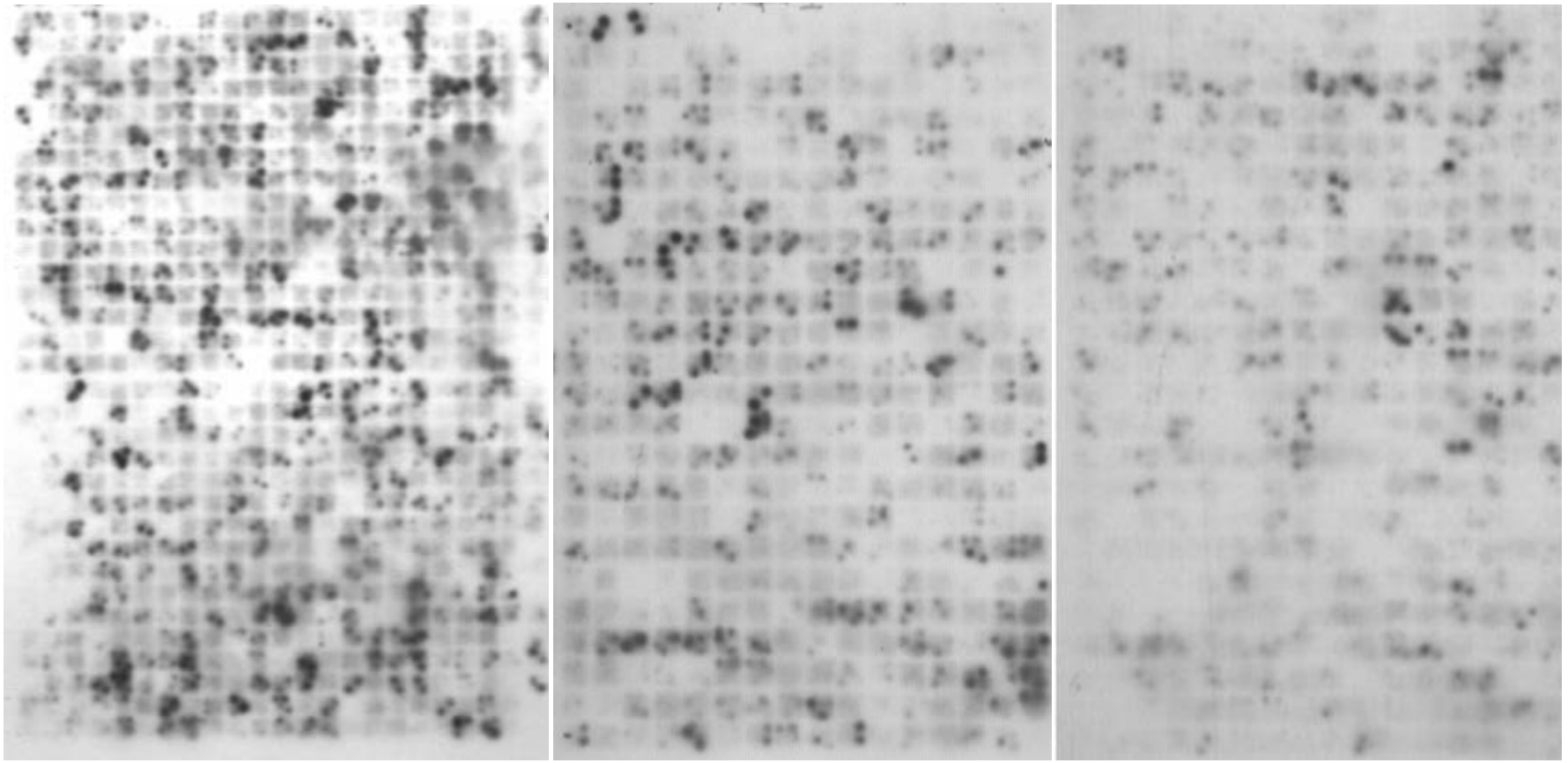


Table 2. Identification of contigs containing genes of interest and verification of the physical map. The positive clones of the genes were identified previously by screening the physical map source BIBAC library using overgo probes designed from the gene sequences (Lee et al. 2011). The clones not bolded in the table indicate singletons.

Gene	GenBank acc. No.	Annotation	Positive clones	Contig identified	Sub-genome assignment
CelA1	HQ143024.1	Cellulose synthase A1	**B130C07, B016N19**	Ctg1108 (556 kb)	Unassigned
CelA3	HQ143030.1	Cellulose synthase A3	**B024E21, B072G17, B086A07, B083I01, B073K01, B073L02, B096F08**	Ctg979 (884 kb) Ctg2187 (1,212 kb)	Unassigned A
CelA6	GQ200733.1	Cellulose synthase catalytic subunit A3	B098I23, **B116B12, B017I05,** B035E07, **B115K16,** B005M05, **B086H21**	Ctg364 (1,367 kb) Ctg2189 (800 kb)	A A
MIC3	GQ231916.1	*Meloidogyne*-induced cotton protein 3	**B099A19, B024H22, B014D05, B027E13, B026F22, B092E14**	Ctg1816(299 kb) Ctg2583 (225 kb) Ctg2716 (271 kb) Ctg2723 (874 kb)	Unassigned A Unassigned A
MIC1-15	EU025993	*Meloidogyne*-induced cotton protein 1-15	**B099A19, B024H22, B014D05, B027E13, B026F22, B092E14**	Ctg1816 (299 kb) Ctg2583 (225 kb) Ctg2716 (271 kb) Ctg2723 (874 kb)	Unassigned A Unassigned A
RDL1	AY633558.1	GaRDL1 gene, promoter region	B175F03, B050B20, **B187C03, B186M07, B016L07, B166C05, B161G09**	Ctg1397 (831 kb)	A
FADO6	Y10112.2	Fatty acid desaturase omega-6	**B138P05, B080A19**	Ctg1156 (2,079 kb)	A
MYBB	AF034130.1	MYB-like DNA-binding domain protein	**B174C01, B192C23**	Ctg1137 (768 kb)	Unassigned
MYBT2	AY366352.1	MYB-like transcription factor 2	**B026F22, B007F03**	Ctg2723 (874 kb) Ctg3247 (831 kb)	A A
GhCesA2	U58284.1	Secondary wall cellulose synthase A2	**B048N17, B085P21, B162F12,** B046G15	Ctg3423 (1,614 kb)	A, D
GhIRX3	DT048689	Irregular xylem 3/cellulose synthase A7	**B070C20**, B173F02	Ctg1090 (2,671 kb)	D
GhCesA3	AF150630.2	Primary wall cellulose synthase	**B075L23, B097O20, B108L10, B146H15**, **B178M15,** B170C06, B009G10, B065P05	Ctg258 (331 kb)	Unassigned
GhCes	AF150630	Unknown cellulose synthase	**B145L18** **B164E04, B165A23,** B008D22	Ctg258 (331 kb) Ctg2187 (1,212 kb)	Unassigned A

Overgo Hybridization Results



pXP128 (A genome) pXP137 (A genome) pXP195 (D genome)

Results: Identification of A- and D-subgenome Physical Maps

Table 3. Summary of the subgenome physical maps.

Probe	Origin of subgenome	No. of positive clones	No. of contigs	Physical Length (Mb)	Genome coverage
pXP128 pXP137	A genome-specific	2,056 1,148	1,063	1,095	49%
pXP195	D genome-specific	523	141	97	4%
	Contamination		7	6	0.27%
		3,727	1,211 (34.9%)		53.27%

3 Results: Identification of A- and D-subgenome Physical Maps

- These results suggest that the physical map contigs were assembled according to their origin of A- or D-subgenome; thus, A- and D-subgenome physical maps have been developed, separately.
- In this study, only three subgenome-specific probes were used to separate the subgenomes, but, surprisingly, over 50% of the physical map has been sorted successfully according to their origin of subgenome. Therefore, the tetraploid Upland cotton physical map could be sorted completely using this strategy if more subgenome-specific probes were used.
- These results provide a line of evidence and a strategy for genome physical mapping of most, if not all, allopolyploids with BACs and/or BIBACs such as wheat, canola and tobacco.

3 Results: Identification of MTP Clones of the Physical Map

We identified a total of **15,277 MTP** clones of the cotton physical map. These MTP clones collectively spanned **1,955 Mb**, approximately 81.5% of the Upland cotton genome.

3 Results: Blat Subgenome-specific BES against *G. raimondii* Genome

Local Blat analysis was carried out:

Query:

AD1-genome BES: 9,711 BESs
(mean = 400 bp)
A-subgenome specific BES: 418 BESs
D-subgenome specific BES: 184 BESs

Database:

G. raimondii (Cotton D V1.0)

Results: Blat Subgenome-specific BES against *G. raimondii* Genome

Table 5 Comparative sequence analysis of Upland cotton BESs with the D genome sequence of *G. raimondii* (DOE Joint Genome Institute, Cotton D V1.0, http://www.phytozome.net/cotton.php)

Inquired Upland cotton BESs	Criteria of the analysis	Sequence similarity					
All AD-subgenome BESs (No. of BESs = 9,711; total sequence length = 3,842,009 bp)	Continuous MinMatch (bp)	100	100	100	100	100	100
	Sequence identity (%)	100	95	90	80	70	60
	No. of BESs aligned	57	4,587	6,588	7,118	7,257	7,277
	% of BESs aligned	0.59	47.24	67.84	73.30	74.73	74.94
	Total sequence length of BESs aligned to the raimondii sequence (bp)	15,539	1,367,385	1,960,413	2,048,311	1,996,820	1,975,511
	% of sequence length of BESs aligned to the raimondii sequence	0.41	35.59	51.10	53.31	51.95	51.42
A-subgenome-specific BESs (No. of BESs = 418; total sequence length = 169,533 bp)	Continuous MinMatch (bp)	100	100	100	100	100	100
	Sequence identity (%)	100	95	90	80	70	60
	No. of BESs aligned	7	205	294	310	317	317
	% of BESs aligned	1.67	49.04	70.33	74.16	75.84	75.84
D-subgenome-specific BESs (No. of BESs = 184; total sequence length = 71,410 bp)	Continuous MinMatch (bp)	100	100	100	100	100	100
	Sequence identity (%)	100	95	90	80	70	60
	No. of BESs aligned	1	88	128	136	139	139
	% of BESs aligned	0.54	47.83	69.57	73.91	75.54	75.54

Results: Blat Subgenome-specific BES against *G. raimondii* Genome

Table 1 DNA sequence correspondences (RSCs) between diploid species/lineage and five polyploid species.

Genome/lineage	TB	GMB	$(AD)_1$ GMB	$(AD)_1$ RSC	$(AD)_2$ GMB	$(AD)_2$ RSC	$(AD)_3$ GMB	$(AD)_3$ RSC	$(AD)_4$ GMB	$(AD)_4$ RSC	$(AD)_5$ GMB	$(AD)_5$ RSC
D_4	1,033	3	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
D_5	1,033	34	4	0.12	5	0.15	6	0.18	3	0.09	4	0.12
D_7	**1,033**	**19**	**14**	**0.74**	**12**	**0.63**	**11**	**0.58**	**14**	**0.74**	**12**	**0.63**
D_8	1,033	6	0	0.00	0	0.00	1	0.17	0	0.00	0	0.00
All D genomes	**1,033**	**6**	**3**	**0.50**	**3**	**0.50**	**4**	**0.67**	**5**	**0.83**	**2**	**0.33**
A_1	490	23	8	0.35	7	0.30	8	0.35	8	0.35	5	0.22
A_2	490	17	1	0.06	3	0.18	3	0.17	2	0.12	5	0.29
E	490	3	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
F	490	4	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
A_1+A_2	**490**	**104**	**61**	**0.59**	**60**	**0.58**	**65**	**0.63**	**62**	**0.60**	**56**	**0.54**
A_1+A_2+F	**490**	**5**	**4**	**0.80**	**4**	**0.80**	**4**	**0.80**	**5**	**1.00**	**5**	**1.00**
A_1+A_2+F+E+B	**490**	**9**	**8**	**0.89**	**8**	**0.89**	**9**	**1.00**	**8**	**0.89**	**8**	**0.89**
C+G+K	490	6	1	0.17	1	0.17	0	0.00	0	0.00	0	0.00
A_1+A_2+F+E+B+ C+G+K	**490**	**6**	**5**	**0.83**	**5**	**0.83**	**5**	**0.83**	**5**	**0.83**	**6**	**1.00**

TB, total number of bands examined; GMB, genome-specific marker bands identified from the bands studied. For the genome/lineage with <3 GMBs are not listed. RSC = the number of GMBs of a diploid presented in a polyploid divided by the total number of GMBs of the diploid. Therefore, the RSC value reflects the relationship between a genome of the polyploid and the genome of the diploid and has a range from 0.00 to 1.00. If none of the GMBs of a diploid is encountered in a polyploid or it is not statistically different from zero ($P > 0.05$), the RSC is 0.00, suggetsing that it is unlikely that the polyploid contains a genome from the diploid. However, if all GMBs of a diploid are encountered in a polyploid, the RSC is 1.00, suggesting that one of more of the genomes of the polyploid likely originated from the diploid.

Results: Blat Subgenome-specific BES against *G. raimondii* Genome

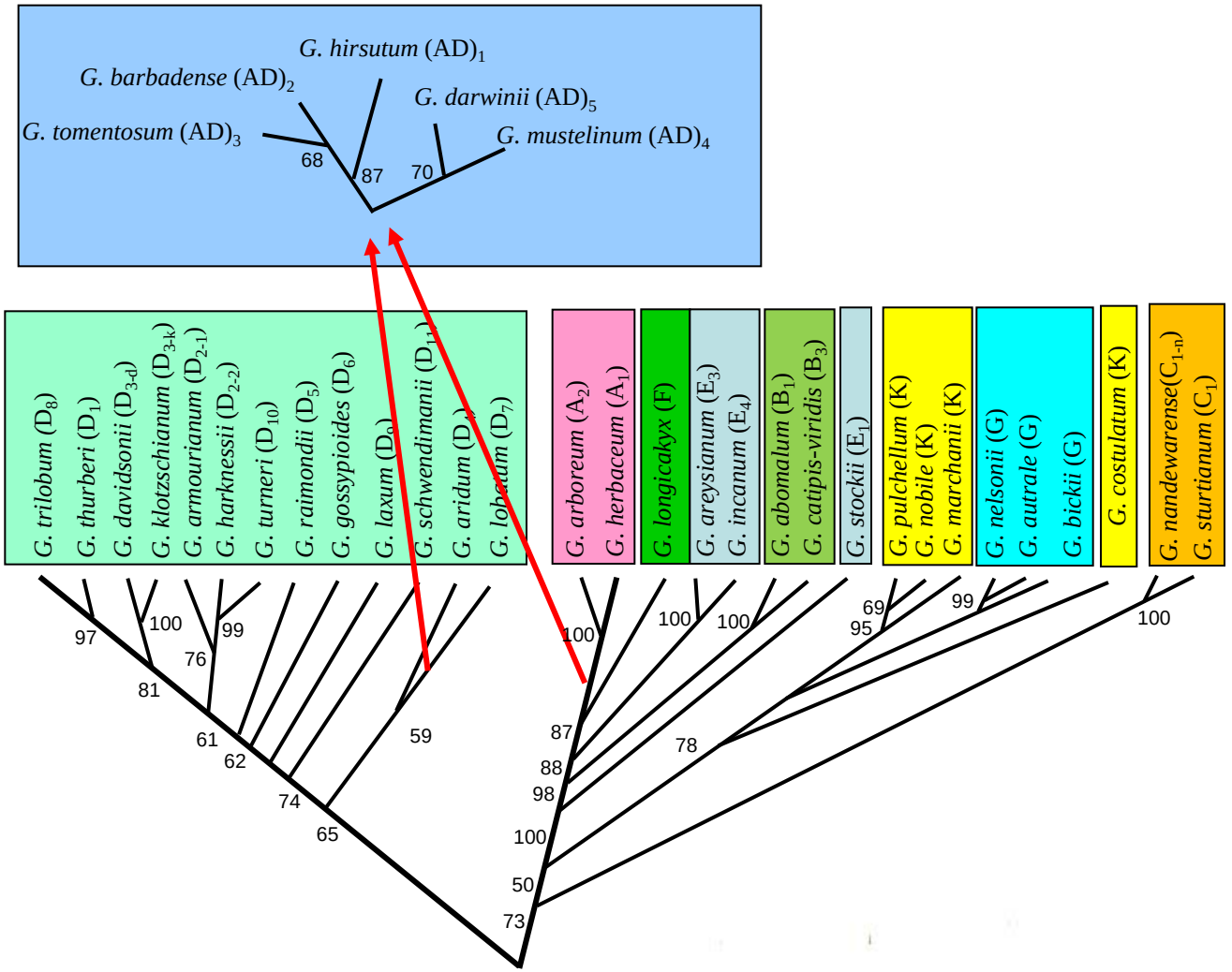


Fig. 3. Phylogeny and evolution of *Gossypium* species. The number below each branch is the percentage of confidence calculated by bootstrap computation.

Conclusions

4 Conclusions

- We have developed a BIBAC physical map of the tetraploid Upland cotton, *G. hirsutum*, cv. TM-1. It consists of 3,450 BIBAC contigs, with an N50 contig size of 863 kb and covering a total length of 2,244 Mb.
- We assembled physical maps for the A- and D-subgenomes of the Upland cotton, separately, isolated A- or D-subgenome-specific contigs spanning 1,192 Mb, sequenced nearly 10,000 BIBAC ends, and identified MTP clones spanning 1,955 Mb, thus providing a platform essential for sequencing the Upland cotton genome using the next-gen sequencing technology.
- The results of this study demonstrated that it is feasible to develop the physical maps of allopolyploids by BAC fingerprinting and contig assembly, thus significantly promoting the genome physical mapping and sequencing of polyploid species.

5

Future Plan

Key words:
Integrated physical and genetic maps, individual Chromosome

5 Future Plan: Individual Chromosome Maps for Accurately Sequencing

In our lab at Texas A&M University

- HiSeq2000 for 198 Recombinant Inbred Lines (RIL) derived from intercross between *G. hirsutum* and *G. Barbadense*.
- Just got the first bunch of Seq. Data back.
- Analyzing: CLC Genomics Workbench.

Publicly available genetic information (SSR, SNP, RFLP…)

- CottonGen
- The Cotton Marker Database
- Cotton Genome Database

5 Future Plan: Individual Chromosome Maps for Accurately Sequencing

Table 3 Summary properties of comparative BAC contig map and rearrangements

MGA chromosome	Number of contigs	GGA orthologue	Total contig length (bp)	Average contig length (bp)	Number of major inversions
MGA1	9	GGA1	198,141,542	22,015,727	0
MGA2	1	GGA3	111,826,550	111,826,550	2
MGA3	1	GGA2q	100,502,741	100,502,741	1
MGA4	3	GGA4q	72,769,490	24,256,497	0
MGA5	2	GGA5	60,352,001	30,176,001	2
MGA6	1	GGA2p	51,981,596	51,981,596	0
MGA7	5	GGA7	36,281,918	7,256,384	1
MGA8	2	GGA6	34,867,379	17,433,690	4
MGA9	1	GGA4p	19,188,313	19,188,313	0
MGA10	2	GGA8	29,163,447	14,581,724	2
MGA11	2	GGA9	23,514,388	11,757,194	1
MGA12	1	GGA10	22,321,123	22,321,123	0-2*
MGA13	1	GGA11	21,290,332	21,290,332	1
MGA14	2	GGA12	19,602,960	9,801,480	1-4
MGA15	2	GGA13	17,856,723	8,928,362	2
MGA16	1	GGA14	15,894,242	15,894,242	1
MGA17	1	GGA15	12,928,248	12,928,248	0
MGA18	1	GGA16	68,068	68,068	ND
MGA19	1	GGA17	10,577,421	10,577,421	0
MGA20	1	GGA18	10,507,821	10,507,821	1
MGA21	1	GGA19	9,897,437	9,897,437	0
MGA22	3	GGA20	13,624,242	4,541,414	0
MGA23	1	GGA21	6,854,714	6,854,714	0
MGA24	3	GGA22	3,657,921	1,219,307	0
MGA25	2	GGA23	5,832,856	2,916,428	0-2*
MGA26	1	GGA24	6,430,646	6,430,646	0
MGA27	5	GGA25	2,143,571	428,714	ND
MGA28	2	GGA26	4,831,899	2,415,950	0
MGA29	3	GGA27	4,641,426	1,547,142	0
MGA30	2	GGA28	4,439,785	2,219,893	ND
MGAZ	11	GGAZ	72,157,792	6,559,799	1**
TOTALS	74		1,004,148,592	13,569,576	20-27

ND indicates that the comparative map and/or chicken sequence assembly are incomplete

*Pattern can be explained by two sequential inversions or centromere replacement

**There may be additional small rearrangements

An integrated physical and genetic map of turkey (Yang et al. 2011, BMC Genomics 12:447)

5 Future Plan: Individual Chromosome Maps for Accurately Sequencing

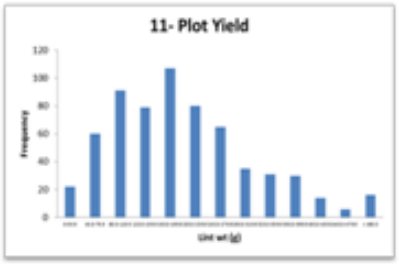


Fig. 1. Plot fiber yield obtained in 2011. The field trial was conducted in completely randomized design with three replicates in the TAMU field experimental station, College Station, Texas. Similar results were obtained in 2009 and 2010.

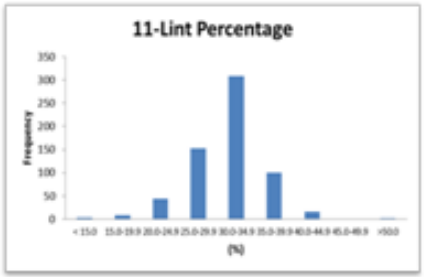


Fig. 2. Lint percentage obtained in 2011. The field trial was conducted in completely randomized design with three replicates in the TAMU field experimental station, College Station, Texas. Similar results were obtained in 2009 and 2010.

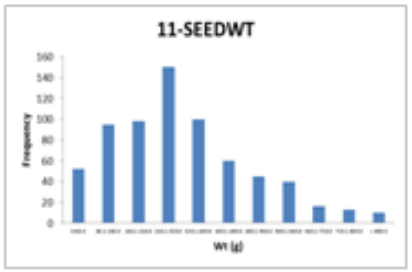


Fig. 3. Plot seed weight (g) obtained in 2011. The field trial was conducted in completely randomized design with three replicates in the TAMU field experimental station, College Station, Texas. Similar results were obtained in 2009 and 2010.

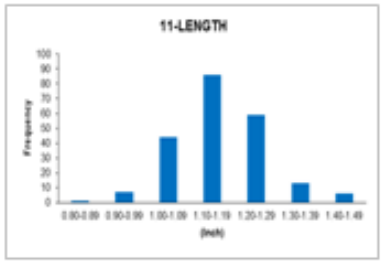


Fig. 4. HVI Fiber length obtained in 2011. The field trial was conducted in completely randomized design with three replicates in the TAMU field experimental station, College Station, Texas. Similar results were obtained in 2009 and 2010.

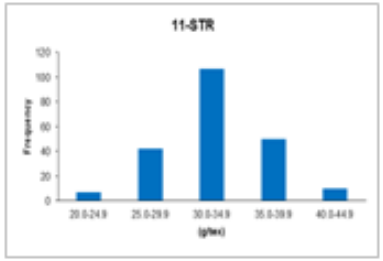


Fig. 5. HVI Fiber strength obtained in 2011. The field trial was conducted in completely randomized design with three replicates in the TAMU field experimental station, College Station, Texas. Similar results were obtained in 2009 and 2010.

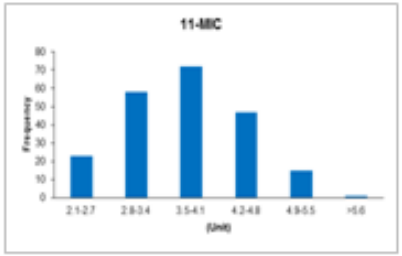


Fig. 6 HVI Fiber micronaire obtained in 2011. The field trial was conducted in completely randomized design with three replicates in the TAMU field experimental station, College Station, Texas. Similar results were obtained in 2009 and 2010.

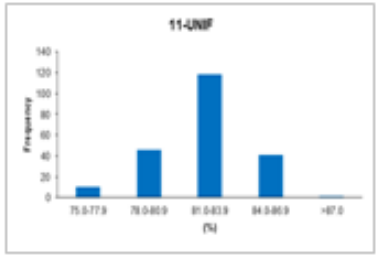


Fig. 7. HVI Fiber uniformity obtained in 2011. The field trial was conducted in completely randomized design with three replicates in the TAMU field experimental station, College Station, Texas. Similar results were obtained in 2009 and 2010.

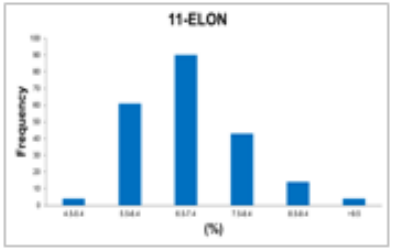


Fig. 8. HVI Fiber elongation obtained in 2011. The field trial was conducted in completely randomized design with three replicates in the TAMU field experimental station, College Station, Texas. Similar results were obtained in 2009 and 2010.

Acknowledgements: This study was supported by the grants from the Texas AgriLife Research Cotton Improvement Program.

AgriLIFE RESEARCH & EXTENSION
Texas A&M System

Thank you!