

Development of high-density 80 K SNP array and its evaluation and application in *G. hirsutum*

Wangzhen Guo

Nanjing Agricultural University





Outline

- **Research background**
- **Development and characteristics of the CottonSNP80K array**
- **Potential utilization**

1. Research background

Brief history for molecular markers

1st: molecular hybridization: **RFLP**

2nd: PCR-based technology:

a. PCR technology: RAPD, ISSR, **SSR** et al.

b. PCR+RE: AFLP, CAPs et al.

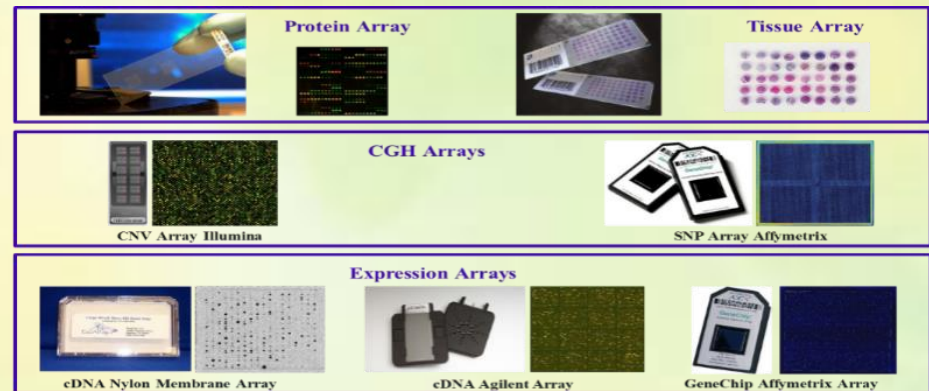
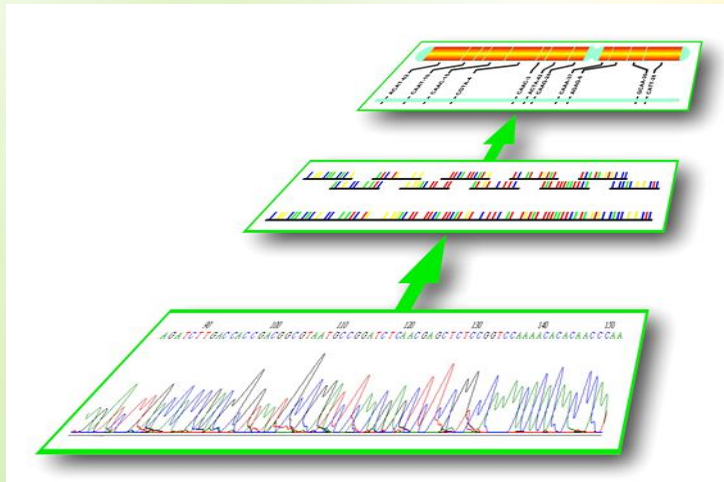
3rd: **SNPs**

Molecular markers have been successfully used in crops genetics and breeding

SNPs are the most abundant DNA sequence variation present in plant genomes

SSR: low-throughput, labor-intensive, time-consuming

SNP: high-throughput, virtually unlimited, evenly distributed along the genome, bi-allelic , co-dominant



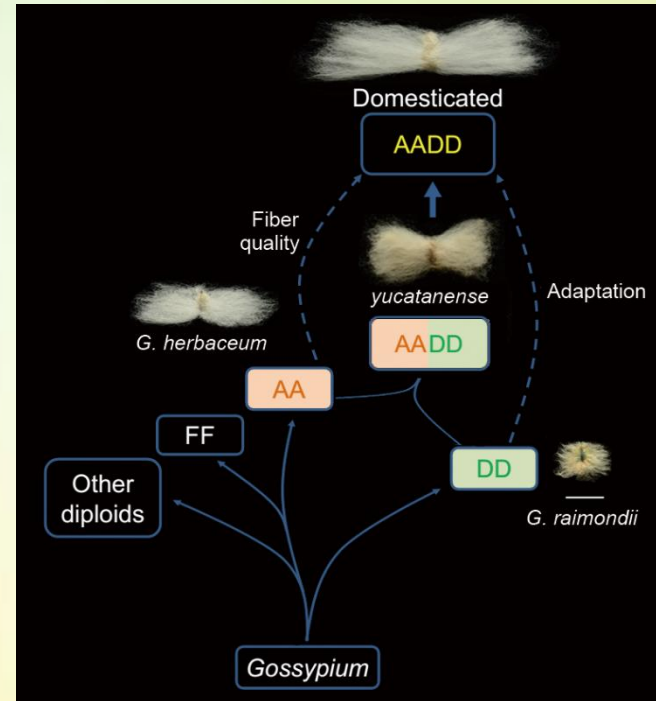
Cotton is the world's most important natural textile fiber and a significant oilseed crop.



Gossypium hirsutum acc. TM-1

Upland cotton has the highest yield, over 95% worldwide cotton production.

Upland cotton is a model for polyploidy crop domestication and transgenic improvement



Modern upland cotton cultivars have narrow genetic diversity, especially in China, their development is from the limited quantity of resources in the United States.

The whole-genome sequencing of four different cotton species has made great progress

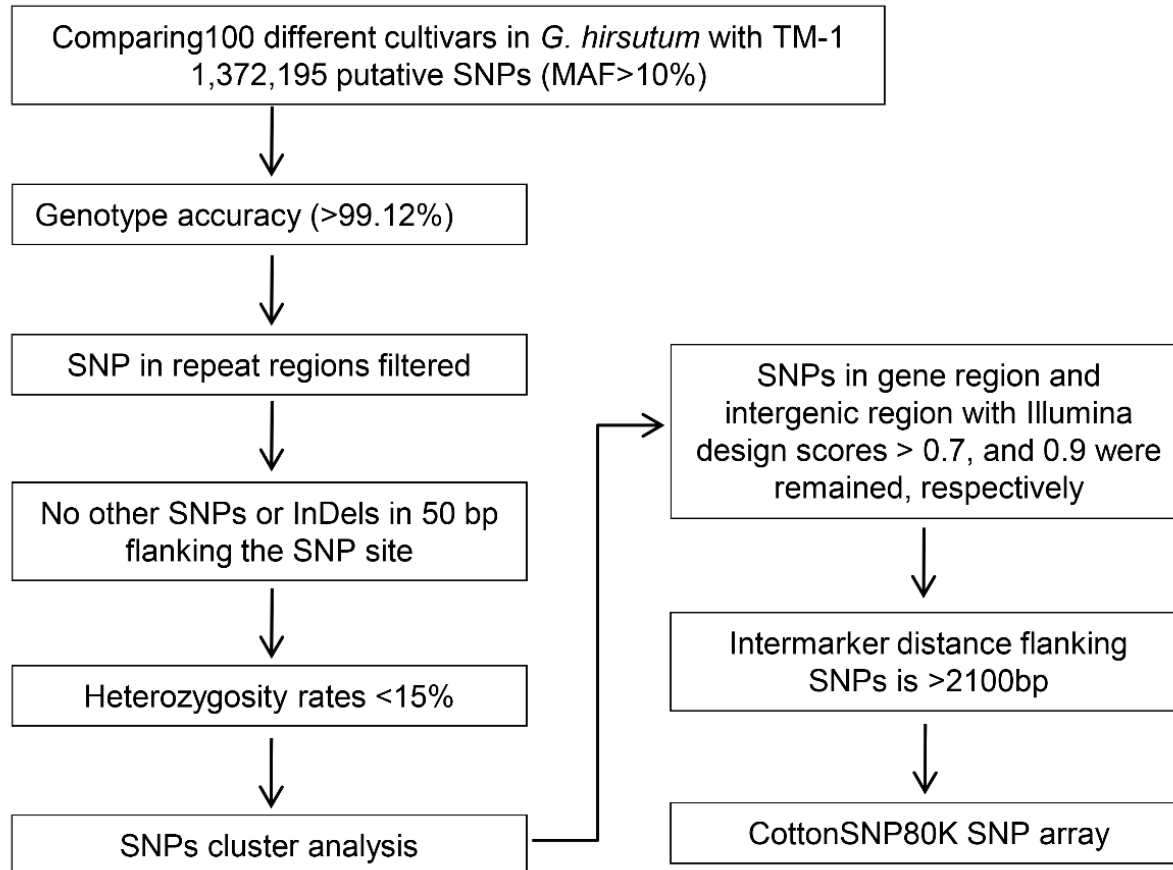
Species	Origin	Assembled size (Mb)	Scaffold N50	Contig N50 (kb)	Genes annotated	Transposable elements (%)	Sequencing technology	Ref
<i>G. raimondii</i>	CAAS	775	2.3 Mb	45	40,976	57	Illumina	Wang et al., 2012 NG
<i>G. raimondii</i>	JGI	761	18.8 Mb	136	37,505	61	Illumina,454, Sanger	Paterson et al., 2012 NATURE
<i>G. arboreum</i> SXY1	CAAS	1,694	666 kb	72	41,330	68.5	Illumina	Li et al., 2014 NG
<i>G. hirsutum</i> TM-1	NAU	2,432	1.6 Mb	34	70,478	64.8	Illumina, Sanger	Zhang et al., 2015 NB
<i>G. hirsutum</i> TM-1	CAAS	2,173	764 kb	80	76,943	66	Illumina	Li et al., 2015 NB
<i>G. barbadense</i> XH21	Esquima	2,171	503 kb	72	77,526	65.1	Illumina,454, Pacbio	Liu et al., 2015 SR
<i>G. barbadense</i> 3-79	HAU	2,573	260 kb	24	80,876	69.1	Illumina	Yuan et al., 2015 SR

The availability of genomic re-sequencing data from different upland cotton accessions made it possible to develop SNP markers at a genome-wide level.

- **321 cultivated accessions 6.9X**
Wang et al. Nat Genetics 2017
- **318 landraces and modern improved cultivars/lines 5X**
Fang et al. Nat Genetics 2017
- **419 core collection of upland cotton accessions 6.55X**
Ma et al. Nat Genetics 2018

Compared to re-sequencing analysis, SNP arrays, with loci known and addressable, can produce large-scale genotyping data through one hybridization procedure at a relatively low cost.

2. Development and characteristics of the CottonSNP80K array



Selection criteria

genome-wide

MAF

LD

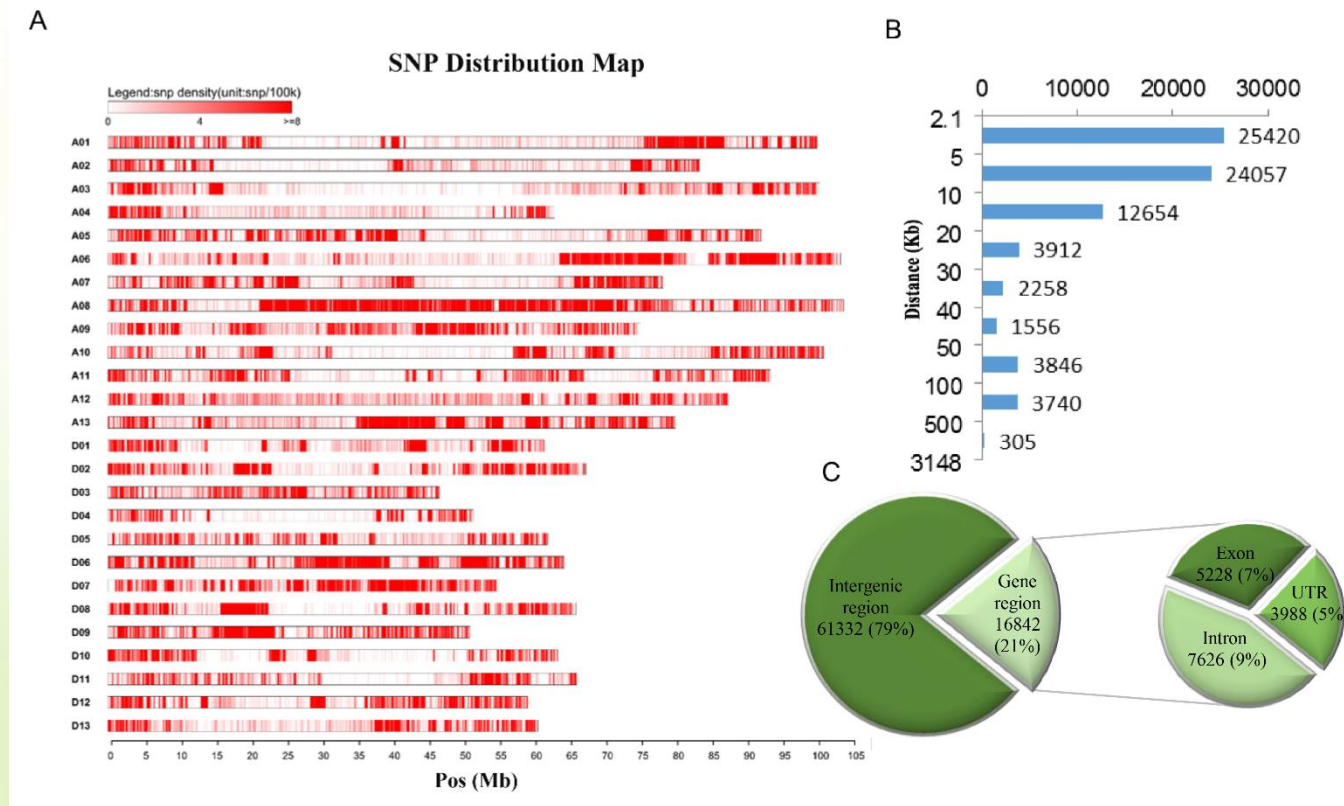
No repeat regions

Inclined in gene regions

Distinguishable homologous loci

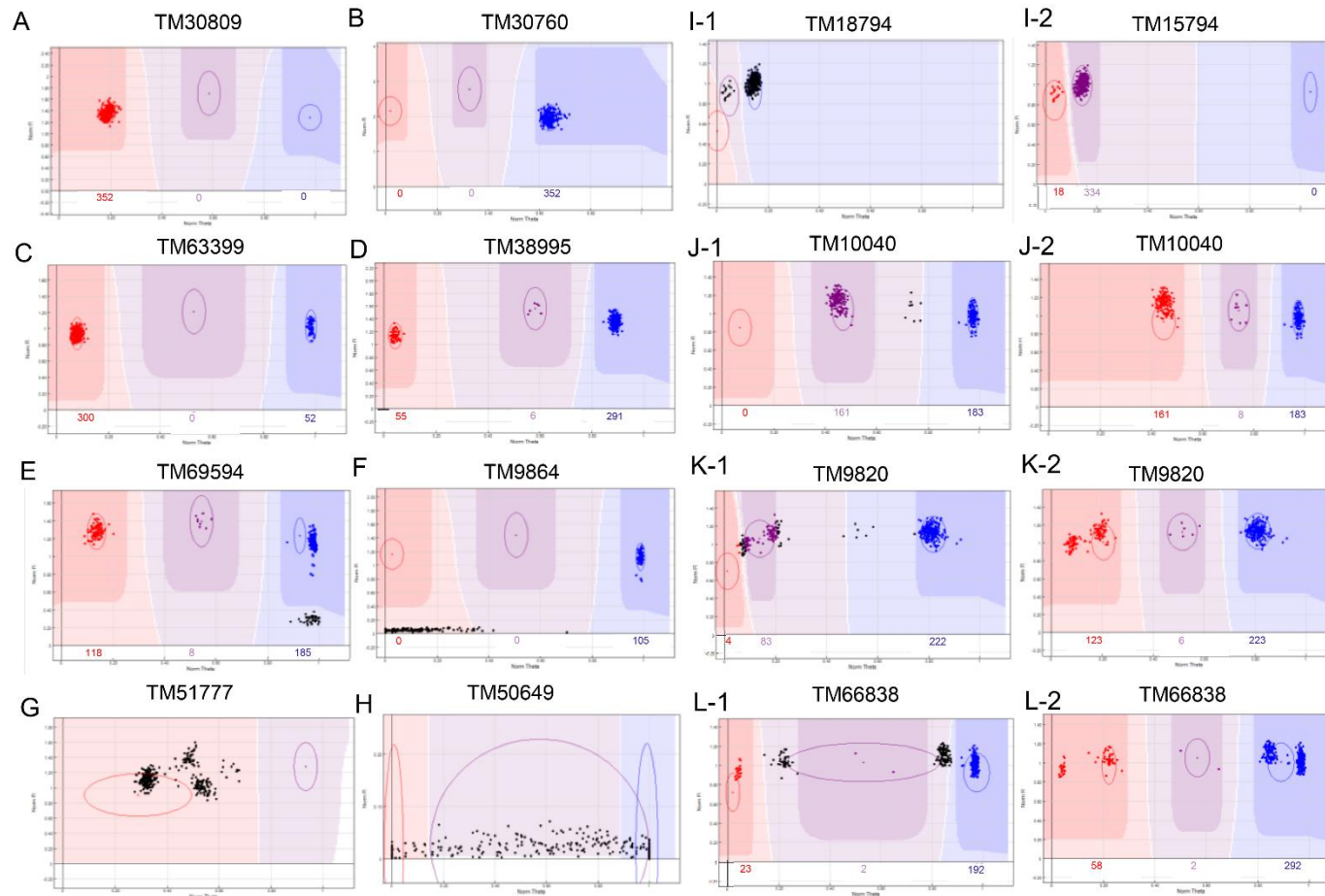
A high-density SNP array for high-throughput intraspecific upland cotton genotyping identification.

Characteristics of the CottonSNP80K array



- **Totally, 77,774 loci were successfully synthesized on the array.**
- **The average distance between adjacent SNPs was 24.9 Kb with 45,183 (58.10%) in the At- subgenomes and 32,591 loci (41.90%) in the Dt- subgenomes.**
- **16,642 SNPs (21.40%) were tagged in the genic region of 9,902 genes, and 61,132 SNPs (78.60%) in intergenic regions.**

SNP genotype calling in allotetraploid cotton



With 352 tested accessions, 77,252 loci (99.33%) had high call rates >95%. Of them, 59,502 SNPs (76.51%) showed **polymorphic loci**, with MAFs>0.05 for 57,071 loci (95.91%) and MAFs>0.1 for 48,940 (82.25%), respectively.

CottonSNP80K array had high genotyping accuracy, good repeatability, and distinguishability

	NGS & SNP Chip
TM-1 (62X)	95.40%
Hai7124 (39X)	97.19%

The SNPs of TM-1 and Hai7124 by comparing their array detection with the re-sequencing results.

	SNP Chip
W0 vs W0	100%
TM-1 vs TM-1	99.99%

Three technical replicates for W0
Two biological replicates for TM-1

	SSR	SNP Chip
XXXJ vs XXXJ mutant		16.11%
XZ142 vs XZ142 mutant		19.04%
7235 vs 7235 mutant	2.17% (Lu QX, 2004)	25.01%
imim vs TM-1	1.28%(Wang C,2012)	21.00%
TM-1 vs SL1-7-1		26.82%
TM-1 vs MD17		30.62%
TM-1 vs N1		22.73%
TM-1 vs n2		25.71%

Several mutants and their corresponding donors with similar genetic backgrounds

3. Potential utilization of the CottonSNP80K array

Fingerprint of varieties

Genome-wide association studies (GWAS)

Genetic diversity analysis

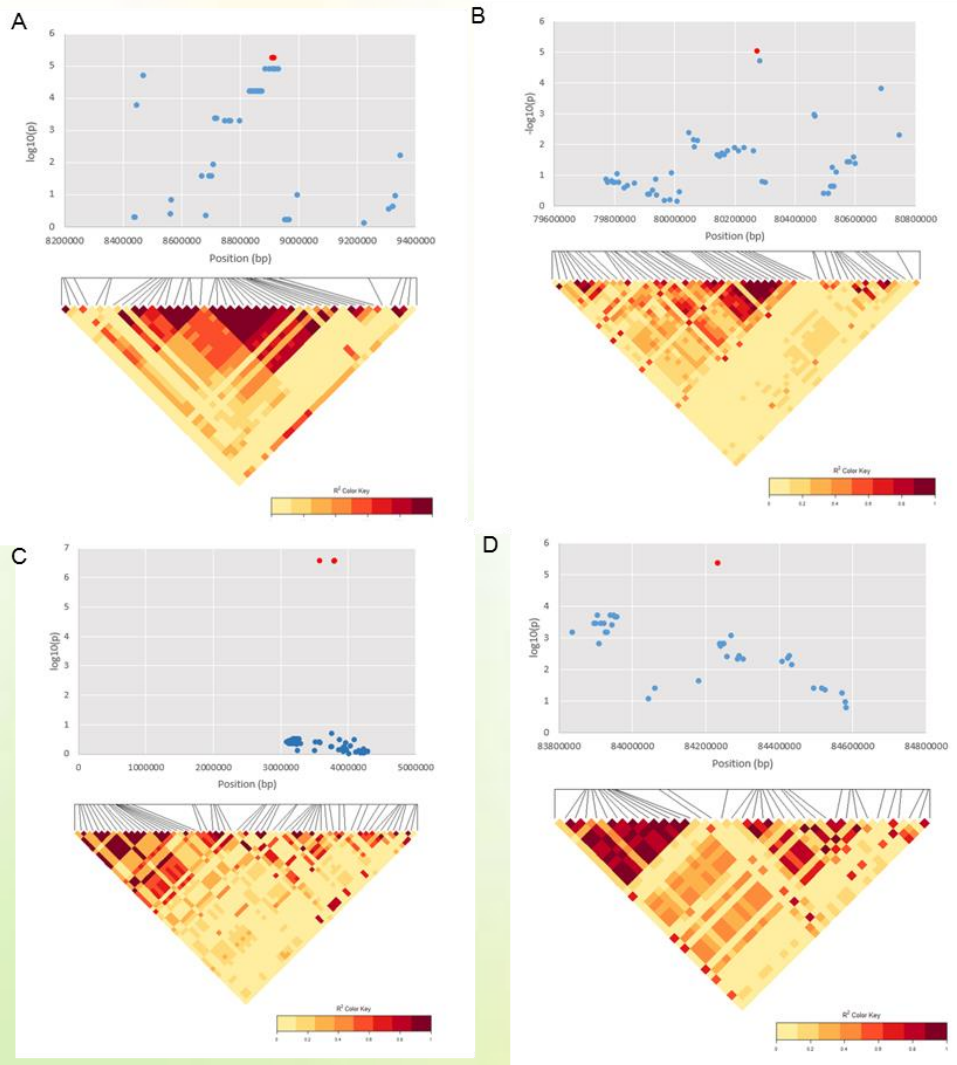
High-density genetic mapping

Quantitative trait loci (QTLs) mapping

Genomic selection (GS)

Molecular breeding by design

Genome-wide association studies for salt stress traits



Eight significant SNPs for three salt stress traits

288 *G. hirsutum* accessions

54,588 SNP polymorphic loci

10 salt-stress phenotyping traits

- **GR**: germination rate ; **GP**: germination percentage
- **PH**: Plant height; **SDM**: shoot dry matter; **RDM**: root dry matter; **CC**: chlorophyll content
- **MDA**: malondialdehyde content; **SOD**: superoxide dismutase; **POD**: peroxidase; **CAT**: catalase enzyme activity

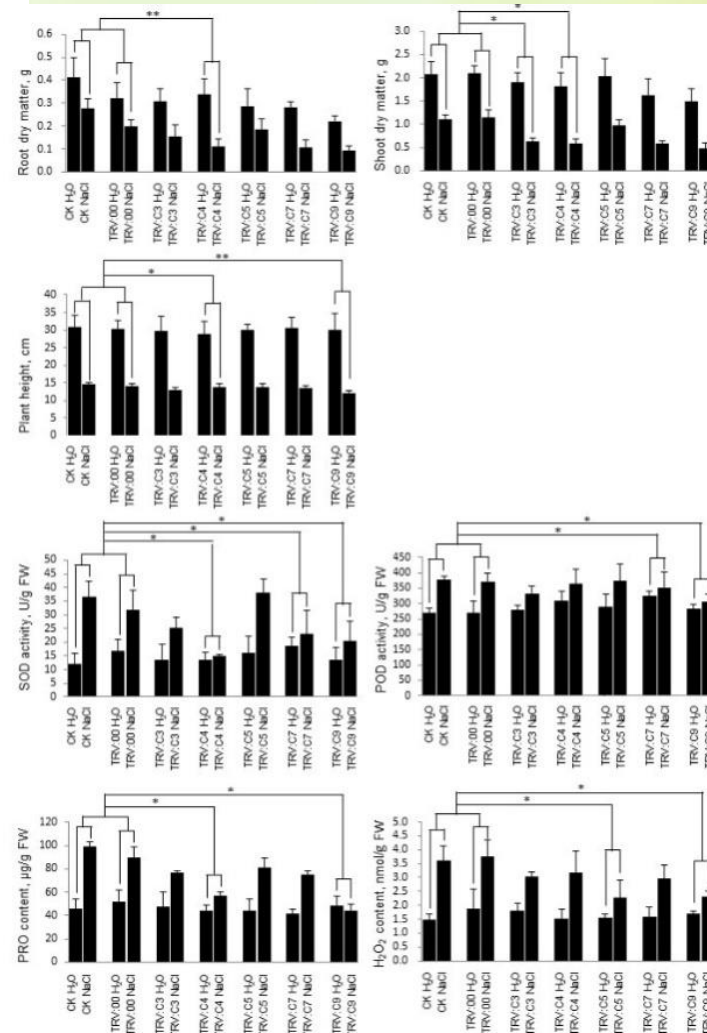
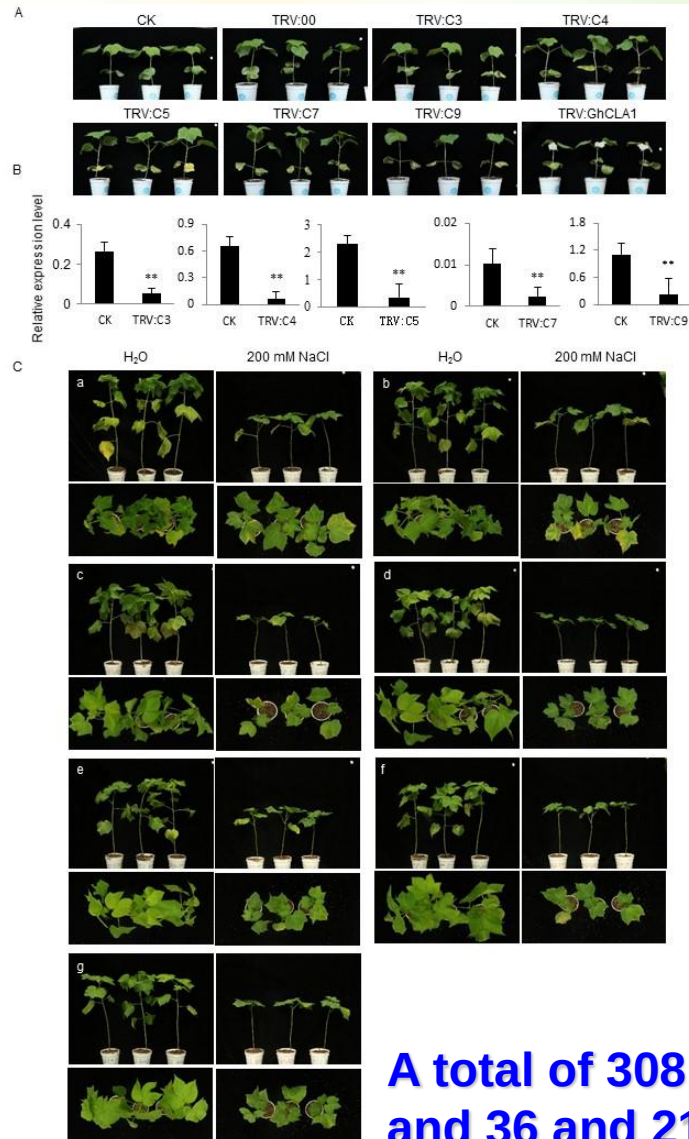
D5: TM57102, TM57104; **RCC**

A2: TM5633; **RMDA**

D9: TM70162, TM70169, TM70170, TM70171; **RMDA**

A12: TM43002; **RGR**

Mining candidate genes associated with salt stress traits



C3	MAPK
C4	P450
C5	WD40
C7	WRKY
C9	PIP

- Silencing each of the five genes can significantly enhance cotton susceptibility to salt stress.
- Function and transgenic lines ongoing

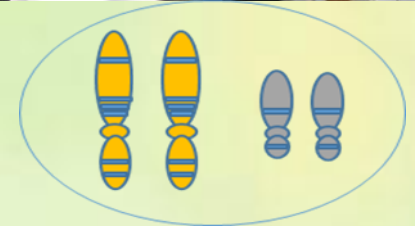
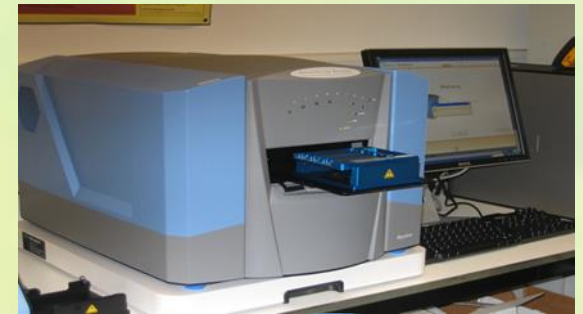
A total of 308 genes were annotated in these different regions, and 36 and 21 genes were annotated as response to stimulus and stress, respectively.

Comparison for different source chips (63K vs 80K)

Type	CottonSNP63K	CottonSNP80K
Number	63K	80K
SNP loci	Intra: 45,104 Inter: 17,954	Intra: 82,259
SNP loci origin	Intra: RNA-seq 25K; gDNA-seq 25K Inter: RNA-seq 15K; gDNA-seq 0.5K	Intra: re-seq 170K
MAF	>5%: 66.8%; >10%: 55.8% >20%: 40.0%	>10%: 100%
Subgenome belongings	unclear	clear
Genome coverage	unclear	2.1K at least; 23.5K flanking SNPs average; even distribution
Polymorphic rate	61.6%/1156 accessions	100%/300+ accessions
Gene %	Int	Higher SNP density with genome-wide distribution, more accurate addressable loci, and higher upland cotton intraspecific polymorphism.
Homoelogenous/ paralogous regions		

Future application

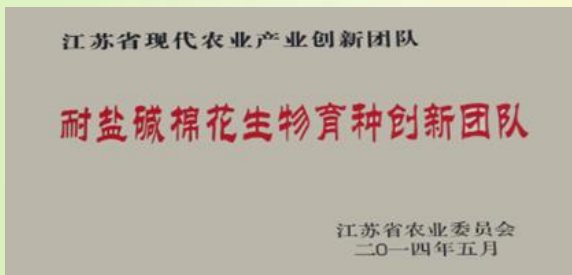
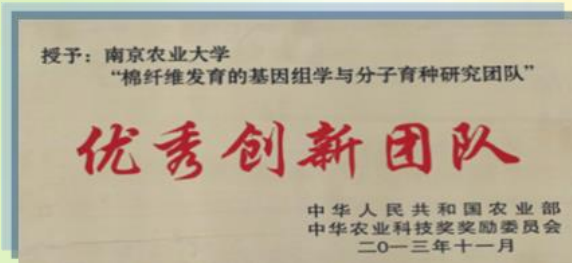
- Genetic diversity analysis
- High-density genetic mapping
- Genome-wide association studies (GWAS)
- Quantitative trait loci (QTLs) mapping
- Genomic selection (GS)
- Fingerprint of varieties
- Molecular breeding by design



Acknowledgment



Research group in NAU



- National Key R & D Program for Crop Breeding
- National Key Technology Support Program of China
- NSFC
- The Fundamental Research Funds by the Central Universities
- The Priority Academic Program Development of Jiangsu Higher Education Institutions
- Jiangsu Collaborative Innovation Center for Modern Crop Production project



Thanks!

谢谢!