



棉花生物学
国家重点实验室

State Key Laboratory of Cotton Biology

中华人民共和国科学技术部
The Ministry of Science and Technology of the People's Republic of China

The Genome of a Diploid Cotton *Gossypium raimondii*

Shuxun Yu, Wuwei Ye

State Key Laboratory of Cotton Biology

Cotton Research Institute, CAAS, China

Email: yew158@163.com

Outline

- **Background**
- **Sequencing and assembly**
- **Gene annotation and analysis**
- **Phylogeny studies, WGD and paleohexaploidization events**
- **Expansion of transposable elements**
- **Genes involved in cotton fiber development**
- **Genes involved in gossypol biosynthesis**
- **Acknowledgment**

Search: with



Cotton, commonly well-known as one of the most important economic crops, with its fiber, as cotton lint, is a principal source for the textile industry world-wide. There are 75 cotton-producing countries located between 32° south and 47° north latitude on the globe, and 33 million ha, or about 5% of the world's arable land.

CGP (Cotton Genome Project), were initiated and performed by [Institute of Cotton Research of CAAS](http://www.most.gov.cn/). Accompanied by BGI, CGP are mainly focused at cotton sequencing and functional analysis.

Mapview

Download

- ☒ Genome assembly
- ☐ Gene annotation
- ☒ Repeats annotation
- ☐ ncRNA annotation

Links

- Ministry of Science and Technology of the People's Republic of China
<http://www.most.gov.cn/>
- Ministry of Agriculture of the People's Republic of China
<http://www.most.gov.cn/>
- Chinese academy of agricultural sciences
<http://www.most.gov.cn/>

<http://cgp.genomics.org.cn/page/species/index.jsp>

“Cotton Genome Project (CGP)” was initiated by CCRI, China, in collaboration with BGI-China and USDA-ARS in 2007, which aimed at Upland cotton sequencing.



Download

Genome assembly:

1. Scaffold (v 1.0).fa.gz [FTP](#)

Gene annotation:

1. Gossypium raimondii L.(v 1.0).cds.gz [FTP](#)
2. Gossypium raimondii L.(v 1.0).pep.gz [FTP](#)

Repeats annotation:

1. denovo.gff(v 1.0).gz [FTP](#)
2. proteinmask.gff(v 1.0).gz [FTP](#)
3. repeatmasker.gff(v 1.0).gz [FTP](#)
4. trf.gff.gz [FTP](#)

ncRNA annotation:

1. miRNA.gff(v 1.0).gz [FTP](#)
2. rRNA.gff(v 1.0).gz [FTP](#)
3. snRNA.gff(v 1.0).gz [FTP](#)

Please [contact us](#) for FTP Username and Password.

BLAST

Blast (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. It is more than a tool to view sequences aligned with each other or to calculate percent homology, but a program to locate regions of sequence similarity with a view to comparing structure and function.

«Back Forward»

Choose program to use and database to search: Program **blastn** Species **Gossypium raimondii L.** Database **Genome(Scaffolds)**

Enter sequence below in FASTA format:

Clear sequence

Search

The query sequence is filtered for low complexity regions by default.

Filter: ☒ Low complexity ☐ Mask for lookup table only

Expect: **10** Matrix **BLOSUM62**

☐ Perform ungapped alignment

Query Genetic Codes (blastx only): **Standard (1)**

Database Genetic Codes (tblast[nx] only): **Standard (1)**

Frame shift penalty for blastx: **No OOF**

Other advanced options:

☒ Graphical Overview

Alignment view: **Pairwise**

Descriptions: **100**

Alignments: **50**

Color schema: **No color schema**

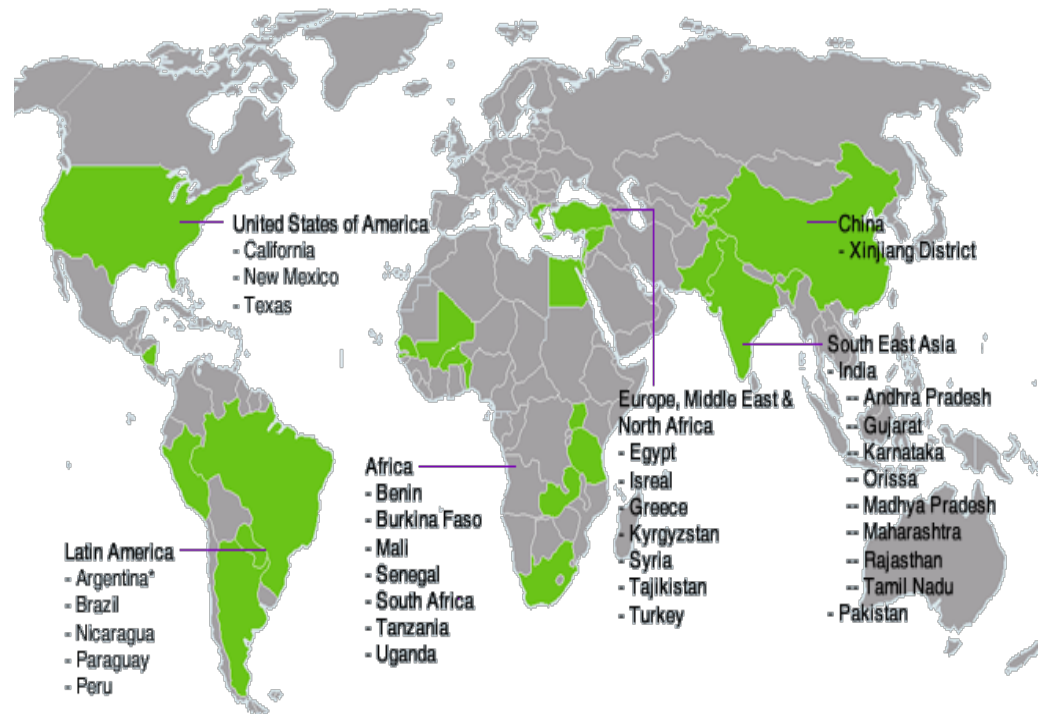
Clear sequence

Search

Background

➤ Cotton is one of the most important economic crops

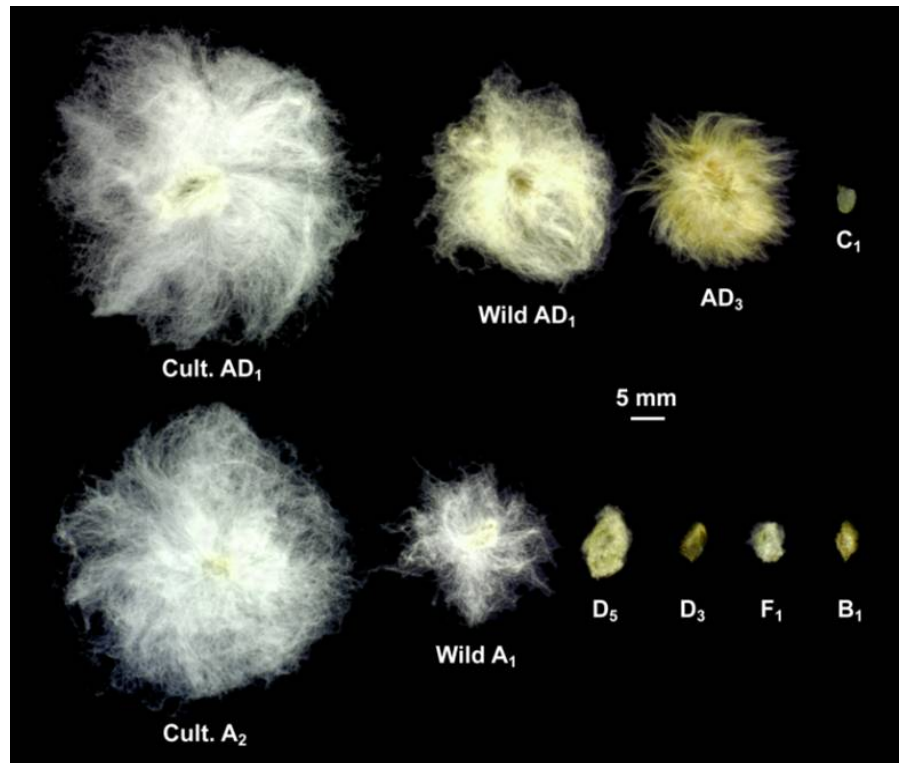
- With 75 cotton-producing countries located between **32° south and 47° north latitude** on the globe
- About **33 million ha or 5%** of the world's arable land is used for cotton planting annually



From <http://farmhub.textileexchange.org/learning-zone/growing-regions>

➤ **Cotton is an excellent model system for studying cell elongation and differentiation**

Gossypium seeds
exhibits
remarkable
variation among
the ~50 wild and
domesticated
species



*Variation in seed trichome (fiber) morphology in wild
and domesticated cottons*

From <http://cottonrevolution.info>

➤ Why *G. raimondii* genome sequencing

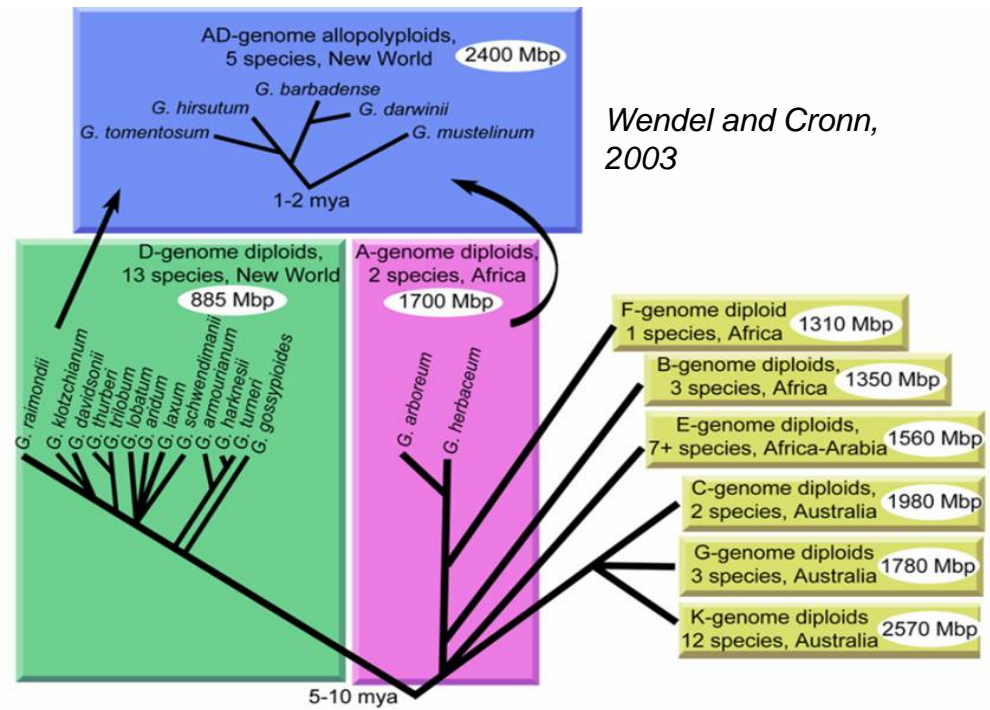
To gain insights into genome fusion and rearrangement in the cultivated **polyploid** genomes, we have sequenced the putative D genome donor, *G. raimondii*.

Gossypium Species:

5 tetraploid and over 45 diploid species

Gossypium Genome Size:

880 Mb (*G. raimondii*) to 2,500 Mb

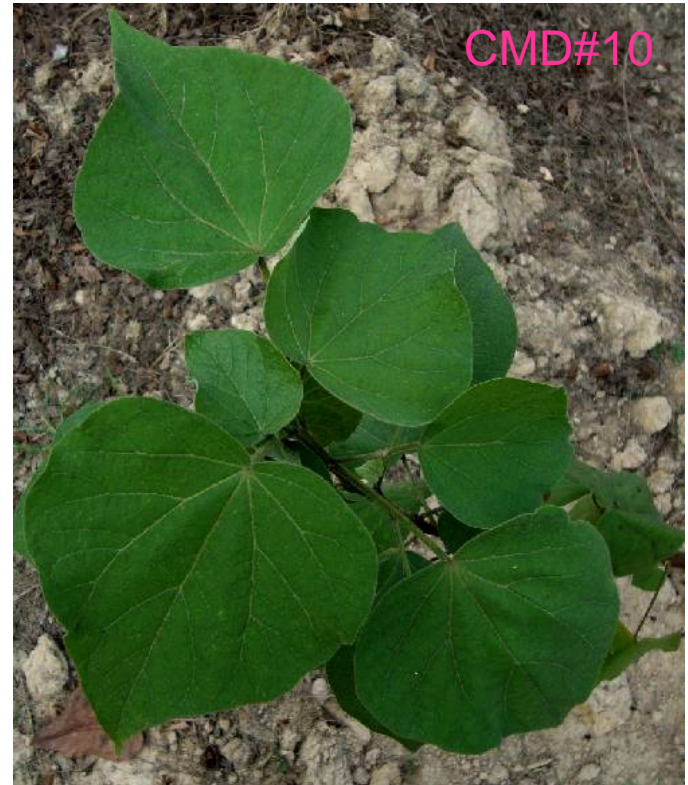


Phylogeny and evolution of Gossypium species

Sequencing and assembly

➤ Material and Methods

- ▶ Strategy: Whole genome shotgun
- ▶ Material: D₅-3 (CMD#10)
- ▶ Sequencer: Illumina HiSeq 2000
- ▶ Assembly soft: SOAPdenovo



➤ Sequenced genome data

- ▶ 78.7 Gb Illumina paired-end reads
- ▶ 103.6 folds of the 775.2-Mb assembled *G. raimondii* genome
- ▶ Pair-end libraries (insert size) from 170bp to 40kb

Global statistics of *G. raimondii* genome sequencing data

Pair-end libraries (insert size)	Total data (Gb)	Reads length	Sequence coverage (×)
170 (bp)	15.0	92_92	19.7
250	28.8	140_140	37.9
500	11.7	91_91	15.4
800	11.9	91_91	15.7
2 (kb)	6.8	49_49	9.0
5	1.9	49_49	2.5
10	1.3	49_49	1.7
20	0.7	49_49	1.0
40	0.6	90_90	0.8
Total	78.7	-	103.6

➤ Genome assembly

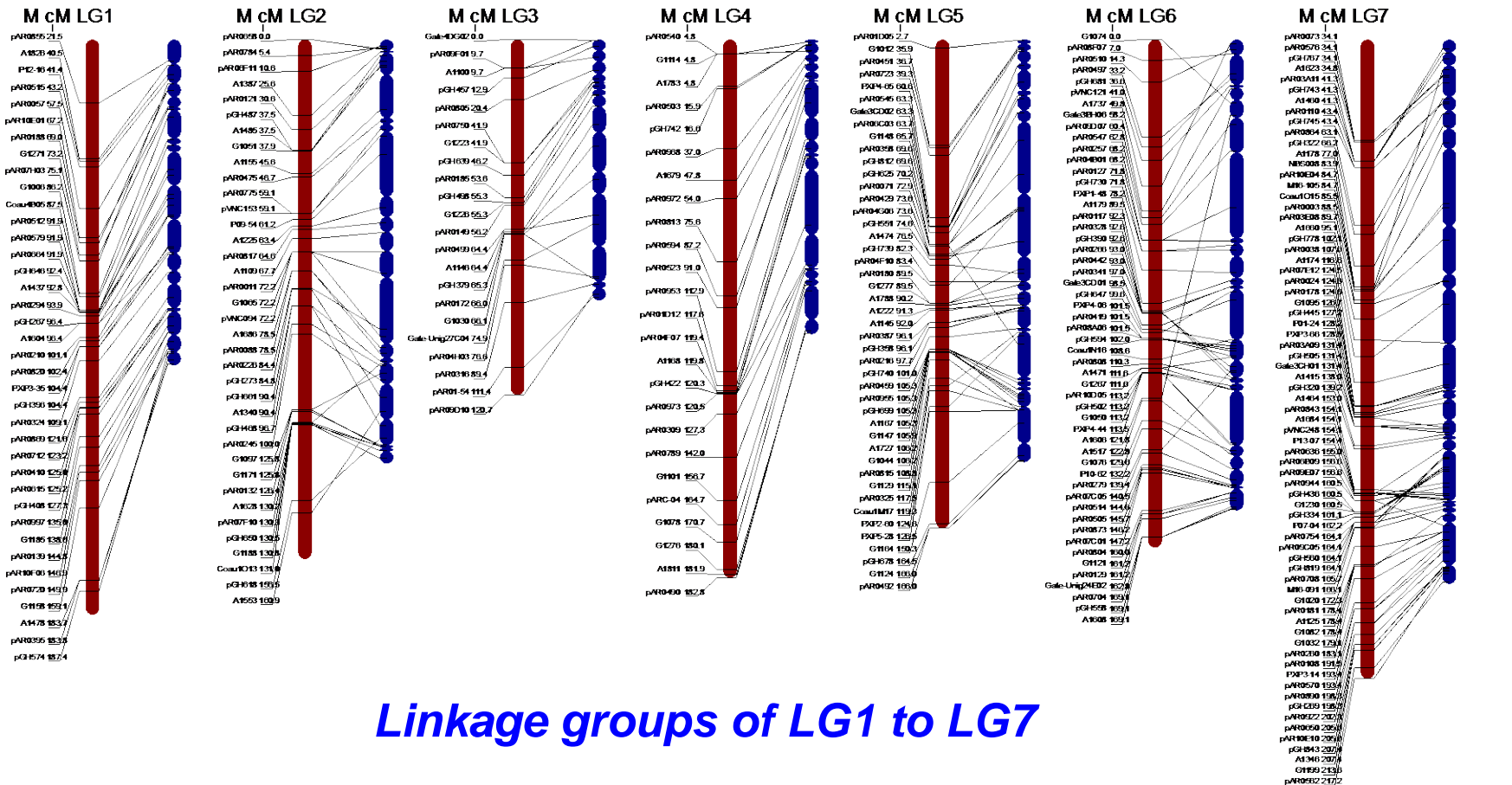
- ▶ The assembly consists of **41,307** contigs and **4,715** scaffolds
- ▶ Those sequence accounts for about **88.1%** of the estimated *G. raimondii* genome
- ▶ The longest contig and scaffold are about **333.6Kb** and **12.8Mb**, respectively
- ▶ The number of N90 and N50 contigs and scaffolds are **337** and **95**, respectively

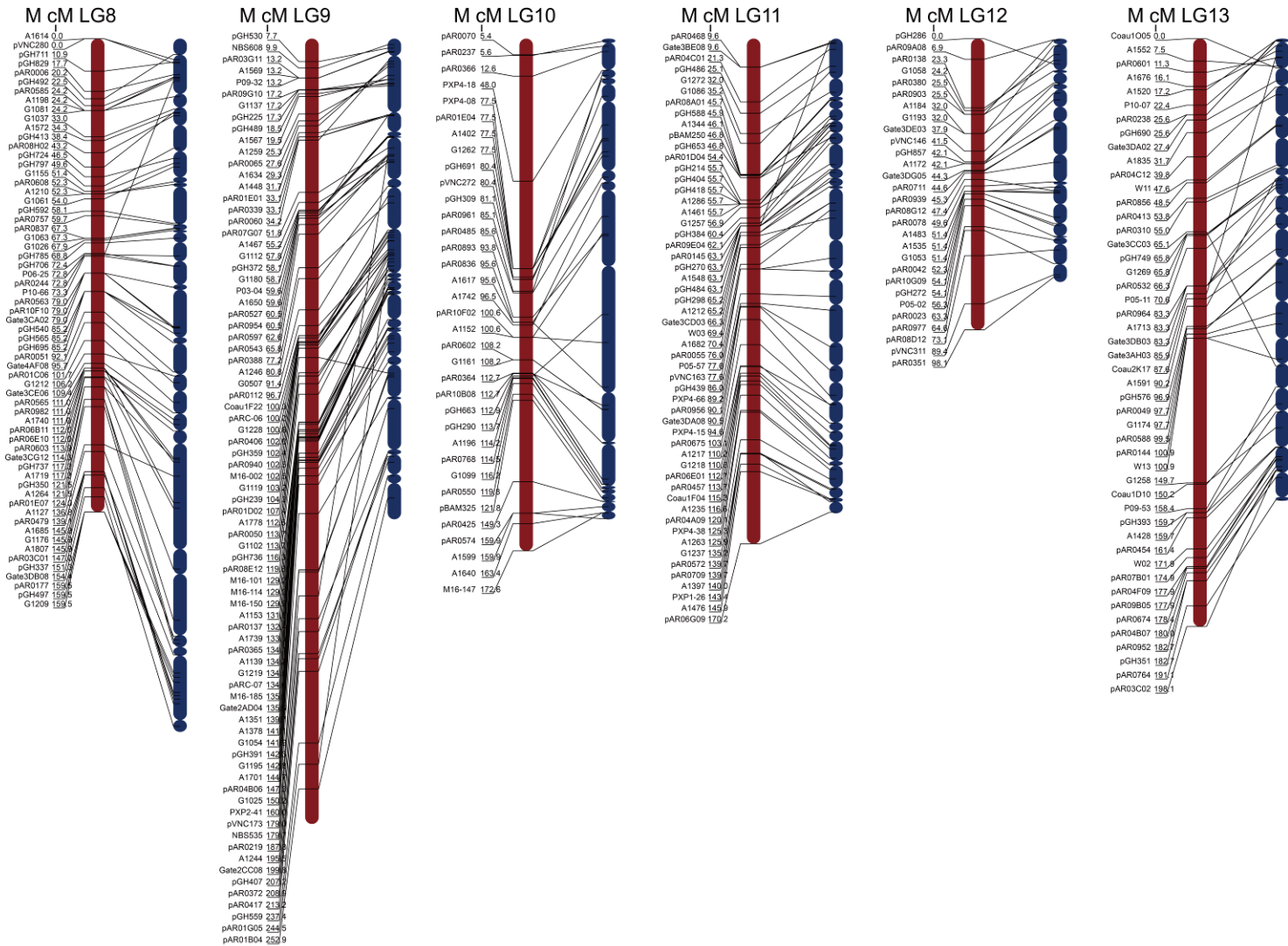
Summary of the *G. raimondii* genome assembly

	Contig		Scaffold	
	Size (bp)	Number	Size (bp)	Number
N90	11,092	17,042	644,316	337
N50	44,853	4,918	2,284,095	95
Longest	333,622	-	12,776,819	-
Total size	732,159,341	-	775,153,485	-
Total number (all)		41,307	-	4,715
Total number (≥2 kb)		27,997	-	1,367

➤ Anchoring the *G. raimondii* genome to the cotton consensus map according to specific markers

- ▶ 567.2 Mb (73.2%) scaffolds were anchored on the map assisted by previously known specific genetic markers

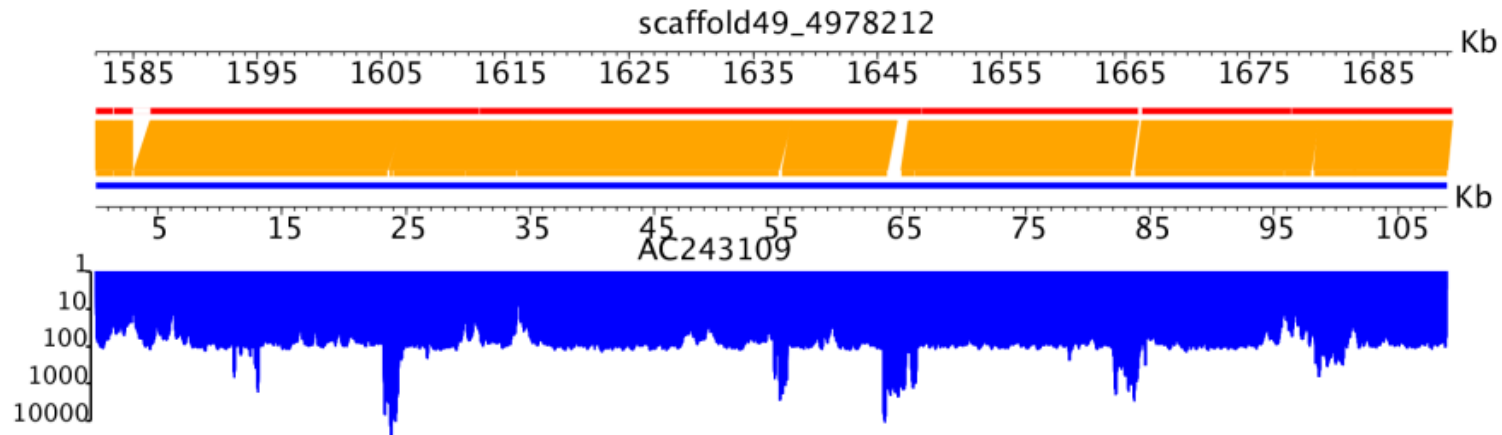




Linkage groups of LG8 to LG13

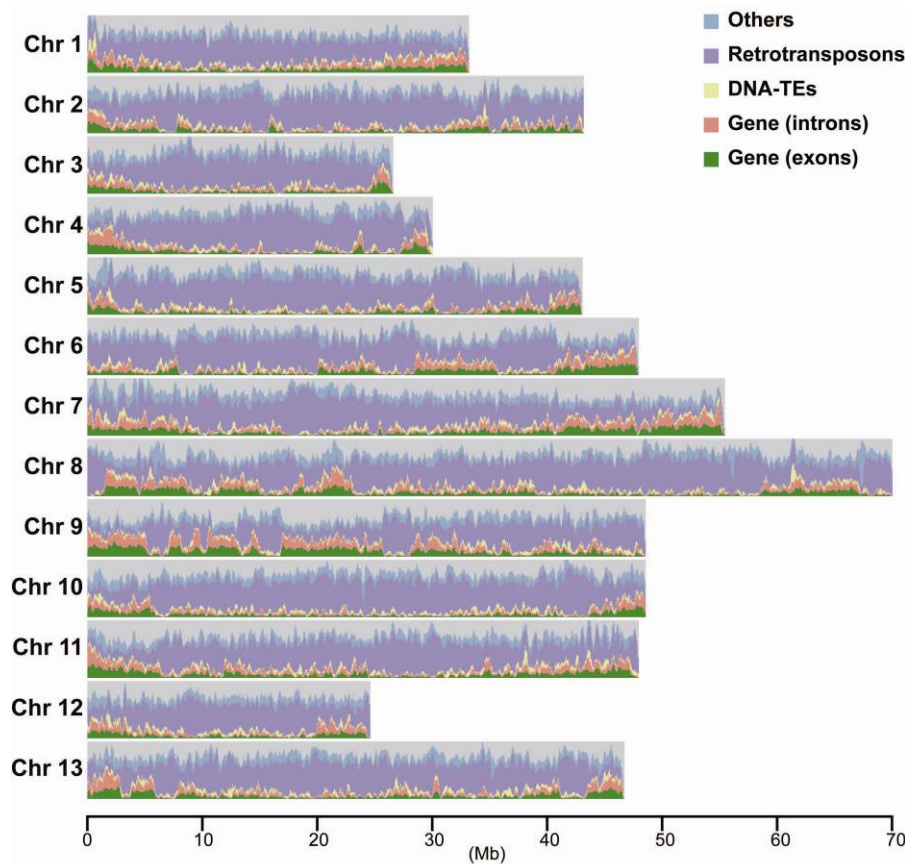
➤ Assembly evaluation with BAC and EST

- ▶ From 58,061 expressed sequence tags reported in *G. raimondii*, 93.4% were identified in the assembly
- ▶ When compared to 25 completely sequenced *G. raimondii* bacterial artificial chromosome clones, 24 can be recovered fully from our assembly.



Assembly evaluation with BAC and EST

➤ Genomic landscape of the assembled chromosomes



Major DNA components are categorized into **exons** (green), **introns** (nattier blue), **DNA-TEs** (DNA transposons, yellow), **LTR** (long terminal repeat retrotransposons, blue) and **other** (repeat sequence other than DNA-TE and LTR, bright green). Grey color represents **unclassified DNA**.

Genomic landscape of the assembled chromosomes

Gene annotation and analysis

➤ Gene Prediction

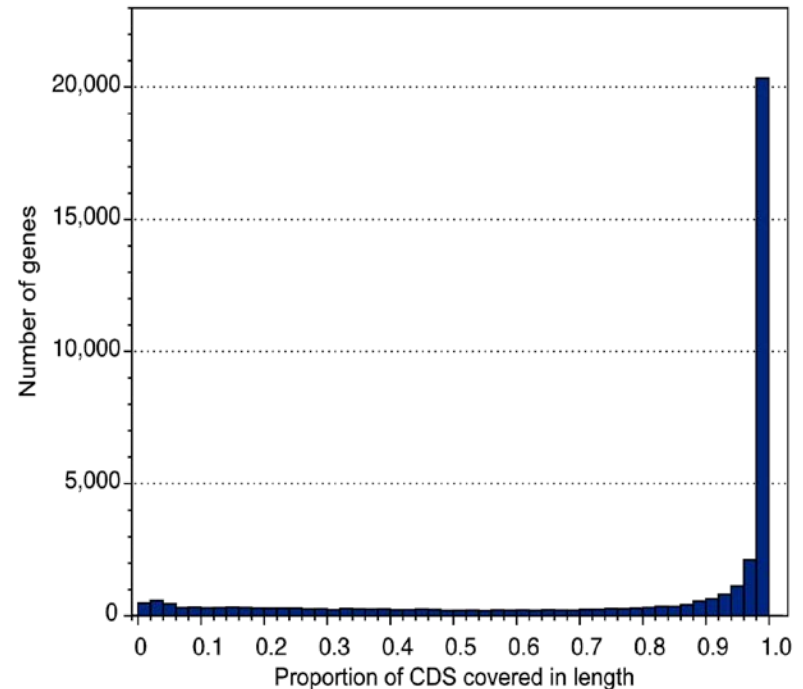
- ▶ Genome annotation combining both **gleaning results** obtained from *ab initio* prediction, homology search and EST alignment
- ▶ **40,976** protein-coding genes in the *G. raimondii* genome
- ▶ An average transcript size of **2,485** bp **4.5** exons per gene

➤ Gene annotation

- ▶ **83.69%** show homology in TrEMBL protein database
- ▶ **69.98%** are identified in InterPro
- ▶ **92.2%** of coding sequences were supported by transcriptome data

Number of genes with homology or functional classifications by different methods

	Number	Percent (%)
Total	40,976	100
Annotated	34,507	84.21
SwissProt	26,587	64.88
TrEMBL	34,288	83.69
InterPro	28,676	69.98
KEGG	23,167	47.99
GO	21,801	53.20
Unannotated	6,469	15.79



Transcriptome data support for gleaned gene models in the assembled G. raimondii genome

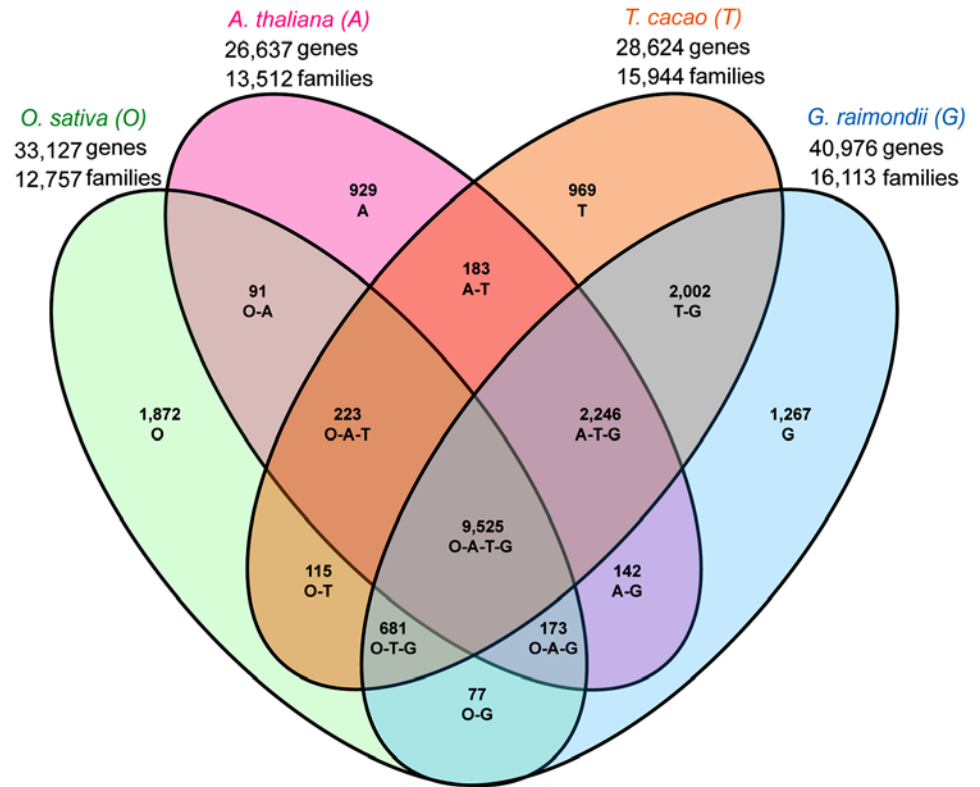
➤ Non-coding RNA genes

- ▶ MiRNA, microRNA: accounts for 0.6% of estimated *G. raimondii* genome
- ▶ tRNA, transferRNA: 1.0% of estimated *G. raimondii* genome
- ▶ rRNA, ribosomalRNA: 1.3% of estimated *G. raimondii* genome
- ▶ snRNA, small nuclear RNA: 1.5% of estimated *G. raimondii* genome

Type	Copy number	Average length (bp)	Total length (bp)	% of genome
miRNA	348	124	43,345	0.006
tRNA	1,041	75	78,062	0.010
rRNA(1.3%)	18S	78	46,701	0.006
	28S	109	12,982	0.002
	5.8S	31	4,705	0.001
	5S	347	38,033	0.005
snRNA(1.5%)	CD-box	935	99,241	0.013
	HACA-box	29	3,569	0.000
	splicing	118	17,784	0.002

➤ Analysis of major gene families

- ▶ Four different plant species possessed similar numbers of gene families with a core set of **9,525** gene families
- ▶ Of the **16,113** *G. raimondii* gene families, all but **1,267** are conserved with at least one other plant genome

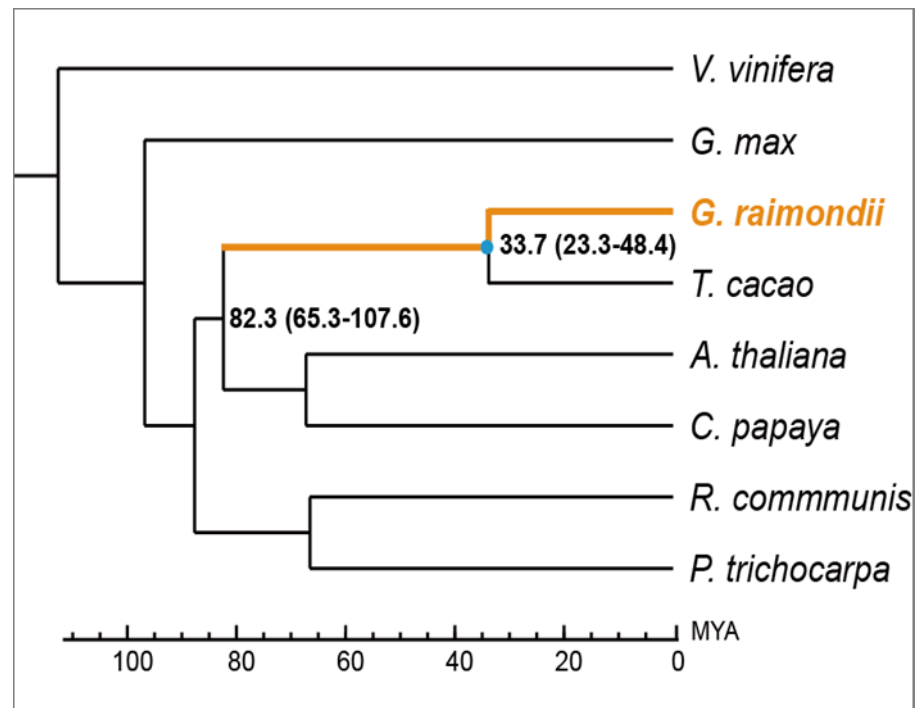


Venn diagram analyses of unique and shared genes or gene families amongst *O. sativa*, *T. cacao*, *A. thaliana* and *G. raimondii*

Phylogeny studies whole-genome duplication

➤ Phylogeny studies

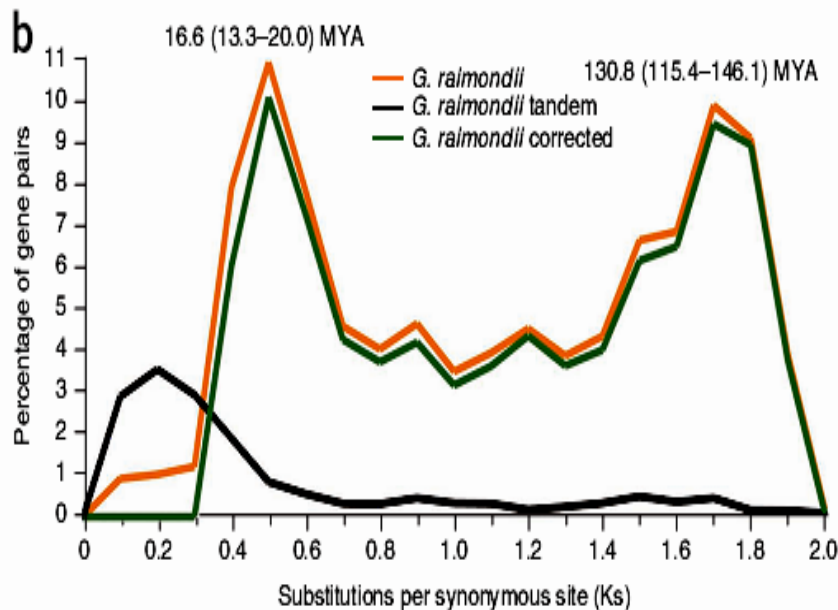
- ▶ *G. raimondii* and *T. cacao* are in the same subclade that most likely diverged at about **33.7 MYA**.
- ▶ *C. papaya* and *Arabidopsis* belong to **another subclade**.
- ▶ The divergence time for two subclades is estimated around **82.3 MYA**.



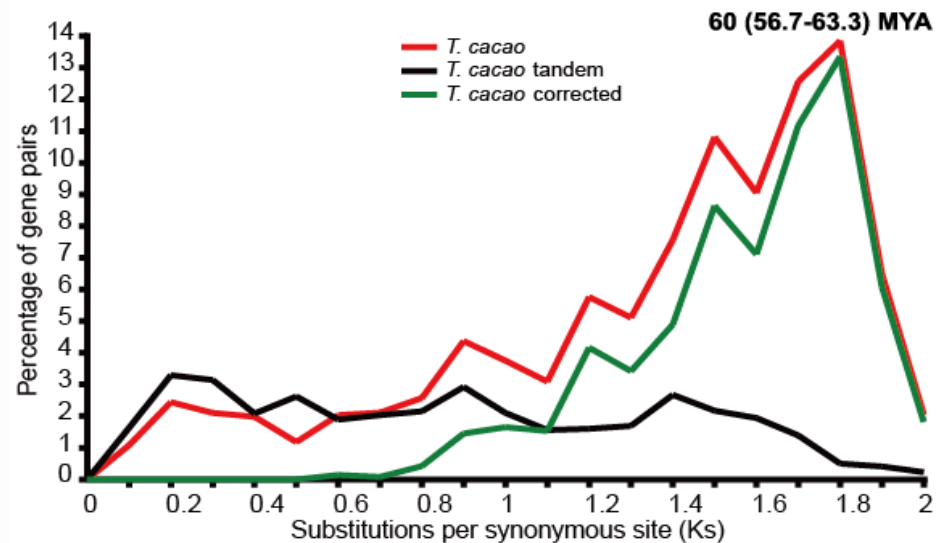
Phylogenetic analysis showed that *G. raimondii* and *T. cacao* were separated about 33.7 MYA
O. sativa was used as the out-group

➤ WGD and paleohexaploidization events

- The hexaploidization event and a cotton-specific WGD events approximately 3–20 MYA was observed.
- The first peak appeared at approximately 16.6 (13.3–20.0) MYA,
- The second peak appeared at approximately 130.8 (115.4–146.1) MYA, corresponding to the paleohexaploidization event shared by the eudicots.



Ks distributions of *G. raimondii*

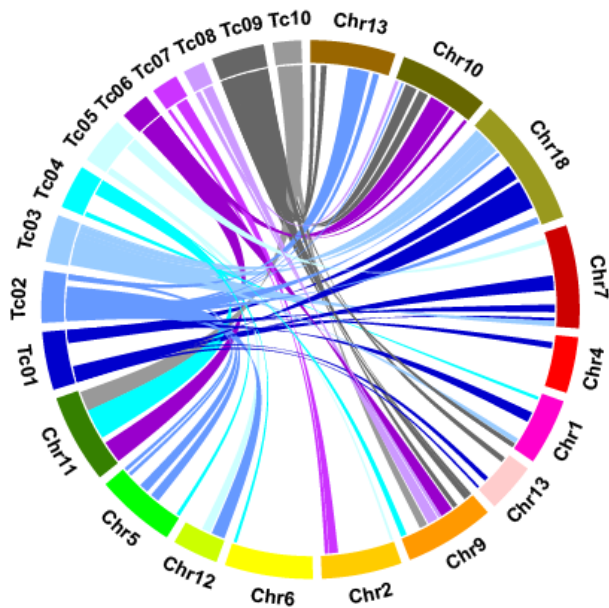


Ks distributions of *T. cacao*

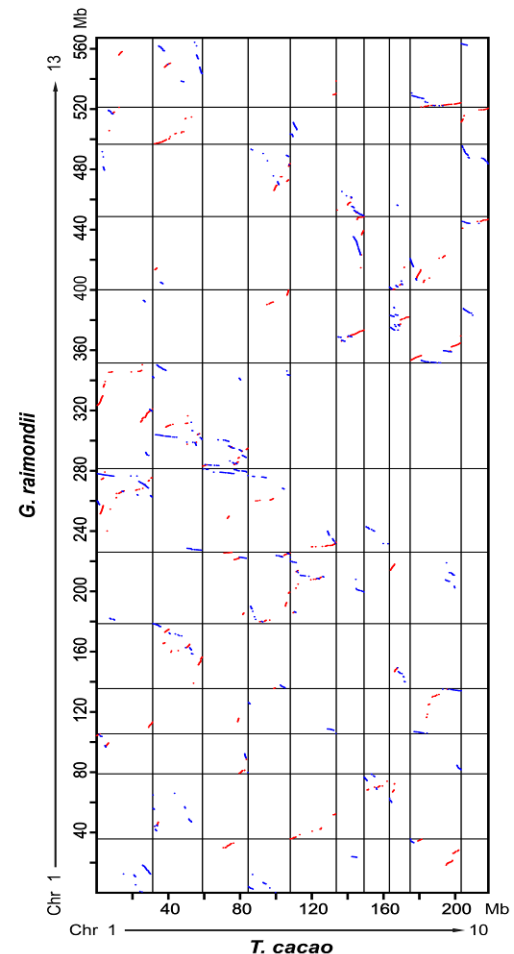
T. cacao has a single peak of WGD event

➤ Synteny blocks between the genomes of *T. cacao* and *G. raimondii*

- ▶ 463 colinear blocks covering **64.8%** and **74.41%** of the assembled *G. raimondii* and *T. cacao* genome respectively



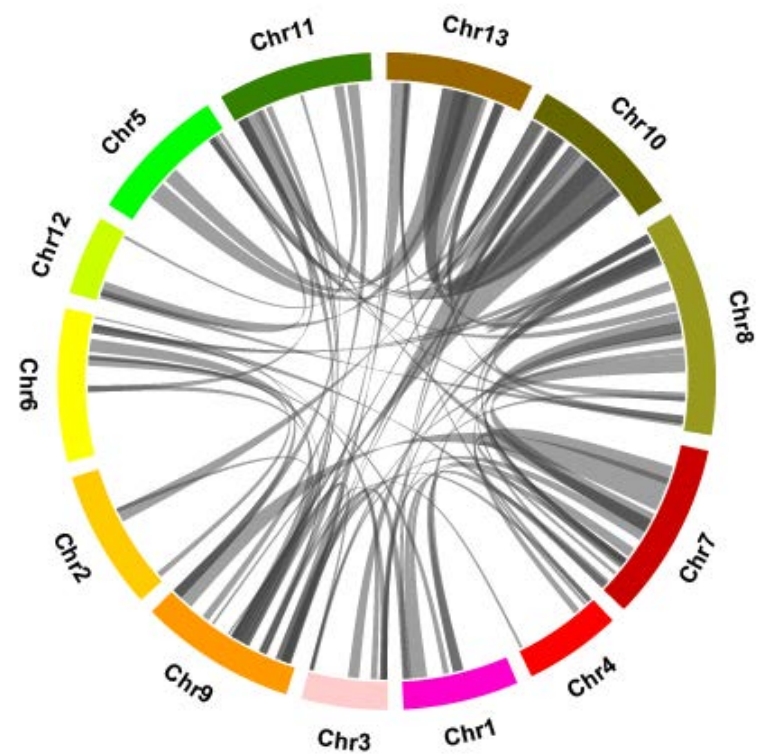
Synteny blocks between *T. cacao* and *G. raimondii*



Analysis of syntenic blocks between *G. raimondii* and *T. cacao* blocks

➤ Synteny blocks among the *G. raimondii* chromosomes

- ▶ 2,355 synteny blocks
21.2% were found involved only in **two** chromosome regions
- ▶ 33.7% spanning **three** chromosome regions
- ▶ 16.2% traversing **four** chromosome regions



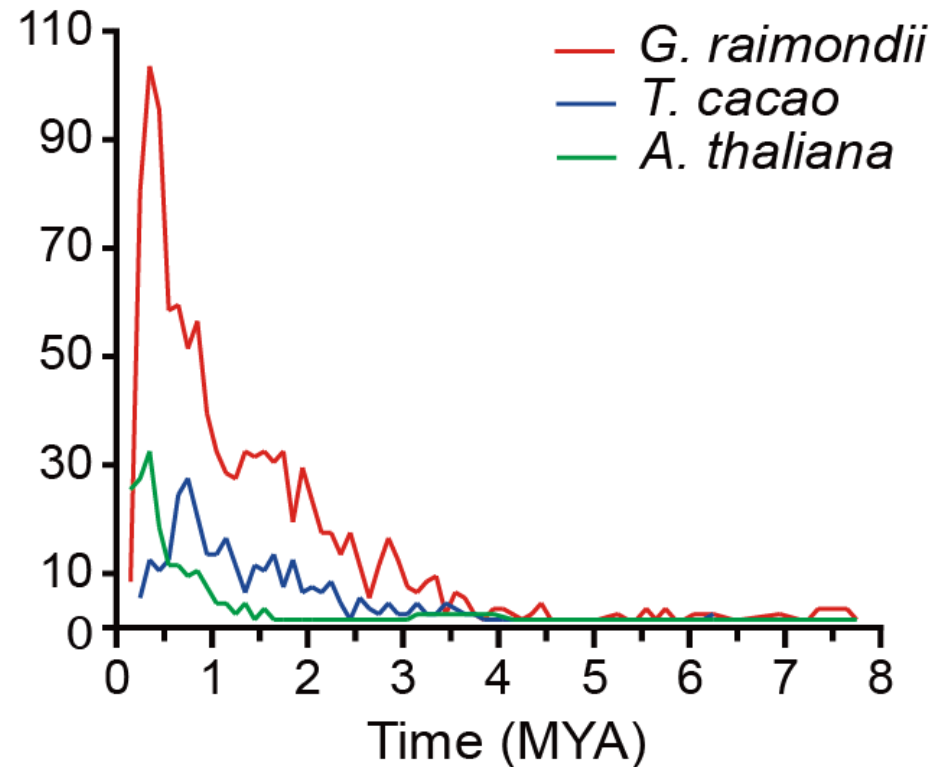
Synteny blocks among different G. raimondii chromosomes

Expansion of transposable elements

- ▶ In *G. raimondii*, transposable elements comprise to about **57%** (441 Mb in total length) of the genome
- ▶ In *T. cacao* and in *A. thaliana*, transposable elements account for 24% and **14%** respectively
- ▶ Suggest that **TEs have significantly proliferated** in the *G. raimondii*, and that their accumulation should partially responsible for *G. raimondii* genome expansion

➤ The Distribution of TE Insertion Time

- ▶ The growth rate of these LTR retrotransposons in *G. raimondii* and *T. cacao* tends to slow down since 0.5 and 0.7 MYA, respectively
- ▶ The number of LTR retrotransposons has been increasing in *A. thaliana* since 1.5 MYA.

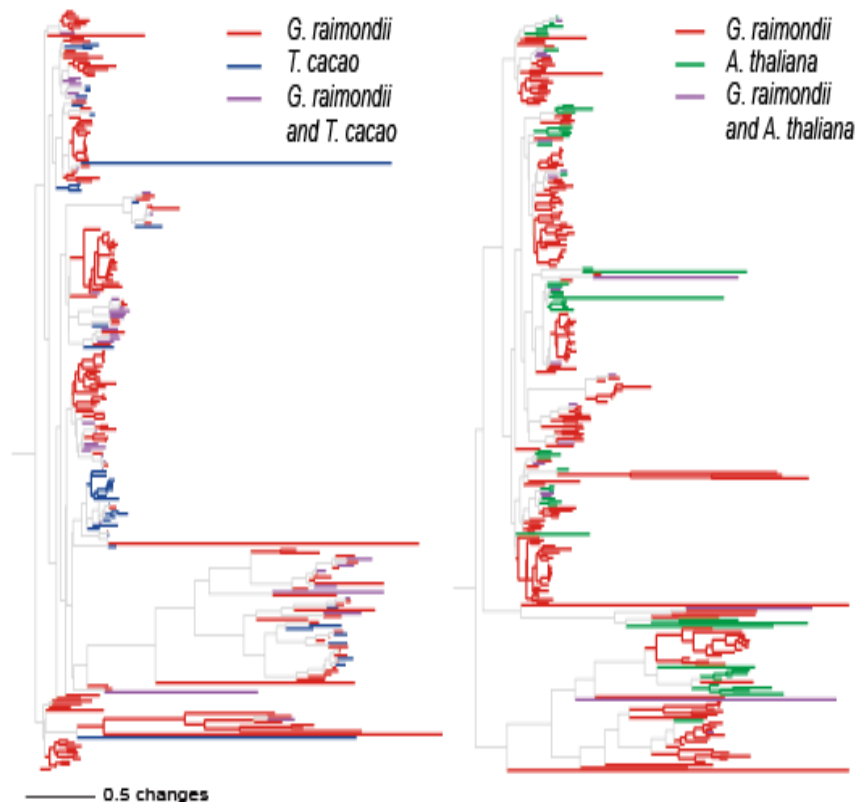


The distribution curve for the number and the insertion time of LTRs in different plant genomes

➤ Phylogeny of LTR retrotransposons in *G. raimondii*, *A. thaliana* and *T. cacao*

► Phylogenetic analysis

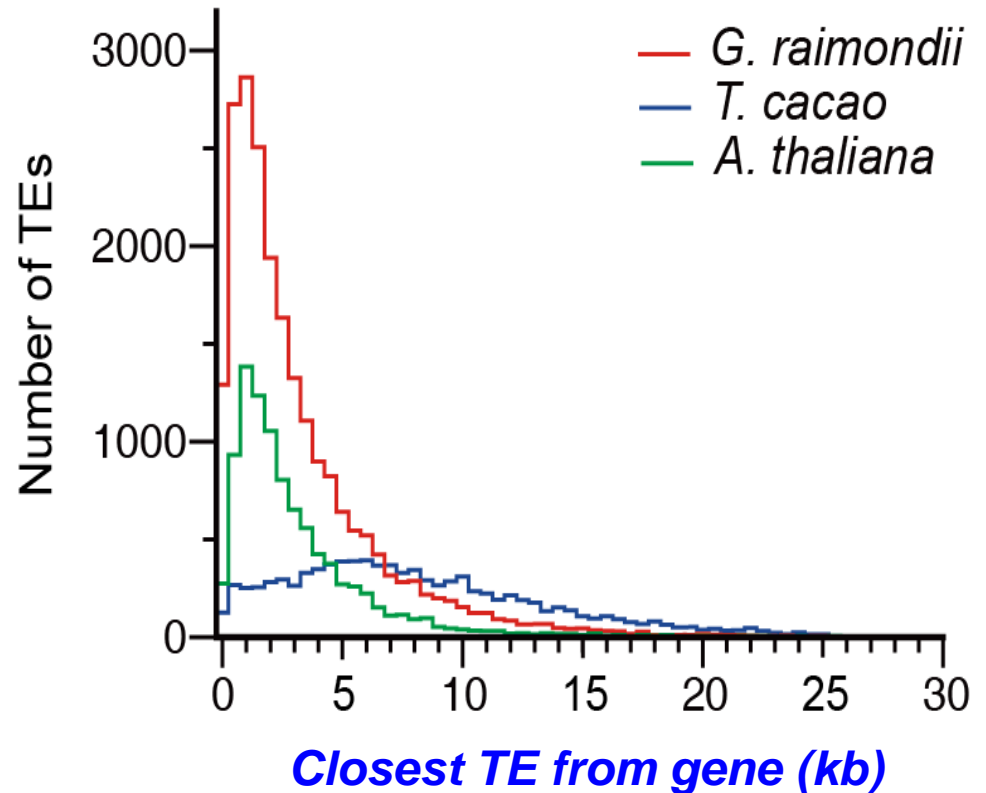
supported the notion that a **greater expansion** of specific LTR retrotransposon clades has occurred in *G. raimondii*, comparisons with TEs in *T. cacao* and *A. thaliana*.



*Phylogeny of LTR retrotransposons in the *G. raimondii*, *T. cacao* and *A. thaliana* genomes*

➤ Distance distributions of nearest TEs from each gene in *G. raimondii*, *A. thaliana* and *T. cacao*

- ▶ *G. raimondii* has a higher proportion of genes with a TE **nearby** than *T. cacao* and *A. thaliana*.
- ▶ *T. cacao* has maintained the **greatest distance** between its genes and its TEs.



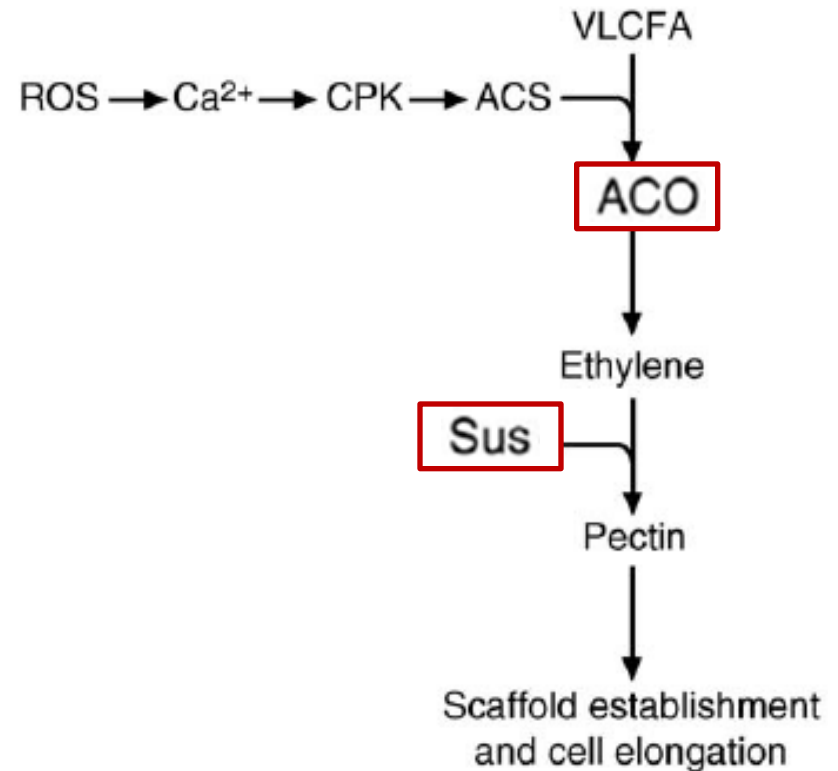
Genes involved in cotton fiber development

➤ Signaling pathway for the linear cell-growth mode

- ▶ **Ethylene** plays a key role in fiber growth.
- ▶ **VLCFAs** promote fiber growth by activating ethylene biosynthesis, whereas ethylene stimulates pectin biosynthesis and scaffold establishment

*Yong-Mei Qin and Yu-Xian Zhu.. Current Opinion in Plant Biology
2011, 14:106–111*

- ▶ We studied the expression level of some fiber development related genes, such as **KCS**, **ACO**, **Sus**




Signaling pathway for the linear cell-growth mode

➤ Twenty-one 3-ketoacyl-CoA synthase (KCS) genes

▶ KCS2, KCS13 and KCS6

were only expressed in *G. hirsutum* while intermediate levels of **KCS7** transcripts were observed both in *G. hirsutum* and *G. raimondii*.

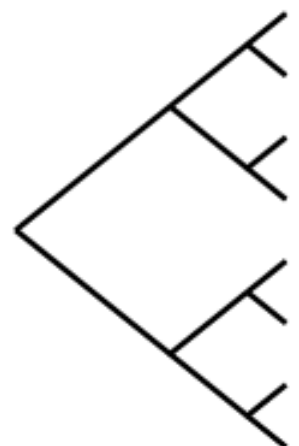


		RPKM	
		G.r	G.h
KCS7	167	22	
KCS2	4	73	
KCS13	48	984	
KCS6	3	288	

*The expression level was estimated by RPKM values
(Reads Per Kilobase of CDS per Million mapped reads)*

➤ 1-aminocyclopropane-1-carboxylic acid oxidase (ACO) genes

- ACO transcripts were recovered from *G. raimondii* at the **3-DPA stage**, suggestive of a major role for the plant hormone ethylene during early fiber cell development.

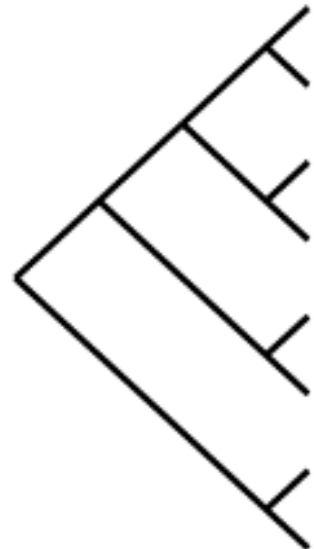


		RPKM	
		G.r	G.h
ACO4	1	36	
ACO3	748	72	
ACO1	4018	13	
ACO2	117	4	

The expression level was estimated by RPKM values (Reads Per Kilobase of CDS per Million mapped reads)

➤ Sucrose synthase (*Sus*) genes

- ▶ Among four *Sus* genes identified in the genome, three (***SusB***, ***Sus1*** and ***SusD***) were expressed significantly higher in *G. hirsutum* when compared with that of the *G. raimondii*.

		RPKM	
		G.r	G.h
	<i>SusB</i>	65	368
	<i>Sus1</i>	119	1609
	<i>SusD</i>	44	178
	<i>SusC</i>	2	4

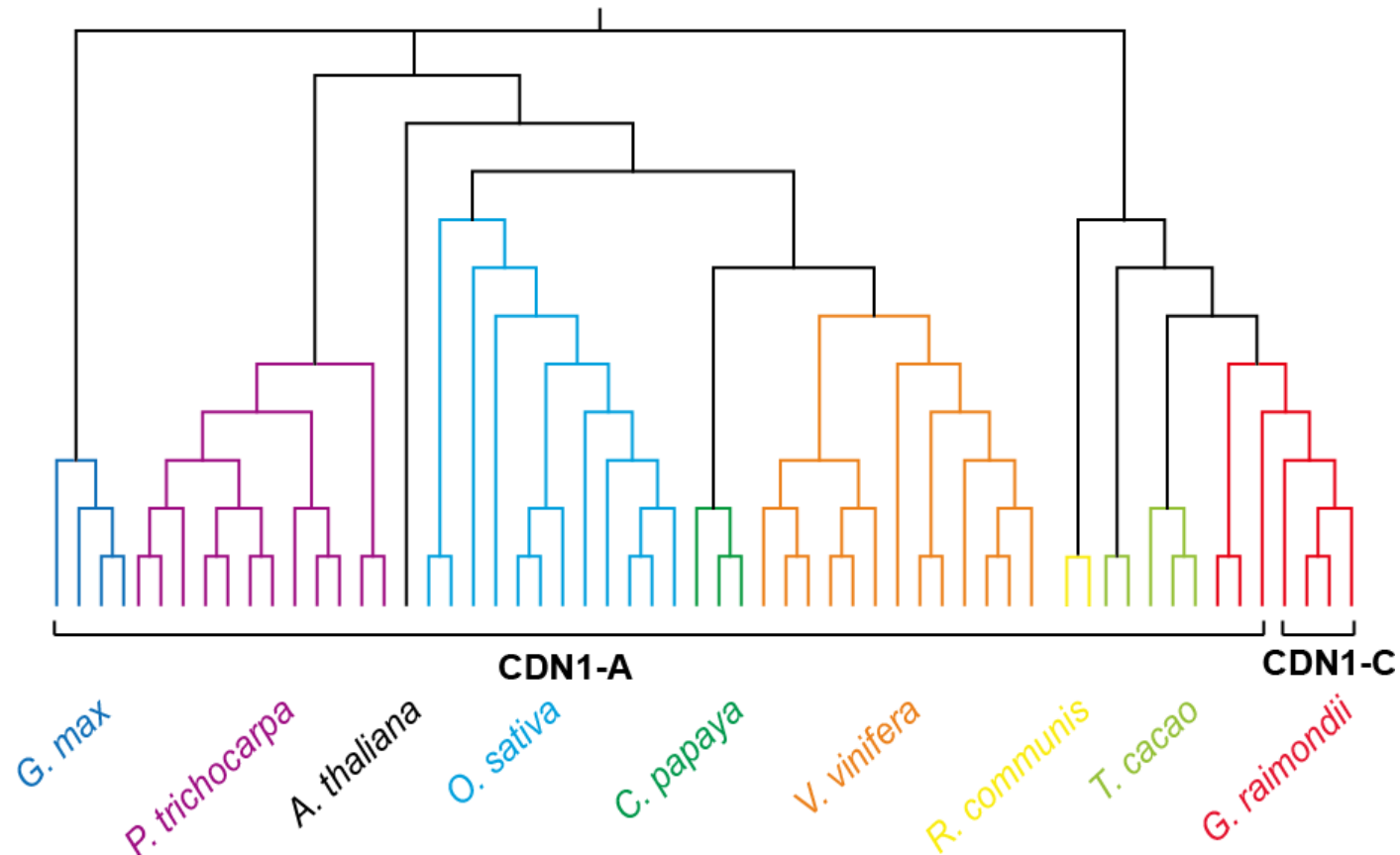
The expression level was estimated by RPKM values (Reads Per Kilobase of CDS per Million mapped reads)

Genes involved in gossypol biosynthesis

- ▶ Cotton is known to produce a unique group of **terpenoids** including desoxyhemigossypol, hemigossypol, gossypol, hemigossypolone and the heliocides
- ▶ Cotton plants accumulate **gossypol** and related sesquiterpenoids in **pigment glands** as their defensive machinery against pathogens and herbivores.

➤ Phylogenetic analysis of the CDN1 gene family

- ▶ **CDN1-C** genes are found only in *G. raimondii* and *T. cacao*



Phylogenetic of the CDN1 gene family

Summary

- An assembly of about 88.1% of the estimated genome size
- Of the 40,976 protein-coding genes identified, 92.2% were confirmed by transcriptome data
- 1,267 unique gene families were found in the *G. raimondii* genome
- The hexaploidization event shared by the eudicots and a WGD event was observed in 3–20 MYA
- *G. raimondii* showed significantly lower gene density with a high proportion of TEs

Acknowledgment

The draft genome of a diploid cotton *Gossypium raimondii*
Nature Genetics, 2012. doi:10.1038/ng.2371.

Kunbo Wang*, Zhiwen Wang*, Fuguang Li*, Wuwei Ye*, Junyi Wang*,
Guoli Song*, Zhen Yue, Lin Cong, Haihong Shang, Shilin Zhu,
Changsong Zou, Qin Li, Youlu Yuan, Cairui Lu, Hengling Wei, Caiyun
Gou, Zequn Zheng, Ye Yin, Xueyan Zhang, Kun Liu, Bo Wang, Chi Song,
Nan Shi, Russell J. Kohel, Richard G. Percy, John Z. Yu, Yuxian Zhu#,
Jun Wang#, Shuxun Yu#.



Thanks