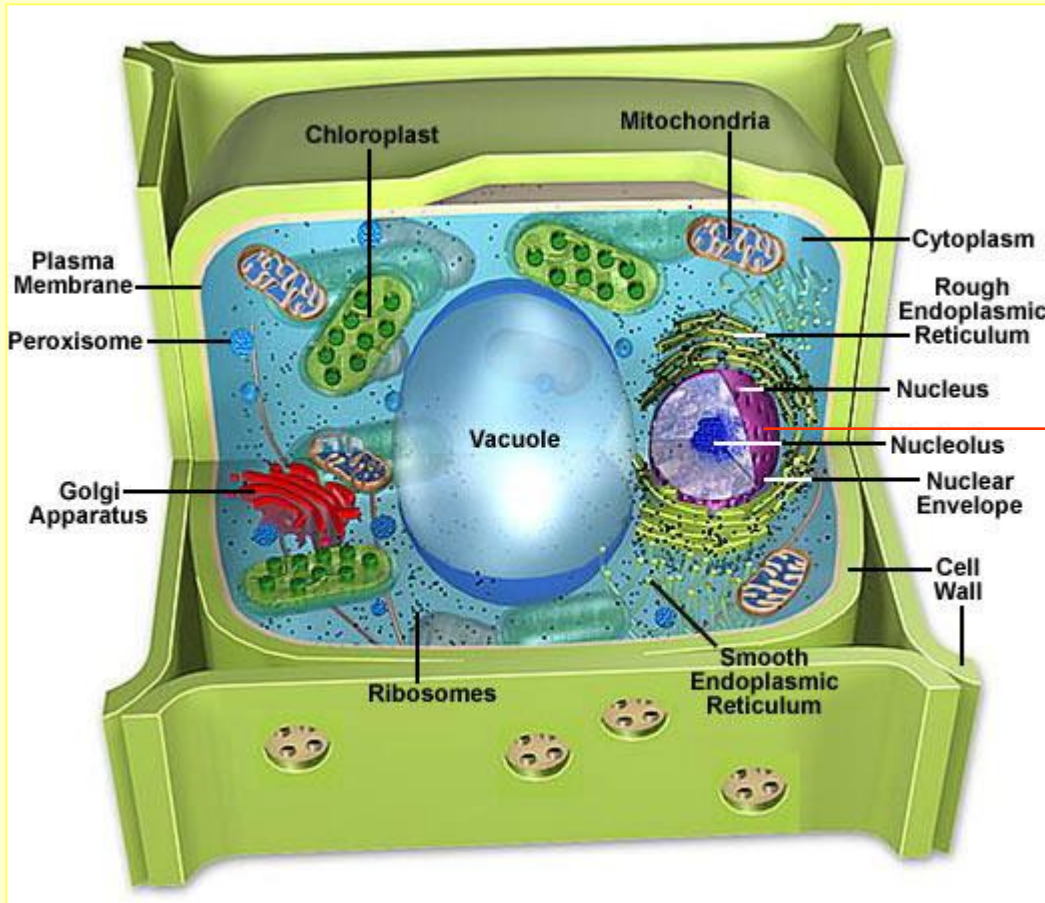


**The *Hevea* Genome Project**  
**Sequencing the whole *Hevea* genome:**  
**strategy & expectations**

Montoro, Chow, Chrestin, Duan, Galand, Garcia, Kolesnikova-Allen, Le Guen, Leclercq,  
Lekawipat, Omokhafa, Pujade-Renaud, Putranto, Sales, Seguin, Teerawatanasuk

**IRRDB NATURAL RUBBER CONFERENCE**  
**Bogor - Indonesia 26-27 October 2009**

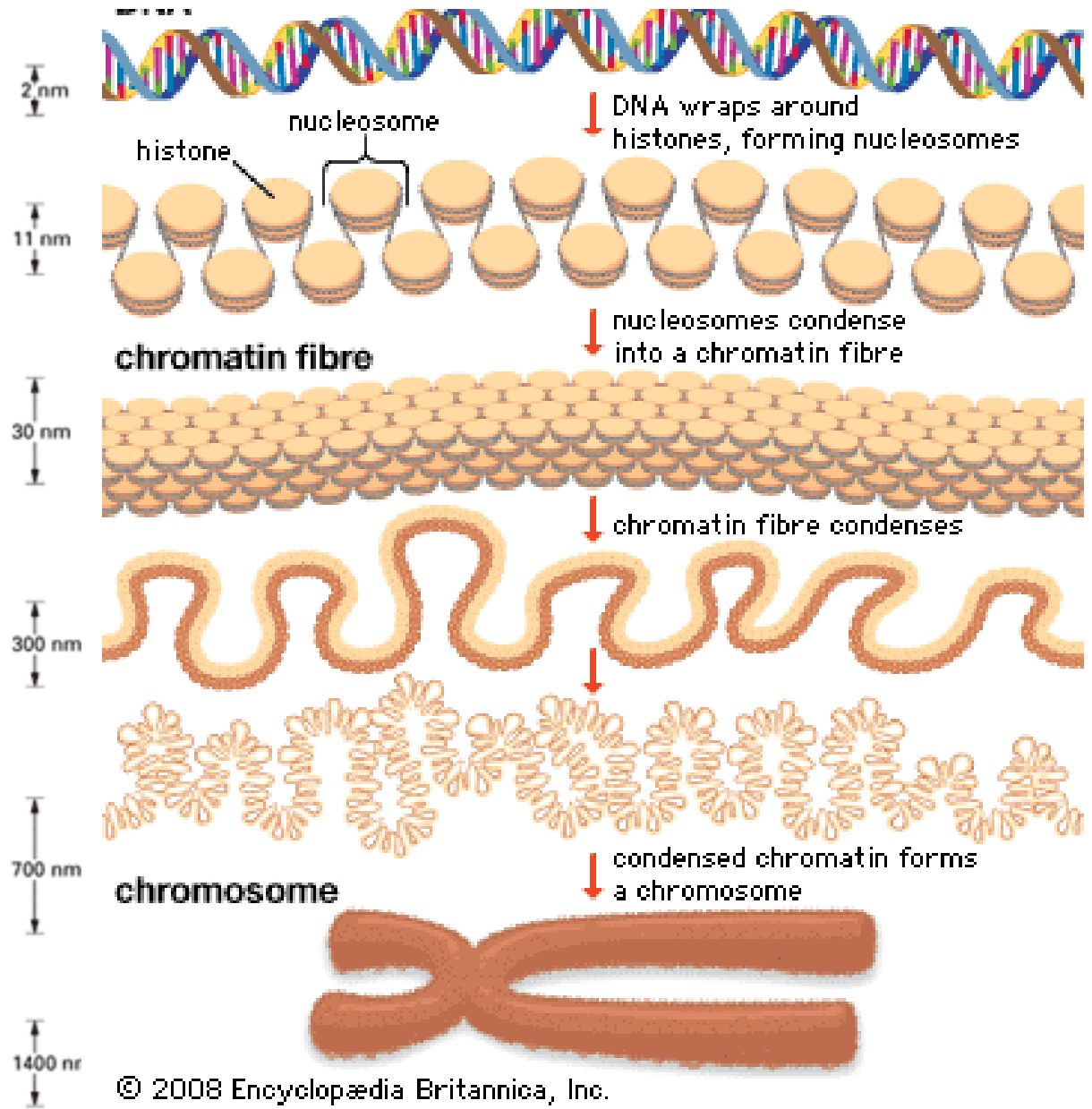
# The cell nucleus contains the majority of the plant cell genetic material



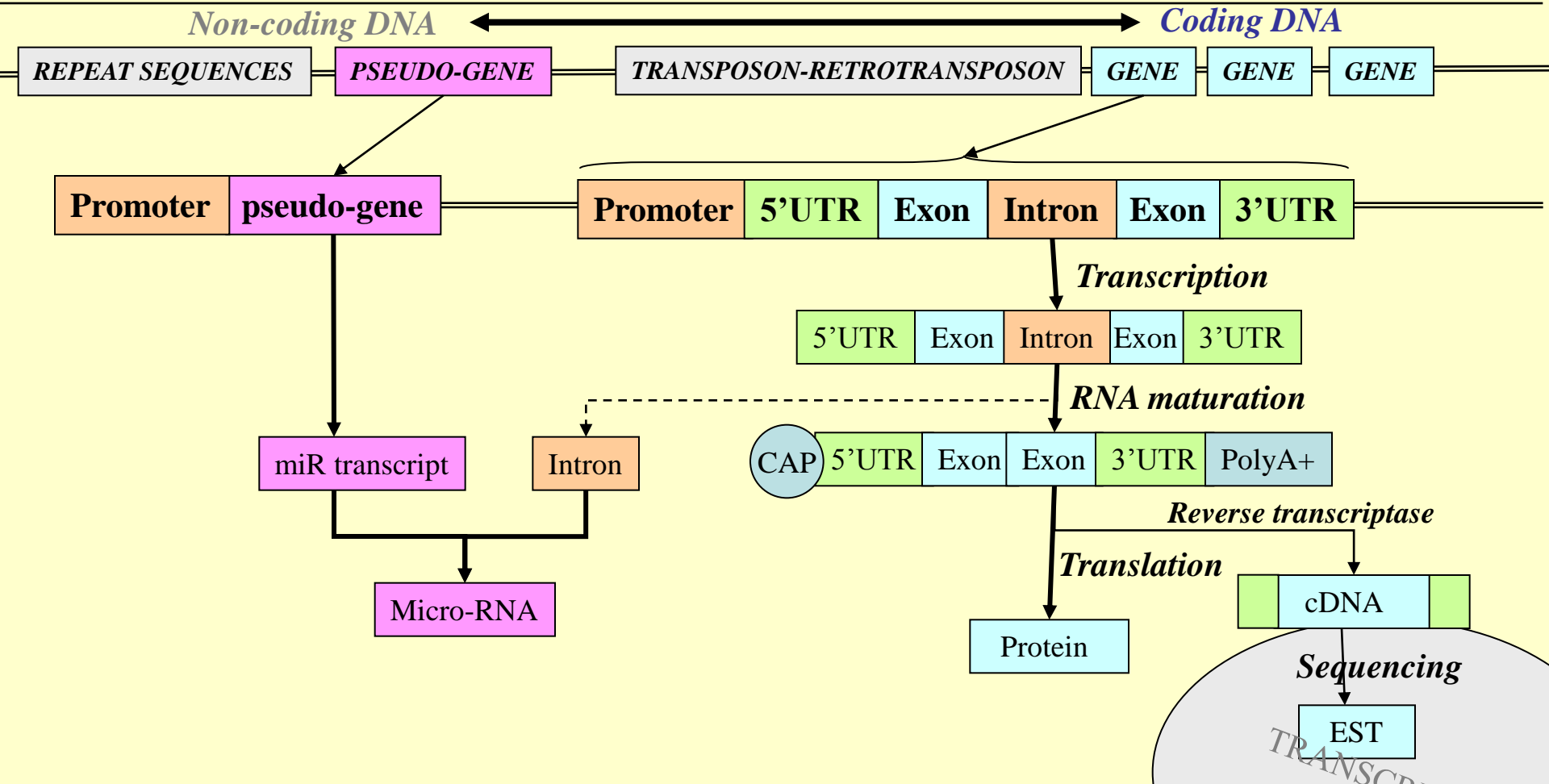
**Mitochondrial genome** (maternal origin, non-Mendelian inheritance, specific DNA code, male sterility): 57 genes, 370000 bp in Arabidopsis.

**Chloroplastic genome** (organisation in replicate nucleoids, linear or ramified)

# Nuclear genome: DNA packaging into chromosome

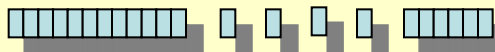












# Components of the genome



- **Transcription regulatory sequences:** Promoter, 5'UTR, 3'UTR, Intron
- **Post-transcriptional & translational regulatory miR:** pseudo-gene & intron
- **Putative function of proteins:** Exon
- **Molecular genetic markers:** SSR in non-coding and coding DNA, SNPs, etc.

# Estimated size of haploid genome in several species

• <i>Triticum aestivum</i>	15966 $10^6$ bases	
• <i>Hevea brasiliensis</i>	2100 $10^6$ bases	
• <i>Elaeis guinensis</i>	1800 $10^6$ bases	
• Cassava	780 $10^6$ bases	
• <i>Vitis vinifera</i>	483 $10^6$ bases	
• <i>Populus</i>	480 $10^6$ bases	
• <i>Oriza sativa</i>	420 $10^6$ bases	
• Castor bean	220 $10^6$ bases	
• <i>Arabidopsis thaliana</i>	150 $10^6$ bases	
• <i>M. grisea</i> (fungus)	40 $10^6$ bases	
• <i>E. coli</i>	5 $10^6$ bases	

# Sequenced Plant Genomes: some examples of initiatives

<b>Species</b>	<b>Status</b>	<b>Consortium</b>
<i>Arabidopsis thaliana</i>	Completed & annotated	International
<i>Oriza sativa</i>	Completed & annotated	International
<i>Elaeis guineensis</i>	<b>Draft &amp; annotated in 2008 *</b>	<b>Malaysia - US *</b>
<i>Populus</i>	Completed, annotation in progress	US-EU
<i>Vitis vinifera</i>	Completed, annotation in progress	EU
<i>Citrus</i>	In progress	Spain, Italy, France
<i>Cassava</i>	In progress	International
<i>Musa</i>	Funded	International
<i>Coffea</i>	Funded	US, Brazil, EU
<i>Theobroma cacao</i>	Funded	US, France
<i>Hevea brasiliensis</i>	Feasability study in June 2009 Announced in July 2009 In progress in 2009 (personal communication)	<b>IRRDB</b> <b>Malaysia MRB</b> <b>Thailand BIOTEC</b>

\* Non published data from Asiatic Centre for Genome Technology (Malaysia) & Synthetic Genomics (US)

# Evolution of Sequencing Technologies

**High-throughput sequencing technologies** are an opportunity for crop species

→ large number of sequences at a low cost but short sequences & quality (GS-FLX)



Applied Biosystems  
ABI 3730XL



Roche / 454  
Genome Sequencer FLX



Illumina / Solexa  
Genetic Analyzer



Applied Biosystems  
SOLiD

Technology	Bases/read	Read/run	Output	Cost/run
	(Nb)	(Nb)	(Mbases)	(€)
<b>Applied Systems ABI Sanger</b>	<b>1,000</b>	<b>1</b>	<b>0.001</b>	<b>3.85</b>
<b>Roche - 454 GS FLX</b>	<b>400 ↑</b>	<b>1,000,000 ↑</b>	<b>400</b>	<b>9 000</b>
<b>Illumina Solexa –Genetic Analyser</b>	<b>76 ↑</b>	<b>10,000,000 ↑</b>	<b>760</b>	<b>4800</b>
<b>Applied Systems Solid</b>	<b>35 ↑</b>	<b>2,000,000,000 ↑</b>	<b>-</b>	<b>-</b>

# The *Hevea* Genome: what is known!

## *Hevea* genus can be considered as diploid

- $2n = 2x = 36$
- 2100 Mbases / 1C in size
- $25,000 \leq \text{genes} \leq 35,000$
- Large region of genome with non-coding sequences & repeat sequences
- High heterozygosity (allogamous species)
- Genetic map with more than 700 markers



# *Hevea* transcriptomes:

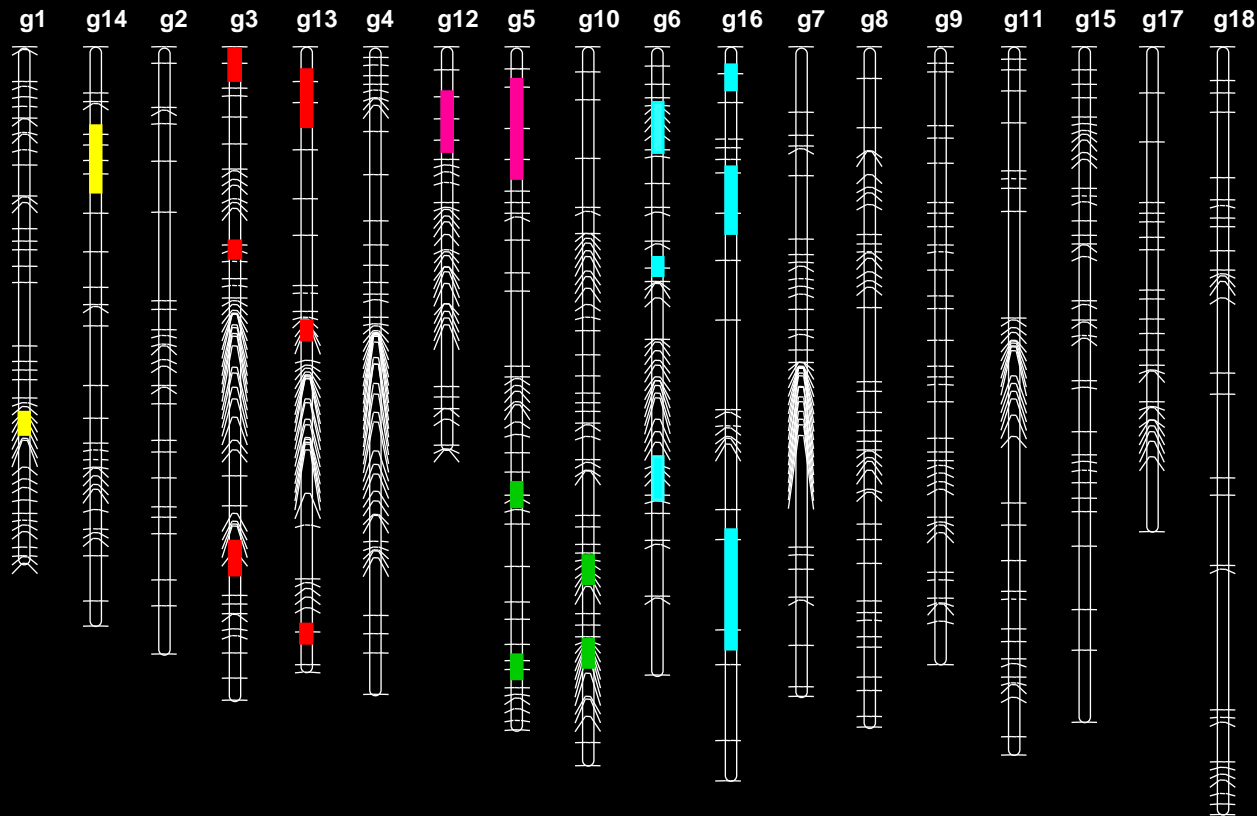
## Generated EST (Expressed Sequence Tags) in June 2009

<b>Institution</b>	<b>Clone</b>	<b>Tissue</b>	<b>Library</b>	<b>Sequencing</b>	<b>ESTs</b>
MRB	RRIM 600	latex		Sanger	35 000
MRB	RRIM600	Cambium		Sanger	500
CATAS	Reyan 973	Latex		Sanger	25 000
CATAS	?	?		Sanger	30 000
CIRAD	PB260	Leaf with Microcyclus	SSH	Sanger	4 000
CIRAD	RO38	Leaf with Microcyclus	SSH	Sanger	4 000
IRD	PB260, PB217	Latex, bark	SSH	Sanger	4 600
IRD	PB260	Bark from TPN-healthy trees	SSH	Sanger	15000
IRD	PB217	Stimulated bark, latex	SSH	Sanger	4 600
CIRAD	PB260	Bark (ethephon)	SSH	Sanger	1 200
CIRAD	PB260	Wounded bark	cDNA	Sanger	2 000
CIRAD	PB260	Bark under stresses	Colligated & nebulized	GS-FLX	35 000
UESC-CIRAD	MDF180	leaf with Microcyclus	SSH	Sanger	4 000
UESC-CIRAD	PB314	leaf with Microcyclus	SSH	Sanger	4 000
<b>TOTAL</b>	<b>&gt; 8 clones</b>	<b>6 tissues</b>			<b>249 900</b>

# Towards a high-density genetic map in *Hevea*

Allelic variation in EST sequences → SNP Single Nucleotide Polymorphisms  
2 SNP/1Mb = 4000 markers/2000 Mb

717 loci (RFLP, AFLP, microsatellites, isozymes): 1C=2000Mb, 2n=36, 2230cM



# Clone PB 260 as the best model clone for de novo sequencing

## Mapping populations

*RRIM600 x PB217* (196 individuals in 16 replicates) from RRIT

*PB260 x FX3899* (~190 individuals in 4 replicates) from CIRAD-Michelin

*PB260 x MDF180* (300 individuals in 8 replicates) from CIRAD-Michelin

*PB5/51 x IAN873* (~60 individuals) from RRIM

## Advantages of clone PB 260

- High-yielding clone
- Genetic map using clone PB 260 as parent
- Micropropagation
- Genetic transformation

# *Hevea* Genome Project: sequencing strategy

- **De novo sequencing of model clone PB 260**

**Recommendations:**

**0- Standard:** ABI Sanger (4X coverage) and GS-FLX (15X)

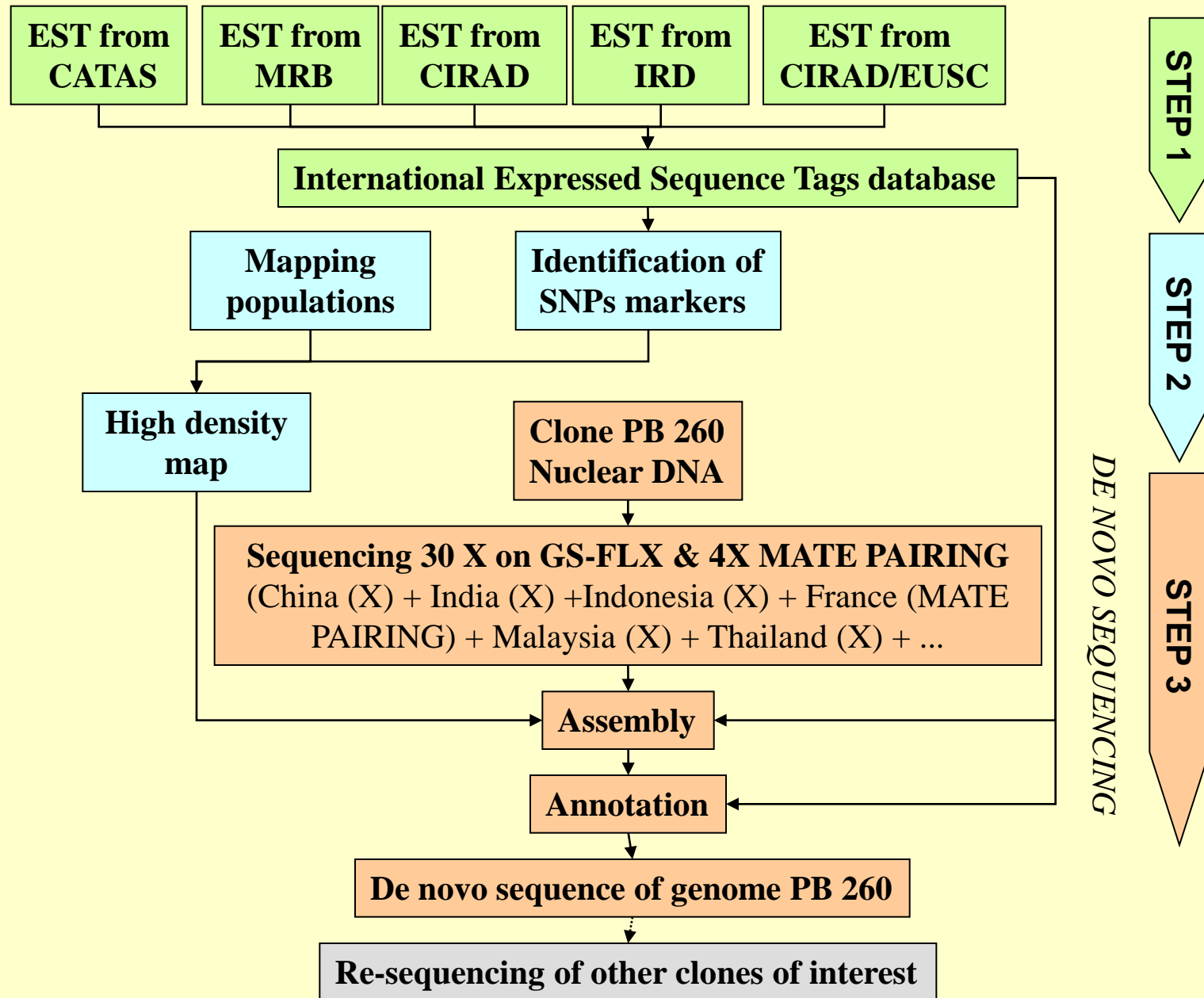
**1- IRRDB Workshop:** GS-FLX (30X coverage)

**2- Recent strategy:** GS FLX (12X coverage: 24 GB, 60 runs) +  
Illumina seq of MATE pair library 3 kb (2x), 10 kb (4x), 20 kb (4x)

- **Assembly of de novo sequences** using **high density map** with 2  
SNP markers for 1 Mb

- **Re-sequencing of selected clones** with GS-FLX (15X coverage) +  
Illumina (10X coverage).

# The strategy of the *Hevea* Genome Project



# The *Hevea* Genome Project

## Scientific outcomes & applications

---

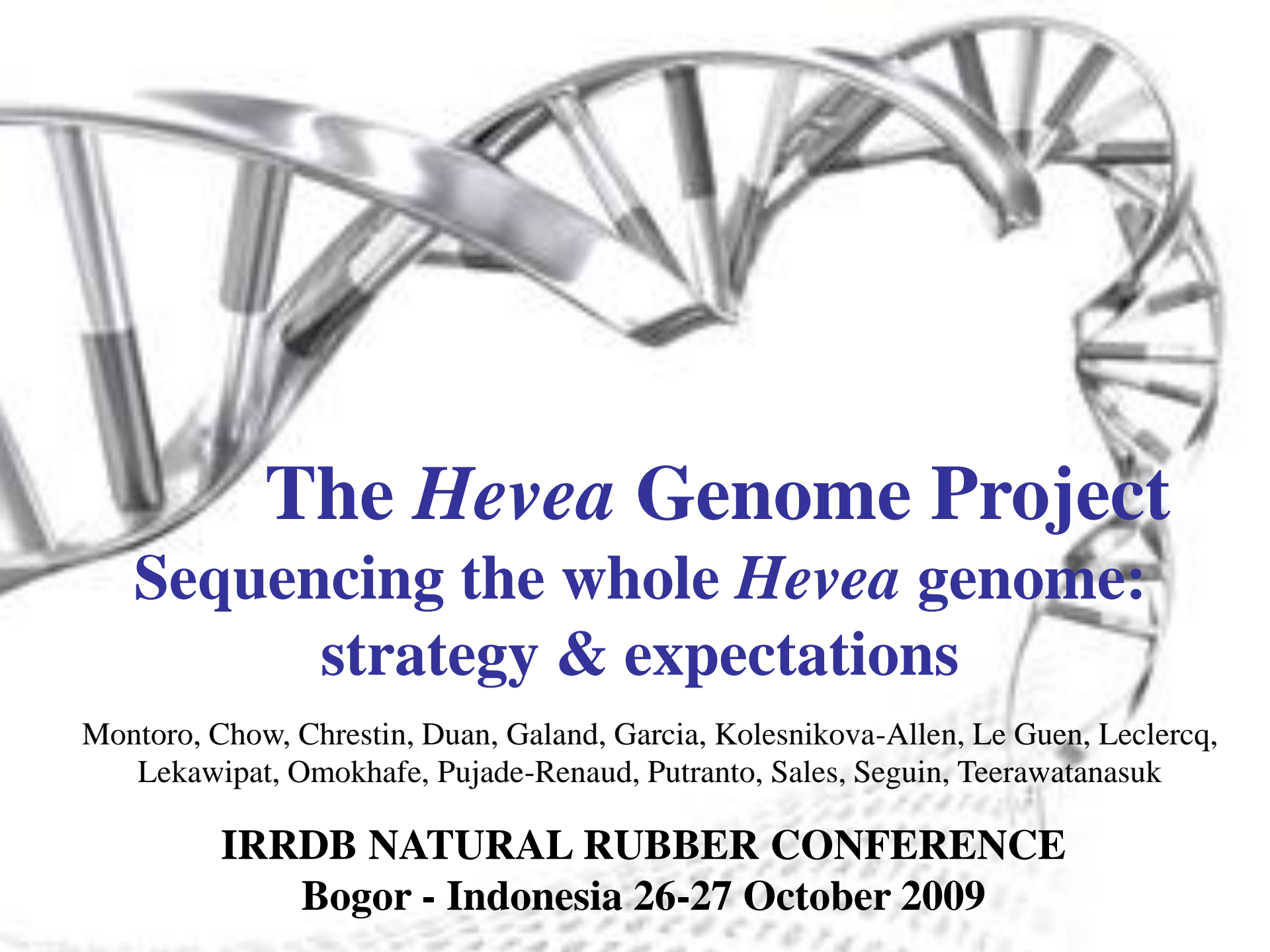
---

- Comparison of genomes
  - Physical genome map
  - Gene discovery
  - DNA chips for high-throughput gene expression analyses
- 
- Dynamics of the genome evolution
  - Model of gene expression control upon plant development & stresses
  - Model of metabolic pathways (rubber biosynthesis, plant defence, etc.)
  - Identification of genes underlying QTLs (Quantitative Traits Loci)
- 
- **Molecular Breeding**
  - **Genetic Engineering**

# The *Hevea* Genome Project

## Questions to be answered

- **Participants to the International Consortium**
  - Scientific & funding organizations
- **Evaluation of the budget**
  - Cost of sequencing is reducing
  - Bioinformatics for the genome assembly & structural and functional annotation
  - Capacity building (training, scientific exchanges, etc.)
- **Define the policy for releasing sequence information (scientific programmes & applications):**
  - Intellectual property rights: public vs private databases?
  - Location and accession to the International EST & genome sequences databases?



**The *Hevea* Genome Project**  
**Sequencing the whole *Hevea* genome:**  
**strategy & expectations**

Montoro, Chow, Chrestin, Duan, Galand, Garcia, Kolesnikova-Allen, Le Guen, Leclercq,  
Lekawipat, Omokhafa, Pujade-Renaud, Putranto, Sales, Seguin, Teerawatanasuk

**IRRDB NATURAL RUBBER CONFERENCE**  
**Bogor - Indonesia 26-27 October 2009**