Survey of BAC clone ends as a preliminary step toward the genome sequencing of coffee tree (*Coffea* L.)

P. Lashermes, IRD - Montpellier

AGENCE NATIONALE DE LA RECHERCHE ANR

Colloque Plant Genomics 2012

Le caféier = un modèle d'étude de grande importance pour le Sud

- Une activité agricole importante (> 10 millions d'hectares) dans plus de 50 pays tropicaux




- Activité familiale (> 70% de la production provient d'exploitations de moins de 5 hectares) et génératrice d'emplois (> 80 millions de personnes)

- En relation avec l'agroforesterie, rôle majeur dans la conservation de la biodiversité, la protection des bassins versants, le bilan carbone, et la mise en valeur des territoires

Le caféier = une culture stratégique dans la France d'outre-mer
(DROM-COM, Départements et régions d'outre-mer - Collectivités d'outre-mer)

DOMAINE IDA-MARC
Café de Nouvelle-Calédonie

- Développement de filières de qualité

WHY DO WE WANT TO SEQUENCE COFFEE GENOME?

✓ Genetic variation in wild coffee accessions is considerable and still largely unexplored

✓ Sequencing the coffee genome will help decipher the genetic and molecular bases of important biological traits

✓ The coffee genome would represent the first Rubiaceae genome sequenced, one of the largest plant family (> 12 000 species)

../..

WHAT ARE THE BENEFITS FROM THIS INITIATIVE TO THE COFFEE SECTOR?

• Decipher the **genetic and molecular bases of important biological traits** in coffee that are relevant to growers, processors, and consumers.

• Development of optimized characterization and conservation strategy for enhanced utilization of *Coffea* germplasm resources in **breeding programs**.

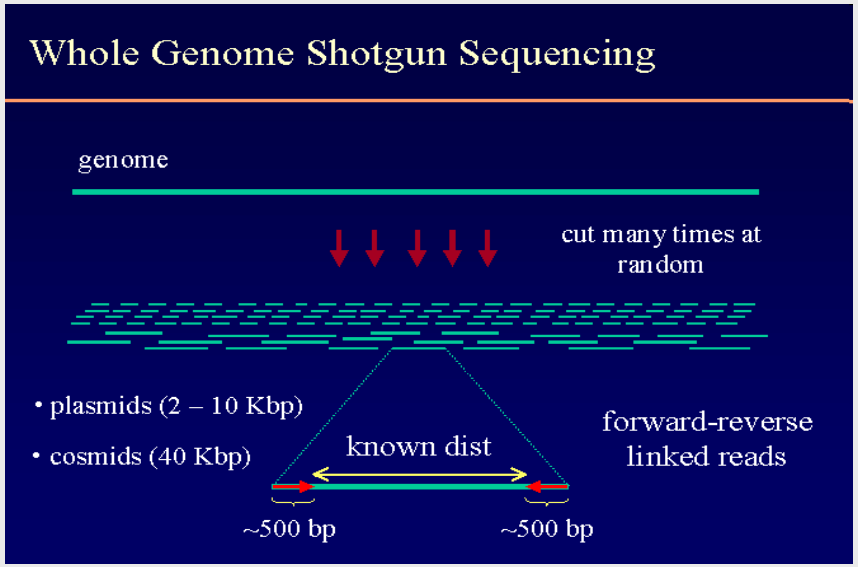
• Ensure long term **sustainable coffee production** (from environmental, social, and economical point of views) in relation to **climate changes**

• Possibility of **innovation** in terms of **enhanced quality** to increase consumer satisfaction and guarantee coffee supply (in term of quality/quantity/cost).

Overall strategy

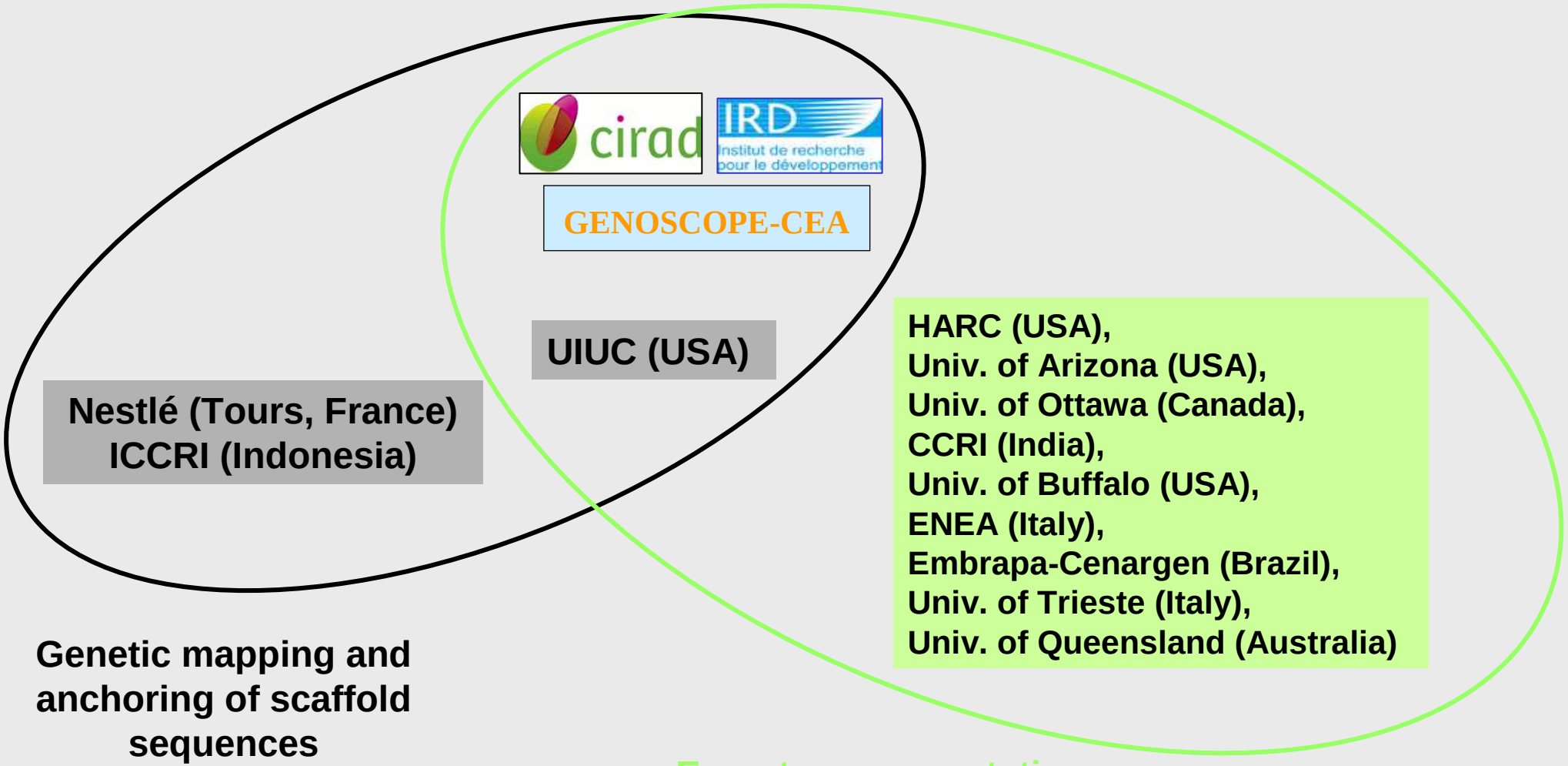
✓ **Use of the Canephora genome has reference sequence**

- Diploid genome of about 710 Mb
- Homozygous genotype (Doubled Haploid 200-94)



✓ **Combine** different high throughput novel **sequencing technologies** (i.e. 454, Illumina) and conventional Sanger technology to generate shotgun reads and **multi-span paired end reads**

An international initiative with a strong French leadership



ANR support to the French participation

Phase 1 – Project « GenomeCafe »

Survey of BAC clone ends as a preliminary step toward the genome sequencing of coffee tree (*Coffea* L.)

***Coord.* P. Lashermes**

UMRs DAP, DIA-PC, GDP, RPB

T. Leroy, A. De Kochko, R. Guyot

Phase 2 – Project « CoffeaSeq »

Sequencing the coffee tree genome (Coffea canephora)

***Coord.* P. Wincker**

UMRs AGAP, DIADE, RPB

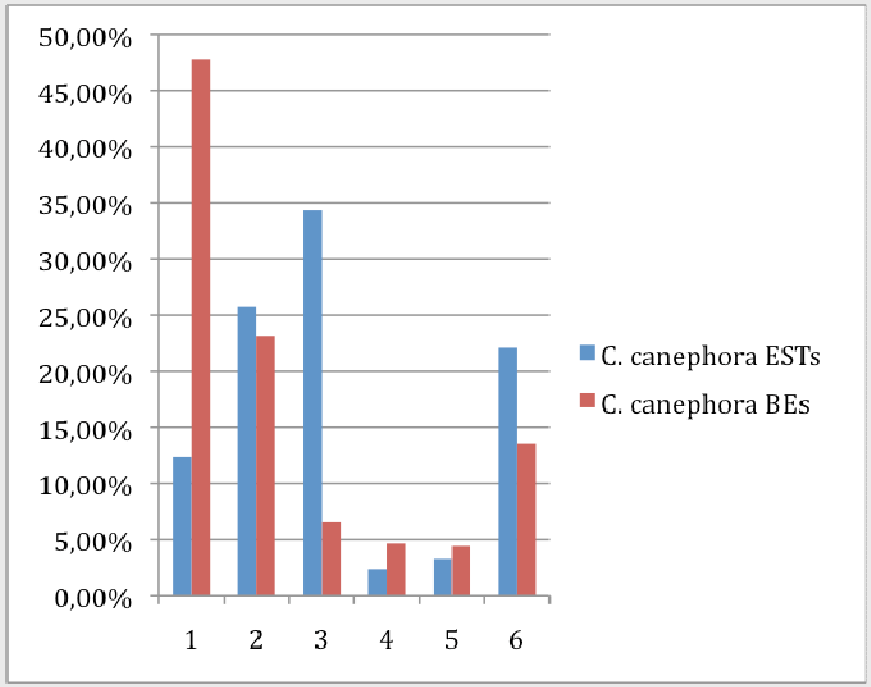
G. Droc, C. Campa, P. Lashermes

Outcome – Project « GenomeCafe »

✓ **Construction of two BAC libraries** (*Hind* III and *Bst*YI) from DH200-94, 73728 clones > 11X

✓ **Production of 131 412** high quality **nuclear sequences** by Sanger BAC-end sequencing (0.14 X)

✓ **Detection of 10 094 SSR** within 6.7% of BESs



SSR motif	*SSR* number	*% SSR*
Dinucleotide	**2333**	**100**
GA-CT	727	31.2
AT-TA	1342	57.5
CA-GT	264	11.3
GC-CG	0	0
trinucleotide	**652**	**100**
CCT-GGA	39	6
AAG-TTC	159	24.4
AGC-TCG	17	2.6
CCA-GGT	23	3.5
AGT-TCA	46	7.1
GAC-CTG	21	3.2
CTA-GAT	33	5.1
GGC-CCG	4	0.6
AAC-TTG	27	4.1
AAT-TTA	283	43.4

../..

✓ **Development of 1420 SSR-based PCR markers** for mapping and genetic diversity analyses

✓ **Coding region estimation using homology with a *Coffea spp.* EST :**

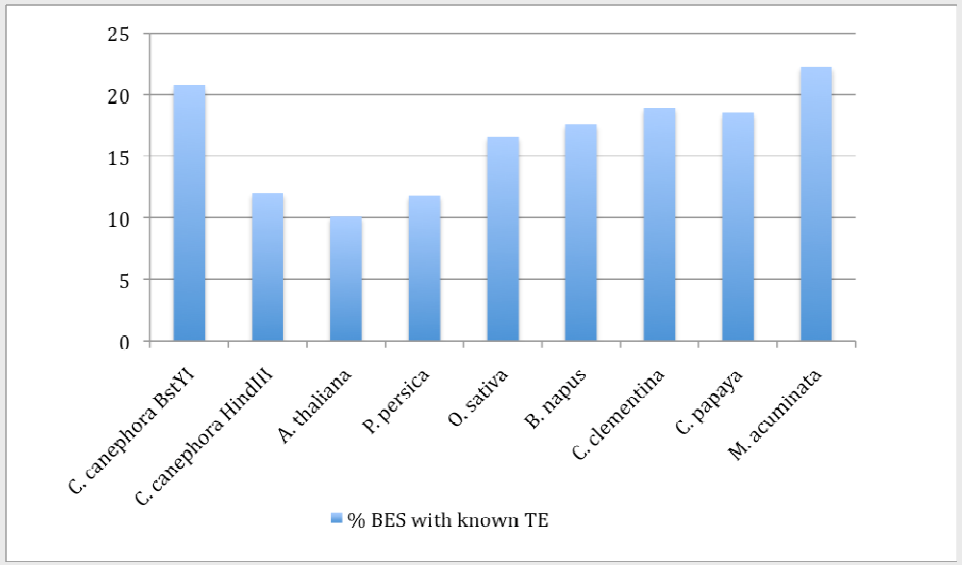
20.3 % of the cumulative length of BESs / estimate of 19800 genes

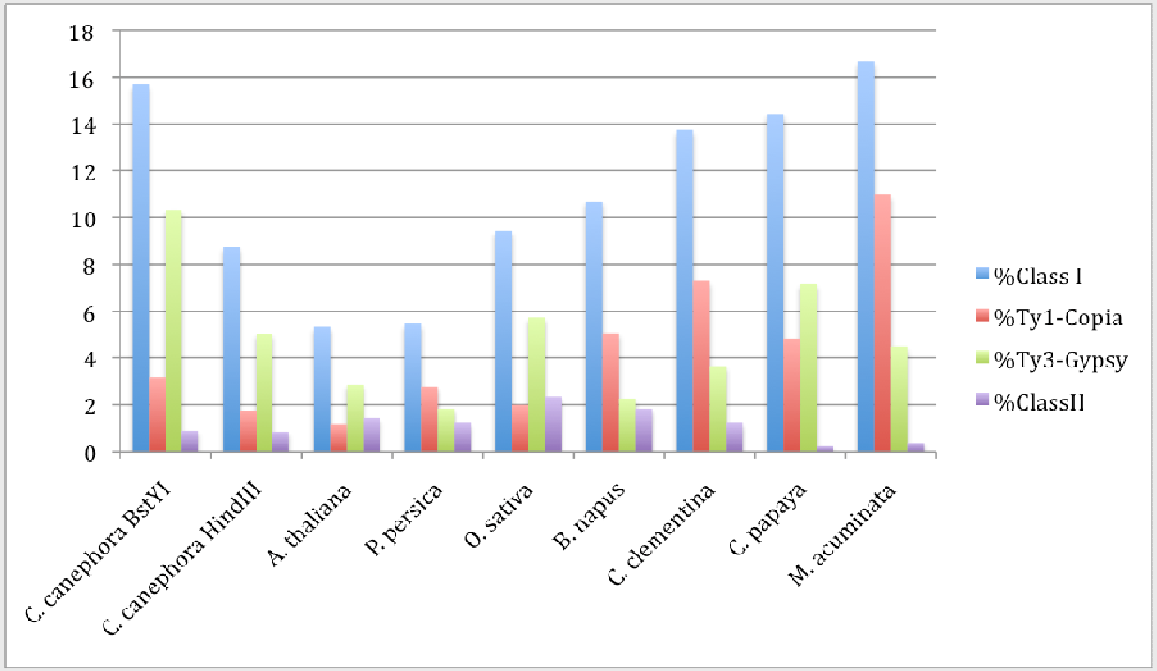
../..

✓ Identification of known repetitive DNA in the BESs

Transposable Elements proteins database from Repbase - Censor (Kohany et al. 2006)

- ➢ 15 % of the genome / 93% belong to the Class I LTR-RT



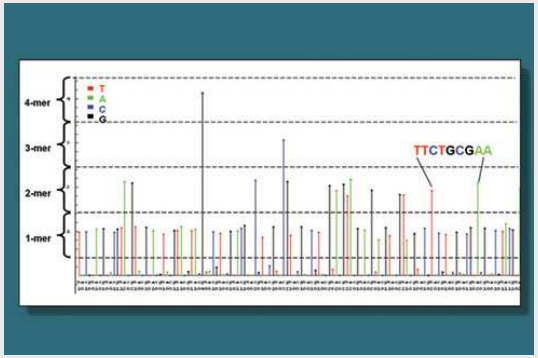


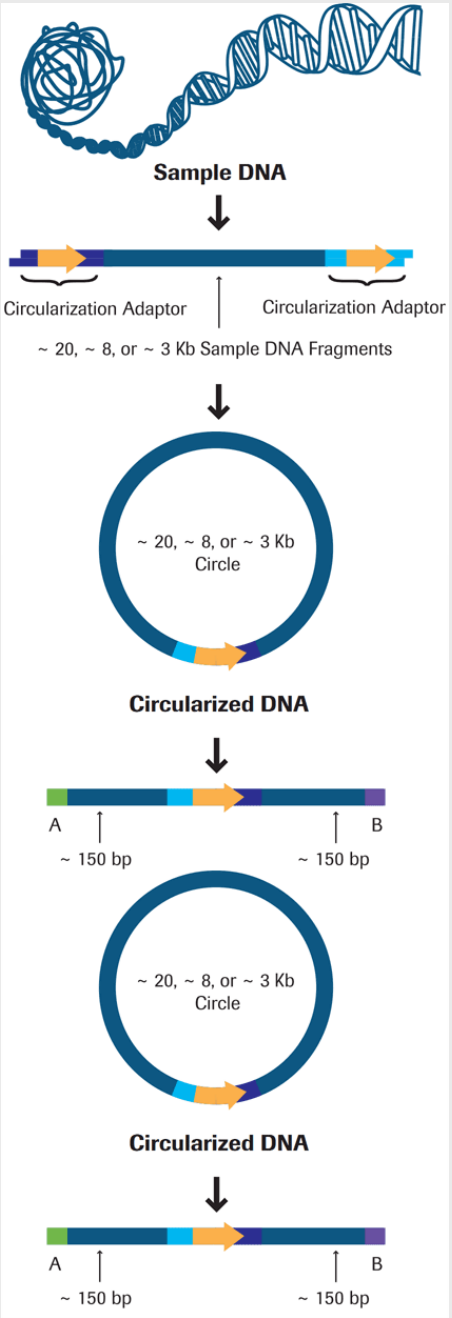
✓ In relation with the ongoing CoffeaSeq project :

- ➢ Development of a *C. canephora* TE database (> 700 contigs)
- ➢ Mapping of 320 SSR markers/SNP derived from BESs on the high-density genetic map

Draft assembly – Project CoffeaSeq

Sequencing data





Séquençage 454 : 28.9 X

Shotgun reads and multi-span paired end reads

– Reads single end : 14.8 X (mean size : 359 bp)

– Long reads single end: 8.2 X (mean size : 462 bp)

– Mate paired reads : 5.8 X (mean size : 252 bp)

2.2 X from library 8 kb
3.6 X from library 20 kb

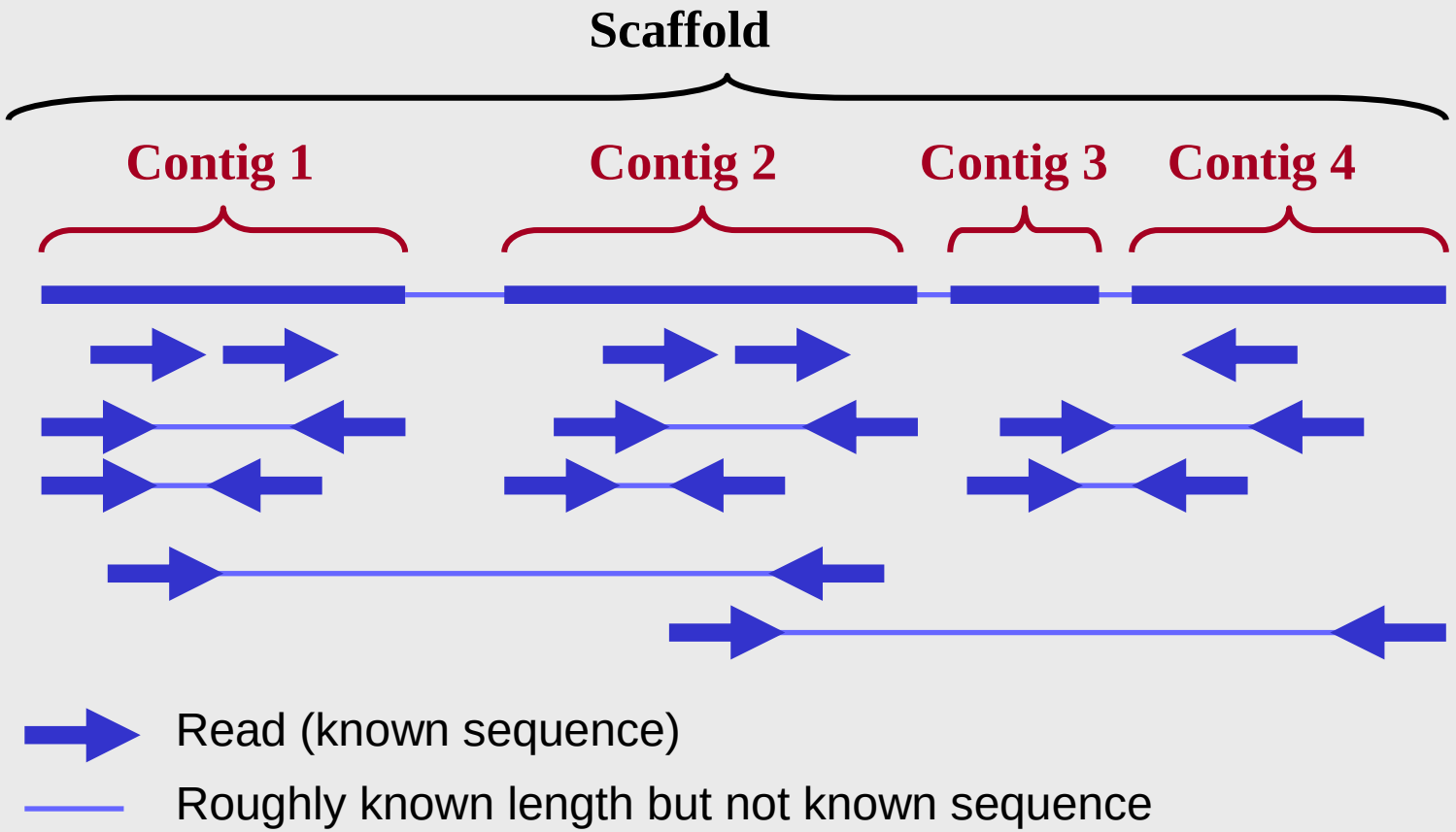
Séquençage Sanger, BAC ends: 0,14 X

From two libraries, *Hind* III and *Bst*YI (73728 clones)

Mean size: 1 350 pb, 63,2kb < insert < 253,6 kb

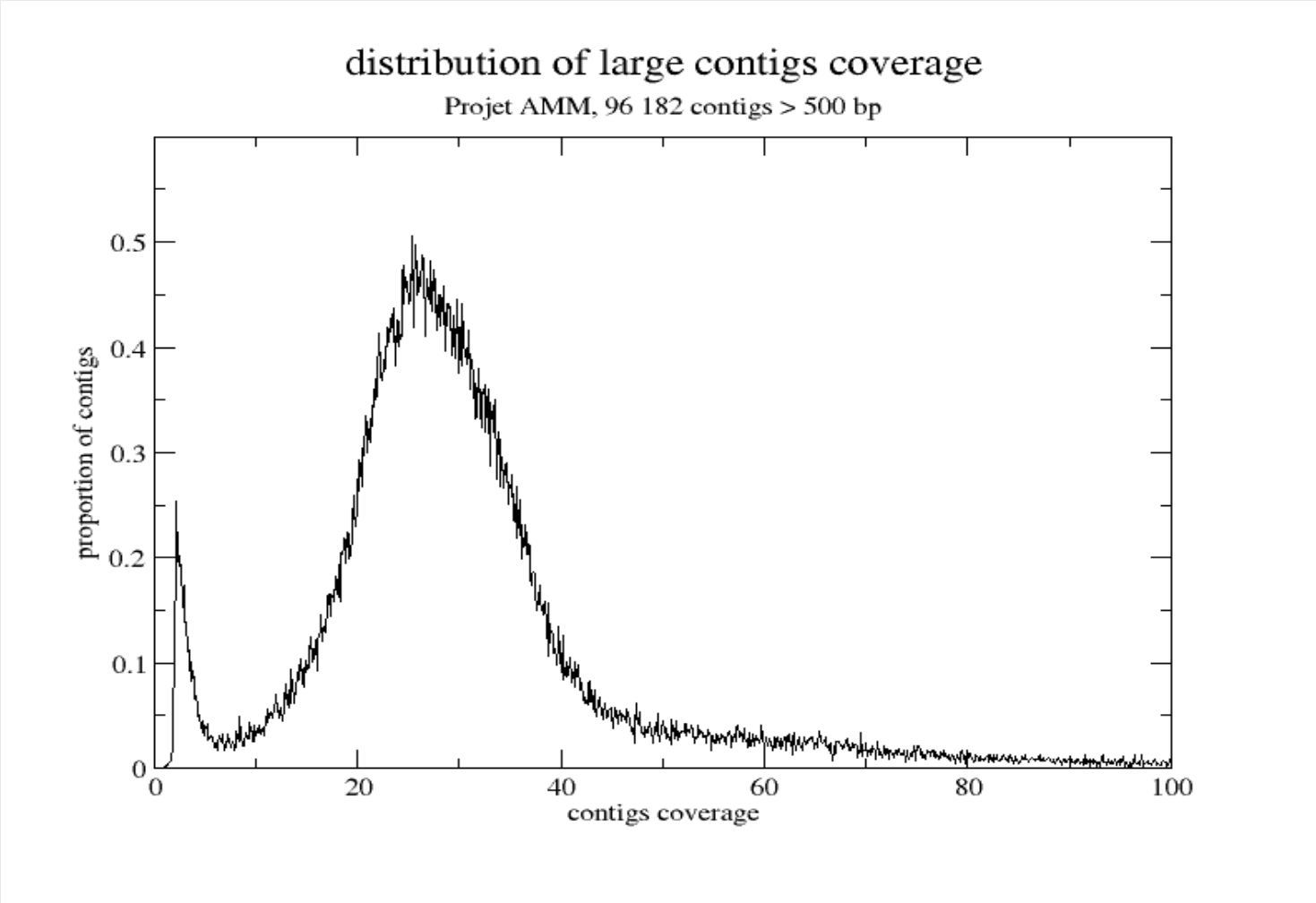
Assembled with Newbler : hybrid assembly combining long GS FLX+ shotgun reads, multi-span paired end reads and short read data

Scaffolds are composed of contigs and gaps



Draft assembly

211 157 contigs including 96 182 contigs > 500 pb



Elimination contigs < 10 X (blast/homology with *Pseudomonas spp.* and *Homo sapiens*)

Assembly correction and gap closure

Sequencing data

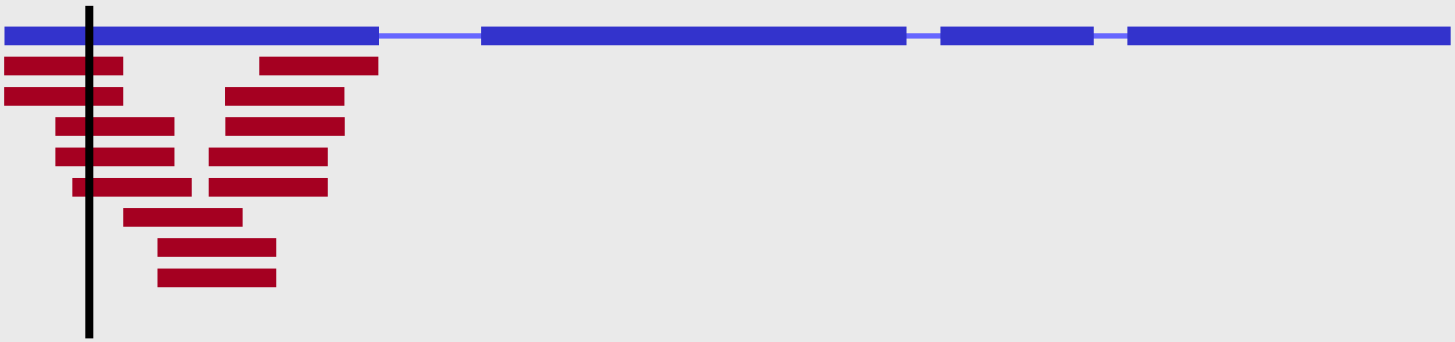
Solexa data : 69.7 X

– Reads single-end : 7.3 X
size : 76 bp (4.8X) or 150 bp (2.5X)

– Paired-end reads : 62.4 X
size : 76 bp (42.4 X) or 108 bp (20 X)

Assembly correction

Mapping with BWA, 8 cycles of error detection

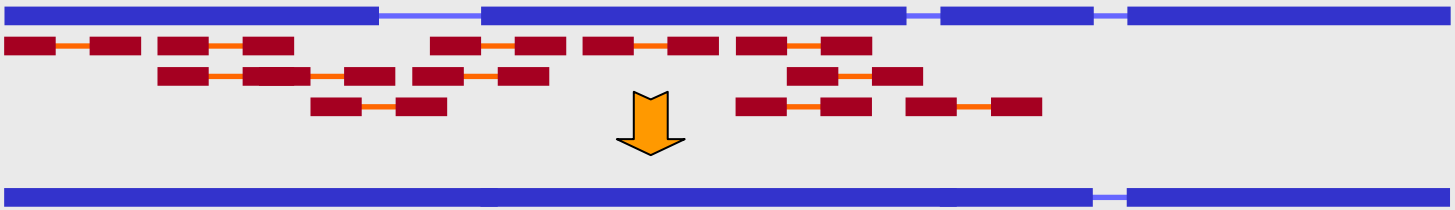


Number of base corrected : 203 018

../..

Gap closure

Gapcloser



✓ Number of contigs ↘ Contig size ↗

	Nb contig	Average size	N50
Cycle 0	43501	9955	18449
Cycle 1	29249	15600	39307
Cycle 2	26685	17458	46294
Cycle 3	25784	18214	48815
Cycle 4	25460	18511	50048

✓ Number of N (undetermined nucleotide) ↘

	Assembly size	N	%
Cycle 0	569 434 129	136 371 931	23.95
Cycle 1	568 745 541	112 442 955	19.77
Cycle 2	568 615 132	102 735 923	18.07
Cycle 3	568 587 531	98 941 642	17.40
Cycle 4	568 577 705	97 270 347	17.11

Assembly statistics

✓ Size statistics of scaffolds

	Nb	Cum	Moy	N50	N80	N90	max
Scaffolds	13 345	568 577 705 (80% genome)	42 606	1 260 636 #108	65 268 #635	21 947 #2 191	9 027 918

✓ Size statistics of contigs

	Nb	Cum	Moy	N50	N80	N90	max
Contigs	25 216	471 313 922 (66% genome)	18 691	51 132 #2 290	15 527 #7 259	6 989 #11 767	817 605

✓ Size statistics of gaps

	Nb	Cum	Moy	min	max
Gaps	11 871	97 263 783 (17% assembly)	8 193	25	817 605

Anchoring of scaffolds to the genetic map

Construction of pseudomolecules for each of the 11 chromosomes

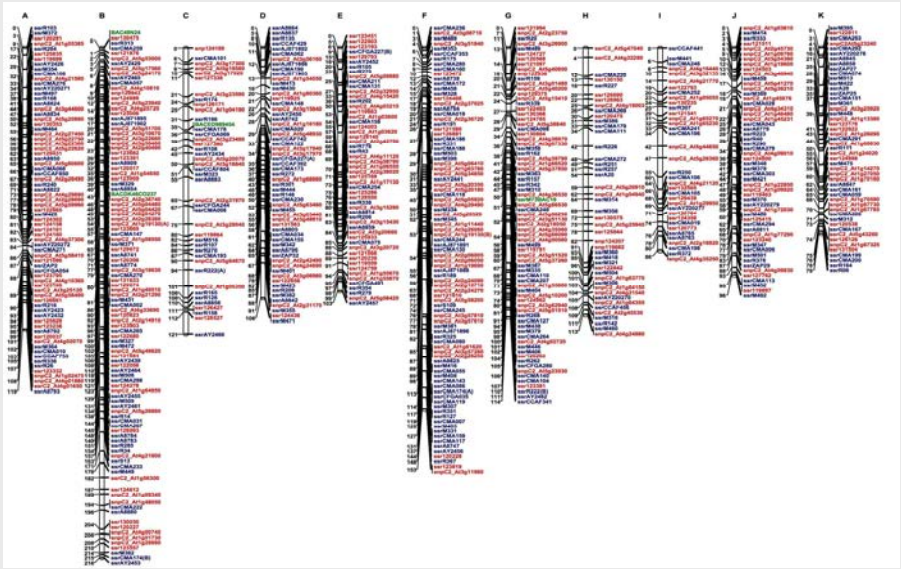
✓ **Based on the population BP409 X Q121 (pseudo-F1 of 93 individuals) developed by the Indonesian Coffee and Cocoa Research Institute (ICCRI)**

✓ **Development of a high-density genetic map**

1200 available markers (Genomic SSR, COSII, EST-SSR, RFLP) from the pioneer work of Nestlé

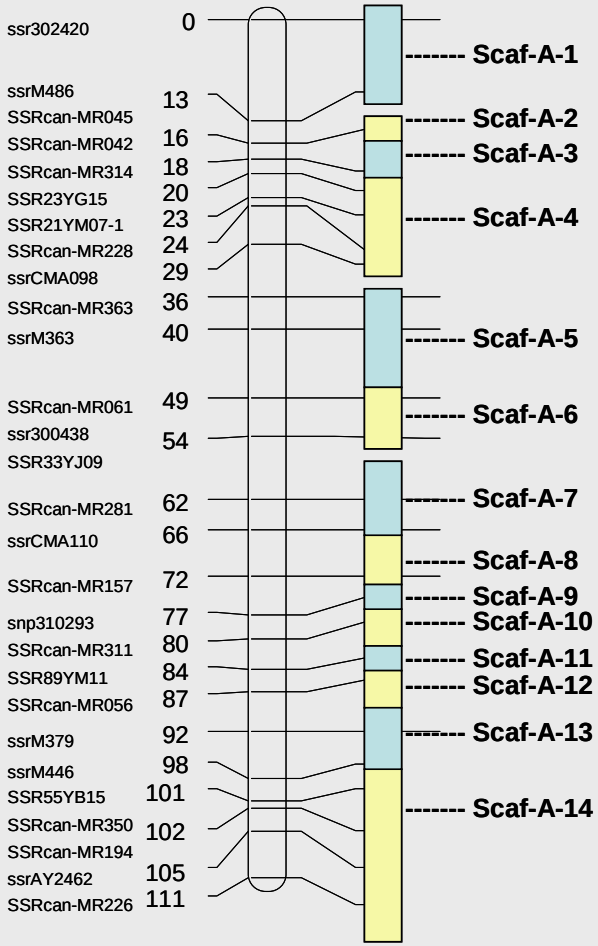
400 markers (SSR, SNP) derived from BAC-ends and targeted scaffold sequences

3200 RAD-seq markers (Double-digest "*Nsi* I and *Mse* I" genome reduction, Illumina sequencing) generated by UIUC-USA



../..

✓ Anchoring of scaffolds on the high-density genetic map



✓ Ongoing work/objectives :

> 80 % of the assembly anchored on chromosomes

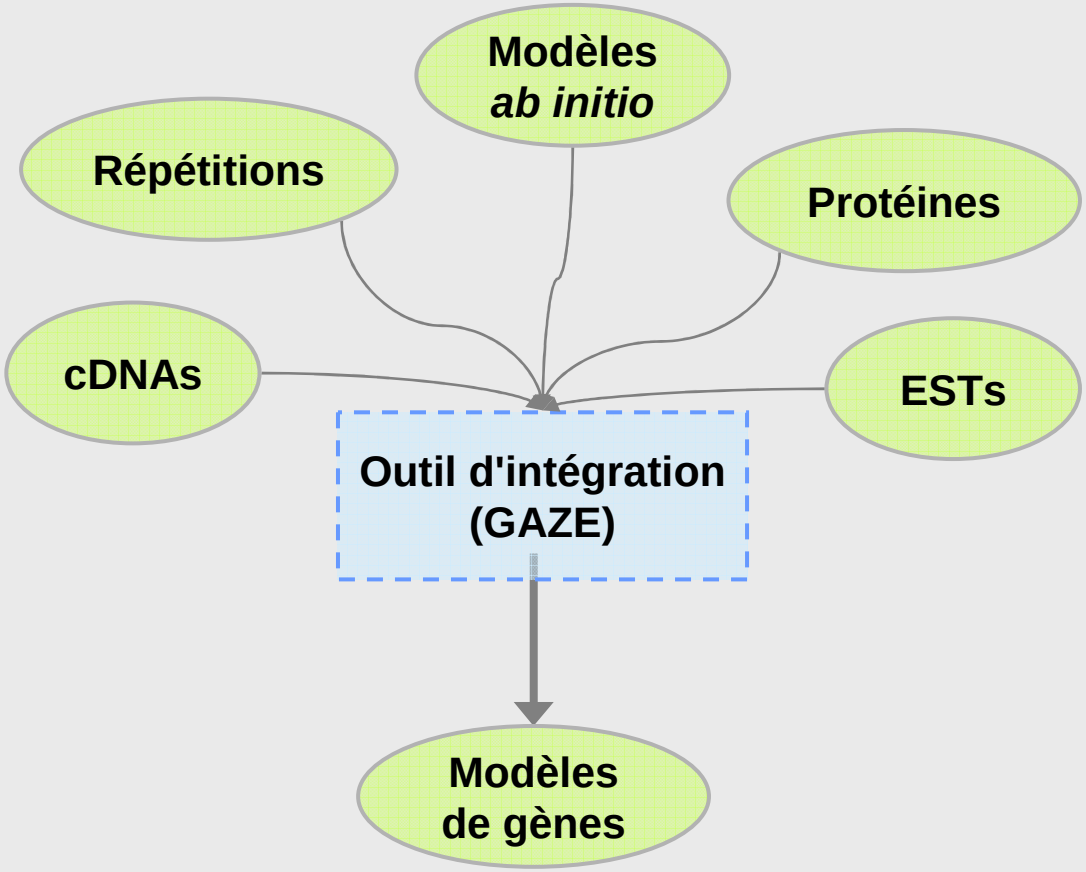
> 60 % of the assembly anchored and oriented on chromosomes

> 95 % of the unigene ressources anchored on chromosomes

L'annotation automatique

Recherche des structures exons-introns sur des séquences issues d'un assemblage dans le but de définir un ensemble de modèles de gènes de référence

Intégration des données : L'objectif est de construire des modèles de gènes qui tiennent compte de l'ensemble des données collectées (cDNAs, ESTs, alignements protéiques, prédictions *ab initio*), selon l'importance de chacune.



GAZE est une méthode généraliste d'intégration de données pour la prédictions de modèles de gènes utilisant la **programmation dynamique** (R. Durbin, K. Howe & T. Chothia).

Un **automate** décrit la structure d'un modèle de gène.

Les données sont représentées suivants 2 types d'éléments :

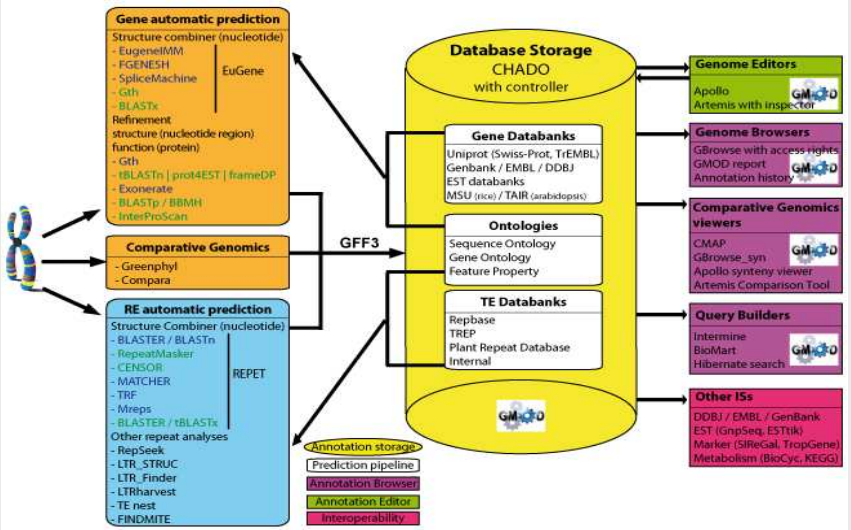
- des **segments** (exons, intron,…)
- des **signaux** (start, GT/AG, stop,…)

Notion de **poids** selon la nature des données.

L'annotation experte

GNPannot

Community annotation system



Ongoing activities

- **Transposable elements, miRNA, overall gene content…**
- **Disease resistance, Lipid biosynthesis, Phenylpropanoid, Ethylene, Caffeine, …**
- **Comparative genomics, Genome evolution and organisation**
- **Relationship between genetic and physical distances**

Acknowledgements

Phase 1 – Project « GenomeCafe » ***Coord.*** **P. Lashermes**

UMRs DAP, DIA-PC, GDP, RPB

T. Leroy, A. De Kochko, R. Guyot

Phase 2 – Project « CoffeaSeq » ***Coord.*** **P. Wincker**

GENOSCOPE-CEA

UMRs AGAP, DIADE, RPB

G. Droc, C. Campa, P. Lashermes

Thank you very much for your attention