Towards a Better Understanding of the Coffea Arabica Genome Structure

The Arabica Coffee Genome Consortium

An international consortium was initiated in 2012 with the goal to perform the sequencing of the *Coffea arabica* genome. This consortium includes 34 researchers, engineers and technicians coming from 13 institutions in six different countries.

In December 2013 the first draft genome of a *Coffea* species was published in Science (Denoeud et al. 2014), it is the genome of the *C. canephora* species, a diploid and the second mainly cultivated species after *C. arabica*. This later species is the only tetraploid of the genus resulting from a recent spontaneous hybridization (± 0,5 Mya) between *C. eugenioides*, a wild species from East Africa, and *C. canephora*, whose genome was sequenced (Lashermes et al. 1999).

Et39 di-haploid (n=22)

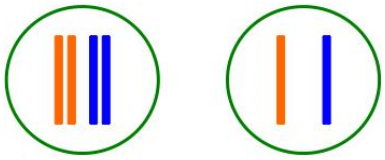

Figure 1. Et 39 dihaploid.

The genome of *C. arabica*, an allotetraploid (2n=4x=44), is about 1.3 gigabase in size (Cros et al. 1995). Its two parental genomes are closely related and their genomes have a high degree of identity. This fact makes a separated assembly of the two sub-genomes quite difficult. In order to overcome in part this difficulty, we chose to sequence a dihaploid plant produced by IRD issued from the Et39 accession.

A dihaploid has only one set of chromosomes from each subgenome (Fig. 1) in this case the sequenced *C. arabica* plant has only 22 chromosomes. This situation allows not to be

embarrassed by the polymorphism that may exist within each subgenome. Choosing this genotype should facilitate the assembly of such a tetraploid genome and differentiate more easily the two subgenomes.

SEQUENCING STRATEGIES

Two main whole genome sequencing (WGS) strategies were used. Both adopted a shut gun approach; using the short reads provided by the Illumina technology for the first one, or taking advantage of the long reads provided by the Pacific Bioscience SMRT sequencing platform for the second one. A very small amount of sequencing was also performed with the Roche 454 chemistry. The final coverage obtained with short reads reached about 164x (Table 1).

Table 1. Sequencing results for short reads.

Technology	Insert Size	Length (bp)	Coverage
Illumina paired-end	400 bp	100 x 2	80 x
Illumina mate pair	3 kb	100 x 2	45 x
Illumina mate pair	8 kb	100 x 2	20 x
Illumina MiSeq	500 bp	~450	17 x
454	500 bp	~393	2 x
Total coverage	-	-	164 x

The long reads from Pacific Bioscience, with two different chemistries (P4C2 and P5C3), gave a total coverage of about 56x , with the longest reads reaching 40,445 bp.

GENOME ASSEMBLY

An assembly was performed on the Pacific Bioscience long reads using the Falcon assembler. The final assembly, made exclusively by contigs, covers almost 80% of the estimated *C. arabica* genome size, i.e. 1.042 Gb out of 1.3 Gb. An independent assembly was also conducted with the Illumina short reads. It appears that, at the contig level, the total proportion of the assembled genome is roughly identical to that obtained with the long reads, but, as expected, the general results are of lower quality (Table 2).

Table 2. Assembly results for short and long reads, independently.

	Short Read	PacBio
Total assembly size (bp)	1,031,405,416	1,042,371,195
# of sequences	119,185	9,840
Longest Sequence	55,056	4,837,614
N50 (bp)	8,810	267,399

GENETIC MAPPING:

A segregation population of 138 F2 plants issued from a cross between two wild Ethiopian genotypes collected by IRD (de Kochko et al. 2010) was used to build a genetic map, which contains 613 SSR markers and covers 4946 cM. Some gaps still remain in the map, these are gradually removed by transposing SSR markers detected in the *C. canephora* genome (Figure 2). This genetic map is under completion using SNP markers and will be used to anchor the genome sequence.

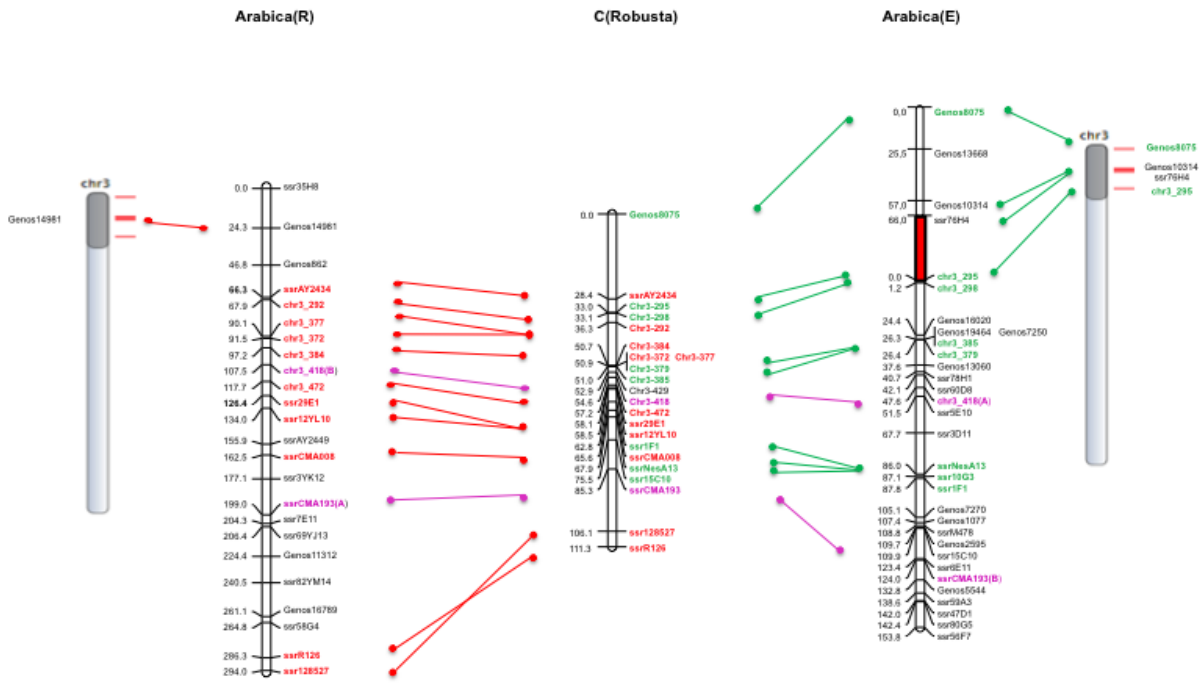

Figure 2. Linkage group C (3) from the *C. arabica* genetic map showing both sub-genomes and compared with the same LG from the *C. canephora* map. Some markers were also positioned on the *C. canephora* physical map.

ORIGIN OF *C. ARABICA*

Previous study based on SSR and GBS analyses of *C. canephora*, *C. eugenioides* and *C. arabica* germplasm indicated that among the genetic diversity groups of *C. canephora*, the most probable *C. canephora* ancestor was originating from the Ugandan genetic group (Poncet et al. personal communication). In the frame of our project, a *C. canephora* genotype from this Ugandan group is also sequenced together with a selected *C. eugenioides* plant.

Comparison between the genomic data from these three *Coffea* species, should give new and precious insights on the exact origin of the tetraploidization leading to the emergence of *C. arabica*.

C. ARABICA RESEQUENCING AND DIVERSITY STUDY

Thirty *C. arabica* accessions, wild and cultivated, selected based upon their genetic diversity and representative of the entire species, were chosen to be re-sequenced using the Illumina short reads technology. This study should reveal eventual neo-diversification emerging in the cultivated varieties vs. their wild relatives.

The goal of this sequencing project is to produce a high quality genome and develop tools that will make the finished genome accessible and useful to breeders and researchers. Results of these efforts will be published and publicly available on a specific website and in the MoccaDB database (http://moccadb.mpl.ird.fr).

The members of the Arabica Coffee Genome Consortium (ACGC) are:

MUELLER Lukas[1], STRICKLER Suzy[1], DOMINGUES Douglas[2], PEREIRA Luiz[2], ANDRADE Alan[3], MARRACCINI Pierre[3], MING Ray[4], WAI Jennifer[4], ALBERT Victor[5], GIULIANO Giovani[6], FIORE Alessia[6], PIETRELLA Marco[6], APREA Giuseppe[6], DESCOMBES Patrick[7], MOINE Déborah[7], GUYOT Romain[8], PONCET Valérie[8], HAMON Perla[8], HAMON Serge[8], TRANCHANT Christine[8], COUTURON Emmanuel[8], de KOCHKO Alexandre[8], LEPELLEY Maud[9], BELLANGER Laurence[9], MEROT-L'ANTHOENE Virginie[9], VANDECASTEELE Céline[9], RIGOREAU Michel[9], CROUZILLAT Dominique[9], PASCHOAL ROSSI Alexandre[10], SANKOFF David[11], ZHENG Chunfang[11], KUHN Gerrit[12], KORLACH Jonas[12], CHIN Jason[12].

[1] Boyce Thompson Institute, Cornell University, US
[2] IAPAR, Brazil
[3] EMBRAPA, Brazil
[4] Illinois University, US
[5] Buffalo University, US
[6] ENEA, Italy
[7] NIHS, Switzerland
[8] IRD, France
[9] Nestlé R&D Center, France
[10] UTFPR, Brazil
[11] University of Ottawa, Canada
[12]Pacific Bioscience, USA.

REFERENCES

Cros J, Combes MC, Chabrillange N, Duperray C, Monnot des Angles A, Hamon S (1995) Nuclear DNA content in the subgenus *Coffea* (*Rubiaceae*): inter- and intra-specific variation in African species. Canadian Journal of Botany 73: 14-20

de Kochko A, Akaffou S, Andrade AC, Campa C, Crouzillat D, Guyot R, Hamon P, Ming R, Mueller LA, Poncet V, Tranchant-Dubreuil C, Hamon S, Jean-Claude K, Michel D (2010) Advances in Coffea Genomics. In: M.Delseny, Kader J (eds) Advances in Botanical Research Vol 53. Academic Press Ltd-Elsevier Science Ltd, pp 23-63

Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, Zheng C, Alberti A, Anthony F, Aprea G, Aury J-M, Bento P, Bernard M, Bocs S, Campa C, Cenci A, Combes M-C, Crouzillat D, Da Silva C, Daddiego L, De Bellis F, Dussert S, Garsmeur O, Gayraud T, Guignon V, Jahn K, Jamilloux V, Joët T, Labadie K, Lan T, Leclercq J, Lepelley M, Leroy T, Li L-T, Librado P, Lopez L, Muñoz A, Noel B, Pallavicini A, Perrotta G, Poncet V, Pot D, Priyono, Rigoreau M, Rouard M, Rozas J, Tranchant-Dubreuil C, VanBuren R, Zhang Q, Andrade AC, Argout X, Bertrand B, de Kochko A, Graziosi G, Henry RJ, Jayarama, Ming R, Nagai C, Rounsley S, Sankoff D, Giuliano G, Albert VA, Wincker P, Lashermes P (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345: 1181-1184

Lashermes P, Combes MC, Robert J, Trouslot P, D'Hont A, Anthony F, Charrier A (1999) Molecular characterization and origin of the Coffea arabica L. genome. Molecular and General Genetics 261: 259-266