Sequencing the Coffee Genome: Overall Strategy and Progress Made in the Frame of ICGN

P. LASHERMES[1], X. ARGOUT[2], D. CROUZILLAT[3], PRIYONO[4], M. YEPES[5], P. WINCKER[6]

[1]IRD, Montpellier, France
[2]CIRAD, Montpellier, France
[3]Nestlé R&D, Tours, France
[4]ICCRI, Jember, Indonesia
[5]Cornell University, Geneva, New York, USA
[6]Génoscope-CEA, Evry, France

SUMMARY

The International Coffee Genomics Network (ICGN) is a worldwide network of scientists. ICGN's ultimate goal is to sequence the coffee genome and decipher through international partnerships the genetic and molecular bases of important biological traits in coffee tree species that are relevant to growers, processors, and consumers. As part of ICGN efforts, an international high density reference genetic map for *Coffea canephora* Pierre is being constructed in collaboration with R&D Nestlé Center and the Indonesian Coffee and Cocoa Research Institute. Furthermore, with funding from the French Agency ANR (Agence Nationale de la Recherche), several institutes (Genoscope-CEA, IRD and CIRAD) are combining their scientific resources and expertise to sequence, assemble, and annotate the entire genome of *C. canephora*. Several others ICGN members are planning to join these efforts particularly for mapping and genome sequencing and annotation. The *C. canephora* genome consists of 11 chromosomes, is about 710 Mb in size, and is being sequenced *de novo* with deep coverage using different sequencing platforms to obtain a reference genome for *Coffea*. The overall sequencing strategy and progress of the project are described.

INTRODUCTION

The International Coffee Genome Network (ICGN) is a worldwide network of scientists from universities, research institutes and industry within the coffee producing and coffee consuming countries (http://www.coffeegenome.org/). Our collaborative network is focused on building the foundation for advancing agricultural research for coffee by developing genomic tools and resources to further our understanding of the coffee genome at the molecular, biochemical, and physiological levels.

ICGN includes more than 100 individual and Institutional members networking scientific groups around the world in Africa, America, Europe, & Asia (http://www.coffeegenome.org/about/members.php). ICGN membership is opened to any individual, laboratory, or institution that can contribute to this effort in genomics resource development, sequencing and genome assembly, annotation, biological scientific expertise, or funding. ICGN is committed to advancing coffee genomic research through international partnerships for sustainable coffee production worldwide.

IMPORTANCE OF SEQUENCING THE COFFEE GENOME

Significant advances in our understanding of the coffee genome and its biology must be achieved in the next decades to increase quality, yield and protect the crop from major losses caused by insect pests, diseases and abiotic stress related to climatic change. Sequencing the coffee genome will help decipher the genetic and molecular bases of important biological traits in coffee that are relevant to growers, processors, and consumers. This knowledge is fundamental to allow efficient use and conservation of coffee genetic resources. Although considerable diversity exists in diploid *Coffea* species, its use in conventional coffee breeding programs has been very limited. *Coffea arabica* is characterized by a very low genetic diversity, which is attributable to its allotetraploid origin, reproductive biology, and evolution. The narrow genetic base of cultivated *C. arabica* has created a bottleneck for coffee breeding and limits cultivar improvement. Similarly, the considerable genetic diversity observed in *C. canephora* is still largely unexploited in the cultivated varieties. In the future, the ability to capture and manipulate genetic diversity and effectively utilize germplasm in traditional coffee breeding programs will be vital for sustainable coffee production.

DEVELOPMENT OF A HIGH-DENSITY GENETIC MAP FOR *COFFEA CANEPHORA*

An identified common objective is the establishment of a high-density genetic map. Ideally for genome assembly, this reference coffee genetic map would need to have 2 sequence-based markers per 1 million bp. To reach this objective, it was decided to take advantage of the initial important effort done by Nestle Tours Centre in collaboration with the Indonesian ICCRI institute. Those groups have already developed a map based on a cross between two highly heterozygous genotypes, a Congolese group genotype (BP409) and a Congolese-Guinean hybrid parent (Q121). The segregating population is composed of 93 F1 individuals. ICGN members will continue to saturate the map by mapping approximately another 1,000 sequence based markers such as SSRs and SNPs into the Indonesian population.

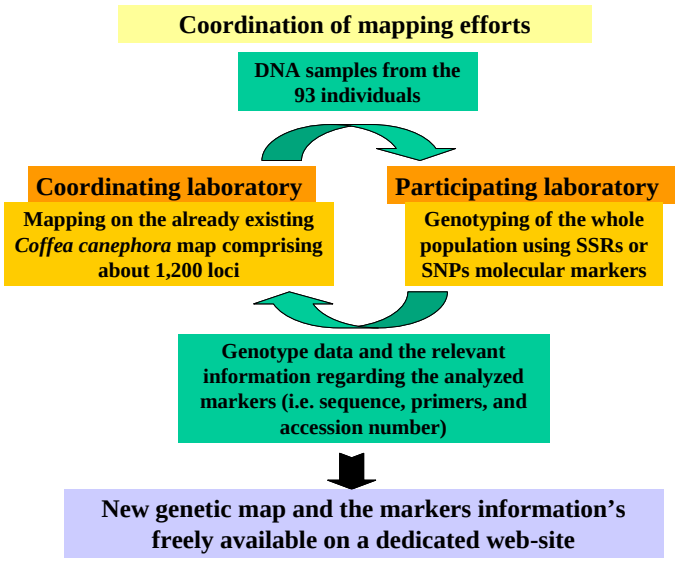

Figure 1. Organisation of the ICGN genetic mapping initiative.

A three-steps strategy has been adopted (Figure 1): 1) Upon request, DNA samples from the 93 individual segregating plants of the population BP409 X Q121 and the two parental clones is sent to the participating ICGN members; 2) After genotyping of the whole population, the

participating ICGN members must send back the genotype data and the relevant information regarding the analyzed markers (i.e. sequence, primers, sequence accession number); and 3) The additional sequence-characterized markers is mapped on the already existing *C. canephora* map to produce a high density reference map. Both, the high density genetic map as well as the marker information's will be freely available on a dedicated web-site (e.g. SOL web site).

ESTABLISHMENT AND ANNOTATION OF A REFERENCE GENOME SEQUENCE FOR COFFEE-TREES

Several institutes are combining their scientific resources and expertise to establish a reference genome sequence for coffee. *C. canephora* was chosen initially for this purpose because it is a diploid species (about 710 Mb in size). Also, *C. canephora* is one of the ancestral progenitors of the widely cultivated, *C. arabica* a recent allotetraploid species formed of the merge of the diploid species *C. canephora* and *C. eugenioides*. The accession DH200-94, a doubled haploid genotype was selected because of its homozygous status to facilitate genome assembly. *De novo* sequencing with deep coverage is being performed using both Roche pyrosequencing (454) and Illumina technologies. Direct whole genome shotgun sequencing and paired end sequencing of large insert libraries are underway; two 8kb and 20 kb insert libraries have been constructed. Furthermore, clones from two *C. canephora* BAC libraries were BAC-end sequenced using Sanger technology.

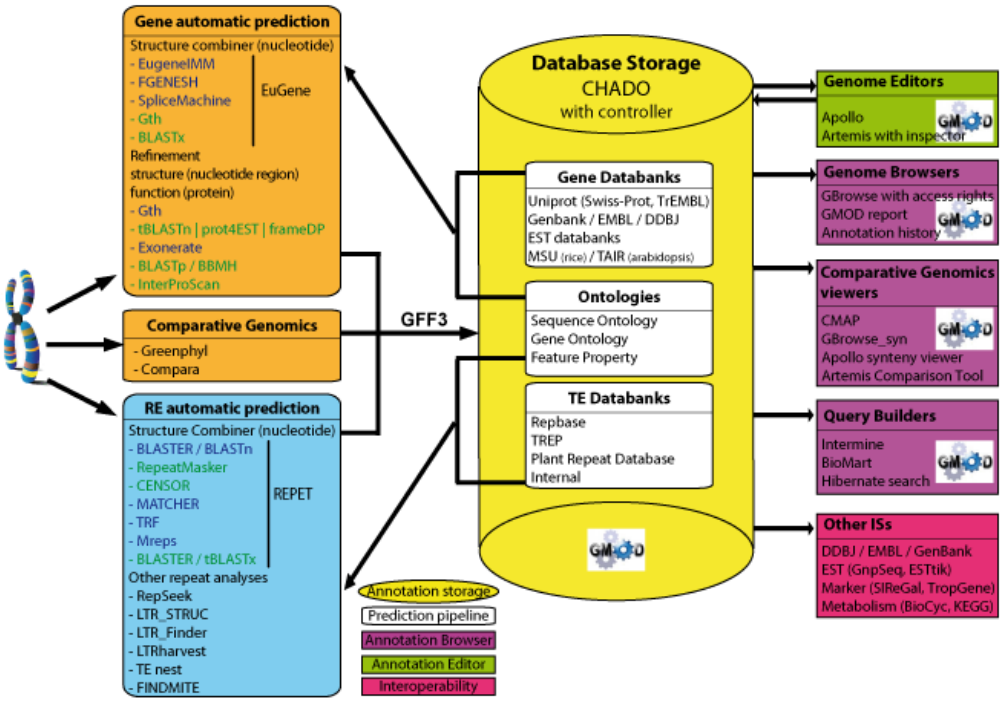

Figure 2. GNPannot, Community system for structural and functional annotation.

In addition to the publicly available ESTs, more transcriptome sequencing for *C. canephora* was done via 454 sequencing to facilitate genome annotation. All available evidences (cDNA sequences, the uniprot database, *ab initio* predictions) will be used for automatic annotation. The whole genome automatic predictions will then be integrated into a community annotation system (CAS) such as GnpAnnot (Figure 2) for gene expert annotation and genome analysis.

PERSPECTIVES

A future goal of the coffee ICGN community will be to establish the complete genome sequence of the allotetraploid *C. arabica* using the diploid progenitor species *C. canephora* and *C. eugenioides* as frameworks. *De novo* sequencing of *C. eugenioides* is expected to start by the end of 2010 with funding from the InterAmerican Development Bank allocated to the Colombian National Research Center and Cornell University. Furthermore and to ensure a full benefit of the generated resources by the coffee sector, funding for long term maintenance of the databases, and for the development of friendly end-user tools as well as training courses will be necessary.