THE INTERNATIONAL COCOA GENOME SEQUENCING CONSORTIUM (ICGS):

a coordinated strategy to sequence and analyse Theobroma cacao genome

White Paper, version 1.1

THE INTERNATIONAL COCOA GENOME SEQUENCING CONSORTIUM (ICGS): a coordinated strategy to sequence and analyse *Theobroma cacao* genome

Poster presented at PlantGEMS 24-27/9 2008 – Albena (Bulgaria)

Lanaud C[1], Argout X[1], Fouet O[1], Allegre M[1], Sidibe-bocs S[1], Ruiz M[1], Kudrna D[2], Jetty SSA[2], Wing R[2], Clément D[3,1], Gramacho K [3], Tahi M[4], Brunel D[5], Berard A[5], Boccara M [1,6], Udall J[7], Guiltinan M[8], Infante D[9], Costet P[10], Zang D[11], Risterucci AM[1], Legavre T[1], Sabau X[1],Wincker P[12]

The ICGS has been founded with representatives of the following institutions (in alphabetical order):

- [2]Arizona Genomics Institute, The University of Arizona, 1657 E Helen Street, Keating BLD , Tucson, AZ 85721- USA
- [3]CEPLAC Km 22 Rod. Ilheus Itabuna, Cx. postal 07, Itabuna 45600-00- Bahia - BRAZIL
- [1]CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), - URM DAP TA A96/03-34398 Montpellier cedex 5 – FRANCE
- [4]CNRA- 01 BP 1740 Abidjan 01 Abidjan - COTE D'IVOIRE
- [6]CRU Cocoa Research Unit, The University of the West Indies, St Augustine - TRINIDAD ET TOBAGO
- [12]GENOSCOPE - Centre National de Séquençage, 2 rue Gaston Crémieux CP 5721 91057 Evry Cedex – FRANCE
- [5]INRA-EPGV, CEA/Institut de Génomique - Centre National de Génotypage, 2 rue Gaston Crémieux, CP 5724, 91057 Evry Cedex – FRANCE
- [8]Penn State University- The Biotechnology Institute-306 Wartik Lab- University Park, PA 16802-5807 – USA
- [7]Plant and Wildlife Science Dept., Brigham Young University, Provo, UT 84602 – USA
- [9]Unidad de Biotecnología de Plantas, Instituto de Estudios Avanzados - Apdo. 17606 Parque Central- Caracas 1015-A, VENEZUELA
- [11]USDA-ARS- 5601 Sunnyside Avenue. Beltsville, MD, 20705-5139- USA.
- [10]VALRHONA - ZA des lots, 26600 Tain l'Hermitage, FRANCE

I. Introduction

In order to foster collaboration and communication between cacao breeders and geneticists, the International Group for Genetic Improvement of Cocoa (INGENIC) (http://ingenic.cas.psu.edu) was created in 1994. It now includes over 300 members, representing 35 developing and developed countries around the world, and now, the cacao genetics research community is well organized. The INGENIC Study Group for Molecular Biology (INGENIC-MOL-BIOL) was formally chartered in October of 2003 (Johnson 2003) to coordinate the activities of the INGENIC members interested in molecular approaches (Guiltinan, 2007). Several molecular resources were produced during the last years and shared between this community, and among them, a large number of SSR markers and a large collection of EST recently produced in the frame of an international project.

Our goal is now to build a community effort to coordinate the production of a complete annotated sequence of the cocoa genome (Criollo variety), facilitating the access of all genes and molecular tools important for cocoa improvement, and ensuring that these sequences will be publically available for all. To this end, the ICGS will refine a genome sequencing strategy to produce an assembly at high genome coverage, integrating already existing molecular resources and planning the production of new ones. To advance this goal, the partners agree to contribute or share resources adapted to develop such a project.

The interests of an International Cocoa Genome Sequencing Project and its current opportunity were discussed during the last International Cocoa Conference held in October 2006 in Costa Rica and during the last months by emails exchanges between the participants to this consortium.

II. Context and justification of cocoa sequencing

Theobroma cacao L. is a diploid tree fruit species (2n=2x=20) with a small genome (380Mb) (Lanaud et al., 1992 ; Figueira et al., 1992), similar to that of rice. It belongs to the Malvaceae family, as do cotton, and is close to the model species *A. Thaliana*. *T. cacao* originated from the tropical rainforest of South America and is one of the major cash crops for many tropical countries. The fruits of *T. cacao* (or pods) contains 20 to 40 beans which are used to produce chocolate and cocoa butter after a post harvest treatment including fermentation, drying and torrefaction steps.

Economic issues

Cocoa is the third product on the world market of raw materials after sugar and coffee. Cocoa is mainly produced on smallholdings, and according to the World Cocoa Foundation (http://www.worldcocoafoundation.org/), 40 to 50 million people depend upon cocoa for their livelihood, worldwide. About three million tons of cocoa are produced annually, from which 70% is contributed by Africa and the demand for cocoa is increasing. This production corresponds to a global market value of $5.1 billion.

A strong disease threat – towards a sustainable cocoa resistance

Producers are faced to an increasing threat due to several diseases and insect attacks which compromise the sustainability of cocoa production with some farms experiencing 100% losses (Keane, 1992; Bowers et al. 2001). Three of the most severe diseases of cacao are black pod caused by several *Phytophthora* species, frosty pod caused by *Moniliophthora roreri* and witches' broom caused by *Moniliophthora perniciosa* (Fulton, 1989; Ploetz, 2006).

Cocoa production is seriously affected by *Phytophthora,* sp., (black pod) which are responsible, worldwide, for 30% of losses. Several species are involved. *P. palmivora* is present in the entire cacao growing area, whereas *P. capsici* and *P. citrophthora* are prevalent in South America. *P. megakarya* is by far the most aggressive species with losses of production up to 50%, but is limited to some countries in West Africa. Harvest

losses due to *Phytophthora* species were estimated to 450000 tons (Bowers *et al*., 2001).
Two basidiomycetes, *Moniliophthora roreri* (frosty pod) and *Moniliophthora perniciosa* (witches' broom) are also responsible of important harvest losses (Ploetz, 2007). In Brazil, *M. perniciosa* was responsible of a drastic yield loss with a fall in production from 405000 tons in 1986 to less than 130000 tons in 1998. *Moniliophthora roreri* causes a very destructive pod rot and has already dramatic effects in some countries such as Ecuador and Costa Rica. *M. roreri* was confined in several countries of Central and South America, but is continuously spreading towards other Central America countries like Mexico or southwar towards countries like Peru.

Chemical control can be effective against *fungus* diseases but is polluting and often too expensive. Integrated pest management centered on the use of resistant material, enhanced with other methods of control (cultural, biological) is probably the best way of combating this pathogen over the long term to ensure a sustainable resistance.
Consequently, disease resistance is the primary trait targeted by cacao breeders. Sources of resistance have been identified for black pod (Iwaro et al, 2006), witches' broom (Umaharan et al., 2005) and frosty pod (Philips-Mora and Wilkinson, 2007). However, the molecular basis of cocoa resistance genes remain unknown. Decoding the cocoa genome will contribute significantly to our understanding of the functional aspect of cocoa resistance. Moreover, a genome sequencing program is currently carried out on *Moniliophthora perniciosa*, the pathogen responsible of witches' broom. The cocoa genome sequencing will provide a good model to study these plant pathogen interactions.

Maintaining and increasing cocoa quality

Cocoa quality improvement is another important cocoa trait for all actors of cocoa and chocolate production. Food quality improvement, nutritional as organoleptic, is now a strong request of society. Fundamental knowledge on quality construction from the genetic origin until post harvest treatements is an important challenge to address this request.
Flavour is among the main criterion of quality for chocolate manufacturers, but these characteristics are largely understudied by the cocoa research and breeding community due to their complexity and a dramatic lack of fundamental knowledge related to these traits. Flavour components depend strongly on conditions of post-harvest processing (environmental effects, storage, fermentation, drying, roasting) (Chanlieu and Cros, 1996). However, it is now well recognized that the genetic origin is also a strong determinant of flavour, independently of the conditions of post-harvest processing (Clapperton, 1994).
Aroma is composed of a bulk of volatiles compounds responsible of aroma perception and belonging to several classes of organic compounds such as hydrocarbons, aldehydes, acids, alcohols, esters, terpens etc.... In cocoa more than 500 volatiles compounds have been detected. However, only some of them could be determinant for specific aroma varieties.
Independently to volatiles compounds, some other biochemical compounds are known to interact with cocoa organoleptic traits. It is the case of polyphenols. Catechin,

epicatechin and procyanidines are the main polyphenols present in cocoa. They have well known antioxidant biological activities and beneficial effects on cardiovascular system (Wollgast et al., 2000; Steinberg et al., 2003; Othman et al., 2007;). Contributing to bitterness and astringency, polyphenols influence cocoa organoleptic quality (Counet et al., 2004). They influence aromatic profiles of cocoa in restricting Maillard's reactions, which generates a majority of the aromatic compounds of cocoa.

Planting material in farmers' plantations is constantly changing. Indeed, all producer countries are obliged to adapt their planting material to changes in threats from diseases and/or pests and in other environmental changes. There is a risk therefore of losing cocoa which has good flavour characteristics as there is no selection pressure to maintain this planting material. Up to now, flavour has always been assessed at the end of the breeding process and no breeding activities have aimed to improve cocoa quality. An improved knowledge of the genomic bases of different flavour attributes will provide plant breeders with tools allowing them to better exploit genetic resources and master quality traits selection with the other selection criteria, such as pest and disease resistance, during the breeding processes.

A strong need for a deeper exploitation of T. cacao genetic resources emphasized by climatic changes

With climatic changes, faster progresses in breeding and selection of new agronomic traits will be needed to produced varieties adapted to new environmental conditions. For example, in Africa, pathogen stress resistances were previously considered as the main traits of primary interest. However, with climatic changes, drought stresses are causing drastic losses of trees and yield in some countries, and new traits such as drought resistance now become important traits to select.

To this goal, a deeper exploitation of genetic resources based on an increased knowledge of gene function and allelic diversity will be needed. Genetic and genomic approaches can be integrated to identify more efficiently the key genes underlying agronomic or metabolite trait variations. In cocoa, a recent review made an inventory of nearly 300 QTL already identified in cocoa (Lanaud et al., unpublished data). However, no gene underlying these QTL has been isolated until now. The availability of a complete sequence of cocoa genome will make positional cloning and candidate gene discovery faster, allowing a more efficient screening of genetic resources made on the genes directly involved in trait variations.

Comparative genomics: a huge amount of genomic information from model plants useful to understand cocoa traits

The availability of a complete cocoa sequence will facilitate our understanding of the genomic organization and allow access to the gene content of the cocoa genome. A huge amount of information has been accumulated these last years on the model plants such as *Arabidopsis thaliana* and rice and is accessible through on-line resource centers such as the Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org; Rhee and Beavis 2003 ; Gramene http://www.gramene.org/; and OMAP http://www.omap.org). The research carried out on these model plants represents an invaluable source of information on gene structure and function which must be used to accelerate our understanding of cocoa gene functions.

By comparing the cocoa sequence with sequences from *Arabidopsis thaliana*, the model species closest to *T. cacao* and in the same rosid lineage, regions of similarity and differences can be identified. Such syntenic analyses will help the analysis of numerous cocoa metabolic pathways. However, *T. cacao* is a tropical tree, and genes more specific to its woody and tropical nature and absent in *Arabidopsis*, must exist.

A recent comparison of cocoa EST with EST from international database suggested that *Theobroma cacao* sequences presented a higher similarity with the proteome of another fruit tree crop: *Vitis vinifera*. (Argout et al., submitted). These findings could be explained by the fact that *Theobroma cacao* and *Vitis vinifera* are both fruit trees. Moreover, secondary metabolites are determinant for the qualities of the final product of each species (chocolate and wine), with common important metabolisms like flavonoids or terpenoid pathways involved in quality products. The genome of Vitis vinifera is now completely sequenced (Jaillon et al., 2007) and will be very useful for the annotation of cocoa genome.

The comparative genomic approaches will enable phylogenomic analysis of gene families and metabolic pathways among dicots and will provide tools to access more rapidly to key cocoa genes involved in important metabolic pathways. They represent also a powerful tool for studying evolutionary changes among organisms.

III. Proposed strategy to sequence the cocoa genome

The cocoa sequencing strategy will rely on a global WGS (whole genome shotgun) approach which will be supported by the integration of several already or newly produced genetic, genomic and bioinformatic resources. Its general steps include :

- the choice of the cocoa genotype to be sequenced
- the production of genomic resources to support the sequencing activities
- the sequencing steps
- the sequence alignment
- the automatic and manual annotation

Cocoa genotype chosen for the sequencing project

The Criollo genetic group corresponds to one of the two possible aromatic cocoa varieties that provides a fine flavor chocolate which is highly appreciated by chocolate manufacturers and which is bought at a superior price in the international market. Thus, it represents, an important economic « niche » for several countries from Central America, Latin America and the Caribbean.

The Criollo Variety was the first to be domesticated more than 2000 years ago by Maya and Aztec people. After Spanish colonization, the production of Criollo was extended to South America and the Caribbean region and open pollinations happened between Criollo and another genetically divergent variety, a Forastero which originated from Lower Amazonia. The vigourous hybrid forms, named Trinitario, gradually spread into original Criollo plantations, and now, most of the modern Criollo varieties, selected for

their quality traits, result from the recombination between ancestral Criollo (which could be identified), and invading Trinitario (Motamayor et al., 2003).
The «ancient» Criollo variety is self compatible and has a very narrow genetic base. It includes individuals with a complete homozygosity. Some of them, collected in oldest plantations, are conserved in the International Cocoa Genebank, Trinidad (ICG,T) which contains more than 2,000 accessions. It is one of the most diverse cacao germplasm collections worldwide and Criollo representation was increased recently with the acquisition of relic populations from Belize (Mooleedhar, 1997).
The complete homozygosity of some Criollo from Belize was already verified with 150 microsatellite markers. One of them was chosen for the construction of a BAC library and for the sequencing activities of the whole genome.

The «ancient» Criollo variety is also susceptible to several diseases. A better knowledge of genetic and molecular bases of traits of interest will accelerate the production of new Criollo varieties, productive, resistant to diseases and keeping the same sensorial qualities than the « ancient » Criollo variety.

Sequencing strategies

The genome sequence will be established using three different DNA sequencing technologies carried out with :
- the 454/Roche GSFLX sequencer. The newly released Titanium kits will be chosen, as they produce read lengths in the range of 400 bases.
- the Solexa/Illumina sequencing platform. These reads will be remapped on the assembly, and an automatic procedure developed at Genoscope will be used to correct the consensus errors, that are expected to be essentially due to homopolymer misinterpretation by the 454 technique.
- A classical Sanger sequencing method. These sequences will provide long-range continuity to the assembly.

Molecular resources needed for the project

Several molecular resources have been constituted these last years or are currently produced, and will be available for this project :

• <u>Mapping populations, markers and genetic maps</u>

Nearly 15 different mapping populations of different sizes have been produced and mapped during the last 10 years, mainly with the objectives to map QTLs. Among them, two mapping populations, with a larger number of individuals could be used for this project to constitute high density maps:

• A progeny of 250 trees, implemented in Côte d'Ivoire, and belonging to the cross UPA 402 x UF676. UPA 402 is a Forastero genotype originated from Upper Amazonia of Peru, UF 676 is a Trinitario corresponding to an hybrid between a Criollo genotype and Forastero genotype from Lower Amazonia in Brazil.
A part of this population has been used until now to establish the CIRAD cocoa

reference map on which all new markers have been successively mapped including AFLP, RFLP, SSR. This map (783 cM) included 465 codominant markers, and among them 268 SSR markers (Pugh *et al*., 2004).

• A progeny of 1500 trees, implemented in Brazil, and corresponding to an F2 population has been established from an F1 hybrid between ICS1 and Scavina 6. ICS1 is a Trinitario genotype (hybrid between a Criollo genotype and a Forastero genotype from Lower Amazonia in Brazil) and Scavina 6 is a Forastero genotype originated from Upper Amazonia of Peru. A part of this population (250 individuals) is currently used to map markers defined in the EST collection.

A project, currently developed by CIRAD and CNG aims to map a set of about 800 new markers defined in genes having a high similarity with known function genes (115 SSR and about 700 SNP) on the 2 mapping populations mentioned above. A large number of these SNP, identified from an EST collection recently produced (see below) correspond to a polymorphism between Criollo and Lower Amazon Forastero

These gene maps will constitute a substantial resource helping the whole genome sequence assembly. If needed , a densification of these maps will be possible by adding other SNP markers defined in the EST collection.

BAC libraries

The sequencing of BAC ends provide another very helpful molecular resource that helps the assembly of sequences provided by a whole genome shotgun approach.

In cocoa, two BAC libraries were already produced, one is from a Forastero genotype (Scavina 6) originated from Peru (Clement *et al*., 2004). This BAC library corresponds to about 11X genome coverage. The other is from a Forastero genotype (LCTEEN 67) originated from Ecuador and represents approximately 11 genome equivalents (USDA / Clemson BAC Resource Center).

However, both libraries were constructed from heterozygous Forastero genotypes and are not adapted for our project. New BAC libraries will be constructed during the next years from the same homozygous Criollo genotype chosen for the whole genome sequencing, and the BAC ends sequenced in the frame of a collaborative ANR project involving the University of Arizona, GENOSCOPE, CRU and CIRAD. The libraries will be constructed with two different restriction enzymes. Each library will correspond to a 6X genome coverage.

EST collection

A large EST collection, enriched in full length sequences is an important resource which will facilitate the annotation process.

Small collections of ESTs have been produced and used to study gene expression related to quality, stress or disease resistance and defense (Jones et al., 2002; Verica et al., 2004). More recently a larger collection of 149650 EST corresponding to 48594 unique transcripts has been constituted with the collaboration of the Genoscope, and in the frame of an international project coordinated by CIRAD (Argout et al., submitted). Fifty six cDNA libraries were constructed from different organs, different genotypes and environmental conditions. Among them, 25 corresponded to cocoa tissues submitted to different biotic stresses, and 11 corresponded to seed development and fermentation stages.

Among the 149650 EST, 2850 sequences were produced from a pure homozygous Criollo and 47800 were produced from hybrids between Criollo and a Lower Amazon Forastero genotype.

This large EST collection already will constitute a solid basis for the annotation of the whole genome sequence. However a deeper transcriptome sequencing made from the Criollo genotype could facilitate the annotations. The reads will be mapped onto the scaffolds and joined into gene models using methods developed at Genoscope.

IV. Databases and Bioinformatic tools

Databases and bioinformatic tools are also a key component of the project, to store, analyze and make the sequences available to all the community.

Several databases or WEB portals exist to store and manage data related to *T. cacao* :

Databases to store and analyse genetic and EST data

• **TropgeneDB** (http://tropgenedb.cirad.fr) is organized on a crop basis with presently nine modules (banana, cocoa, coconut, coffee, cotton, oil palm, rice, rubber tree and sugarcane). The most common data stored in TropgeneDB are genetic and physical maps, marker information, Quantitative Trait Loci (QTLs), sequence data, and molecular data on genetic resources. The CMAP tool is a Web-based tool that allows users to view genetic and physical maps and to observe synteny relationships by building bridges between linkage groups. This database will be used to store all information on markers and genetic maps used for the project

• **ESTtik :** The Expressed Sequence Tag Treatment and Investigation Kit tool (EST*tik*) was initiated to analyze and store results from processing of cDNA. To this end, a semi automatic pipeline for analysis and annotation of sequences, as well as a relational database developed for the storage of information related to the processing, were deployed. The EST*tik* pipeline programs are a set of Perl packages which contain a main program related to 9 modules in charge to complete different processing. The pipeline successively performs base calling, vector trimming, assembly and functional annotation. The non redundant data are then mined for Simple Sequence Repeat (SSR) markers and three primer pairs for each SSR found are designed. Single Nucleotide Polymorphisms (SNPs) are also detected. Data from the pipeline are systematically stored in a relational database, searchable *via* a local Web browser-based interface. This tool allows complex queries of the data to explore the expressed genes.

EST*tik* will be used to store and annotate the EST sequences obtained from a deeper Criollo transcriptome sequencing.

• **CocoaGen DB,** a Web portal on cocoa, that comprises molecular genetic, genomic and phenotypic data, was initiated through a collaborative project involving CIRAD, University of Reading (School of Plants Sciences, UK) and USDA/ARS (United States

Department of Agriculture, USA). This Web information system combines molecular genetic information from TropgeneDB (http://tropgenedb.cirad.fr) and phenotypic data from ICGD (International Cocoa Germplasm Database, http://www.icgd.reading.ac.uk). Users have the possibility to perform complex queries combining genetic and phenotypic information Public data produced by the ESTtik pipeline will be available through CocoaGen DB.

The genetic information available using the CocoaGen DB Web portal comprised around 1500 clones with their genotypes at various markers. Six genetic maps, a lot of information about 950 markers, 98 QTLs and 250 sequences data are also accessible.

CocoaGenDB is available through internet at the URL http://cocoagendb.cirad.fr.

• *whole genome automatic annotation pipeline*

The automatic annotation will be performed with three data types: *Ab initio* gene predictions, protein homologies with known genes and cocoa and other plants transcripts will be mapped onto the scaffolds . In particular we will make use of new RNA-Seq data obtained from the Criollo genotype chosen for the sequencing, and of a large collection of cocoa ESTs.

All data will be used to produce gene models using a reconciliation procedure successfully used to annotate many eukaryote genomes, including *Tetraodon* (Jaillon et al. 2004), *Paramecium* (Aury et al. 2006) and the grapevine (Jaillon et al., 2007).

When the predicted protein set is judged definitive, a series of tasks will be automatically performed, as domain predictions, GO assignments, orthologous and paralogous relationships (Best BLAST Mutual Hit (BBMH) and EC number determination. Synteny relationships with other sequenced plant genomes, and potential duplications of the cocoa genome, will be explored using tools previously developed (Jaillon et al. 2007)

• *whole genome manual or more specific annotation*

GnpAnnot Information system

A current GnpAnnot project (ANR GENOPLANTE edition 2007) coordinated by Cirad and INRA aims at developing a system of structural and functional annotation supported by comparative genomics and dedicated to plant and bio-aggressor genomes, allowing both automatic predictions and manual curation of genes and transposable elements. This system will allow hosting of the whole cocoa genome sequence and the annotations at the end of the project to facilitate its availability for users.

GnpAnnot system comprises a database, a set of sequence analysis methods and workflows and user interfaces. GnpAnnot is mainly based on components of the Generic Model Organism Database project.

The results of analyses of genomic sequences are stored in the Chado database, for instance:

- **Repetitive elements** predicted by transposable element combiners: ab initio repeat search, consensus identification and classification and extrinsic sequence similarity search
- **Protein coding genes** predicted by gene combiners which combine all available evidences: *ab initio* gene finders and extrinsic similarity search

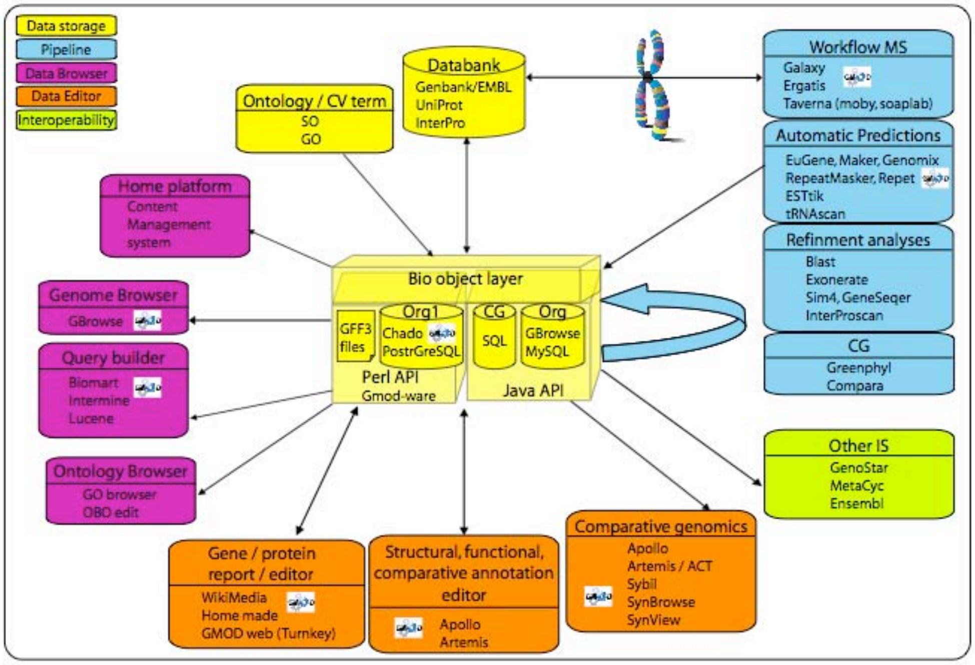
- **Protein non-coding genes** predicted by ncRNA gene finders using existing sequence alignments or structural alignments.

Web graphic interfaces allows the users to visualize (GBrowse), to query, to compare and to annotate the results of the methods.

Whole genome comparative genomics

Monocotyledon (*Oryza, Sorghum, Brachypodium, Maize*) and dicotyledon (*Arabidopsis, Poplar, Medicago, Vitis, Citrus, Solanum*) whole genome comparisons will help to better understand evolution of angiosperm phyla. To achieve this goal, we will assign angiosperm sequences to gene families to carry out phylogenetic analyses. Phylogenomic results will help to annotate manually the structure and function of the genes and to identify candidate genes for any plant genome.

Community manual annotation system



These relationships between gene pairs (BBMH or phylogenetic) will allow to reconstruct synteny groups between the reference genome *T. cacao* and the other plant genomes (*i.e.* to reconstruct groups of orthologous genes which are colocalised along the chromosomes) (Boyer, Morgat et al. 2005). These synteny groups will help to the manual annotation in particular to the comparative annotation and also to understand the genome evolution.

V. Summary of the main goals of the cocoa genome sequencing project:

Short term goal

- complete the molecular resources needed to increase the efficiency of sequences assembly, and more specifically, produce a very high density genes map based mainly on SNP markers defined in genes, and produce BAC libraries and pair ends sequencing.
- sequence the cocoa genome (Criollo) and perform the sequence assembly integrating all resources
- perform a complete and detailed annotation of the cocoa genome sequence to trigger gene discovery and facilitate map based cloning strategies.
- establish a performant database to manage and exploit the annotated sequences data, with links with others molecular resources as linkage maps, BAC ends , EST...

Long term goal

- integrate genetic and genomic sequences data to identify key genes involved in traits variations
- provide a general knowledge on the cocoa genome organization and gene content, and benefit, by comparative genomics approaches, of the advances made in model plants to identify and understand more easily the function of key cocoa genes
- provide new genomic performant tools to facilitate and stimulate germplasm characterization and allelic diversity of key genes
- increase breeding efficiency by marker assisted selection based on the genomics tools provided by this project.
- participate to the global understanding of plant evolution and complex biological processes

VI. Data management and release policies

All traces will be made available through the Trace Repository. The validated assemblies will be made available in the EMBL/Genbank WGS section when produced.

VII. Meetings and coordination

Regular meetings will be held to follow the advance of the genome sequencing and of the parallel anchoring of scaffolds to the genetic map. A specific meeting will be held once the automatic annotation is completed to define guidelines for the manual annotation tasks.

VIII. Conclusion

With the fast evolution of biotechnology tools, the availability of molecular and genetic resources already produced, and the large background of information accumulated in model plants, the conditions are met to progress quickly in the cocoa genome sequencing and fundamental knowledge on cocoa gene functions. These major progress will accelerate drastically cocoa improvement for resistance to biotic and abiotic stress, and quality components ; they will provide « diagnostic or predictive tools» to screen and optimize genetic resources exploitation.

Due to its small genome size, about three times those of *Arabidopsis*, its facility to reproduce vegetatively and sexually, and the richness of its fruits in secondary metabolisms, *T. cacao* L. represents a good model to study perennial woody fruit crops.

References

• Argout et al. (in submission). Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions. Submitted to BMC genomics.

• Aury JM et al. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. Nov 9;444(7116):171-8.

• Bowers JH, Bailey BA, Hebbar PK, Sanogo S, Lumsden RD (2001) The impact of plant diseases on worldwide chocolate production (http://www.plantmanagementnetwork.org/pub/php/review/cacao/)

• Boyer F., Morgat A., Labarre L., Pothier J., Viari A. Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. (2005). Bioinformatics, 21, 4209-4214.

• Chanliau S, Cros E. (1996). Influence du traitement post-récolte et de la torréfaction sur le développement de l'arôme cacao. *12th Alliance's Inter. Cocoa Conf.*, Salvador de Bahia (Brazil), Novembre 1996 : 959-964.

• Clapperton, J.F., Yow, S.T.K., Chan, J., Lim, D.H.K. (1994). Effects of planting materials on flavour. *Cocoa Growers' Bulletin* 48: 47-59.

• D. Clément, C. Lanaud, X. Sabau, O. Fouet, L. Le Cunff, E. Ruiz, A.M. Risterucci, J.C. Glaszmann, P. Piffanelli. (2004). Creation of BAC genomic resources for cocoa (Theobroma cacao L.) for physical mapping of RGA containing BAC clone. Theoretical Applied Genetics, vol 108, n°8, 1627-1634.

• Counet, C.; Ouwerx, C.; Rosoux, D.; Collin, S. (2004). Relationship between Procyanidin and Flavor Contents of Cocoa Liquors from Different Origins. J Agric Food Chem, 52, 6243-9.

• Figueira A, Janick J, Goldsbrough P (1992) Genome size and DNA polymorphism in *Theobroma cacao*. J. of Amer. Soc. Hort. Sci. 17 : 673-677.

• Guiltinan M (2007) Cacao. In: Pua EC, Davey MR (eds) Biotechnology in Agriculture and Forestry - Transgenic Crops VI. Springer-Verlag, Berlin Heidelbelg. Vol 60 p. 497-518.

• Iwaro AD, Bekele FL, Butler DR (2003) Evaluation and utilisation of cacao (Theobroma cacao L.) germplasm at the International Cocoa Genebank, Trinidad. Euphytica 130: 207-221.

• Iwaro AD, Butler DR, Eskes AB. (2006). Sources of resistance to *Phytophthora* pod rot at the International Cocoa Genebank, Trinidad. Genetic Resources and Crop Evolution 53: 99-109.

• Jaillon et al. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 2004 Oct 21;431(7011):946-57.

• Jaillon et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. Sep 27;449(7161):463-467.

• Johnson L (2003) INGENIC Workshop, Cocoa genomics group. Gro Cocoa 4: 4-5.

• Jones PG, Allaway D, Gilmour DM, Harris C, Rankin D, Retzel ER, Jones CA (2002) Gene discovery and microarray analysis of cacao (Theobroma cacao L.) varieties. Planta 216: 255-264.

• Keane PJ (1992) Diseases and pests of cocoa: An overview. Cocoa pest and disease management in Southeast Asia and Australasia, FAO Plant Production and Protection Paper 112: 1-12.

• Lanaud C., Hamon P., Duperray C., (1992). Estimation of nuclear DNA content of Theobroma cacao L. by flow cytometry. *Café Cacao Thé*, vol. 36, n. 1: 3-8.

• Mooleedhar V. (1997). A study of the morphlogical variation in a relicCriollo cacao population from Belize. In Annual report of the Cocoa Research Unit, p 5-14.

• Motamayor JC, Risterucci AM, Heath M, Lanaud C (2003). Cacao domestication II: progenitor germplasm of the Trinitario cacao cultivar. Heredity 91:322–330.

• Othman A, Ismail A, Ghani NA, Adenan I (2007): Antioxydant capacity and phenolic content of cocoa beans. *Food Chemistry*, 100:1523-1530.

• Phillips-Mora, W., and Wilkinson, M. J. (2007). Frosty pod of cacao: A disease with a limited geographic range but unlimited potential of damage. Phytopathology 97:1644-1647.

• Ploetz, R. C. (2007). Cacao diseases: Important threats to chocolate production worldwide. Phytopathology 97:1634-1639.

• Pugh T, Fouet O, Risterucci AM, Brottier P, Abouladze M, Deletrez C, Courtois B, Clement D, Larmande P, N'Goran JAK, Lanaud C. (2004). A new cocoa linkage map based on codominant markers: Development and integration of 201 new microsatellite markers. Theoretical Applied Genetics, vol 108, n°6, 1151-1161.

• Steinberg FM, Bearden MM, Keen CL (2003): Cocoa and chocolate flavonoids: implications for cardiovascular health. *Journal of the American Dietetic Association*, 103(2):215-223.

• Umaharan R., Thévenin J.M., Surujdeo-Maharaj S., Butler D.R. (2005). Identification of resistance to witches broom disease in the International Cocoa Genebank, Trinidad. In : *14th International Cocoa Research Conference. Proceedings* p. 161-169, Accra, Ghana.

• Verica JA, Maximova SN, Strem MD, Carlson JE, Bailey BA, Guiltinan MJ (2004) Isolation of ESTs from cacao (Theobroma cacao L.) leaves treated with inducers of the defense response. Plant Cell Rep 23: 404-413.

• Wollgast J, Anklam E (2000). Polyphenols in chocolate: is there a contribution to human health? *Food Research International*, 33:449-459.

Abstract

The goal of the International Cocoa Genome Sequencing Consortium (ICGS) is to build a community effort to coordinate the production of a complete annotated sequence of the cocoa genome (Criollo variety), facilitating the access to all genes and the development of molecular tools important for cocoa improvement, and ensuring that these sequences will be available to all. To this end, the ICGS will refine a genome sequencing and annotation strategy, integrating already existing genetic, genomic and bioinformatic resources, which provide a solid basis to complete such a project, and planning the production of new ones.