A LARGE EST RESOURCE FOR THEOBROMA CACAO INCLUDING CDNAS
ISOLATED FROM VARIOUS ORGANS AND UNDER VARIOUS BIOTIC AND
ABIOTIC STRESSES

Claire Lanaud1, Olivier Fouet1, Karina Gramacho2, Xavier Argout1, Thierry Legavre1, Xavier Sabau1, Ange Marie
Risterucci1, Patrick Wincker1, Corinne Da Silva1, Gaston Loo1, Ulisson Lopes1, Julio Cascardo1, Brigitte Courtois1,
David Kuhn1, Ray Schnell1, Bryaïn Bailey1, Regis Babin8, Olivier Souingeot8, Michel Ducamp9, Didier Paulin9,
Peninna Deherdt10, Joseph Vertica11, Marc Guiltrain11, Laurence Alemanno12, Regina Machado13, Wilberth Phillips14,
Fabienne Michi11, Abelmon da Silva Gestem1, Didier Clement15, Siela Maximova11, Michel Bocca11,15
David Butler15, Eric Rosenquist1, Martin Gilmour15, Jean Christophe Glasmann1

1CIRAD - URM PIA TA40/03-34398 Montpellier cedex 5 - France
2CEPLAC Km 22 Rod. Ilheus Itabuna, Cx. postal 07, Itabuna 45600-00- Bahia BRAZIL
3GENOSCOPE - 2 rue Gaston Crémieux (CP 5721) 91057 (Ervy Cedex - France
4INIAE - EET-Pichilingue. Código Postal 24 Km 5 via Quevedo El Empalme, Ecuador.
5UESC-Laboratório de Genômica e Expressão Gênica Rodovia Ilhéus-Itabuna, Km 16- Ilhéus-BA – (BRAZIL)
6USDA-ARS, 13601 Old Cutler Rd. Miami, Florida - USA
7USDA-ARS- 5601 Sunnydale Avenue. Beltsville, MD, 20705-5139- USA.
8IRAD - Nkolbisson, BP 2067 Yaoundé Cameroon
9CIRAD- UPR TA 80/02 - 34398 Montpellier Cedex 5 - France
10CIRAD- UMR BGPI TA41/K 34398 Montpellier France
11Penn State University- The Biotechnology Institute-306 Wartik Lab- University Park, PA 16802-5807 USA
12CIRAD- UMR BEPC TA 80/03 - 34398 Montpellier Cedex 5 - France
13MASTERFOODS Almirante – Brazil
14CATIE P.O.Box 7170, Turrialba, Costa Rica
15CRU Cocoa Research Unit, The University of the West Indies, St Augustine
Trinidad et Tobago
15Mars Inc. Dundee Road, Slough, SL1 4JX, UK.

SUMMARY

An international project aiming to sequence a large collection of 180,000 cDNA cocoa clones enriched in full length
cDNAs representing genes expressed in T. cacao L. was carried out. These clones were isolated mainly from two genotypes:
Scavina6 and ICS1, and from different organs with and/or without treatment with various biotic and abiotic stresses.
About 40 cDNA libraries were constructed from a wide panel of organs: flowers (self pollinated and cross pollinated),
flower cushions, seeds at different stages of development and during fermentation, cherels, pod cortex, shoot, wood,
root, germinated seeds, and embryos from vitro culture. Librarias were also constructed from organs subjected to biotic
stresses: leaves and pods inoculated with Phytophthora palmivora and megakarya, shoots and pods inoculated with
Crinipellis pericosa, pod inoculated with Moniliophthora roreri, shoot inoculated with Ceratocystis fimbriata,
and shoots attacked by mirids. Suppression subtractive hybridisation (SSH) libraries were also generally constructed from
these plant pathogen interactions to facilitate further identification of resistance or defense genes expressed in cocoa
during pathogen infections. Also, two SSH libraries were made from flowers pollinated by selfcompatible or self
incompatible pollen. Libraries have also been constructed from cuttings (leaves and roots) under drought conditions.

The construction and management of cDNA libraries, picking and replicating, was performed with the help of the robotic
platform, part of the “Montpellier Languedoc-Roussillon GENOPOLE”. All the sequencing work was performed by
GENOSCOPE (Ervy), French National Center of Sequencing.

To efficiently store and mined the gene information, a bioinformatic pipeline (ESTuk) was used which automatically
processes the sequences, assembles, annotates them, and integrates the results into a web-based database, allowing
researchers browse and query of the results. After annotation and comparison of cocoa sequence data with international
sequence database (NCBI) about 62% of cocoa sequences showed significant similarity to gene sequences from other
species. These comparisons allowed the annotation of potential gene function of many of the sequences, as well as a
general classification of cocoa cDNA sequences according to the gene ontology system.
This new resource of cacao ESTs makes possible the identification of hundreds of new microsatellite markers and thousand of SNPs (Single Nucleotide Polymorphisms) markers, suitable for all genetic analyses. This cDNA resource will allow to identify a large unigene set suitable to analyze, by functional genomics, molecular mechanisms underlying resistance, quality or other traits of interest for cocoa breeding.

**UNE VASTE RESSOURCE DE Séquences EST POUR THEOBROMA CACAO COMPRENANT DES ADNC ISOLES À PARTIR DE DIVERS ORGANES ET DANS DIVERSES SITUATIONS DE STRESS BIOTIQUE ET ABIOTIQUE**

**RÉSUMÉ**

Un projet international visant au séquençage d’une importante collection de 180 000 clones ADNC de cacao enrichis en pleine longueur, représentant des gènes exprimés dans *T. cacao* L. a été réalisé. Ces clones ont été isolés principalement à partir de deux génotypes, Scavina6 et ICS1, et à partir de différents organes avec et/ou sans traitement avec divers stress biotiques et abiotiques. Environ 45 bibliothèques d’ADNC ont été élaborées à partir d’une large gamme d’organes : fleurs (auto-pollinisées et avec pollinisation croisée), coissins floraux, graines à différents stades de développement et durant la fermentation, cerises, cortex de cabosse, pousses, bois, racines, graines germées et embryons de culture in vitro. Des bibliothèques ont également été construites à partir d’organes soumis à des stress biotiques : feuilles et cabosses inoculées avec *Phytophthora palmivora* et *Megakarya*, tiges et cabosses inoculées avec *Cremiptella perniciosa*, cabosses inoculées avec *Montoiliophthora roseri*, tige inoculée avec *Ceratocystis fimbriata*, cabosses inoculées avec des endophytes utilisés pour la lutte biologique, et tiges attaquées par des mirides. Des bibliothèques d’hybridations soustractives (SSH) ont aussi été généralement construites à partir de ces interactions avec des pathogènes végétaux pour faciliter l’identification ultérieure de gènes de résistance ou de défense exprimés dans le cacaoyer au cours d’infections par les pathogènes. Par ailleurs, deux bibliothèques SSH ont été réalisées pour les fleurs en utilisant les deux conditions à la fois comme sondes et comme pilotes. Des bibliothèques ont également été établies à partir de boutures dans des conditions de sécheresse.

La construction et la gestion des bibliothèques d’ADNC, le “picking” et la réplication ont été réalisés avec l’aide de la plate-forme robotique, au sein du « GENOPOLE Languedoc-Roussillon ». Tous le travail de séquençage a été réalisé par le GENOSCOPE (Ivry), le Centre national français de séquençage.

Pour stocker et exploiter efficacement les informations sur les gènes, un pipeline bioinformatique (ESTk) a été utilisé, qui traite automatiquement les séquences, les assemble, les annoté, et compile les résultats dans une base de données sur Internet, ce qui permet aux chercheurs de consulter les résultats et d’effectuer des requêtes. Après l’annotation et la comparaison des données des séquences de cacaoyer avec la base de données internationale des séquences (NCBI), environ 68 % des séquences de cacaoyer semblaient appartenir à une similitude significative avec les séquences génétiques d’autres espèces. Ces comparaisons ont permis l’annotation de la fonction génétique de beaucoup des séquences, ainsi qu’une classification générale des séquences d’ADNC de cacaoyer en fonction du système d’ontologie génétique.

Cette nouvelle ressource de séquences EST de cacaoyer rend possible l’identification de centaines de nouveaux marqueurs microsatellites et de milliers de SNP (Polymorphismes d’un seul nucléotide), de nouveaux marqueurs très efficaces, la recherche de gènes candidats de cacaoyer homologues à des gènes spécifiques d’intérêt qui ont été caractérisées chez d’autres espèces (impliqués dans la résistance, la réaction de défense, la qualité…), et l’identification d’un vaste ensemble unigène convenant à l’analyse, par génomique fonctionnelle, des mécanismes moléculaires sous-tendant la résistance, la qualité ou d’autres caractères intéressants pour la sélection génétique du cacaoyer.

**UM RECURSO ALARGADO EST PARA THEOBROMA CACAO INCLUINDO CDNAS ISOladOS DE DIVERSOS ÓRGÃOS E SOB DIVERSOS ESTRESSES BIOtICOS E ABIÓTICOS**

**SUMARIO**

Foi realizado um projeto internacional com o objetivo de sequenciar uma grande coleção de 180,000 cDNA clones de cacau enriquecidos em cDNAs de comprimento total, representando genes expressos em *T. cacao* L. Estes clones foram isolados principalmente a partir de dois génotipos: Scavina6 e ICS1, de diferentes órgãos com e/ou sem tratamento com diversos estresses bióticos e abióticos. Foram construídas cerca de 45 bancos de cDNA a partir de uma vasta gama de órgãos: flores (de polinização cruzada e de auto-polinização), coxins florais, sementes em diversos estágios de desenvolvimento e durante a fermentação, frutos (ceres), córtex da vagem, rebentos, madeira, raiz, sementes germinadas
e embriões de cultura in vitro. Foram também organizados bancos a partir de órgãos sujeitos a estresses bióticos: folhas e frutos inoculados com *Phytophthora palmivora* e *megakarya*, rebentos e frutos inoculados com *Crinipellis perniciosa*, frutos inoculados com *Montlipithora roreri*, rebento inoculado com *Ceratocystis fimbrata*, frutos inoculados com entófitos utilizados para controlo biológico, e rebentos atacados por miríades. No geral, também foram organizados bancos de hibridação subtractiva por supressão (SSH) a partir das interações dos seus patógenos, para facilitar uma maior identificação de genes de resistência ou de defesa presentes no cacau durante as infeções patogénicas. Foram também preparados dois bancos SSH para flores, utilizando ambas as condições como testadores e controladores. Foram também elaborados bancos a partir de podas sob condições de seca.

A elaboração e gestão dos bancos de cDNA, colhendo e replicando, foi realizada com a ajuda de uma plataforma robótica, parte do «Pólo Genético Languedoc-Roussillon». Todo o trabalho sequencial foi realizado por Genoscope (Evry), no Centro Nacional Francês de Sequenciamento.

Para armazenar e realizar a mineração da informação genética foi utilizado um tubo bioinformático (ESTtik), que processa automaticamente as sequências, instala-as e anota-as, e depois integra os resultados numa base de dados em rede, o que permite aos investigadores procurar e questionar os resultados. Após anotação e comparação dos dados sequenciais do cacau com uma base de dados sequenciais internacional (NCBI), cerca de 68% da sequência de cacau apresentava uma semelhança significativa com sequências genéticas de outras espécies. Estas comparações permitiram a anotação de potenciais funções genéticas de muitas das sequências, bem como uma classificação geral de sequências cDNA do cacau, de acordo com o sistema ontológico dos genes.

Este novo recurso de ESTs do cacau torna possível a identificação de centenas de novos marcadores de microsatélites e de milhares de SNPs (Polimorfismos Nucleotídicos Simples), novos e muito eficientes marcadores, a procura de candidatos de genes de cacau homólogos a genes específicos de interesse, identificados noutras espécies (ligados a resistência, resposta defensiva, qualidade, etc...), e a identificação de um grande conjunto de unigens para analisar, por meio de genómica funcional, mecanismos moleculares a resistência, qualidade ou quaisquer outros traços subjacentes, com interesse para o cultivo do cacau.

UN GRAN RECURSO EST PARA EL *THEOBROMA CACAO* INCLUYENDO CADN AISLADO DE VARIOS ÓRGANOS Y BAJO VARIOS ESTRESES BIÓTICOS Y ABIÓTICOS

**RESUMEN**

Se llevó a cabo un proyecto con el objetivo de secuenciar una gran colección de 180.000 cADN de clones de cacao enriquecidos en cADN de largo complejo representando genes expresados en *T. cacao* L. Estos clones fueron aislados principalmente de dos genotipos, Scavina y ICS1, y de diferentes órganos con y/o sin tratamiento con varios estresses bióticos y abióticos. Cerca de 45 bibliotecas de cADN fueron construidas a partir de un amplio panel de órganos: flores (autopolinizadas y con polinización cruzada); hojas de flores; semillas en diferentes etapas de desarrollo y durante la fermentación; cerejas; cortez de la vaina; brotes; madera; raíces; semillas germinadas y embriones de cultivo in vitro. También se construyeron bibliotecas de clones sujeto a estresses bióticos: hojas y vainas inoculadas con *Phytophthora palmivora* y *megakarya*, brotes y vainas inoculadas con *Crinipellis perniciosa*, vaina inoculada con *Montlipithora roreri*, brote inoculado con *Ceratocystis fimbrata*, vainas inoculadas con endofitos usados para control biológico y brotes atacados por *mirtidae*. Las bibliotecas de hibridación subtractiva de supresión (SSH) también son construidas generalmente a partir de estas interacciones de patógeno de la planta para facilitar la identificación posterior de genes de resistencia o de defensa expresados en el cacau durante infecciones de patógenos. También se hicieron dos bibliotecas de SSH para flores usando ambas condiciones como testes y drivers. Además se construyeron bibliotecas de cortes bajo condiciones de sequía.

La construcción y administración de bibliotecas de cADN, seleccionando y duplicando, fue realizada con la ayuda de la plataforma robótica, parte del "Languedoc-Roussillon GENOPOLE". Todo el trabajo de secuenciamento fue realizado por GENOSCOPE (Evry), Centro Nacional de Secuenciamento de Francia.

Para almacenar y extraer eficientemente de la información del gene, se usó un pipeline bioinformático (ESTtik) que procesa automáticamente las secuencias, ensambla, las anota e integra los resultados en una base de datos basada en la Web, permitiendo que los investigadores naveguen y consulten los resultados. Luego de la anotación y la comparación de los datos de secuencia del cacau con la base de datos de secuencia internacional (NCBI), cerca del 68% de la secuencia del cacau mostró una similaridad significativa a las secuencias de genes de otras especies. Estas comparaciones permitieron la anotación de funciones de gen potenciales de varias de las secuencias, así como también una clasificación general de las secuencias de cADN del cacau de acuerdo con el sistema de ontología de gen.
INTRODUCTION

Genetic improvement of cocoa, particularly for resistance, productivity and quality traits, will be enhanced by the availability of characterised genetic resources and modern tools to be used in “marker assisted selection” strategies. In the last years, efforts have been made by the international community to develop molecular resources aiming to provide a better knowledge of genetic diversity and genetic determinism of useful traits in cocoa.

More recently, an international project aiming to sequence a large collection of 180,000 cDNA cocoa clones enriched in full length cDNAs representing genes expressed in *T. cacao* L. was initiated by CIRAD (France) in collaboration with the GENOSCOPE, (French National Center of Sequencing – Evry) and several partners from Brazil (CEPEC, UESC), USA (USDA – Penn State University), Equator (INAP), Cameroon (IRAD), Costa Rica (CATIE), Trinidad (CRU) and Masterfoods. We report here the work already done to produce the cDNA libraries and analyze the first sets of sequences provided by the GENOSCOPE.

MATERIAL AND METHODS

MATERIAL

Cocoa clones used for cDNA libraries construction:

SCA6 and ICS1 were the main clones used for library construction. Some other clones were used for specific libraries construction and included P7, UF 273, UPA134, IMC67, UF676, B240, JACA and B97 C-C-2.

Types of constructed libraries and cocoa clones used for cDNA libraries construction:

About 40 cDNA libraries enriched in full length cDNA were constructed from a wide panel of organs at different stages of development and in some case submitted to biotic and abiotic stresses. Suppression subtractive hybridisation (SSH) libraries were also generally constructed from plant pathogen or insect interactions to facilitate further identification of resistance or defense genes expressed in cocoa during pathogen infections or mirids attacks. Two SSH libraries were also made for flowers pollinated by compatible or incompatible pollen. SCA6 and ICS1 were the main clones used for library construction. The other clones used for specific libraries construction included P7, UF 273, UPA134, IMC67, UF676, B240, JACA and B97 C-C-2.

The following libraries were constructed:

**Organs in development:**

- ICS1 - flowers (different stages of development; incompatible and compatible pollinations)
- B97 C-C-2 - cushions
- SCA6 - stems
- SCA6 - leaves at different stages of development
- SCA6 - roots
- SCA6 - cortex (internal and external part)
- SCA6 - wood
- ICS1 (Self Pol.) - seedlings (1 week and 2/3 weeks, epicotyle and hypocotyle, cotyledons)
- SCA6 - cherels (first week stages, one week to one month stages of development)
- P7 - cherels in wilt
- ICS1 - ovules (2 to 3 months)
- ICS1/B240 - seeds in development (3/4 month; 4/5 months, mature seeds; cotyledons and testa)
- ICS1/B240 - fermented seeds (1 to 4 days, cotyledons and testa)
- SCA6 - embryogenic tissue from vitro culture

**Organs submitted to stresses:**

--- abiotic stress:

- SCA6 - shoot and roots submitted to drought stress

--- biotic stress:

- SCA6 - UPA134 - pods submitted to *Phytophthora palmivora* or *Phytophthora megakarya*
- SCA6, PNG progenies - leaves submitted to *Phytophthora palmivora* and *Phytophthora megakarya*
- P7, UF273 - pods submitted to *Moniliophthora rorieri*
- SCA6 - pods submitted to *Crinipellis perniciosa*
- SCA6 - Shoots submitted to *Crinipellis perniciosa*
- JACA - shoots submitted to *Ceratocystis fimbriata*
- UF676 - shoots submitted to mirids (*Sahlbergella*)

Protocol used for cDNA cloning

RNA extraction from cacao tissues were performed according to Da Silva Gesteira et al. (2003) using tert-butanol. First strand were performed using the Clontech BD SMART PCR cDNA Synthesis KIT (cat No 634902) as recommended by the supplier. 0.05–1 µg of total RNA was incubated at 72°C for 2 min with 1 µl 3' BD SMART CDS Primer II A (12 µM) and 1 µl BD SMART II A Oligonucleotide (12 µM) in a total volume of 5µl, then 2 µl 5X First-Strand Buffer, 1 µl dTT (20 mM), 1 µl dNTP Mix (10 mM of each dNTP), 1 µl BD PowerScript Reverse Transcriptase was added and the mix was incubated at
42°C for 1 hr in an air incubator. In order to eliminate the unfinished strands, a protocol was applied according to Glen K Fu (2003): 3 μl Biotin-dATP (Invitrogen), 3 μl Biotin-dCTP (Invitrogen), 1 μl NV primer 30 μM (30ng), 2 μl 5X First-Strand Buffer, 1 μl BD PowerScript Reverse Transcriptase was added and the mix was kept at 42°C for 30min. For capture of the unfinished strand the reaction was mixed with 600 μl of Streptavidine MagneSphere Paramagnetic Particles (Promega) and eluted as recommended by the supplier. Two μl aliquot from the first-strand synthesis was used for the cDNA Amplification by LD PCR (Clontech). Each reaction was performed with 80 μl Deionized H2O; 10 μl 10X BD Advantage 2 PCR Buffer, 2 μl 50X dNTP Mix (10 mM of each dNTP), 4 μl 5’ PCR Primer II A (12 μM), 2 μl 50X BD Advantage 2 Polymerase Mix in a 98 μl total volume. Between 17 and 23 PCR cycle 95°C 15 sec, 65°C 30 sec, 68°C 6 min, following a final extension at 70°C 10min.

Then cDNA Size Fractionation was performed as recommended by the supplier (Clontech). One μl of the second strand product was cloned in pGEMT (Promega) and transformed by electroporation in the DH10B T1 resistant strain of Escherichia Coli—(In Vitrogen).

Picking and replication of libraries
This step was made in the CIRAD/GENOPOLE languedoc-Roussillon plateform.

Sequencing steps
The sequencing steps were realized at the GENOSCOPE (Evry—France). A part of the clones were sequenced from both 5’ and 3’ ends of the transcripts, but for most of them a unique sequence was made from 3’ or from 5’ end.

BIOINFORMATIC ANALYSES

Sequences storage
To efficiently store and mine the gene information, a bioinformatic pipeline (ESTtk) was constructed, which automatically processes the sequences, assembles, annotates them, and integrates the results into a web-based database, allowing researchers browse and query of the results (for details, see poster Argout et al., in this conference).

Sequences annotation
The annotation and comparison of cocoa sequence data was made with the international sequence database (NCBI). These comparisons allowed the annotation of potential gene function of the sequences. A general classification of cocoa cDNA sequences according to the Gene Ontology system (Conesa et al., 2005) was also made in comparison with Arabidopsis known function genes.

Full length average evaluation
An estimation of the average of full length sequences was made in comparison with Arabidopsis sequences. Indeed, in the phylogenic tree of angiosperm (AGP2, 2003), T. cacao is relatively close to Arabidopsis (in the same rosid 2 group) and a high global gene structure similarity could be expected.

The cDNA cloning was not oriented, so, the sequences, limited in length, could have been initiated, theoretically with a same probability, from the 5’ end or from the poly A end. As all the cDNA synthesis were initiated from the polyA part of the mRNA, we could hypothesis that the sequence having the 5’ end, opposite to the polyA end could represent half of the potential full length cDNA present in the libraries. The 5’ ends were identified in comparison with Arabidopsis genes, with an accepted variation of 3 amino acids. An average of full length cDNA was estimated using these comparisons.

PRELIMINARY RESULTS

Number of clones produced until now
The complete list of libraries mentioned above has been constructed during the last two years. Until now, 108622 cDNA clones have been sent for sequencing.

Sequence annotation and assembly
A part of the clones (9302) were sequenced by both 5’ and 3’ end sides. the remaining clones were sequenced only from one side (5’ or 3’ end side).

Until now 117924 sequences have been made. A set of 93373 valid sequences were isolated et started to be subjected to bioinformatic analyses.

A set of 22888 singletons and 8526 contigs has been identified representing an unigene set of 31414 sequences.

After annotation and comparison of cocoa sequence data with international sequence database (NCBI), about 62% (11955 singletons + 7466 contigs) of cocoa sequence showed significant similarity to gene sequences from other species (e values < 10⁻¹⁰).

A general classification of cocoa cDNA sequences according to their molecular function or to their biological process apparenance was made using the gene ontology system (Conesa et al., 2005) in comparison with Arabidopsis gene sequence homology.

A set of 9076 unique EST could be classified using this tool. The classification of these sequences show that all the several classes of genes are represented in this collection, with the most frequent ones belonging to the binding and catalytic activities for their molecular function, and belonging to the physiological process and metabolism for their biological processes.

Among the 9076 EST, a set of 244 sequences corresponding to genes expressed after a stimuli (for example biotic or abiotic stresses) could be identified.
Evaluation of full length cDNA average:
A set of 19071 cocoa sequences having a hit with Arabidopsis proteins were selected for this analysis. Among them, the estimated average of full length cDNA present in the totality of the libraries is 46%. Ten % of the sequences analyzed present homologies with both 5' and 3' ends of Arabidopsis cDNA and reflect the smaller genes which could be completely sequenced.

SSR analyses
The 31414 unique sequences have been screened for SSR. This unigene represent 17,07Mbp. A set of 1711 SSR have been identified, corresponding to an average of one SSR each 10kbp.

In total, 1537 EST (4.89%) contain at least one microsatellite. Among them, 154 EST contain more than one microsatellite.

Several types of microsatellites have been observed, the dinucleotide repeats being the most frequent class:
- dinucleotide repeats: 903 (52.8%)
- trinucleotide repeats: 714 (41.7%)
- tetranucleotide repeats: 59 (3.8%)
- pentanucleotide repeats: 23 (1.5%)
- hexanucleotide repeats: 12 (0.8%)

Among the di and tri nucleotide repeat types (considering sequence complementary), the following classes were observed:
- AC/GT: 45
- AG/CT: 642
- AT/AT: 216
- AAC/GTT: 47
- AAG/CTT: 266
- AAT/ATT: 85
- ACC/GGT: 53
- ACG/CTG: 34
- ACT/ATG: 39
- AGC/GTG: 43
- AGG/CCT: 70
- AGT/ATC: 67
- CCG/CGG: 10

Among the dinucleotide repeats, the most frequent class of SSR is represented by the AG or CT complementary motifs. Among the trinucleotide repeats, the AAG/GTT motif is the most frequent one.

DISCUSSION - CONCLUSION
EST assemblies have been already reported in cocoa (Jones et al., 2002, Verica et al., 2004). However, these EST collections corresponded to a limited number of EST and libraries constructed. Here, we present the construction and partial sequencing of a large number of cDNA libraries corresponding to a wide panel of cocoa organs submitted to various situations of biotic and abiotic stresses.

This partial collection of 117924 cocoa EST, corresponding to 31414 unique sequences, will be completed in the next months. This collection already represents an important set of genes expressed in cocoa during its development and useful for genetic and genomic analyses. However, the 31414 unique sequences could be an overestimation of the real gene number. Indeed, gene number overestimation may be an artifact during EST assemblies, particularly for larger size genes for which several EST could be generated without sequences overlapping.

In cotton, a related species belonging to the same family (Malvaceae), an unigene set comprising 33665 unique sequences was isolated from a collection of 185000 EST. In Arabidopsis, the number of genes was estimated to 28952 (Wortman et al., 2003).

Many applications could be provided by these EST collection and unigene set.

The first one is the support for all functional genomic studies. The unigene set derived from the global sequences assembly can be used as templates for microarray designs. More specific applications could be made, for example, many libraries have been constructed from biotic stresses; an unigene set could be constructed specifically from all sequences belonging to this unigene set and could serve as template for all cocoa-pathogen interactions.

This project provides also a large resource to generate genetic markers (microsatellites, SNP...) within potentially functional genes, facilitating the construction of a transcript map based and the functional diversity study by the role that they could have in gene expression or function. Due to their localization within genes, these markers will also facilitate genome comparisons between species belonging to related families.

ACKNOWLEDGEMENTS
We are grateful to GENOSCOPE for its support for all sequencing steps of the project and to the Montpellier Languedoc Roussillon Genopole for the availability of the robotic plateform used to manage all the libraries' construction. We also thank MASTERFOODS and USDA-ARS for their financial contribution in this project.

REFERENCES
2. Conesa A, Goetz S, Garcia-Gomez JM, Terol J, Talon M, Robles M.


