Statement of transcriptomics and bioinformatics analyses conducted at CIRAD in rubber tree:

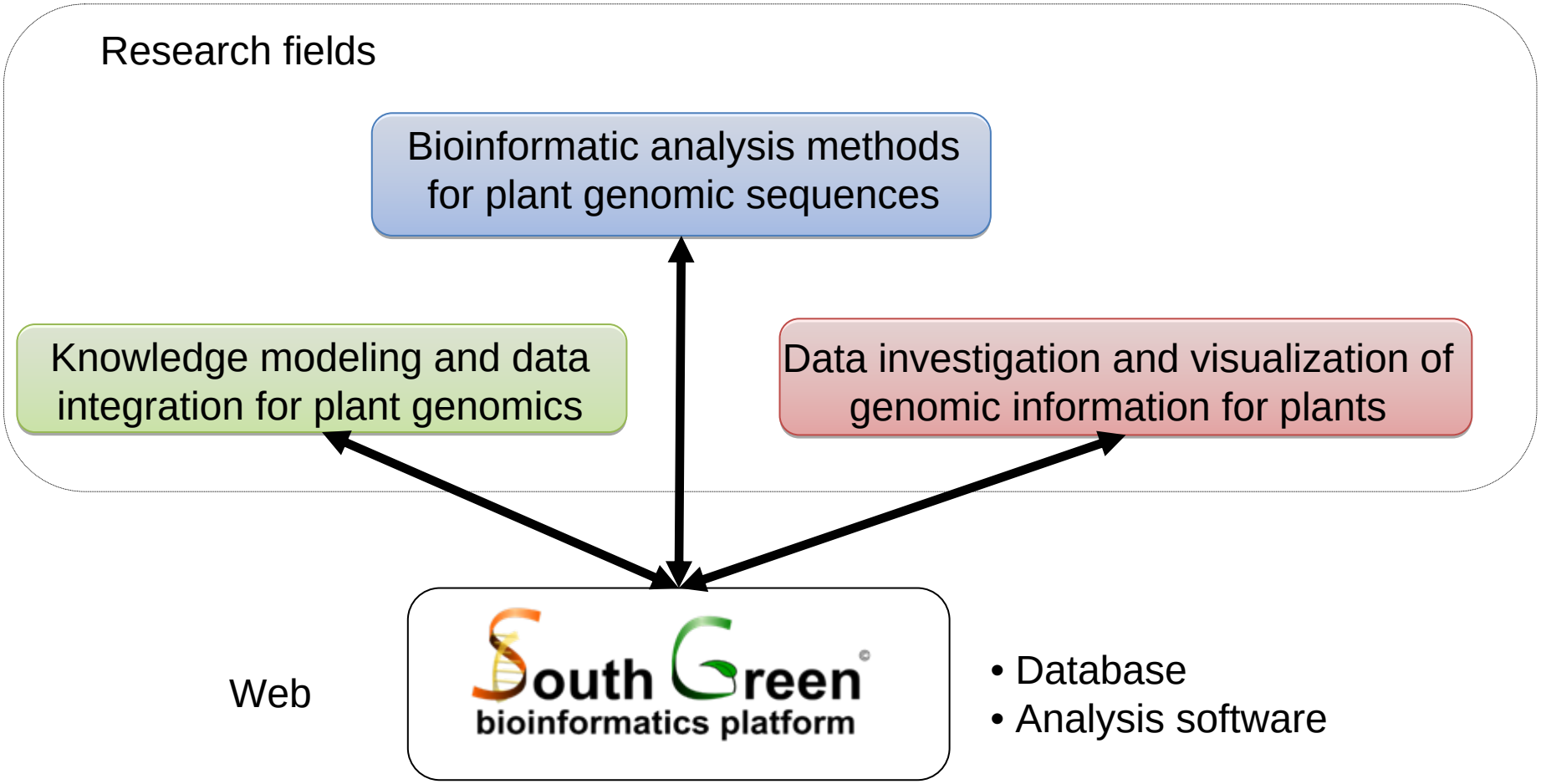
Towards the Genome Analysis

Xavier Argout

04 June 2009

Who are we?

Data Integration team part of the **Research Unit Plant Development and Genetic Improvement**



South Green
bioinformatics platform

http://southgreen.cirad.fr/

Web portal information systems (IS)

TropGENE Database

GenDiversity

CocoaGen DB

EST TDM

SAT SSR Analysis Tool

OryGenesDB

GreenPhyl

Oryza Tag Line

GMOTIS

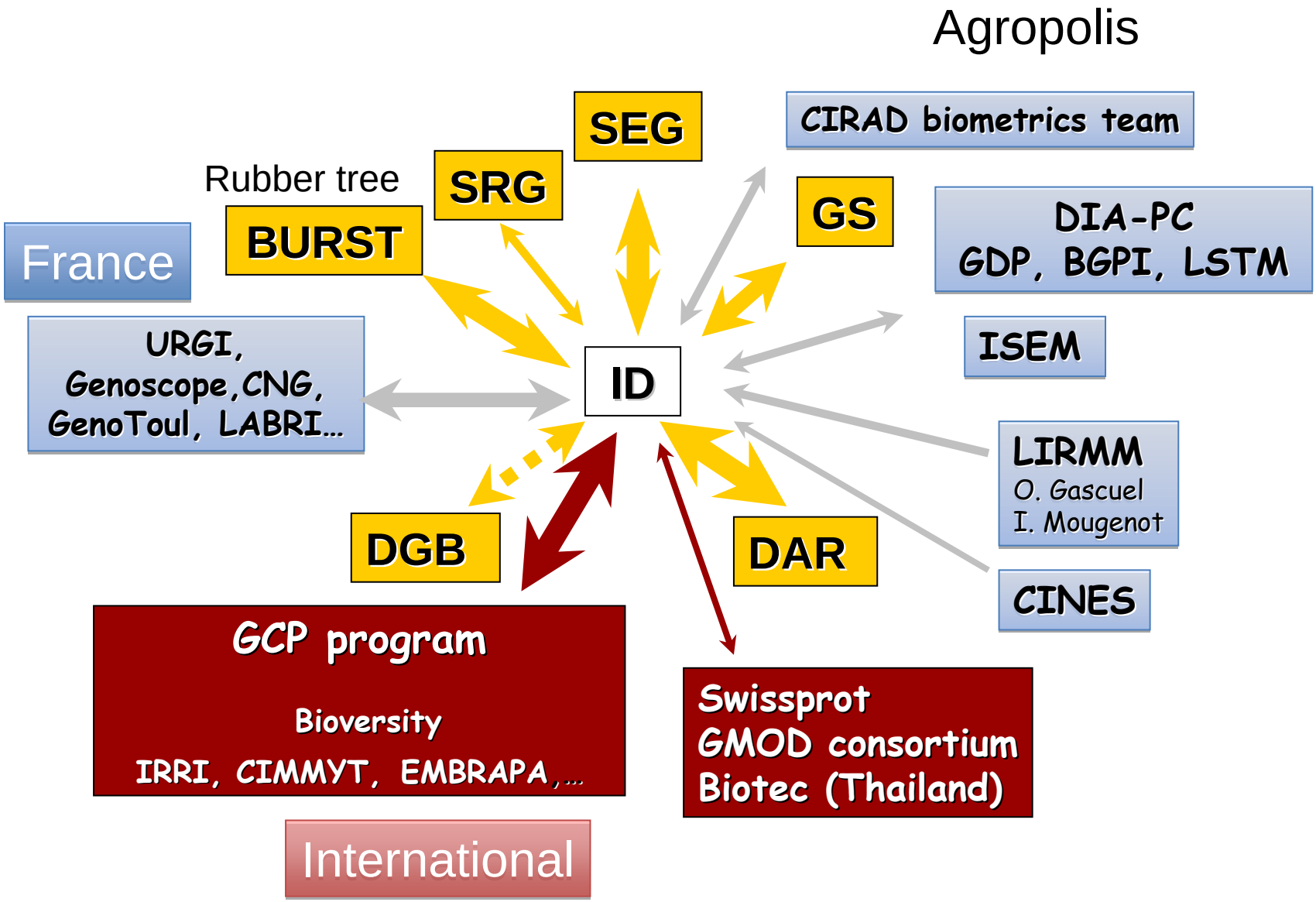
BIBLIOTROP

IS *in development*

Haplophyle

MS-DMind

Partnership



Bioinformatic Analysis methods and Transcriptomic data available for Rubber Tree

ESTtik

A semi-automatic DNA sequence analysis and annotation pipeline for cDNA generation

http://esttik.cirad.fr

esttik@cirad.fr

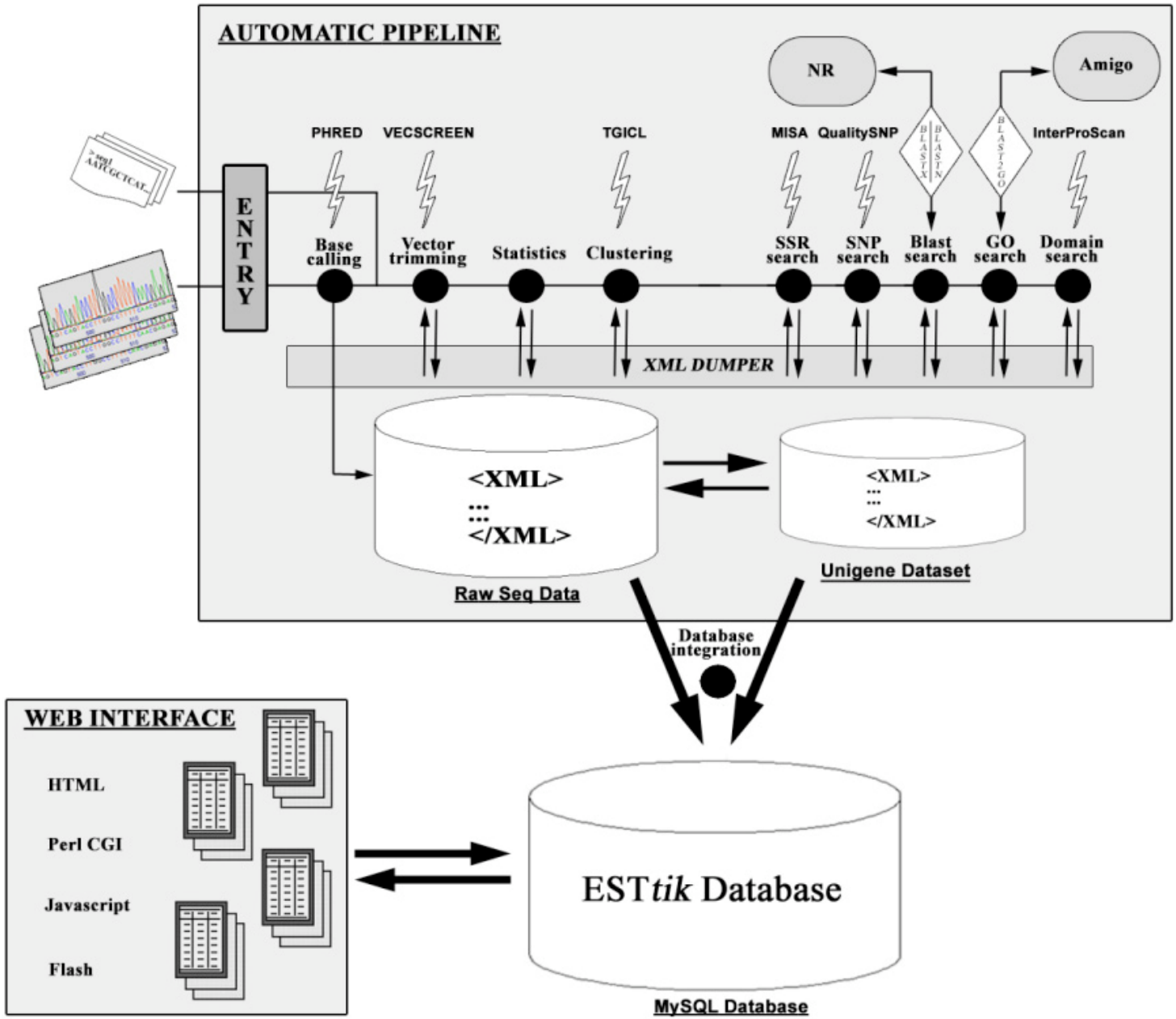
Beginning

- 2 projects :
  - ✓ CITRUS : in 2004, 54 000 ESTs from 9 standard cDNA libraries derived from 4 genotypes
  - ✓ COCOA : in 2005, 150 000 full length cDNA from 56 libraries derived from 14 genotypes

- No automatic tools available for the analysis of these huge data

Project

- Automatic pipeline analysis
  - ✓ Chromatogram input
  - ✓ Vector, adapters and contaminant trimming
  - ✓ Assembly
  - ✓ Annotation
- Database integration
- User friendly interface


AUTOMATIC PIPELINE
NR
Amigo
PHRED
VECSCREEN
TGICL
MISA
QualitySNP
BLASTX
BLASTN
BLASTGO
InterProScan
ENTRY
Base calling
Vector trimming
Statistics
Clustering
SSR search
SNP search
Blast search
GO search
Domain search
XML DUMPER
<XML>
...
</XML>
Raw Seq Data
Unigene Dataset
Database integration
WEB INTERFACE
HTML
Perl CGI
Javascript
Flash
ESTtik Database
MySQL Database

Results

Aeschynomene	19468
Theobroma cacao	183361
Musa	105494
Hevea	31236
Citrus	54000
Tilapia	5250
Total	398809

Publications :

1. Luro FL et al.: Transferability of the EST-SSRs developed on Nules clementine (Citrus clementina Hort ex Tan) to other Citrus species and their effectiveness for genetic mapping. BMC Genomics 2008, 9:287.

2. Argout X et al. : 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. BMC Genomics 2008.

3. Terol J et al.: Analysis of 13000 unique Citrus clusters associated with fruit quality, production and salinity tolerance. BMC Genomics 2007, 8:31.

Actual and future developments

March – September 2009

- 454 technology input module
- New high performance annotation module
- New high performance protein predictions : Prot4EST
- Integration of Blast2GO pipe into the automatic prediction pipeline
- New InterProScan module based on protein predictions
- MicroRNA target prediction module (MIRANDA software?)

- Database integration
- Interface modification
- Modification of “Virtual macroarray tool”

Publication and distribution of this tool before the end of 2009

Hevea transciptomic data

1. Hevea leaves infected by Microcyclus ulei : MDF180 and PB314

Table 1. Summary of the cDNA libraries MDF180 and PB314 created in this study with 6 hours to 58 hours post-infected leaves.

Library	No. of sequences generated	No. of sequences analysed[1]	Singleton (%)[2]	Contigs[3]	Unigene size (%) [4]	Mean size of the sequences (bp)[5]	Redundancy (%)[6]
MDF180 – 6 to 72 hpi	1776	1081 (61)	206 (19)	146	352 (33)	401	67%
MDF180 – 4 to 28 dpi	1790	809 (45)	508 (63)	111	619 (77)	302	23%
PB314 – 6 to 72 hpi	1849	1076 (58)	626 (58)	125	751 (70)	339	30%
PB314 – 4 to 28 dpi	884	715 (80)	39 (5)	58	97 (14)	352	87%
PB314 – 34 to 58 dpi	1728	591 (34)	11 (2)	24	35 (6)	437	94%
Contigated sequences	8027	4272	1165	458	1623	346	62%

Dominique Garcia

Hevea transciptomic data

2. Hevea leaves infected by Microcyclus ulei : R038 and PB260

Banque SSH	Effectif total	Séquences utilisables	Singletons	Contigs	Set unigène 1 (Singletons + contigs)	Séq. Ribosomales	Set unigène 2 (set unigène 1 moins seq. ribosomales)
		(% effectif total)	(% seq. utilisables)		(% seq. utilisables)	(% seq. utilisables)	(% seq. utilisables) total
A	1536	1416 (92 %)	1050 (74%)	97	1147 (81%)	209 (15%)	938 (61%)
B	1536	1435 (93 %)	1005 (70%)	127	1132 (79%)	222 (15%)	910 (59%)
C	1536	1424 (93 %)	892 (63%)	160	1052 (74%)	199 (14%)	853 (56%)
D	1440	1308 (91 %)	970 (74%)	107	1077 (82%)	202 (15%)	875 (67%)
E	1536	1388 (90 %)	991 (71%)	110	1101 (79%)	248 (18%)	853 (56%)
F	1536	1429 (93%)	1135 (79%)	95	1230 (86%)	231 (16%)	999 (65%)
Moyenne		92%				15,5%	

Valérie Pujade

Hevea transciptomic data

3. Public available EST data

Extracted from European EMBL database :

- 10847 ESTs annotated
- 2206 singletons and 1301 contigs generated

- 79 % of redundancy in public database

Marc Seguin

Hevea data

4. High throughput 454 project : cDNA isolated from bark

a half run with a mix of :
- Mature tree (Control, Ethefon, Tapping)
- Juvenile plant (Control, Ethylene, Wounding)

3 main objectives :

Collection of genes expressed in bark

Found target Ethylene Response Factor genes

Study the expression using SOLEXA

Pascal Montoro

Knowledge modeling and data integration for plant genomics

TropGENE Database

a database that manages genetic and genomic information about tropical crops

http://tropgenedb.cirad.fr/

Version 1.0

- **genetic map**
- **QTL data**
- **marker : RFLP, RAPD, SSR, etc.**
- **genotype data**
- **phenotype data**
- **germplasm data**

Banana • Cocoa • Coconut • Coffee • Cotton • Oil Palm • Rice • Rubber Tree • Sugarcane •

Hevea data

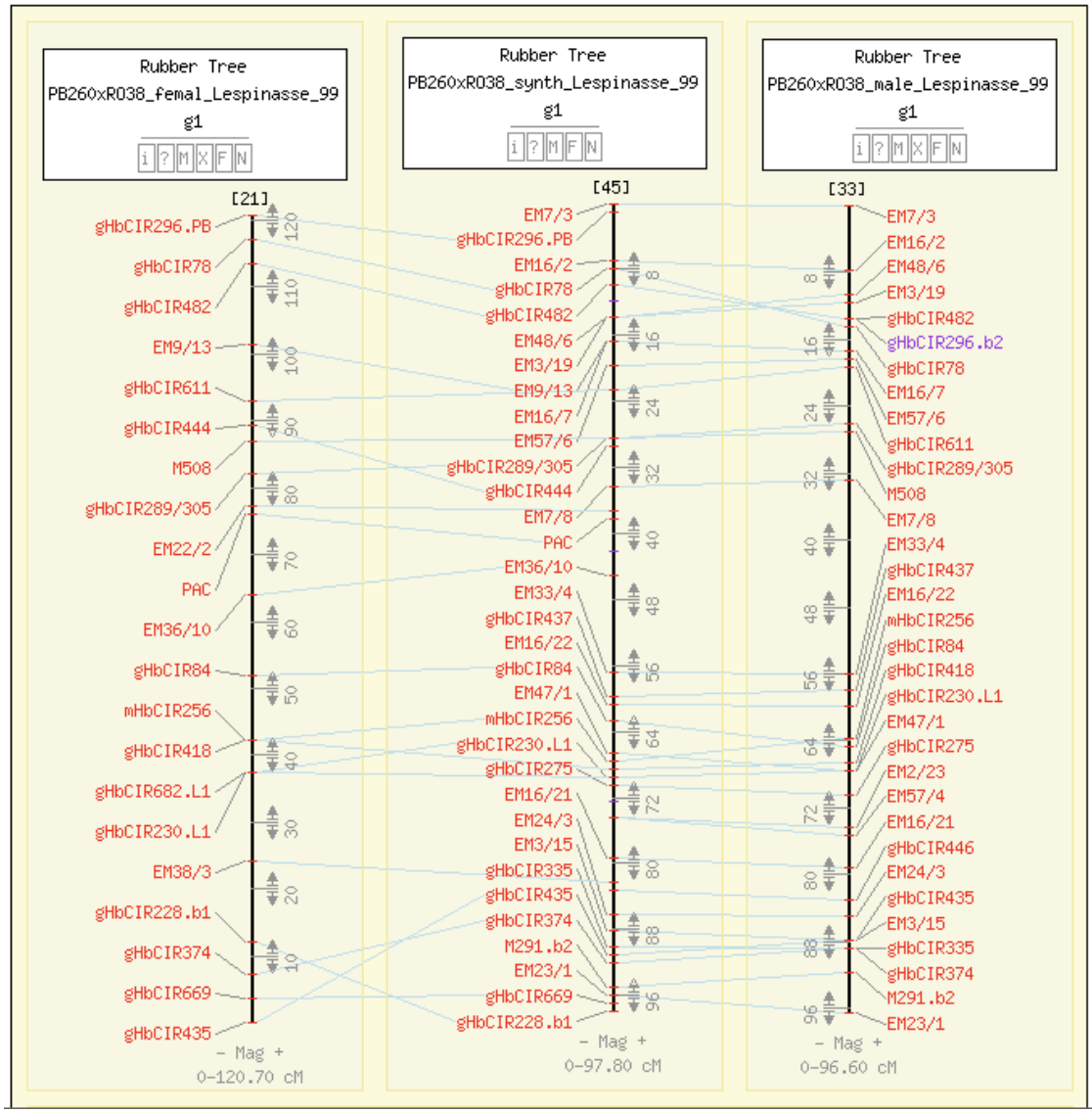
1 synthetic map from PB260xRO38 derived from 2 maps :

- Female (PB260) map PB260xRO38 based on segregation data of markers heterozygous in PB260
- Male (RO38) map PB260xRO38 based on segregation data of markers heterozygous in RO38

Built with :

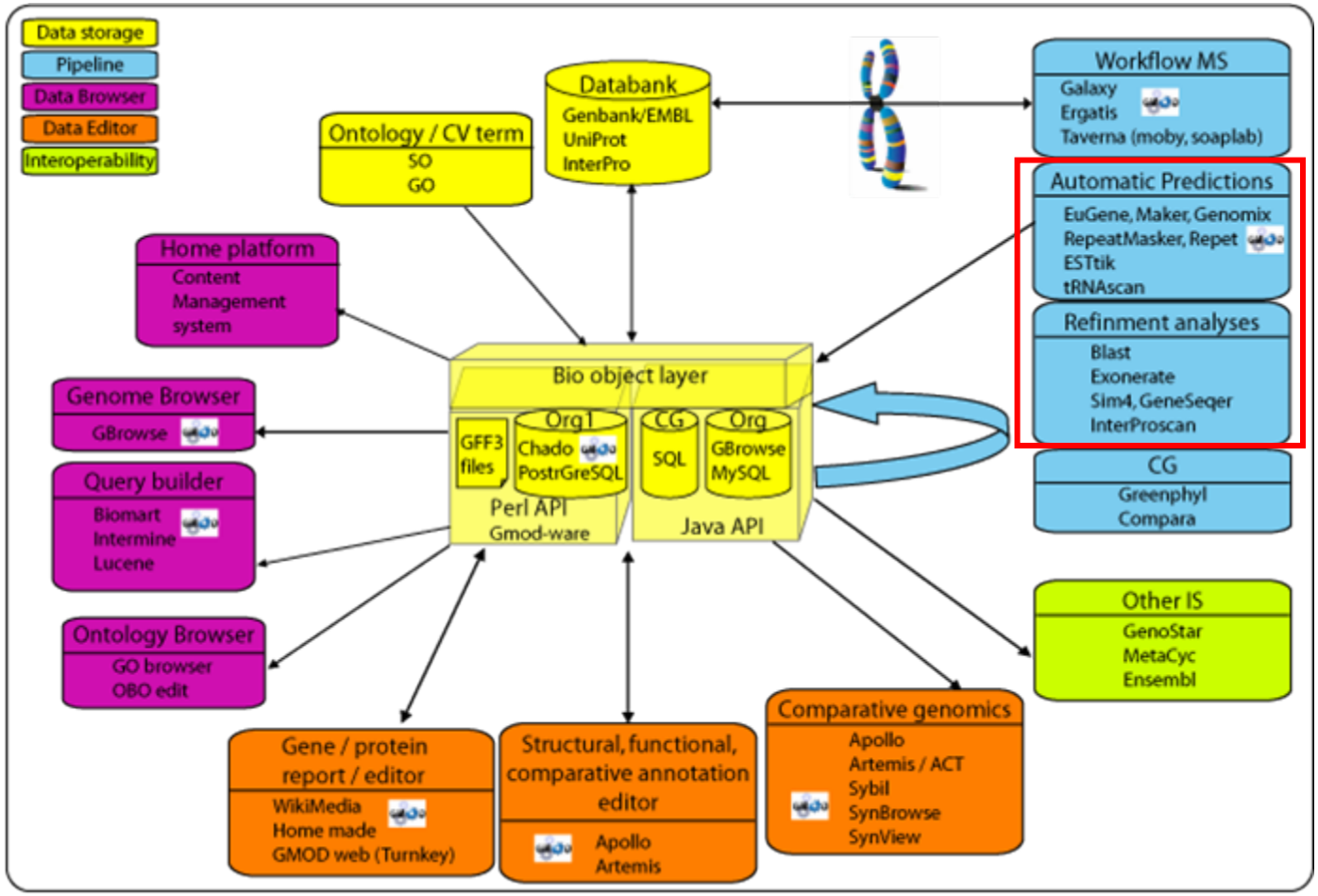
- AFLP
- RFLP
- 18 SSRs
- 10 Isosymes

Marker data also available through TropGeneDB


Rubber Tree
PB260xR038_femal_Lespinasse_99
g1
i ? M X F N
[21]
gHbCIR296.PB
gHbCIR78
gHbCIR482
EM9/13
gHbCIR611
gHbCIR444
M508
gHbCIR289/305
EM22/2
PAC
EM36/10
gHbCIR84
mHbCIR256
gHbCIR418
gHbCIR682.L1
gHbCIR230.L1
EM38/3
gHbCIR228.b1
gHbCIR374
gHbCIR669
gHbCIR435
120
110
100
90
80
70
60
50
40
30
20
10
- Mag +
0-120.70 cM
Rubber Tree
PB260xR038_synth_Lespinasse_99
g1
i ? M F N
[45]
EM7/3
gHbCIR296.PB
EM16/2
gHbCIR78
gHbCIR482
EM48/6
EM3/19
EM9/13
EM16/7
EM57/6
gHbCIR289/305
gHbCIR444
EM7/8
PAC
EM36/10
EM33/4
gHbCIR437
EM16/22
gHbCIR84
EM47/1
mHbCIR256
gHbCIR230.L1
gHbCIR275
EM16/21
EM24/3
EM3/15
gHbCIR335
gHbCIR435
gHbCIR374
M291.b2
EM23/1
gHbCIR669
gHbCIR228.b1
8
16
24
32
40
48
56
64
72
80
88
96
- Mag +
0-97.80 cM
Rubber Tree
PB260xR038_male_Lespinasse_99
g1
i ? M X F N
[33]
EM7/3
EM16/2
EM48/6
EM3/19
gHbCIR482
gHbCIR296.b2
gHbCIR78
EM16/7
EM57/6
gHbCIR611
gHbCIR289/305
M508
EM7/8
EM33/4
gHbCIR437
EM16/22
mHbCIR256
gHbCIR84
gHbCIR418
gHbCIR230.L1
EM47/1
gHbCIR275
EM2/23
EM57/4
EM16/21
gHbCIR446
EM24/3
gHbCIR435
EM3/15
gHbCIR335
gHbCIR374
M291.b2
EM23/1
8
16
24
32
40
48
56
64
72
80
88
96
- Mag +
0-96.60 cM

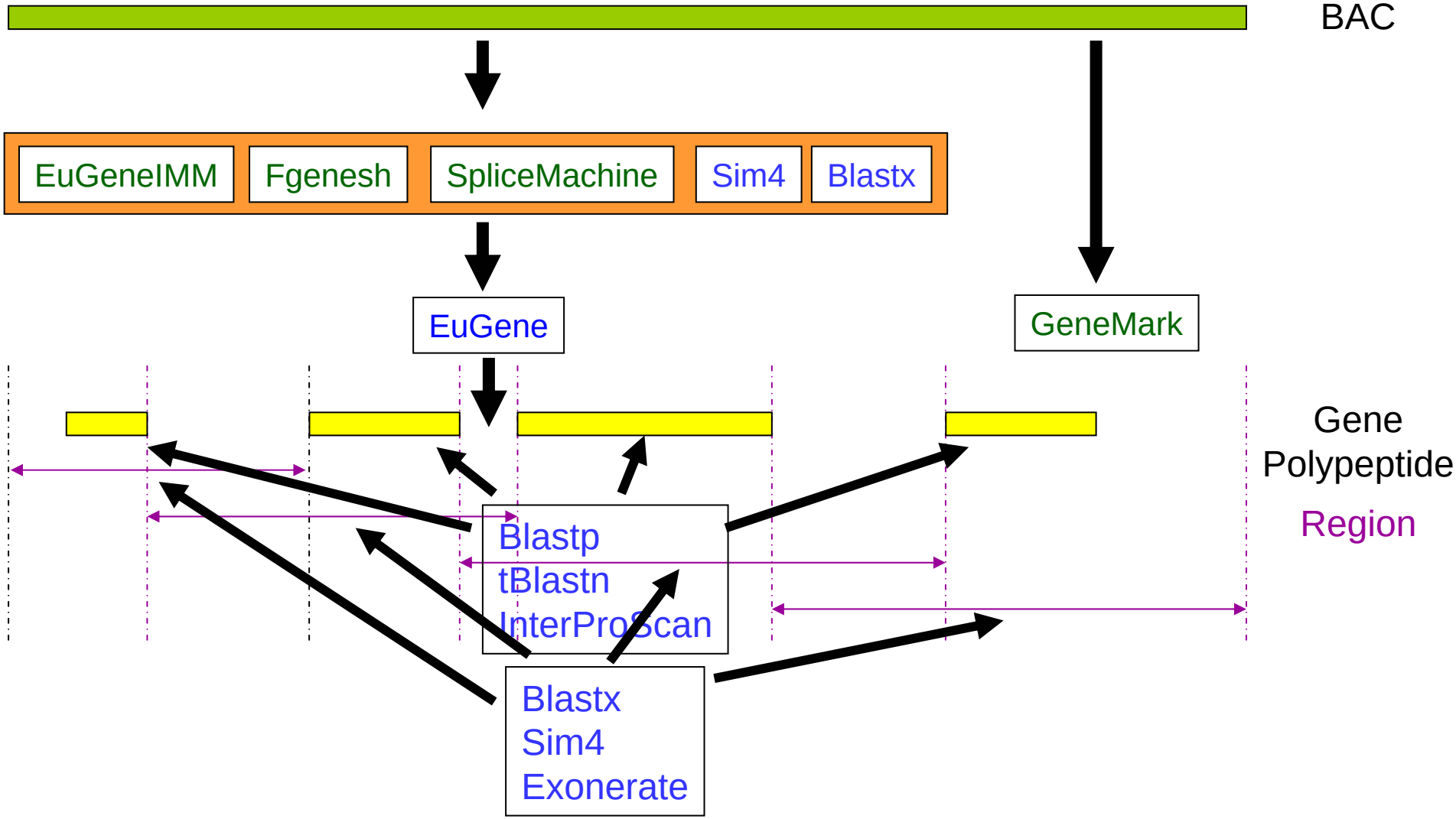
Towards Hevea genome analysis?

A platform of structural and functional annotation dedicated to plant and bio-aggressor genomes supported by comparative genomics



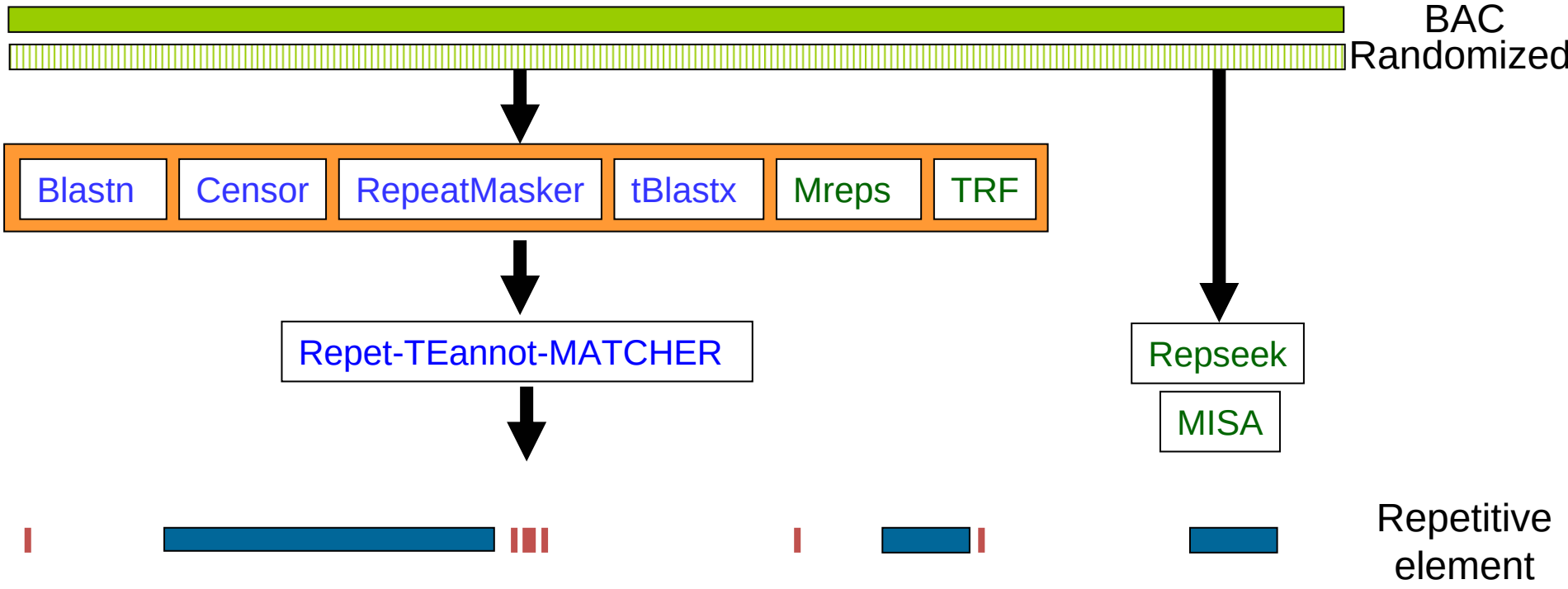
EuGene combiner & annotation refinement

After parameter optimization phase



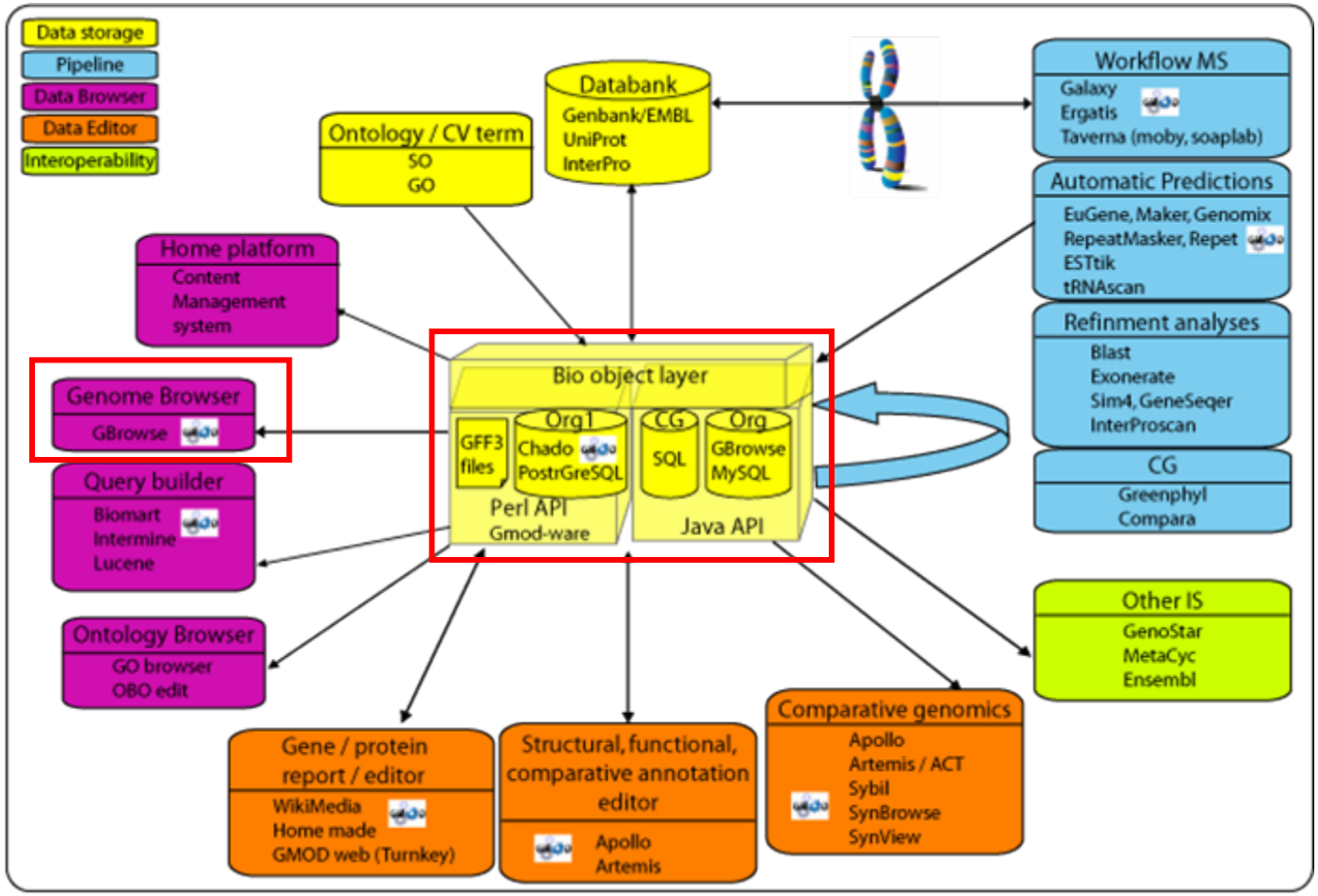
Repet combiner & other repeat analyses

I) parameter optimization phase (TEdenovo)
II) TE prediction (TEannot)

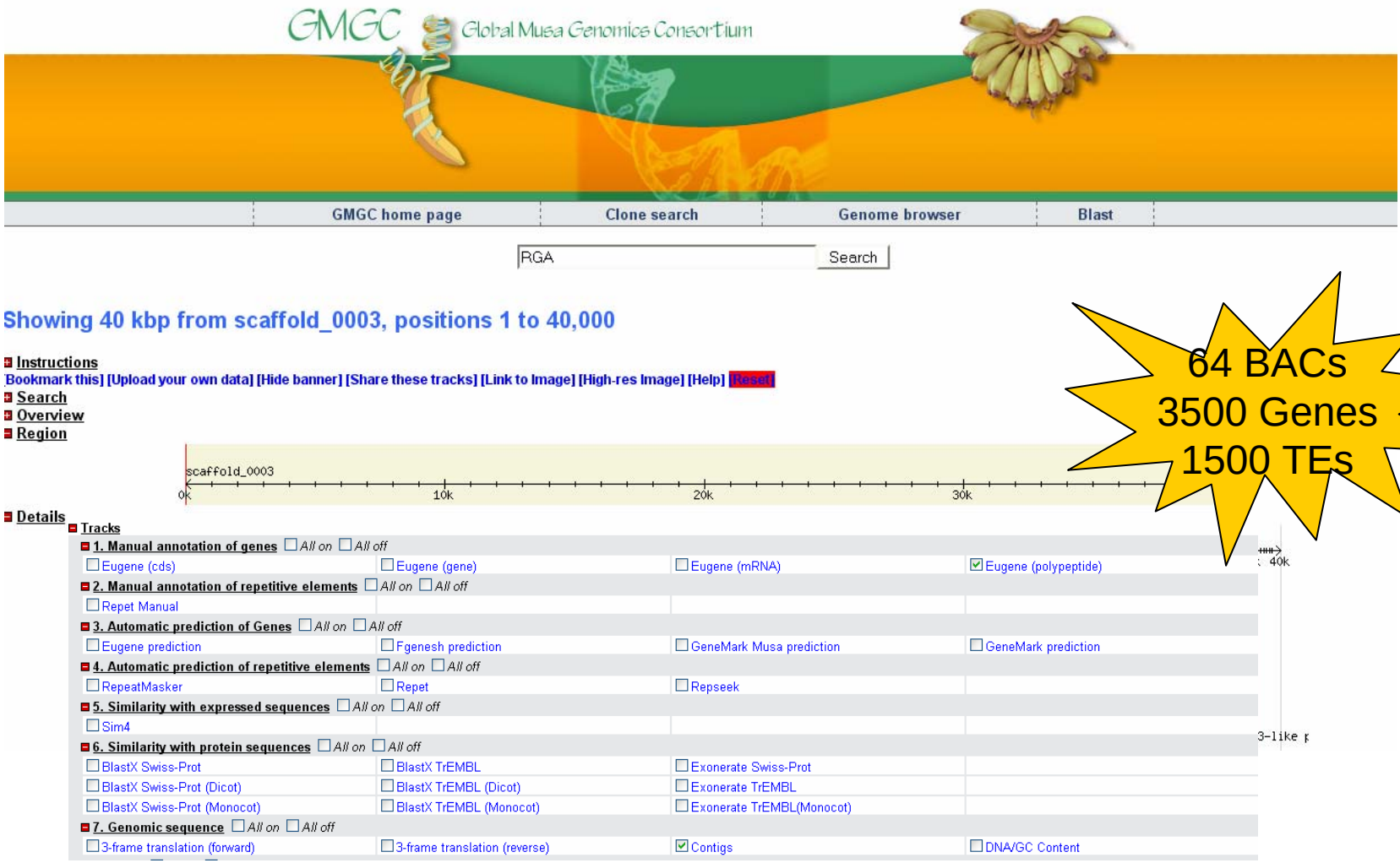


GnpAnnot

A platform of structural and functional annotation dedicated to plant and bio-aggressor genomes supported by comparative genomics



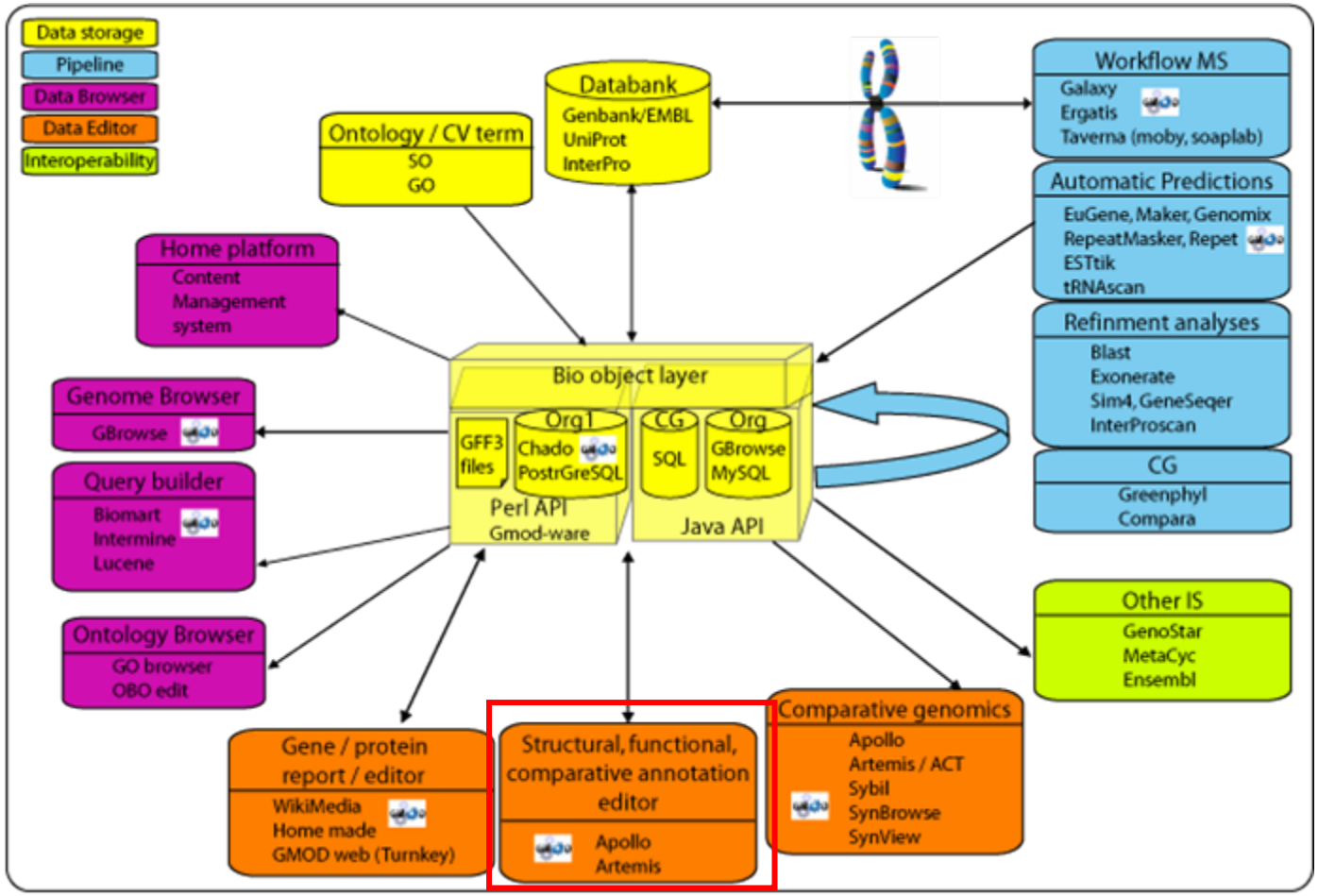
Genome Browser

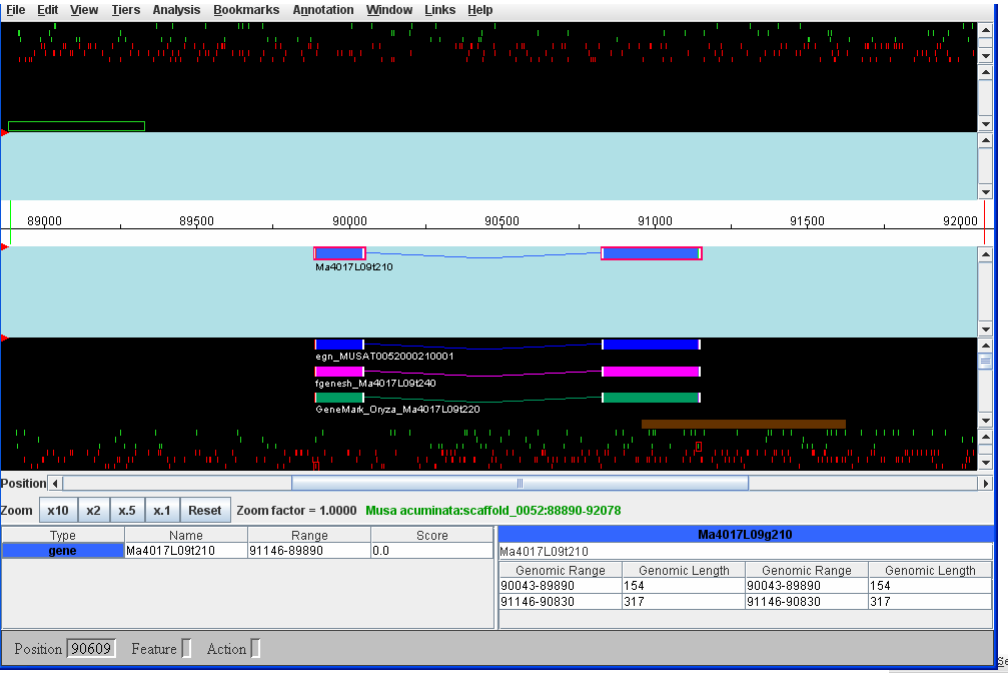


64 BACs
3500 Genes
1500 TEs

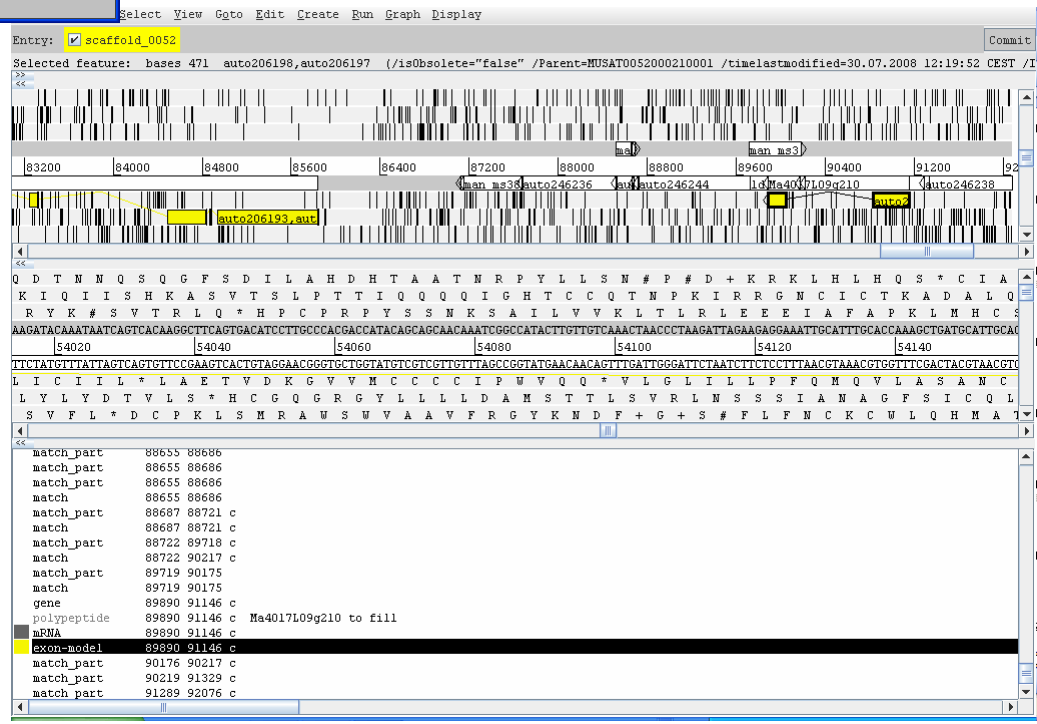
GnpAnnot

A platform of structural and functional annotation dedicated to plant and bio-aggressor genomes

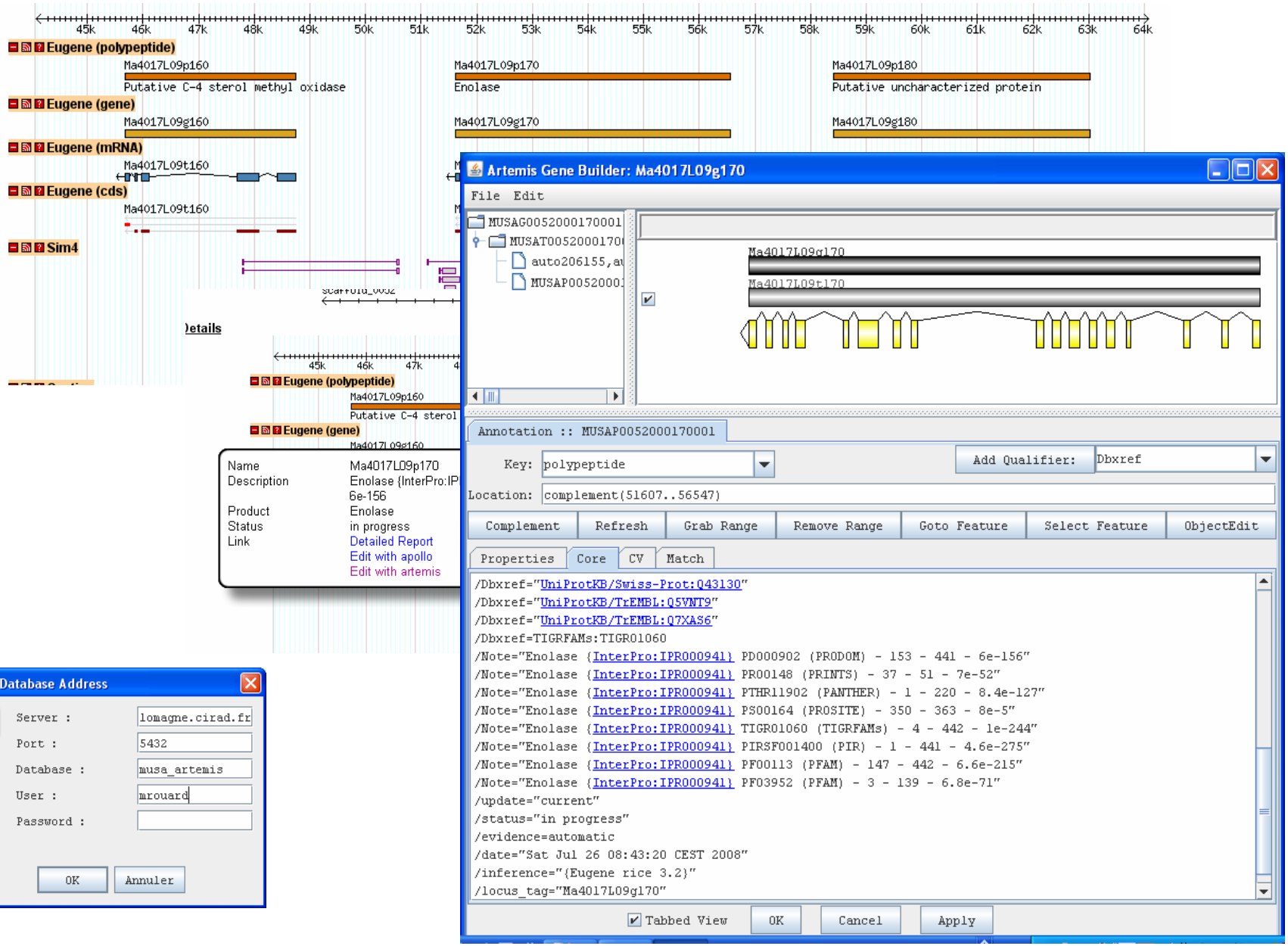


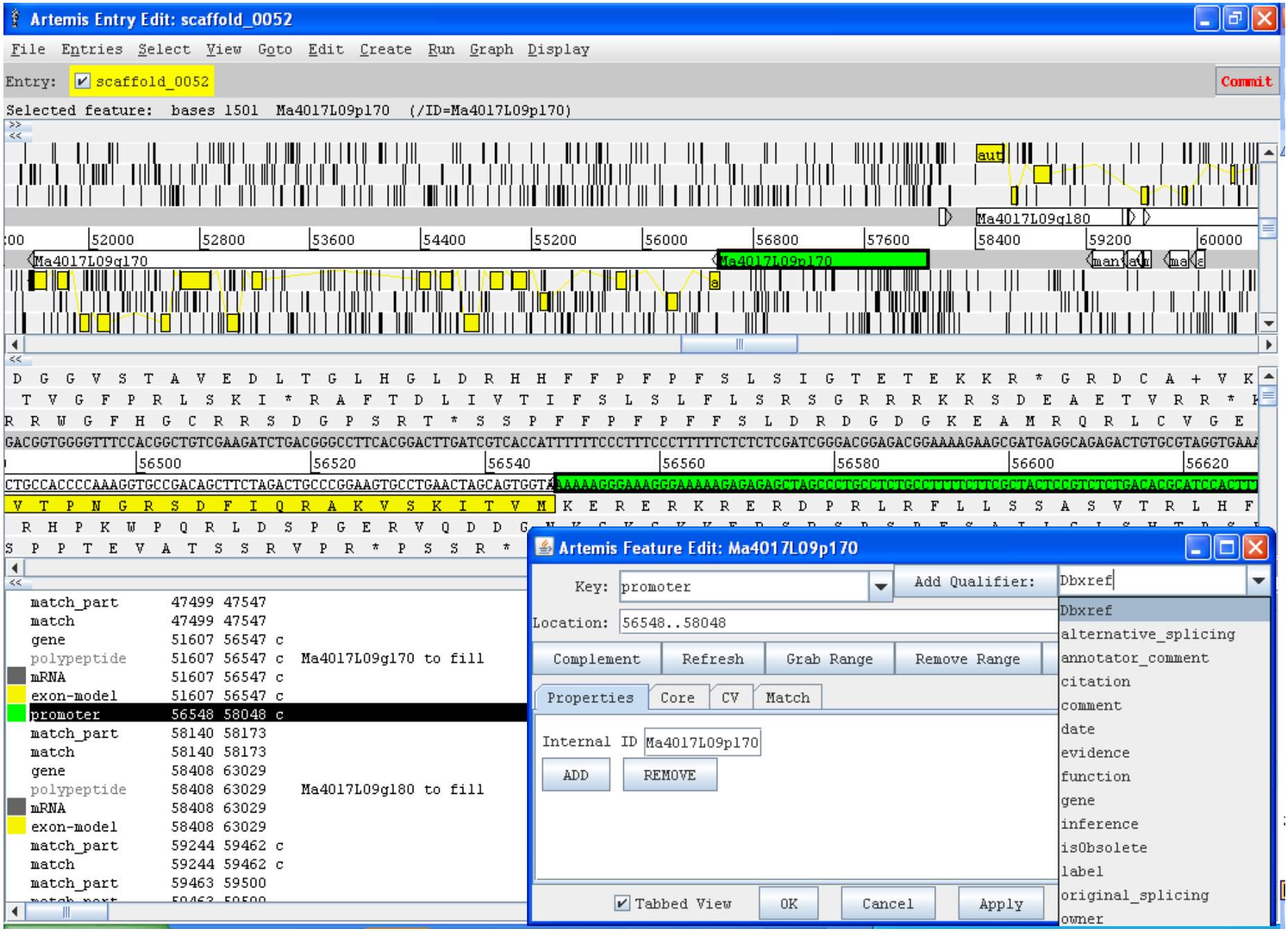


Apollo



Artemis


45k 46k 47k 48k 49k 50k 51k 52k 53k 54k 55k 56k 57k 58k 59k 60k 61k 62k 63k 64k
Eugene (polypeptide)
Ma4017L09p160
Putative C-4 sterol methyl oxidase
Ma4017L09p170
Enolase
Ma4017L09p180
Putative uncharacterized protein
Eugene (gene)
Ma4017L09g160
Ma4017L09g170
Ma4017L09g180
Eugene (mRNA)
Ma4017L09t160
Eugene (cds)
Ma4017L09t160
Sim4
Name Ma4017L09p170
Description Enolase {InterPro:IP
6e-156
Product Enolase
Status in progress
Link Detailed Report
Edit with apollo
Edit with artemis
Enter Database Address
Server : lomagne.cirad.fr
Port : 5432
Database : musa_artemis
User : mrouard
Password :
OK
Annuler
Artemis Gene Builder: Ma4017L09g170
File Edit
MUSAG00520001700001
MUSAT00520001700
auto206155,a
MUSAP0052000
Ma4017L09g170
Ma4017L09t170
Annotation :: MUSAP0052000170001
Key: polypeptide
Add Qualifier: Dbxref
Location: complement(51607..56547)
Complement
Refresh
Grab Range
Remove Range
Goto Feature
Select Feature
ObjectEdit
Properties
Core
CV
Match
/Dbxref="UniProtKB/Swiss-Prot:Q43130"
/Dbxref="UniProtKB/TrEMBL:Q5VNT9"
/Dbxref="UniProtKB/TrEMBL:Q7XAS6"
/Dbxref=TIGRFAMs:TIGR01060
/Note="Enolase {InterPro:IPR000941} PD000902 (PRODOM) - 153 - 441 - 6e-156"
/Note="Enolase {InterPro:IPR000941} PR00148 (PRINTS) - 37 - 51 - 7e-52"
/Note="Enolase {InterPro:IPR000941} PTHR11902 (PANTHER) - 1 - 220 - 8.4e-127"
/Note="Enolase {InterPro:IPR000941} PS00164 (PROSITE) - 350 - 363 - 8e-5"
/Note="Enolase {InterPro:IPR000941} TIGR01060 (TIGRFAMs) - 4 - 442 - 1e-244"
/Note="Enolase {InterPro:IPR000941} PIRSF001400 (PIR) - 1 - 441 - 4.6e-275"
/Note="Enolase {InterPro:IPR000941} PF00113 (PFAM) - 147 - 442 - 6.6e-215"
/Note="Enolase {InterPro:IPR000941} PF03952 (PFAM) - 3 - 139 - 6.8e-71"
/update="current"
/status="in progress"
/evidence=automatic
/date="Sat Jul 26 08:43:20 CEST 2008"
/inference="{Eugene rice 3.2}"
/locus_tag="Ma4017L09g170"
Tabbed View
OK
Cancel
Apply


Artemis Entry Edit: scaffold_0052
File Entries Select View Goto Edit Create Run Graph Display
Entry: scaffold_0052
Commit
Selected feature: bases 1501 Ma4017L09p170 (/ID=Ma4017L09p170)
Ma4017L09g180
Ma4017L09g170
Ma4017L09p170
match_part 47499 47547
match 47499 47547
gene 51607 56547 c
polypeptide 51607 56547 c Ma4017L09g170 to fill
mRNA 51607 56547 c
exon-model 51607 56547 c
promoter 56548 58048 c
match_part 58140 58173
match 58140 58173
gene 58408 63029
polypeptide 58408 63029 Ma4017L09g180 to fill
mRNA 58408 63029
exon-model 58408 63029
match_part 59244 59462 c
match 59244 59462 c
match_part 59463 59500
Artemis Feature Edit: Ma4017L09p170
Key: promoter
Add Qualifier: Dbxref
Location: 56548..58048
Complement
Refresh
Grab Range
Remove Range
Properties
Core
CV
Match
Internal ID Ma4017L09p170
ADD
REMOVE
Tabbed View
OK
Cancel
Apply
Dbxref
alternative_splicing
annotator_comment
citation
comment
date
evidence
function
gene
inference
isObsolete
label
original_splicing
owner

Conclusion

- We conducted research projects on bioinformatics

- We can provide support for :
  - Transcriptomic and genomic sequences
  - Database integration

- We developed a plateform for manual annotation of genomes feature : banana, cocoa, coffea, oil palm, sugarcane...

CIRAD can provide bioinformatic tools for Hevea genome analyses

Acknowledgments

Pascal Montoro

Stephanie Sidibe-Bocs

Manuel Ruiz

Marc Seguin

Dominique Garcia

Valérie Pujade

Mathieu Rouard