

**GNPannot**  
A community annotation system applied to sugarcane sequences

**Oliver GARSMEUR**  
ICSB workshop

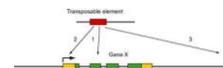
International Plant & Animal Genome XXI January 12-16, 2013 - San Diego, CA, USA

**Introduction to concepts and methods**

**Two main concepts:**

Identify the different elements of the genome, (location and structure) :

**1 - structural annotation**



Attribute a biological information to these elements : **2 - functional annotation**

- Amino acid biosynthesis
- Cellular processes
- DNA metabolism
- Biosynthesis of cofactors, prosthetic groups, and carriers
- Central intermediary metabolism
- Energy metabolism
- Cell envelope
- Disrupting mating barriers
- Fatty acid and phylogenetic metabolism
- Hypothetical proteins
- Pathogen responses
- Purines, pyrimidines, nucleosides, and nucleotides
- Hypothetical proteins (conserved)
- Protein fate
- Regulatory functions
- Mobile and extrachromosomal element functions
- Protein synthesis
- Signal transduction
- Transcription
- Transport and binding proteins
- Unknown function
- Viral functions
- Unannotated

**Automatic annotation – Gene prediction**

**Intrinsic methods (ab-initio)**



- Only based on computational analysis using statistical models.
- Probabilistic models like *Hidden Markov models* (HMM) for discriminating coding and non-coding region of the genome.

The sequences of exons, splice sites and introns have different statistical properties:

- GC% → Introns are AT rich and
- splicing site is almost GT / AG (for plants 95%)

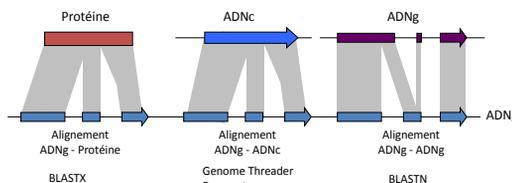
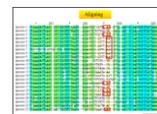
-Need a training set genes. (learning)

The learning set is composed of several hundred of gene-sequences manually annotated derived from cDNAs / genomic alignments. Ideally, these genes represent the diversity of the genes that can be found in the genome.

**Automatic annotation – Gene prediction**

**Extrinsic methods**

Based on **comparative approaches**  
= sequence similarities  
The sequence to annotate is compared with databases.



**Automatic annotation : Extrinsic methods**

**conserved protein domains = signatures**



database of predictive protein "signatures" can be used for the classification and automatic annotation of proteins.

Interproscan classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains and important sites.

**Domain databases** used by interproscan:

- Prosite patterns
- GENE3D
- Pfam
- HAMAP
- ProDom
- PANTHER
- Superfamily
- PIRSF
- TIGRFAMS

**Automatic annotation : Pipeline for functional annotation**



<http://www.blast2go.com/b2ghome>

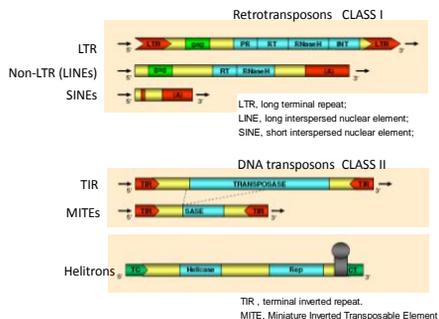
Blast2GO is a bioinformatics tool for functional annotation of sequences, primarily based on the gene ontology (GO) vocabulary.

Transfers the function from homologous sequences through an elaborate algorithm that considers similarity, the extension of the homology, the database of choice, the GO hierarchy. Its also supports InterPro, enzyme codes and KEGG pathways.

QID	Seq	Mapping	Annotations	Statistics	Score	Value	bits
Q12757	Dma 12757.1.32.4	protein	128	11	100.00	100%	0
Q12758	Dma 12758.1.32.4	NON-BIOMAT	144	20	100.00	100%	0
Q12759	Dma 12759.1.32.4	unknown protein	125	10	100.00	100%	0
Q12760	Dma 12760.1.32.4	unknown protein	125	10	100.00	100%	0
Q12761	Dma 12761.1.32.4	unknown protein	144	20	100.00	100%	0
Q12762	Dma 12762.1.32.4	unknown protein	125	10	100.00	100%	0
Q12763	Dma 12763.1.32.4	Protein of unknown fun.	125	10	100.00	100%	0

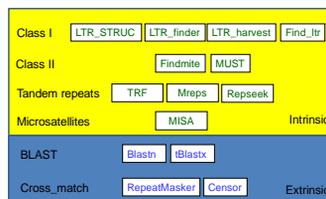
## Transposable Elements (TEs)

Genes represent only a little part of the genome. Some regions can be gene-rich but some other can contain a majority of repeated elements.



## Annotation of transposable elements : tools

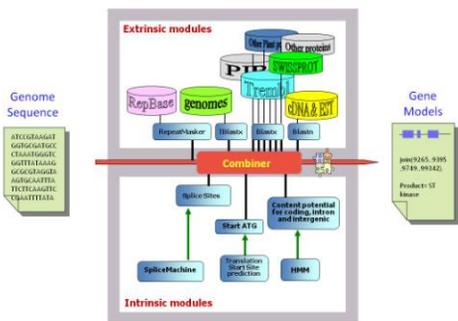
-Several ab-initio programs can be used to detect structures of TEs = intrinsic  
 -Comparisons with databases can be used to classify the elements = extrinsic



- Pipeline to detect and classify TEs = REPET pipeline (Flutre et al, 2011 PLOS one)

## Integrative method – The combiner

Integrative methods = combine ab-initio (intrinsic) + comparative approaches (extrinsic) improving significantly the annotation.



## GNPAnnot Project

GNPAnnot

GNPAnnot is a community system for structural and functional annotation dedicated to plants, insects and fungus genomes allowing both automatic predictions and manual curations of genomic objects.

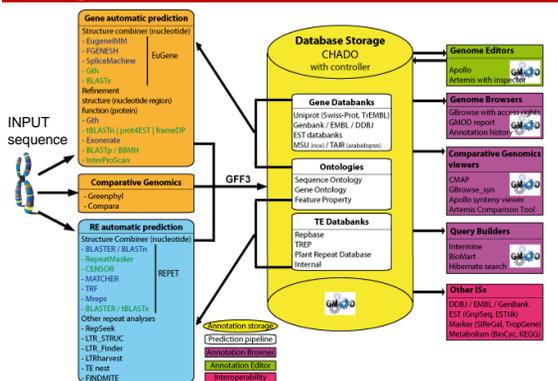


The system is currently being used for various plants, insect and fungus species.

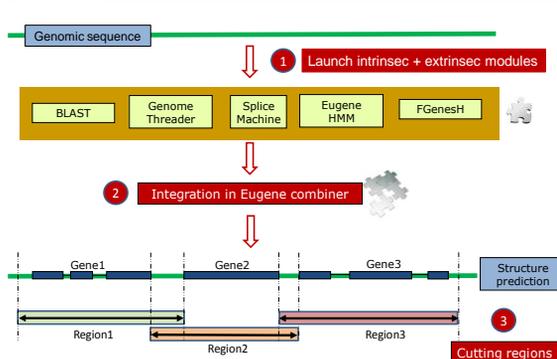
Each specie has a personalized pipeline

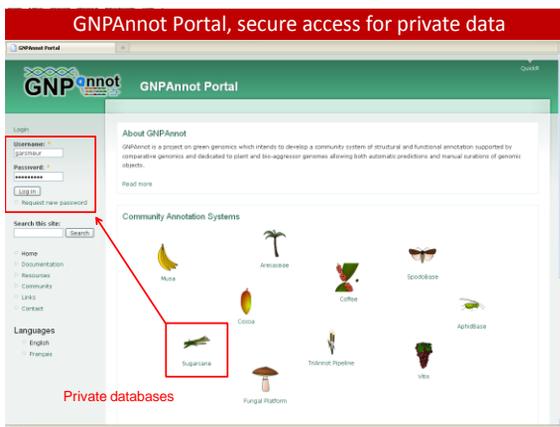
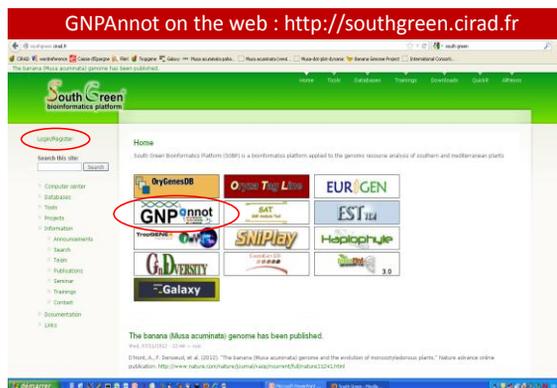
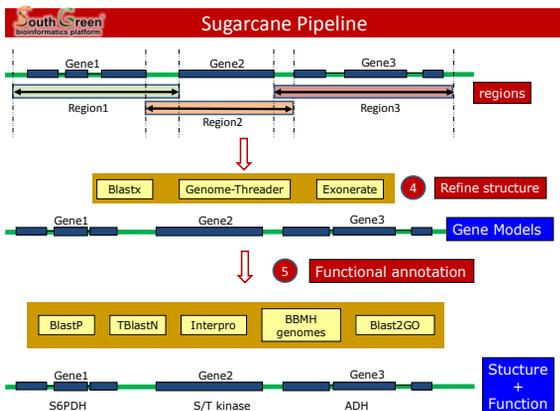
<http://southgreen.cirad.fr/>

## GNPAnnot platform



## Sugarcane Pipeline

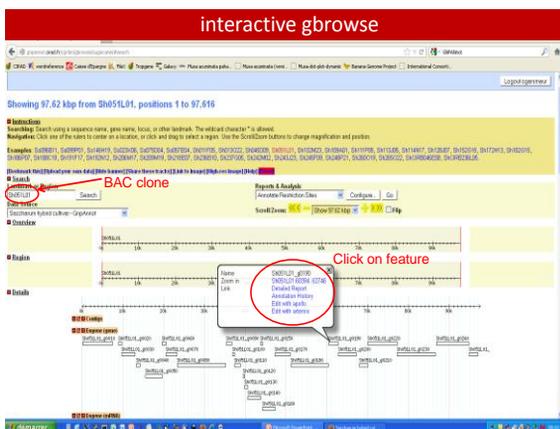




**Tables with links to browse**

The screenshot shows a table of Poaceae statistics. The table has columns for Unique Name, Clone Name, Alias, Accession Number, Length (bp), Predicted Gene Count, Current Gene Count, Curated Gene Count, Observed Gene Count, Current Gene Density (by/region), and %Curated Gene (scufflid length). A red circle highlights a row with the unique name 'SP011331' and the clone name 'SP011331'. A red arrow points to the 'Click on the link to access to the Gbrowse' text below the table.

Unique Name	Clone Name	Alias	Accession Number	Length (bp)	Predicted Gene Count	Current Gene Count	Curated Gene Count	Observed Gene Count	Current Gene Density (by/region)	%Curated Gene (scufflid length)
SP011331	SP011331		125231	36	11	23	12	11895	45%	24%
SP011332	SP011332		137505	51	5	5	0	27801	14%	14%
SP011333	SP011333		65821	22	22	0	0	2592	83%	8%
SP011334	SP011334		101616	31	7	7	0	14519	24%	34%
SP011335	SP011335		137078	42	42	0	0	3284	78%	0%
SP011336	SP011336		95939	23	23	0	0	4148	55%	0%
SP011337	SP011337		94493	20	6	17	11	15749	48%	70%





### Sorghum Gbrowse with R570 BAC clone locations

**Annotations:**

- Sorghum BAC clones:** SORGL11, SORGL12, SORGL13, SORGL14, SORGL15, SORGL16, SORGL17, SORGL18, SORGL19, SORGL20, SORGL21, SORGL22, SORGL23, SORGL24, SORGL25, SORGL26, SORGL27, SORGL28, SORGL29, SORGL30, SORGL31, SORGL32, SORGL33, SORGL34, SORGL35, SORGL36, SORGL37, SORGL38, SORGL39, SORGL40, SORGL41, SORGL42, SORGL43, SORGL44, SORGL45, SORGL46, SORGL47, SORGL48, SORGL49, SORGL50, SORGL51, SORGL52, SORGL53, SORGL54, SORGL55, SORGL56, SORGL57, SORGL58, SORGL59, SORGL60, SORGL61, SORGL62, SORGL63, SORGL64, SORGL65, SORGL66, SORGL67, SORGL68, SORGL69, SORGL70, SORGL71, SORGL72, SORGL73, SORGL74, SORGL75, SORGL76, SORGL77, SORGL78, SORGL79, SORGL80, SORGL81, SORGL82, SORGL83, SORGL84, SORGL85, SORGL86, SORGL87, SORGL88, SORGL89, SORGL90, SORGL91, SORGL92, SORGL93, SORGL94, SORGL95, SORGL96, SORGL97, SORGL98, SORGL99, SORGL100.

### Link to sugarcane GNPAnnot Gbrowse with BAC annotations

**Annotations:**

- BAC clones:** SORGL11, SORGL12, SORGL13, SORGL14, SORGL15, SORGL16, SORGL17, SORGL18, SORGL19, SORGL20, SORGL21, SORGL22, SORGL23, SORGL24, SORGL25, SORGL26, SORGL27, SORGL28, SORGL29, SORGL30, SORGL31, SORGL32, SORGL33, SORGL34, SORGL35, SORGL36, SORGL37, SORGL38, SORGL39, SORGL40, SORGL41, SORGL42, SORGL43, SORGL44, SORGL45, SORGL46, SORGL47, SORGL48, SORGL49, SORGL50, SORGL51, SORGL52, SORGL53, SORGL54, SORGL55, SORGL56, SORGL57, SORGL58, SORGL59, SORGL60, SORGL61, SORGL62, SORGL63, SORGL64, SORGL65, SORGL66, SORGL67, SORGL68, SORGL69, SORGL70, SORGL71, SORGL72, SORGL73, SORGL74, SORGL75, SORGL76, SORGL77, SORGL78, SORGL79, SORGL80, SORGL81, SORGL82, SORGL83, SORGL84, SORGL85, SORGL86, SORGL87, SORGL88, SORGL89, SORGL90, SORGL91, SORGL92, SORGL93, SORGL94, SORGL95, SORGL96, SORGL97, SORGL98, SORGL99, SORGL100.

**Collaborative Network:**

- CIRAD:** Bioanalysis, bioinformatics & Experiments
  - O. Garsmeur
  - C. Charron
  - C. Hervouet
  - S. Sidibe-Bocs
  - G. Droc
  - M. Ruiz
  - A. D'Hont
- GENOSCOPE:** BAC Sequencing
- BECKMAN COULTER GENOMICS**
- INRA CNRGU**
- Bioversity International:** V. Guignon, M. Rouard