GENOME ANNOTATION

INTRODUCTION TO CONCEPTS AND METHODS

Olivier GARSMEUR

Training course in Bioinformatics applied to *Musa* genome
18-22 November 2013

cirad

agap

Bioversity International

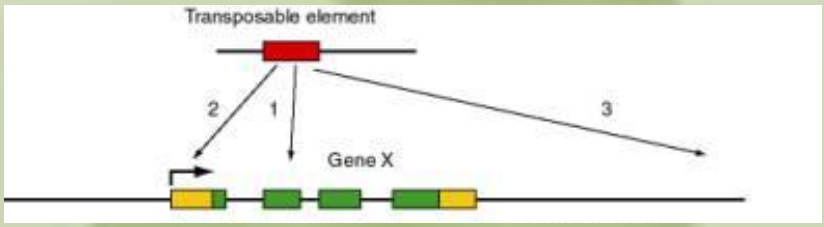
South Green

Introduction

two main concepts:

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

structural annotation



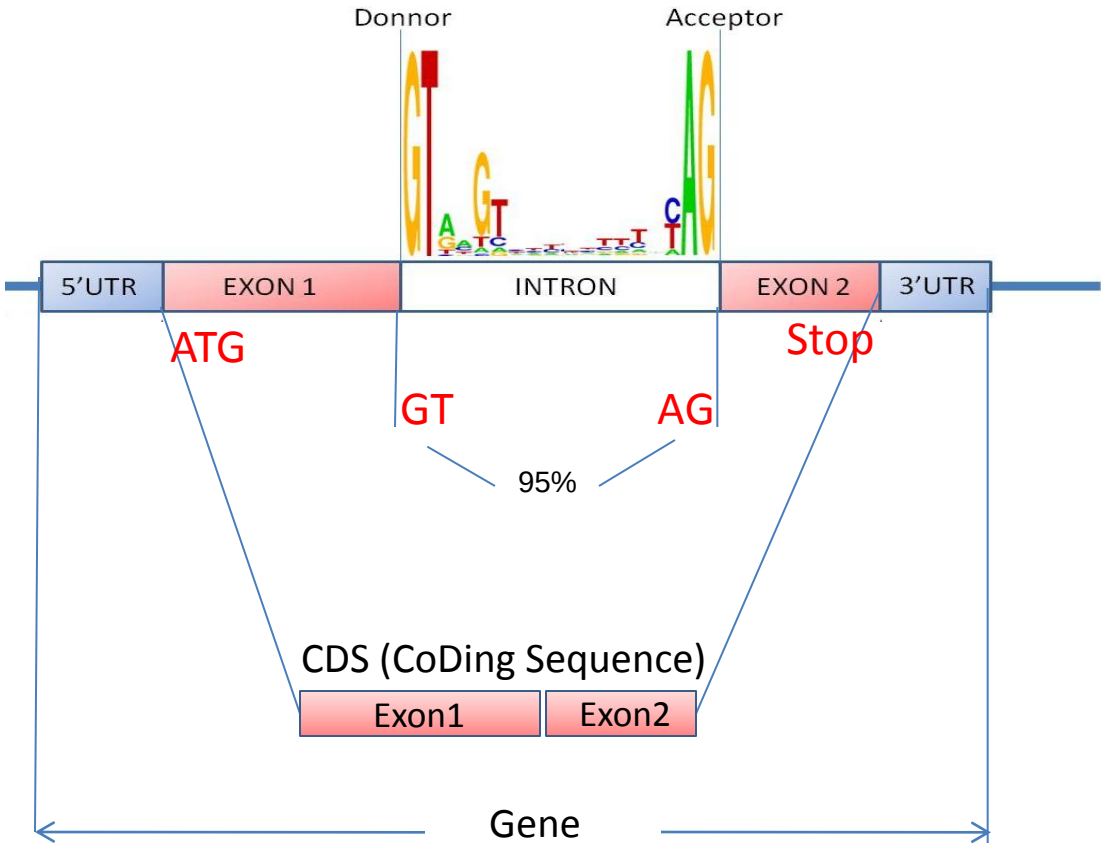
Attribute a biological information to these elements : functional annotation

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

Gene structure

Where are the genes on the sequence ?

Genes predictions (gene-finder softwares) are based on the structure (intron, exon, splice site, UTR).



Predicted functional gene
→ Structure is complete

Starts with M (ATG)
Stop codon TAA, TAG or TGA
GT (GC) / AG splicing site
No stop in exon frame

Pseudogene (not functional)
→ Structure is not complete

missing_acceptor
missing_donor
missing_start_codon
missing_stop_codon
multiple_stop_in_frame

Automatic annotation

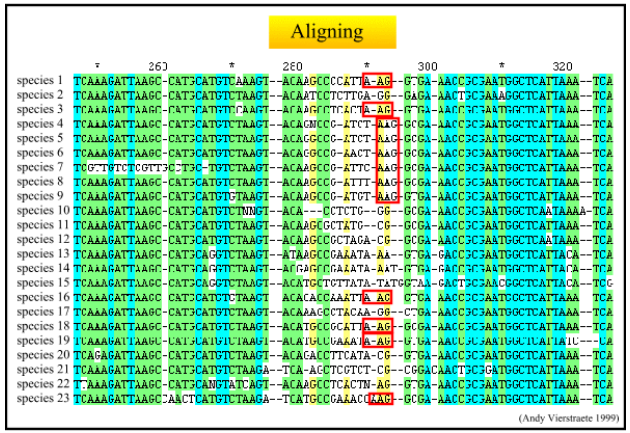
Two methods for automatic predictions

Intrinsec methods (ab-initio)

statistics

Extrinsec methods

Sequence comparisons



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.

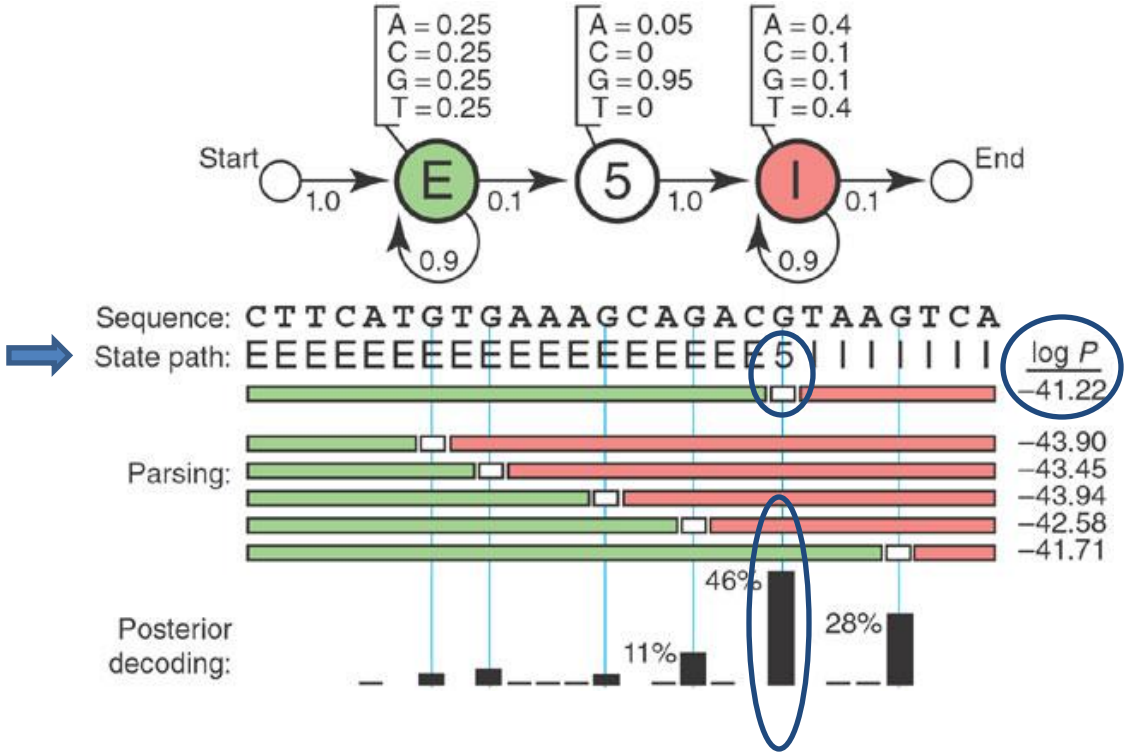
-Need a training set genes. (learning)

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.

Intrinsic methods : Hidden Markov Model = HMM

The sequences of exons, splice sites and introns have different statistical properties, such as GC%. Introns are AT rich and splicing site consensus is almost GT (95%)

Example: Where is the 5' splicing site (5) corresponding to the transition between the exon (E) and the intron (I). ➡ 3 states : E, I, and 5

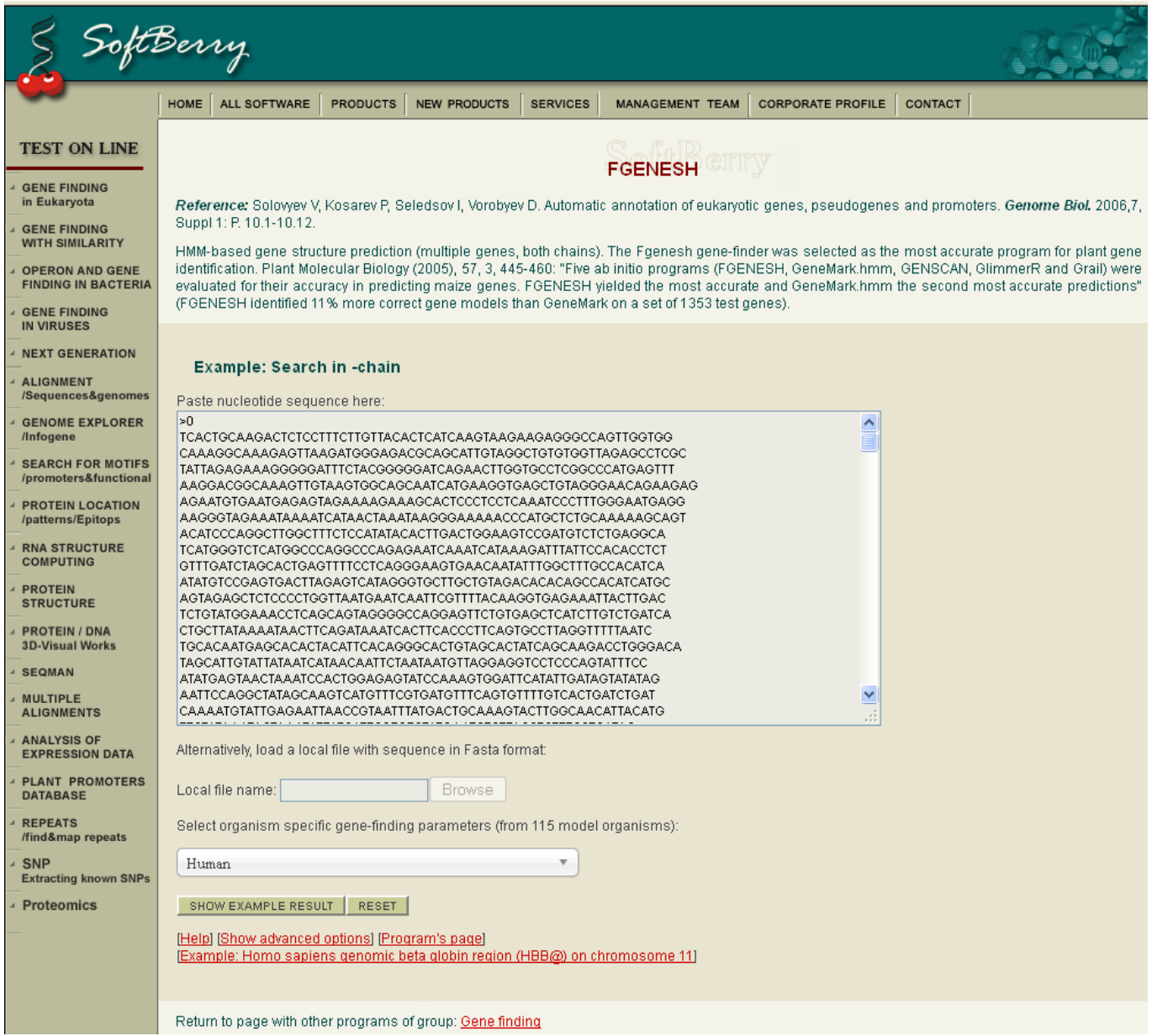


Probabilities of moving from a state to another state

The transition probabilities describe the linear order in which we expect the states to occur: Exons, Introns, Splicing site.

HMM based gene structure prediction : FGENESH

http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind



Automatic annotation : tools

Example of prediction softwares

Plants

GeneMark.HMM
GeneFinder
Eugene-HMM
FgeneSH
GlimmerA
Augustus

GNPannot

Animals

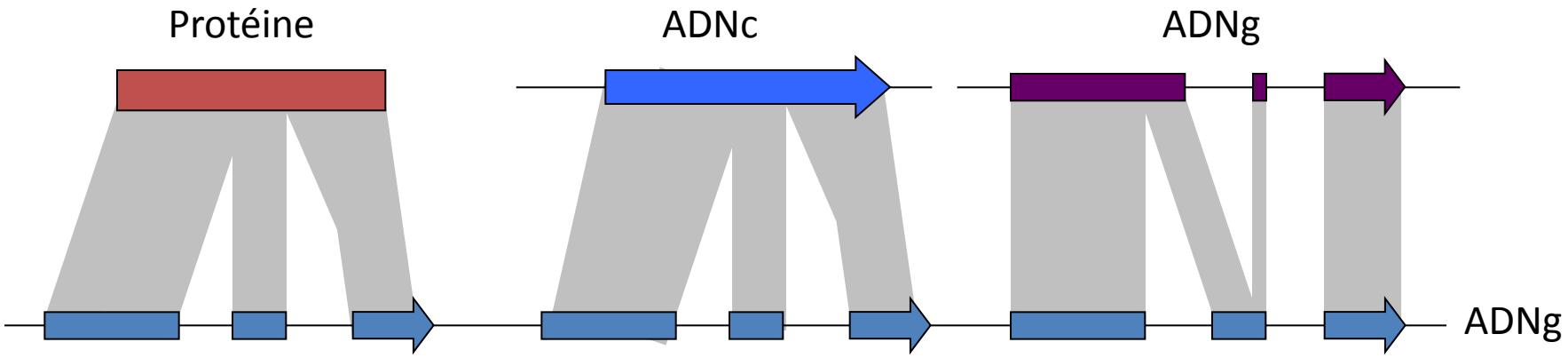
Genie
HMMgene
MagPie
GeneID
Grail

Human

GeneScan
GeneFinder
GeneWise

Extrinsic methods

Comparative approach based on sequence similarities.
The sequence to annotate is compared with databases.



Alignement
ADNg - Protéine

Alignement
ADNg - ADNc

Alignement
ADNg - ADNg

Extrinsic methods

	sequence (query)	target (Subject)	database
BLASTN	nucleotide	nucleotide	NR , EST, genomes
BLASTX	translated nucleotide	protein	Swissprot-Trembl
BLASTP	protein	protein	Swissprot-Trembl
TBLASTX	translated nucleotide	translated nucleotide	NR , EST, genomes
TBALSTN	protein	translated nucleotide	NR , EST, genomes

Comparisons with existing genes or proteins is helpful to refine the structure of predictions.

The Functional annotation will be always deduced of the similarities (homology) of the predicted elements with databases.

Other alignment tools

Genome Threader

Use a similarity-based approach where cDNA/EST and/or protein sequences are compared to the genomic sequence

Useful to predict the splice sites

www.genomethreader.org/

Exonarate

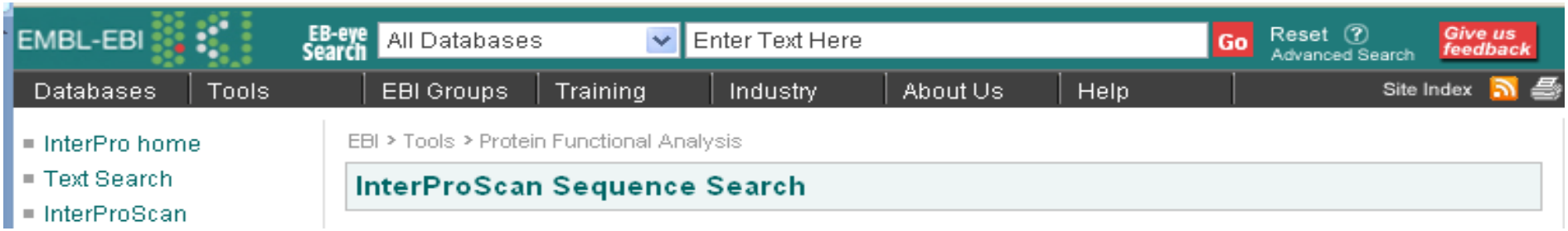
This is a generic tool for pairwise sequence comparison using a many alignment models

This tool is used to refine the structure of gene models (based on cDNA alignments)

http://www.ebi.ac.uk/~guy/exonerate/

Search for conserved protein domains

INTERPROSCAN http://www.ebi.ac.uk/interpro/



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	

database limits

Proteins databases are almost derived from the automatic translation of nucleic sequences. For example, TrEMBL is the automatic translation of the nuclein EMBL (NR) database. (No biological evidence)

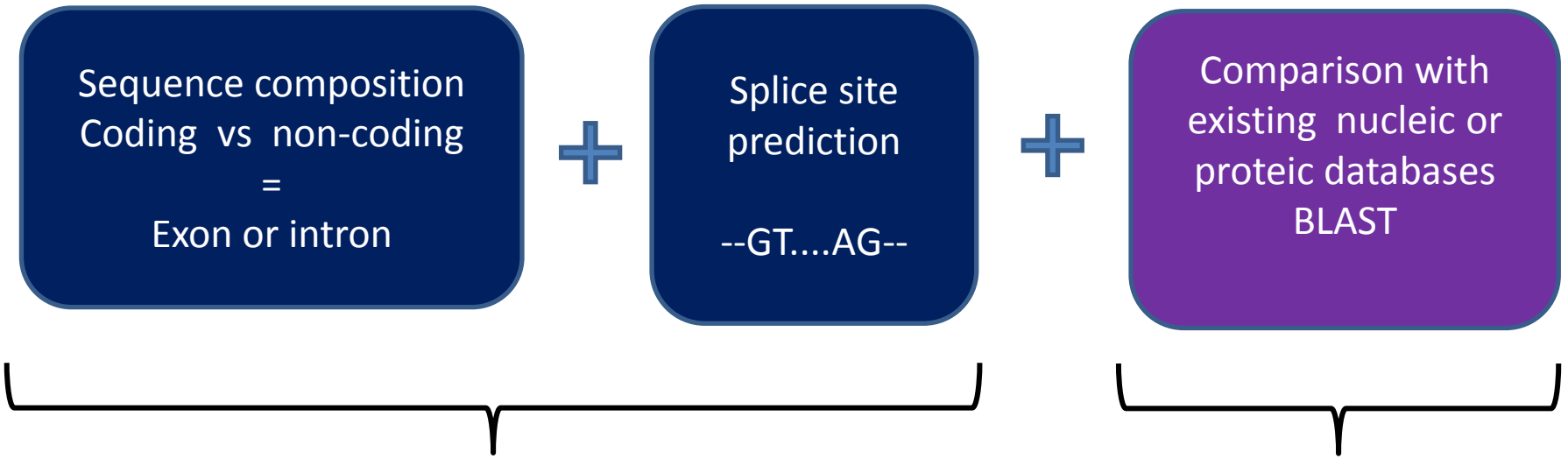
→ SWISSPROT, is another protein database that contains less sequences than TrEMBL, but sequences are manually curated by biologists.

ESTs sequences are often produced from single strand sequencing and contain mistakes.

→Cluster of ESTs are available on a lot of species and represent genes that are really expressed.

Genome sequences are useful, but gene predictions are based on automatic annotations. Taking in account genome evolutions, only close relative species can be used to transfer the annotation.
→For sugarcane annotation, the sorghum sequence is available and represents a good model for comparison.

Integrative method – The combiner



intrinsec methods

Ab initio - HMM
with training step

Extrinsec methods

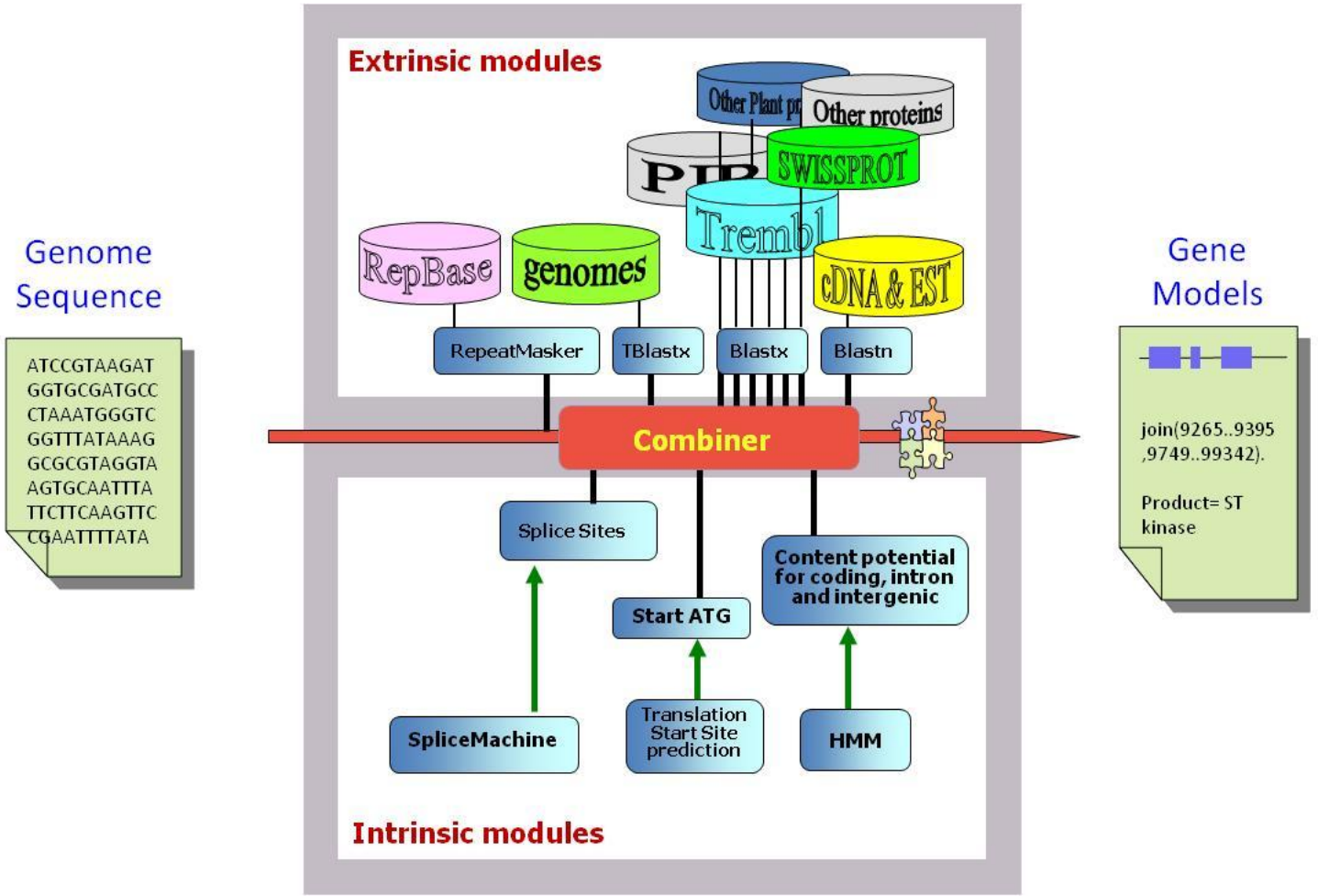
By similarity
= sequence comparisons

Integrative method – The combiner

Integrative methods = ab-initio + comparative approaches

Predictions of ab-initio gene finders combined with the database similarities improve significantly the annotation.

Intrinsic and extrinsic method complement each other.



Automatic annotation tool : Eugene (combiner)

http://eugene.toulouse.inra.fr/

EuGène : A Gene Finding Software

P. BARDOU · M.J. CROS · S. FOISSAC · *J. GOUZY* · A. MOISAN · T. SCHIEX

Home
Plugins
Organisms
References
Documentation
Download
GForge
Links - Contacts

Organisms

- Arabidopsis thaliana :

Arabidopsis thaliana (L.) Heynh. mouseear cress	
Group	Dicot
Family	Brassicaceae
Growth Habit	Forb/herb
Duration	Annual
Links	
USDA, NRCS. 2002. The PLANTS Database, Version 3.5.	
ARS Germplasm Resources Information Network (GRIN)	

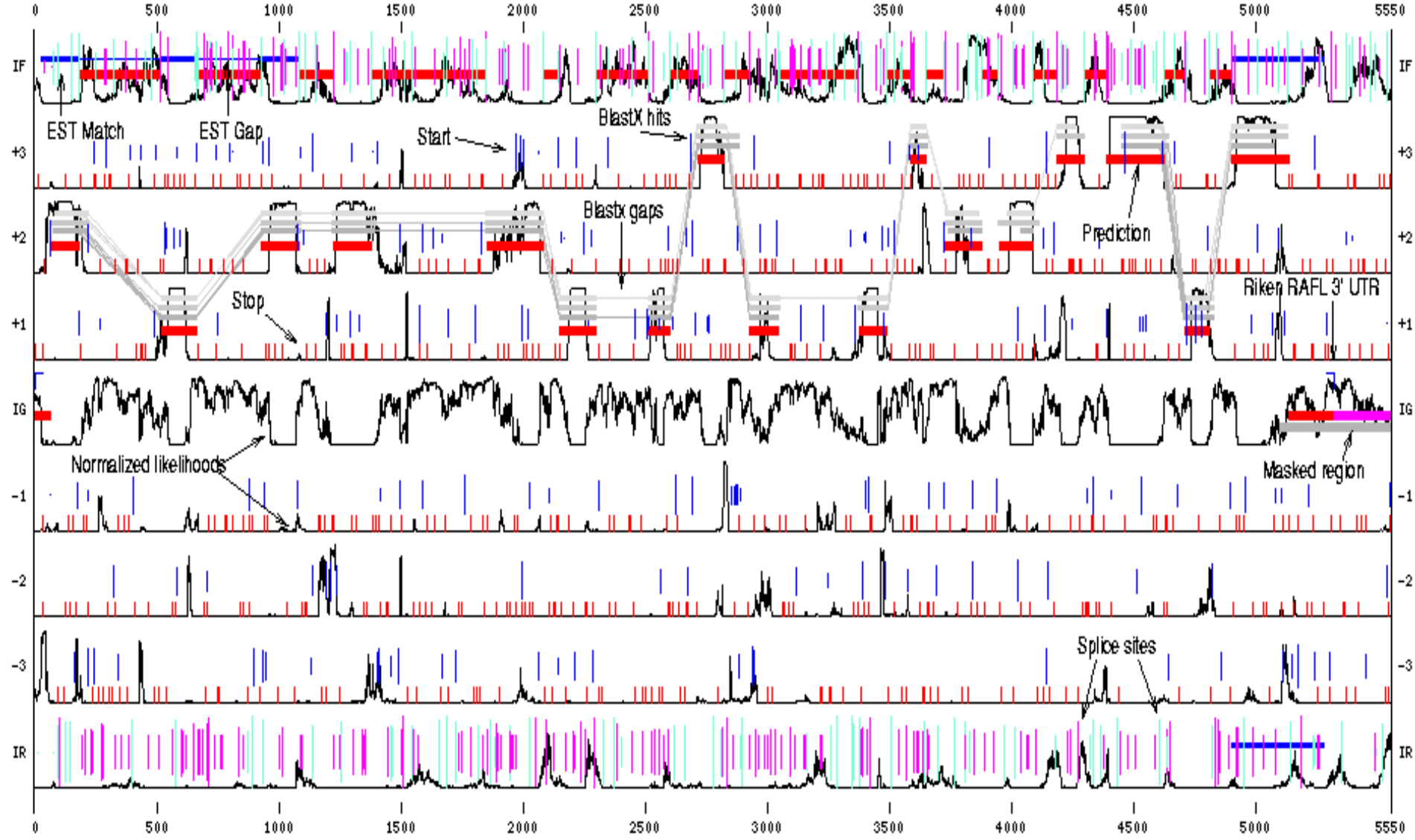
An evaluation of EuGène and several gene prediction programs was performed.

- Oryza sativa :

Oryza sativa L. rice	
Group	Monocot
Family	Poaceae
Growth Habit	Graminoid
Duration	Annual
Links	
USDA, NRCS. 2002. The PLANTS Database, Version 3.5.	
ARS Germplasm Resources Information Network (GRIN)	

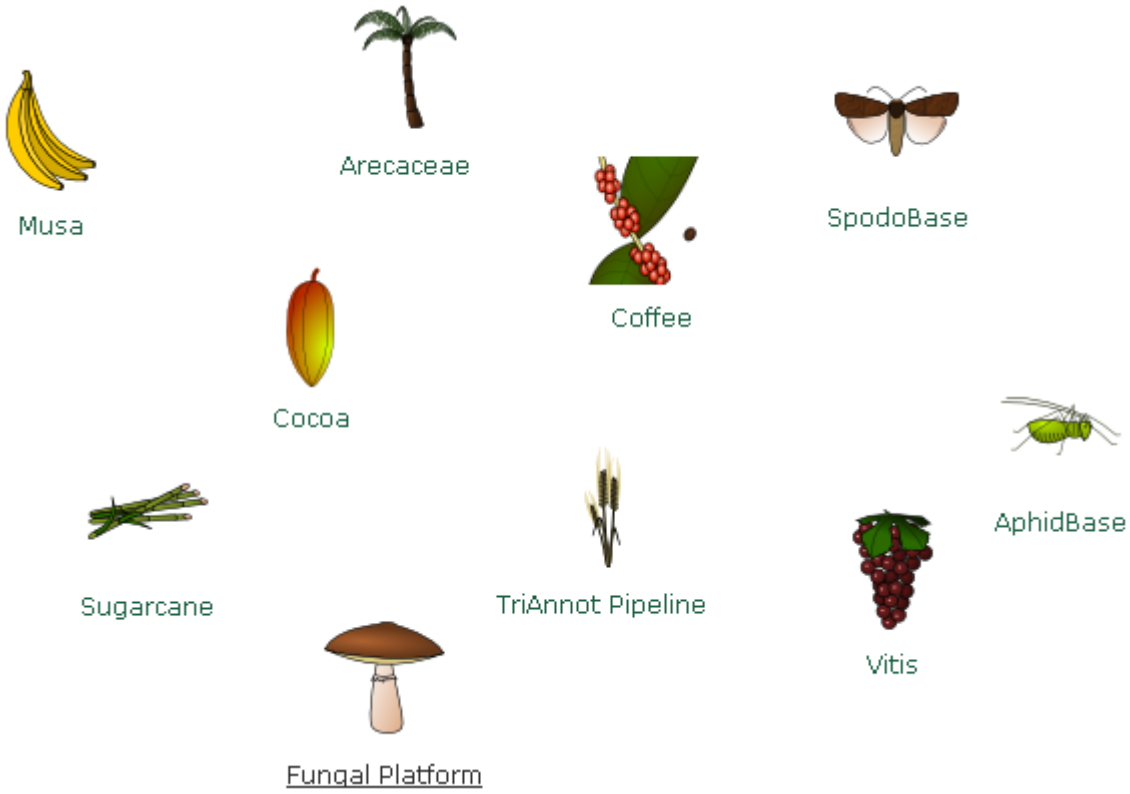
- Medicago truncatula :

Eugène graphical output



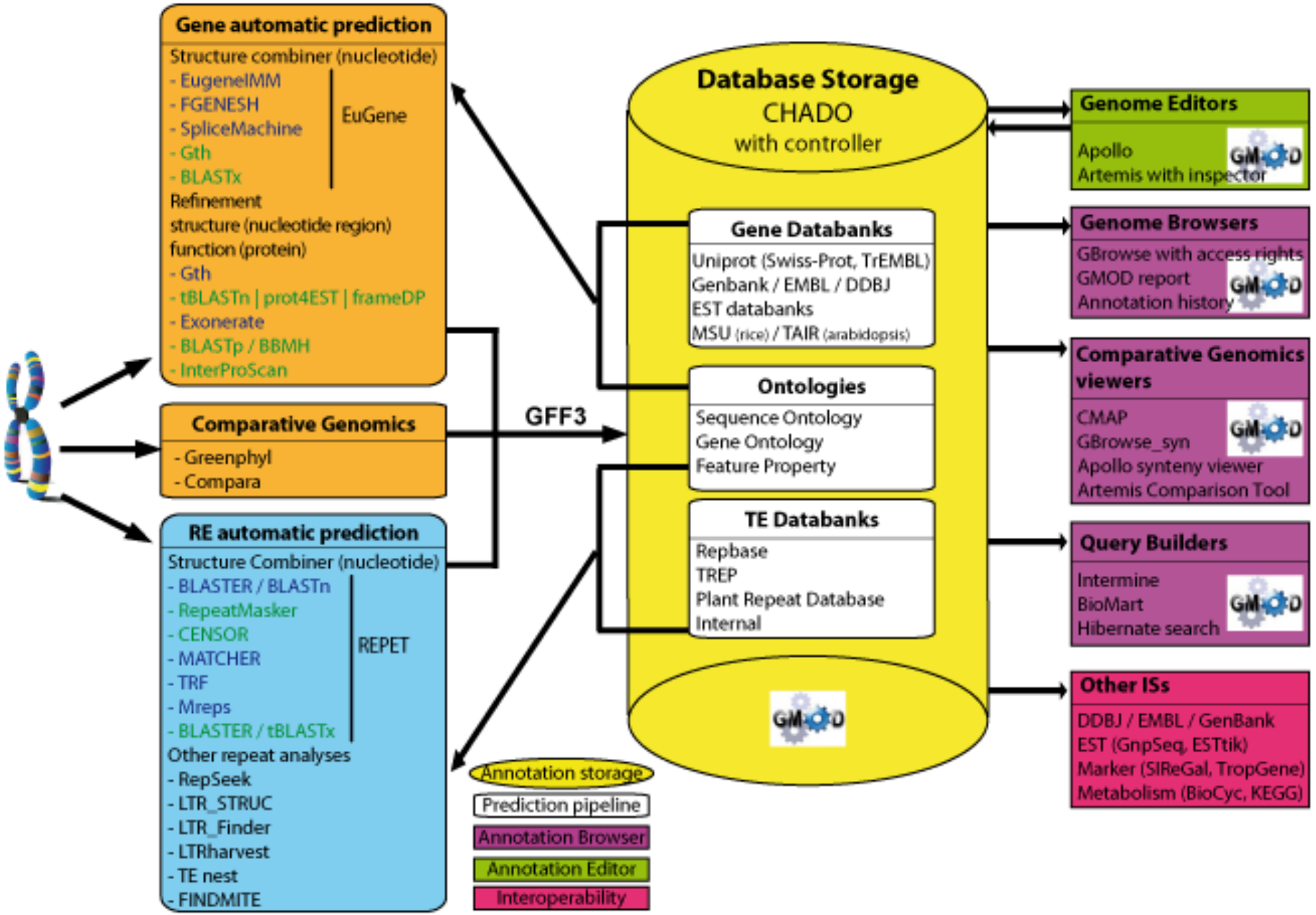
GNPAnnot Project

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.

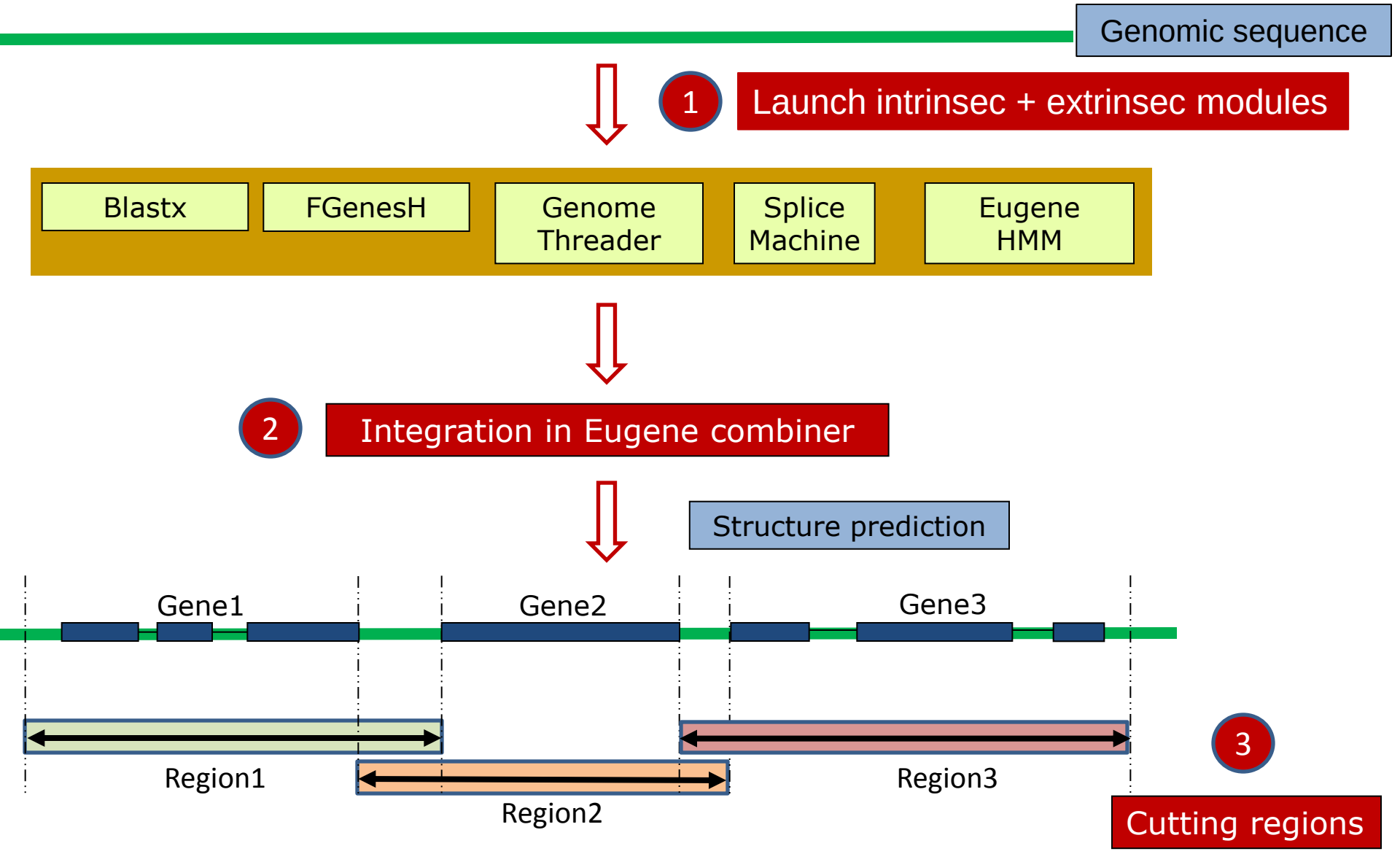


http://southgreen.cirad.fr/

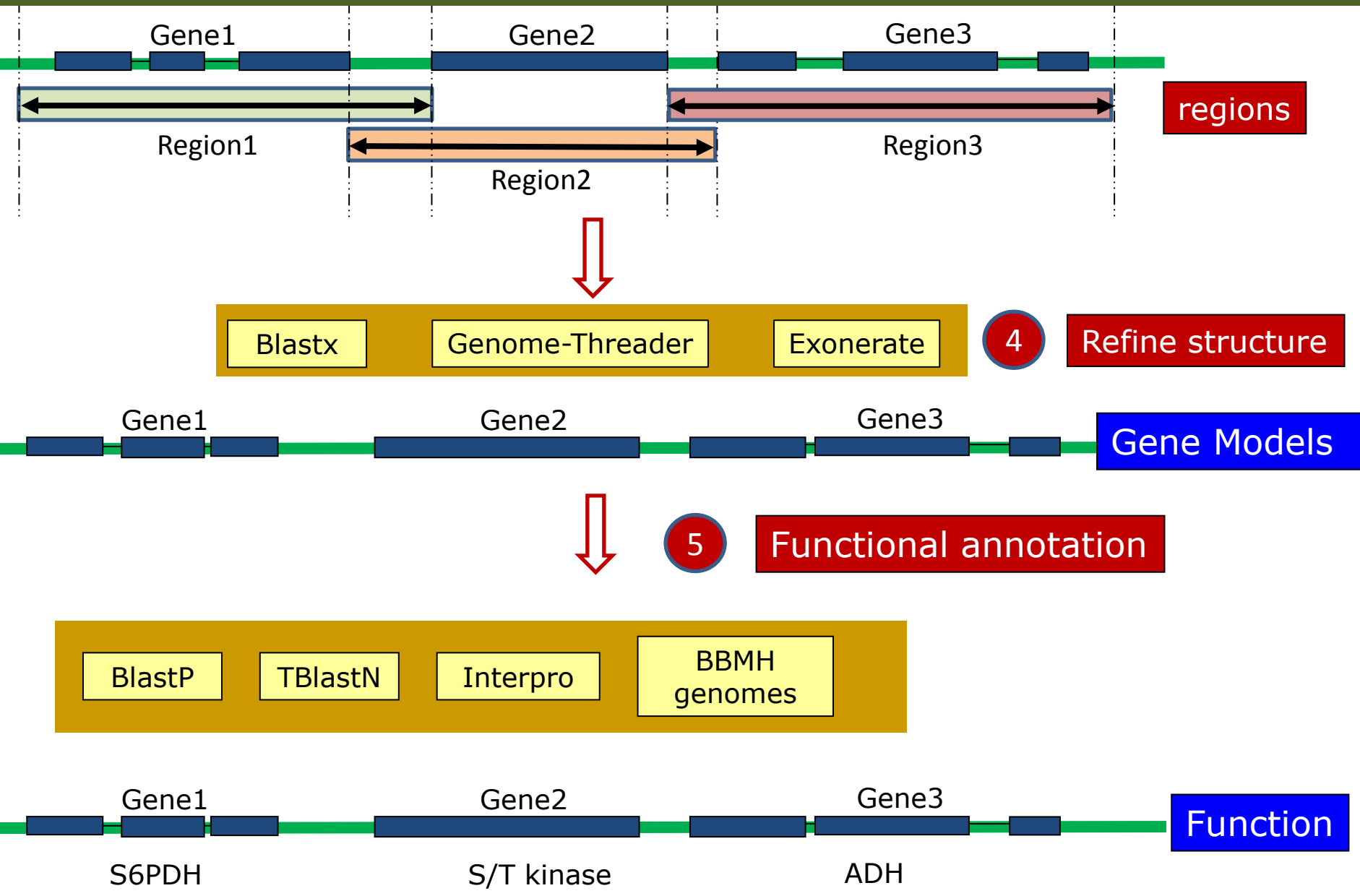
GNPAnnot platform



Gene annotation Pipeline



Annotation Pipeline



Evaluation of the relevance of annotations

VP (true positives) = predicted genes that are really presents = good prediction
FP (False positives) = predicted genes that are normally absents = over prediction
FN (False negatives) = genes that are missed by the automatic annotation = under prediction
VN (True negatives) = genes not predicted that are really absents = good prediction

The relevance is estimated with the ratio Sensibility/spécificity: **Sn/Sp**
The calculation is performed with a set of manually annotated genes

$$\text{Sensibility} = \frac{VP}{VP + FN}$$

$$\text{Spécificité} = \frac{VN}{VN + FP}$$

Gene finder	Sn N	Sp N	Sn E	Sp E	Sn G	Sp G
GenScan+	83.2	98.2	69.6	78	25.8	29
GenMarkHMM	89.9	94.8	73.1	76.6	32.4	31.6
FgenesH-At	95.1	93	85.3	81.4	47	46.5
FgenesH-Mt	97.6	92.1	85.1	80.7	52.8	47.8
EGN	93.7	95	84.7	85.4	55.5	50.5
EGN+FgenesH	97.8	94.2	90	86.9	63.2	56.4
EGN+FH+AA	98.6	93.9	92.4	88	69.2	61.8
EGN+FH+AA+ EST	98.2	99.9	94.4	94.6	80.2	79.4

Transposable Elements (TEs)

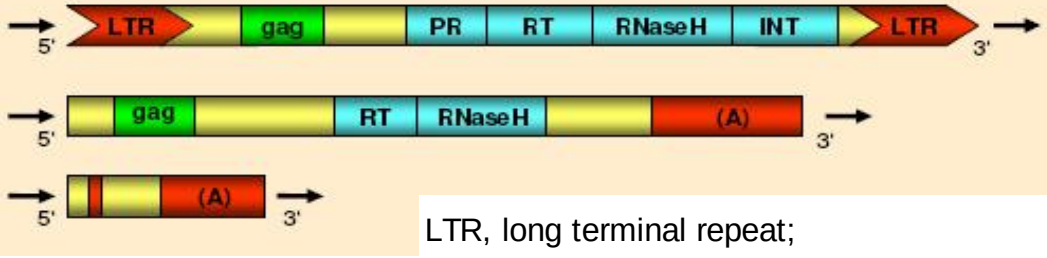
Genes represent only a little part of the genome. Some regions can be gene-rich but some other can contain a majority of repeated elements.

Retrotransposons CLASS I

LTR

Non-LTR (LINEs)

SINEs



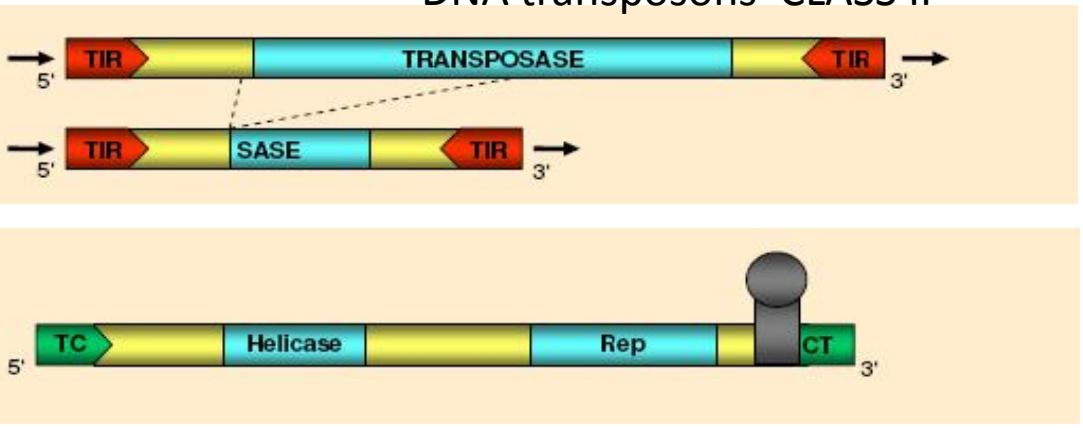
LTR, long terminal repeat;
LINE, long interspersed nuclear element;
SINE, short interspersed nuclear element;

DNA transposons CLASS II

TIR

MITEs

Helitrons



TIR , terminal inverted repeat.
MITE, Miniature Inverted Transposable Element

Transposable Elements

Classification from Wicker et al (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet, 8, 973-982.*

Class	Order	Superfamily	Family	Code / Label
Retrotransposon	LTR	Copia	opie	RLC
		Gypsy	maggy	RLG
		Unclassified		RLX
	LINE	L1		RIL
		Unclassified		RIX
	SINE	Alu		RSA
		Unclassified		RSX
DNA transposons	TIR	CACTA		DTC
		Mutator		DTM
	MITE	Stowaway		DTT
		Tourist		DTH
	Helitron	Helitron		DHH

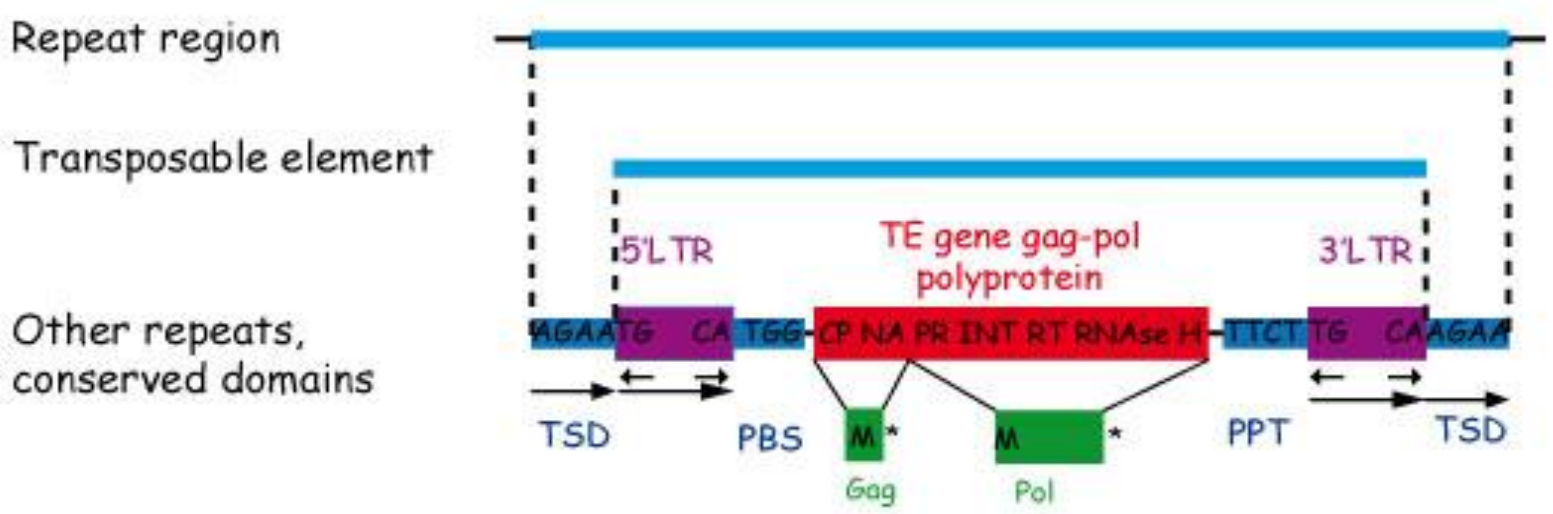
Retrotransposon, LTR, Copia, opie = RLC_opie

Classification of TEs based on sequence similarities : the rule 80-80-80

LTR-Retrotransposons, Ty1 (*copia*) & Ty3 (*gypsy*)

In plants, the most frequent mobile elements are LTR retrotranposons

2 super families of LTR-retrotransposons : *Copia* and *Gypsy*



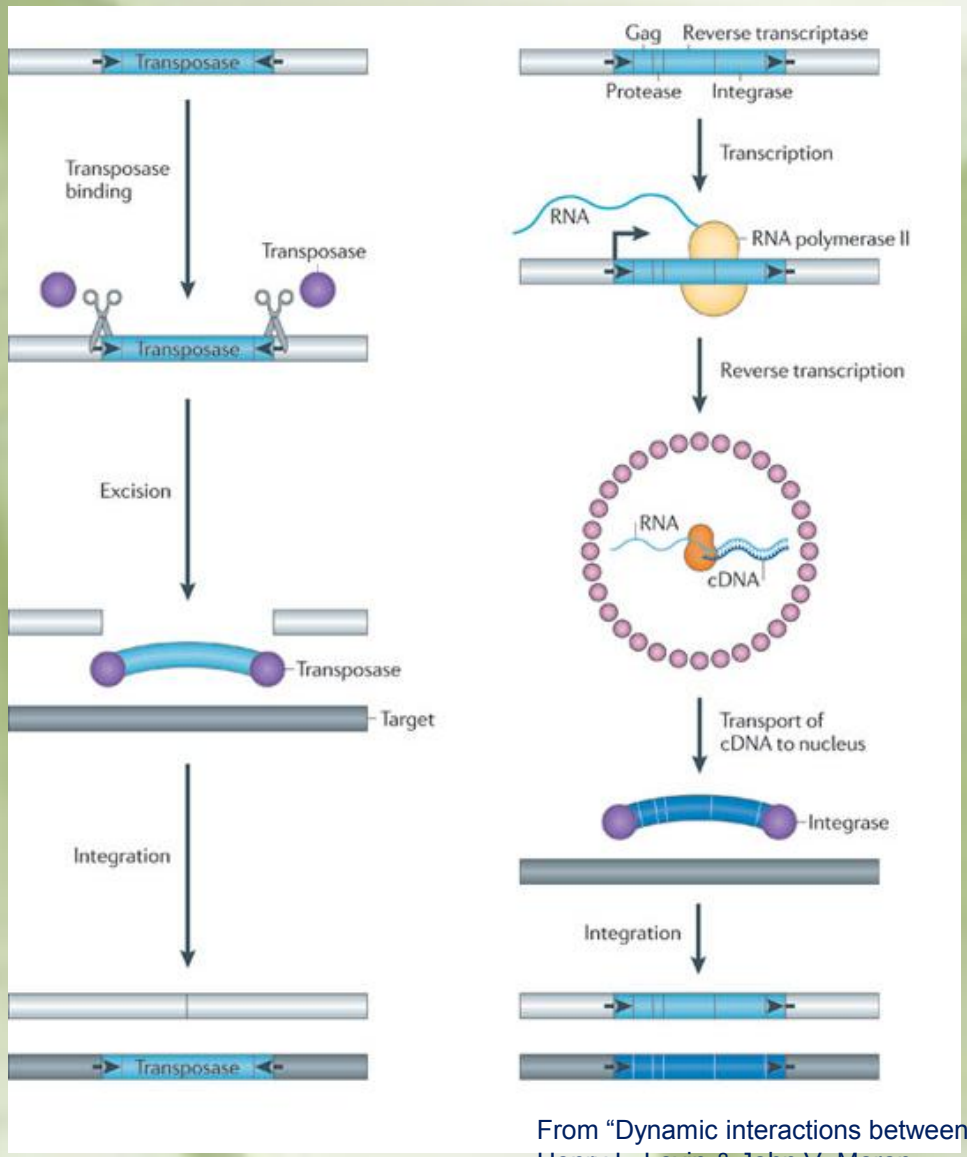
Example of LTR retrotransposon Ty1-copia

In Ty3-gypsy superfamily, the order of the protein domain differ: CP-NA-PR-RT-RNAseH-INT

LTR, long terminal repeat; TSD, target site duplication; PBS, primer binding site;
PPT, polypurine tract (RR tract); CP, capsid-like protein; NA, nucleic acid binding moiety;
PR, protease; INT, integrase; RT, reverse transcriptase.

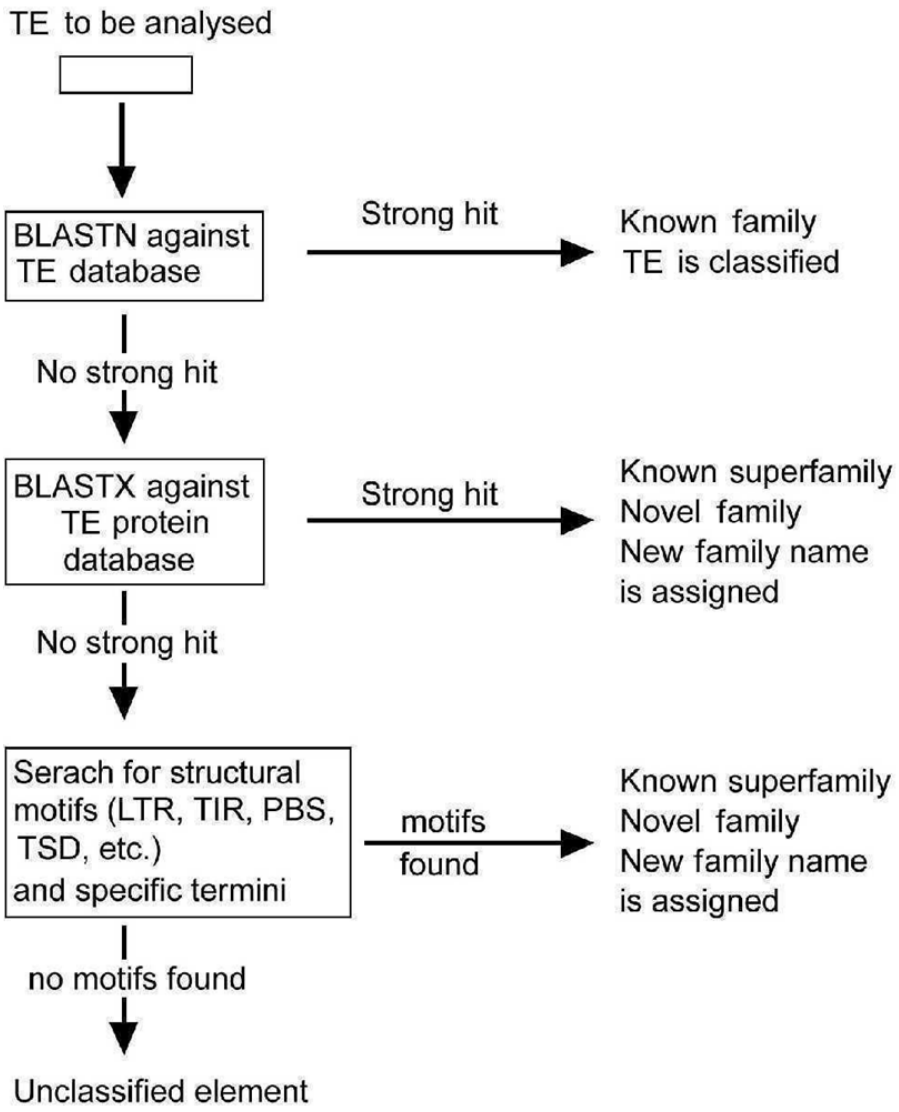
Dynamics of TEs

DNA transposons **LTR retrotransposons**



From “Dynamic interactions between transposable elements and their hosts”
Henry L. Levin & John V. Moran
Nature Reviews Genetics 12, 615-627 (September 2011)

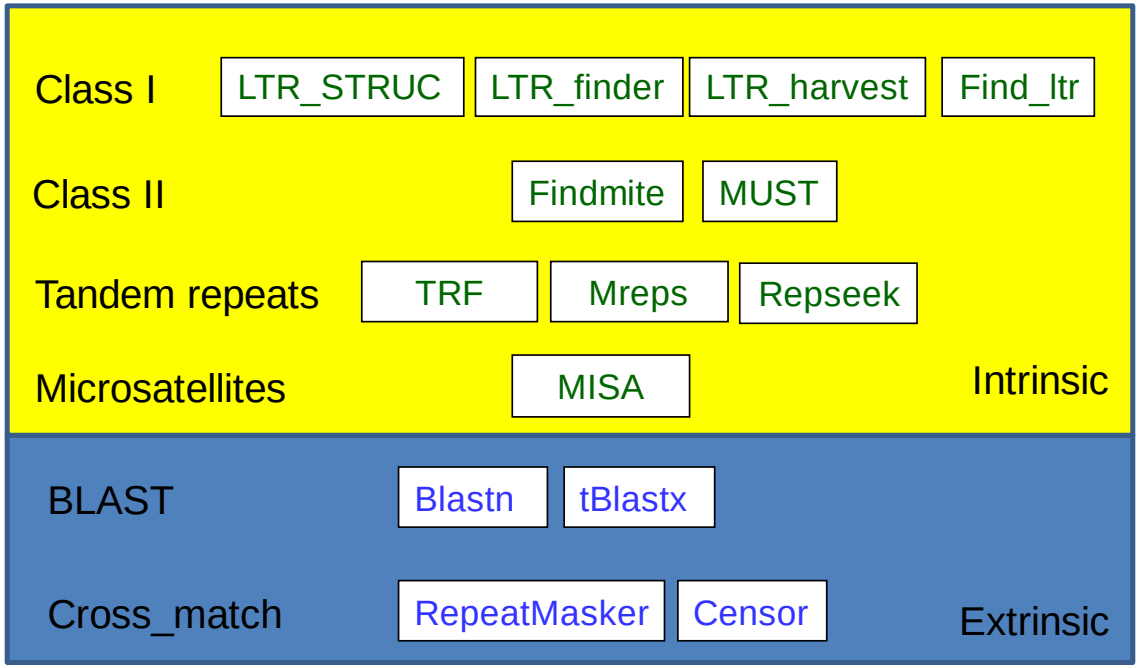
Annotation of TE, general approach



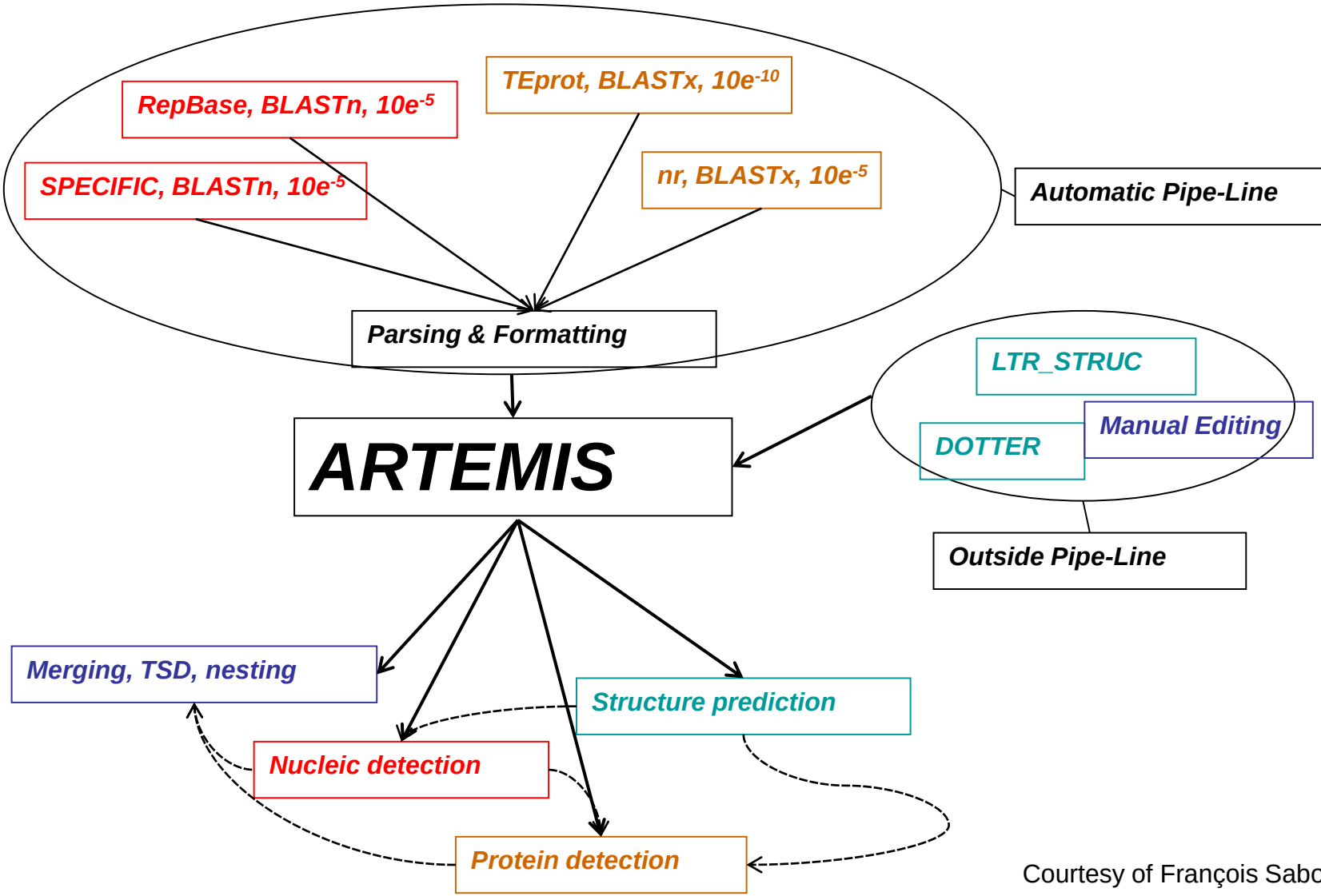
Courtesy of François Sabot.

Annotation of transposable elements : tools

Several ab-initio programs can be used to detect structures of Tes
Comparisons with databases can be used to classify the elements.



Annotation of transposable elements : methods



Courtesy of François Sabot.

Refine the structure of TE search for LTR and IR

DOT PLOT http://www.helmholtz-muenchen.de/icb/gepard

Gepard

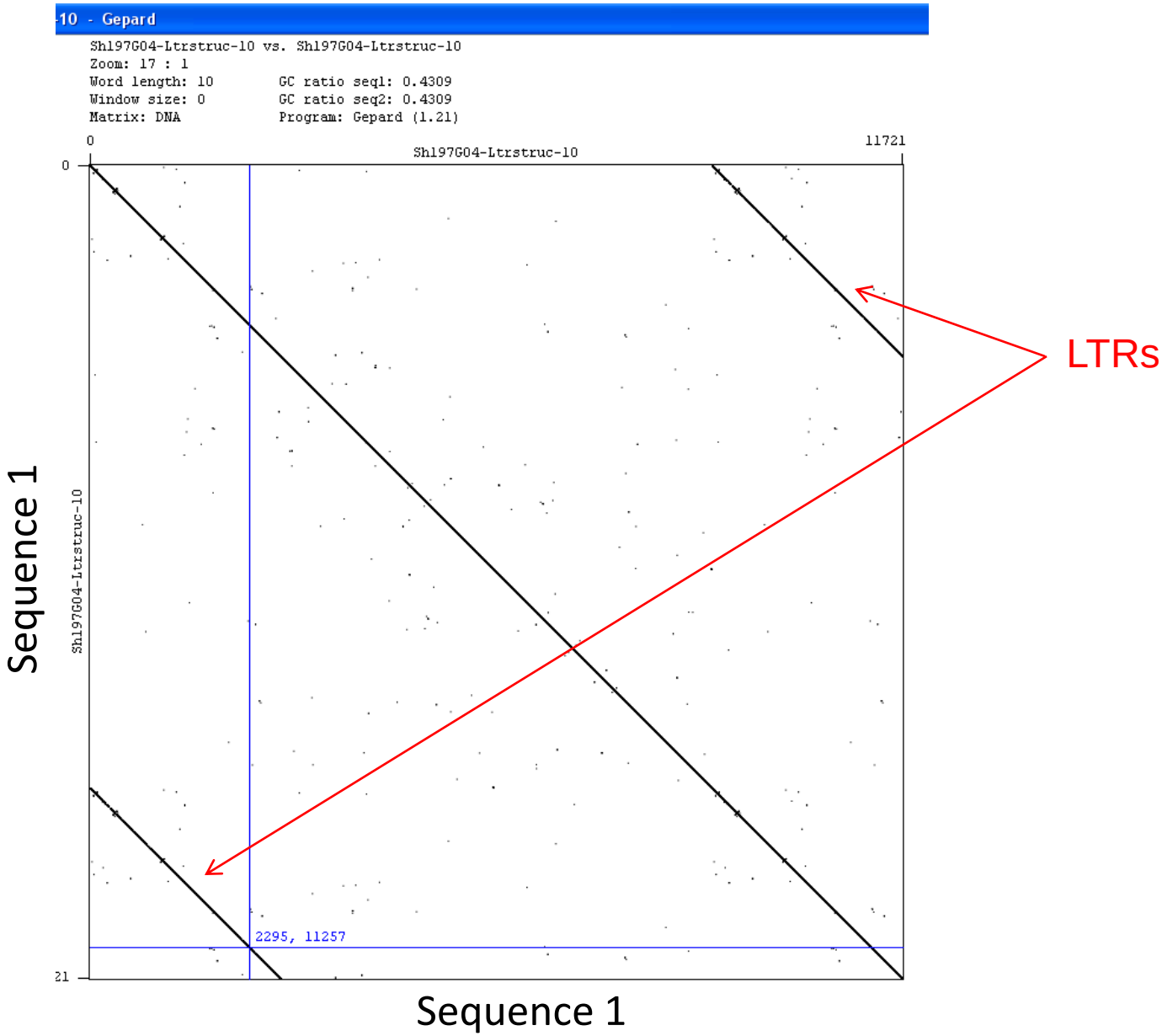
Gepard (German: "cheetah", Backronym for "GEnome PAir - Rapid Dotter") allows the calculation of dotplots even for large sequences like chromosomes or bacterial genomes.

Reference:
Krumsiek J, Arnold R, Rattei T. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 2007; 23(8): 1026-8. PMID: 17309896

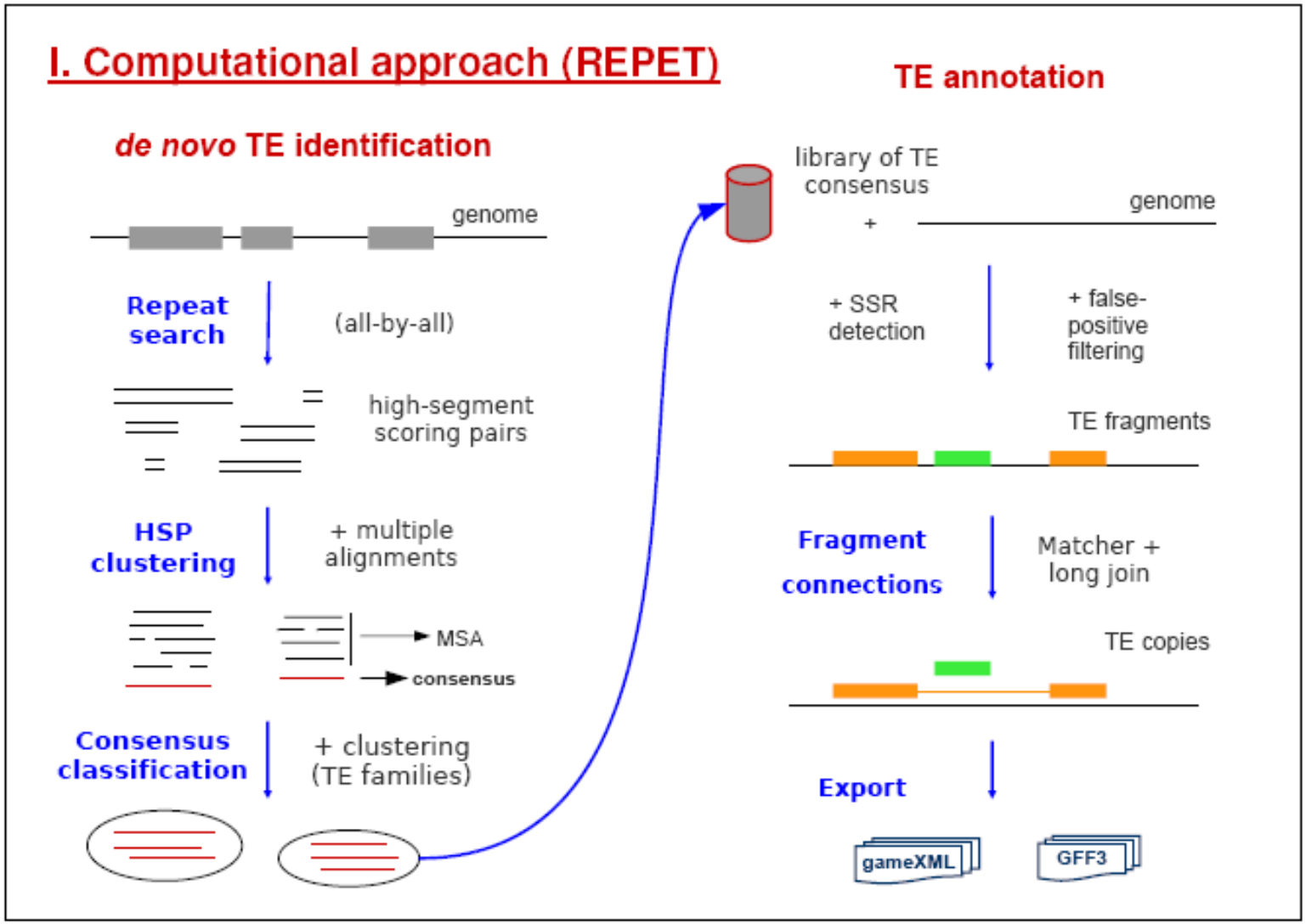
- **Use cases** +
- **Features** +
- **Screenshot** +
- **System requirements** +
- **Download** +
- **Bugs** +
- **Source code** +
- **Tutorial** +
- **Method** +
- **Memory issues / Vmatch support** +
- **Contact** +

Refine the structure of TE, e.g. LTR-retrotransposon

Gepard dotter



Pipeline for Classification, Screening and Mapping of TEs



80 – 80 rule = 80% identity on 80% of the lenght

Example: Annotation of the LTR retrotransposons in the musa genome

Assembled sequence (genome)

Gene models (Fgenesh) unmasqued sequence

STEP 1 :
Reference library of LTR retrotransposons

Comparison with databanks

TBLASTN

BLASTP

gag-pol data set from previously musa annotations

RepBase

64 Musa BAC sequences
NGS-454 low depth sequencing (Hribova et al, 2010 BMC Plant Biol)

Exonerate

gag-pol identification

Clustering of gag-pol 80% id and 80% coverage

BLASCLUST

Gypsy and Copia gag-pol (proteic)

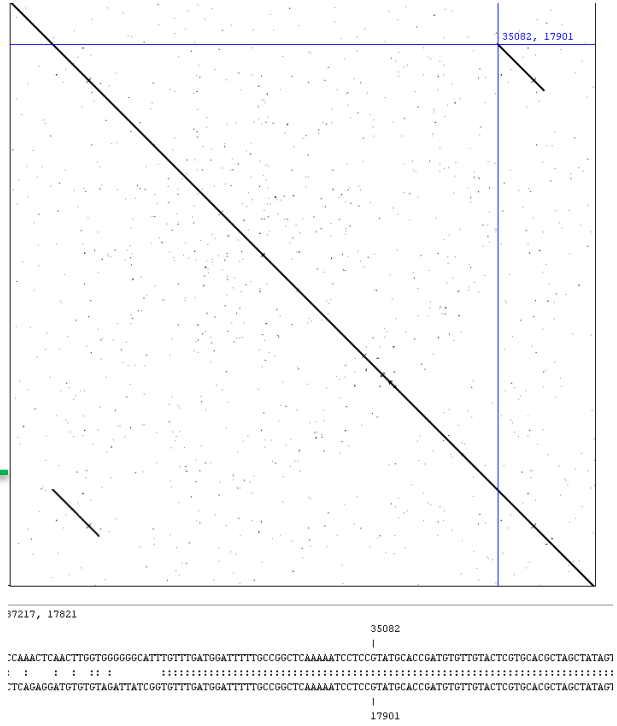
Extraction of genomic regions

In- house Perl scripts

Extension to LTR sequences

Gepard Dotter

Gypsy and Copia nucleic sequences



Classification, screening and mapping of TEs onto the genome

STEP 1 : continued

Gypsy and Copia nucleic sequences

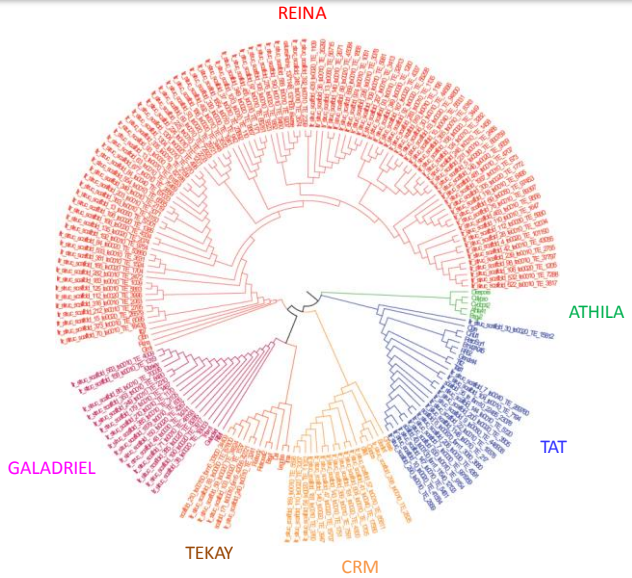
+

LTR_STRUC predictions

Clustering
80% id
80%
coverage

TGICL
Tclust
CAP3

Classification phylogenetic analysic of RT domains



Gypsy and Copia => TE reference library

STEP 2: Screening with REPET (TEannot)

TE reference library
- LTR retrotransposons
- Non-LTR retrotransposons
- DNA transposons

Screening
of the assembly

REPET
(TEannot)

NESTED
OVERLAPPING TEs

Priority rule

DTA > DTX >
RLC > RLG > RLX
> RPP > RIL >
RIX > DHH

TE Mapping

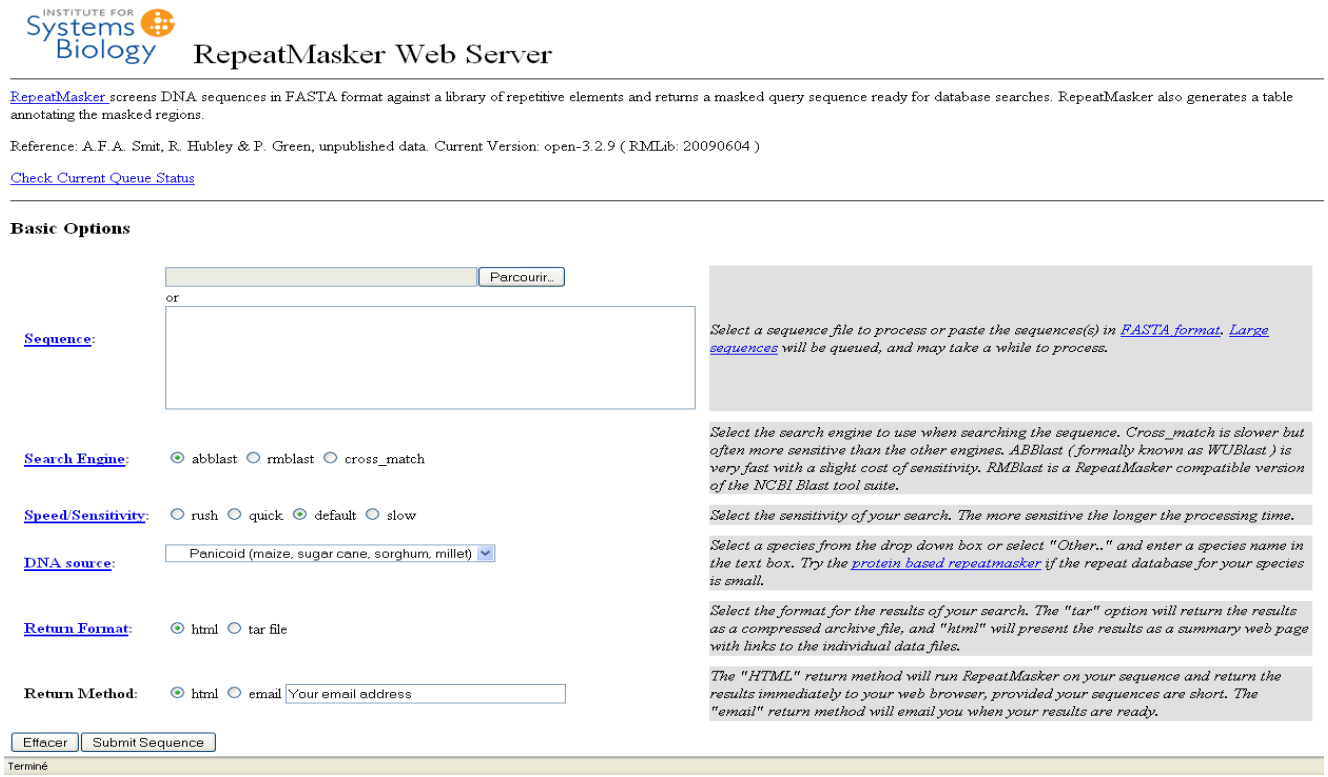
Before gene annotation, mask repeat sequences

Gene modelling is generally done after filtering out repeats because this makes gene finding even more difficult :

TEs may look like protein encoding genes (transposase and remains)

TEs perturb the structure of gene models, e.g. by inserting in introns (long insertion, and gene model is cut in many parts)

→Repeat masker is the tool the mostly used to mask repeats.

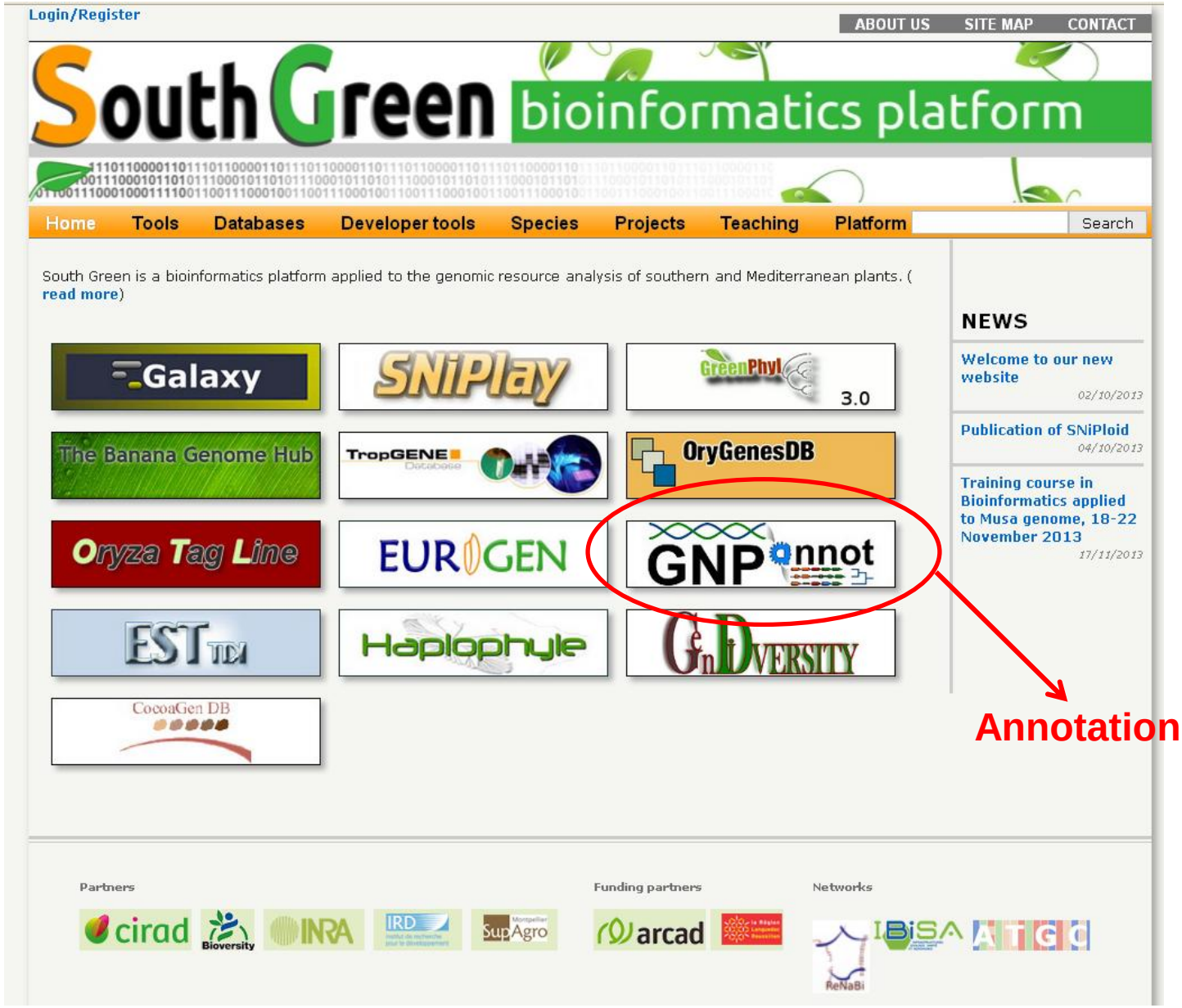


http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker

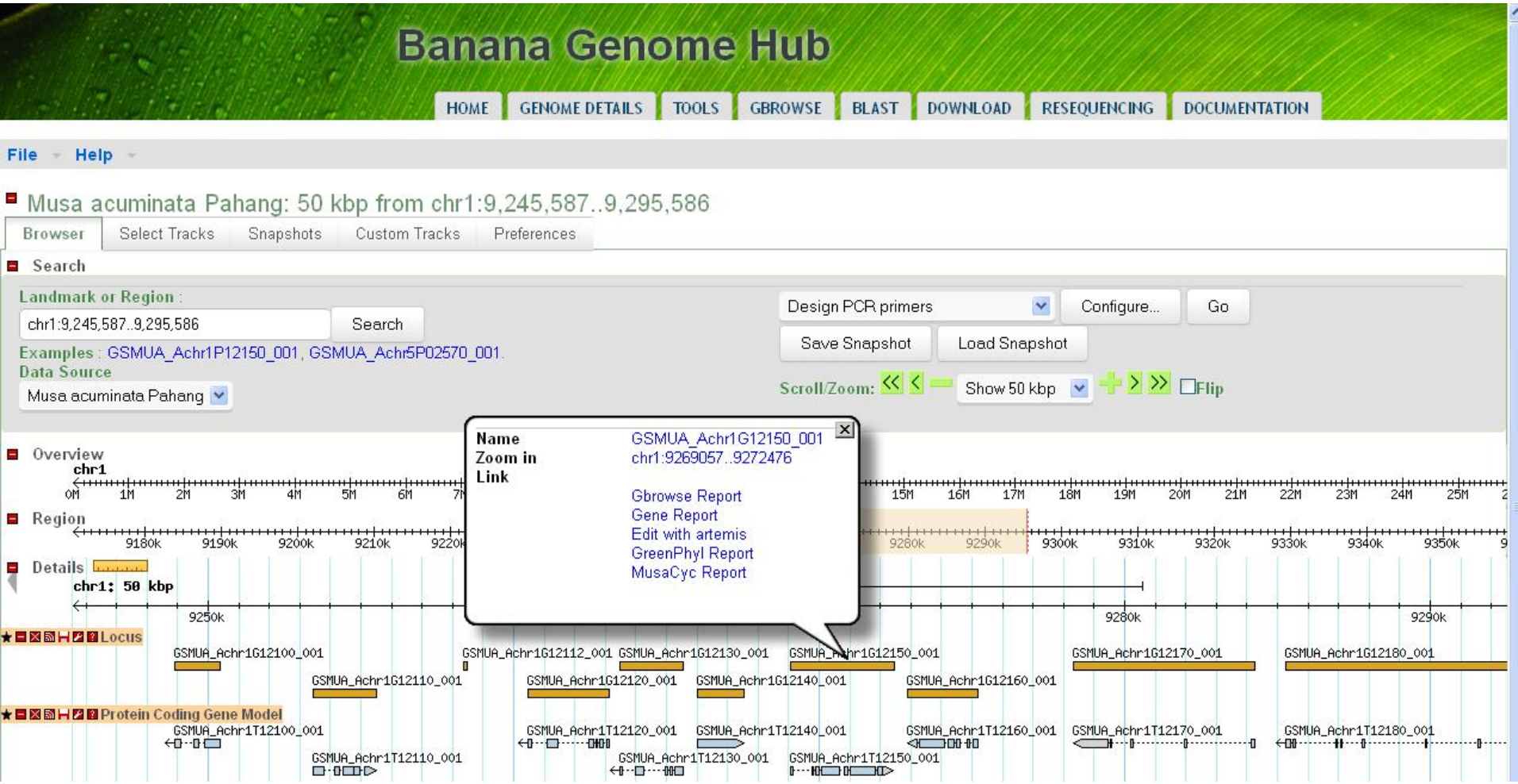
Manual curation of annotations

artemis : to see, to check, to correct, to validate, to store

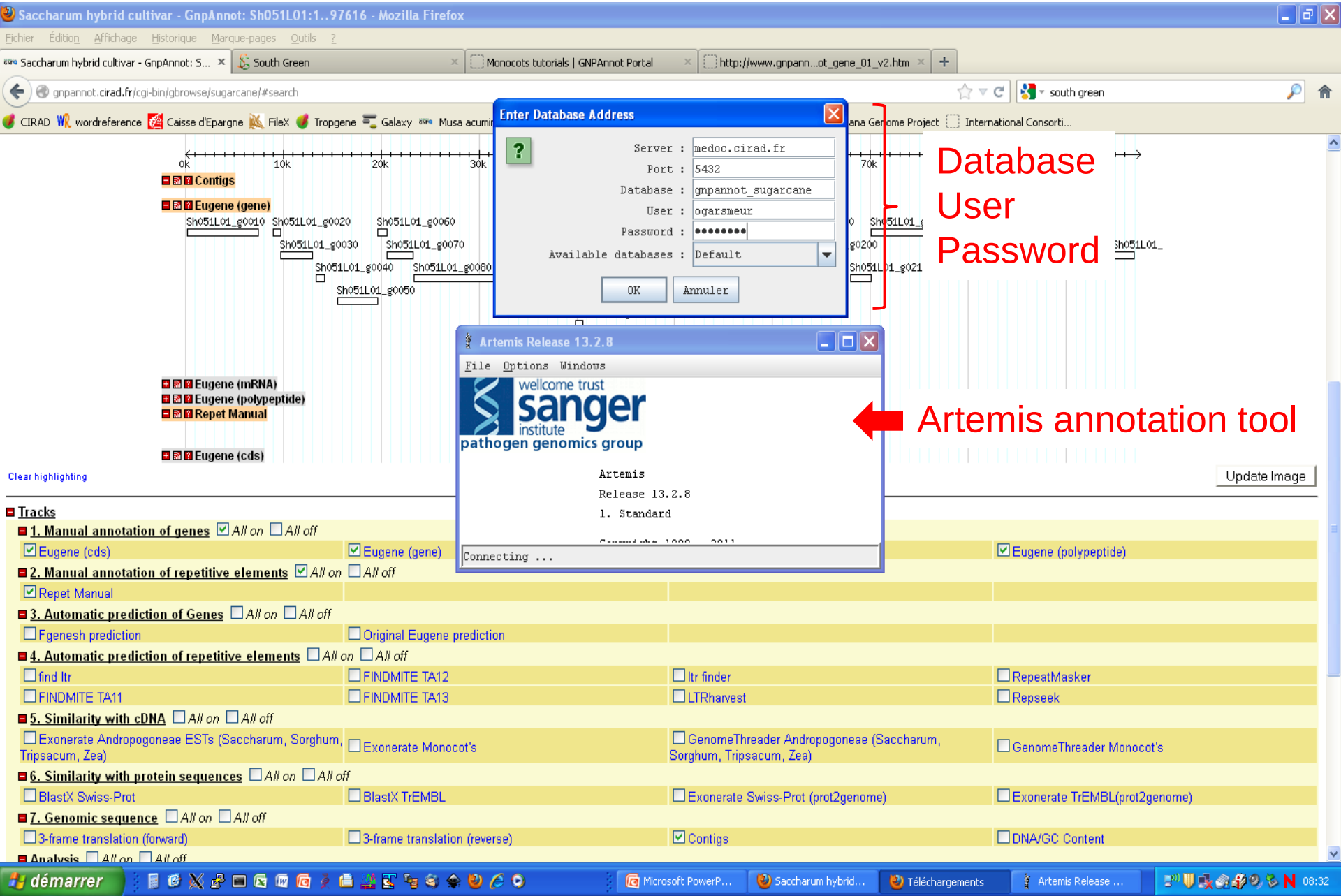
GNPAnnot on the web : http://southgreen.cirad.fr



Manual curation of annotations

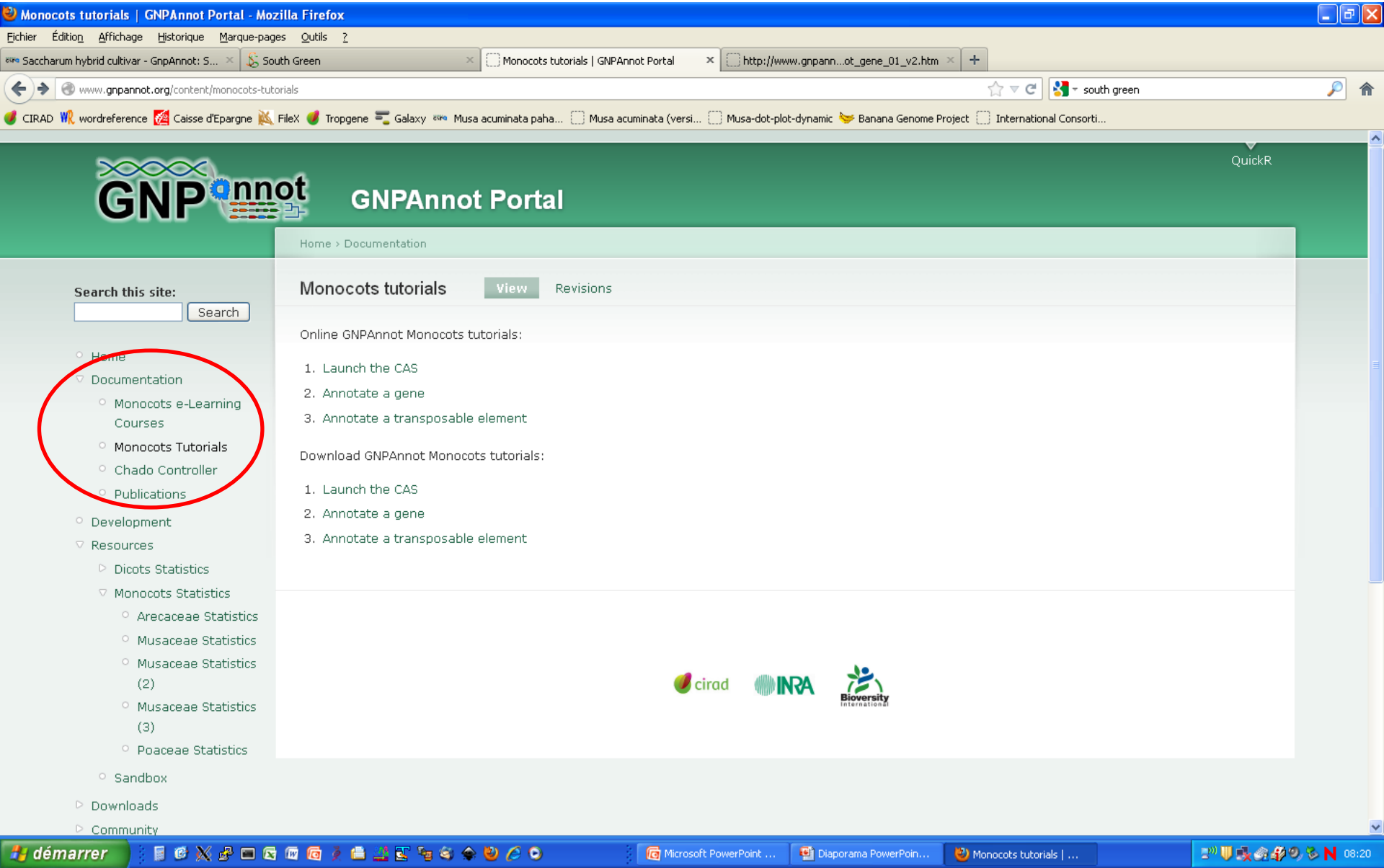


Gbrowse is connected to artemis through the chado database



Tutorials available on the portal

[e.g. tutoriel_gnpannotmonocot_gene]



ARTEMIS

http://www.sanger.ac.uk/resources/software/artemis/