Equipe ID

Gene structural annotation with EuGène based Eukaryotic Pipeline

PGAA Course

Septembre 2018

Task 10.3 Capacity Building

S. Bocs & E. Sallet

stephanie.sidibe-bocs@cirad.fr

eugene-help@groupes.renater.fr

Course plan in five parts

1. **EuGène**
2. **EuGène for model plant genomes**
3. **EuGène based Eukaryotic Pipeline (EGN-EP)**
4. **EuGène & EGN-EP for sugarcane complex genome**
5. **Conclusion**

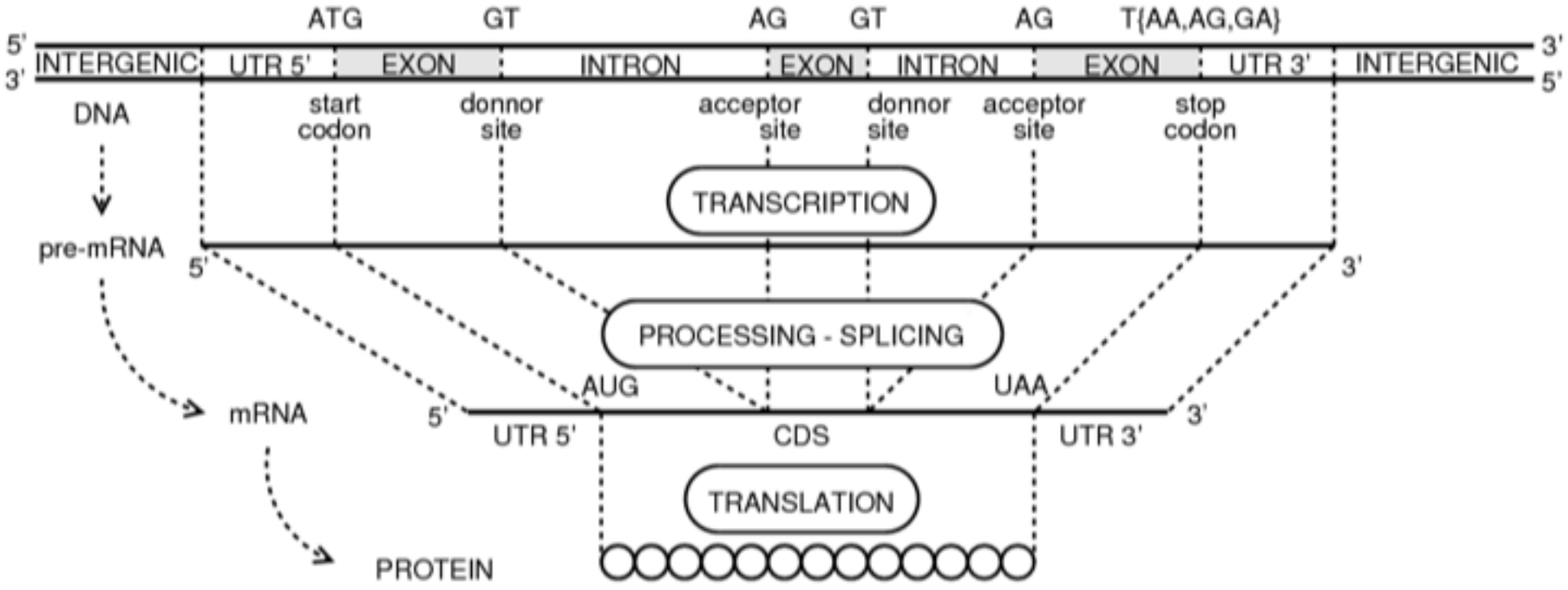
EuGène: http://eugene.toulouse.inra.fr

- Genome Annotation in Plants and Fungi: EuGène as a Model Platform (V3.5; Foissac et al., 2008 – Current bioinformatics)
- Structural annotation of eukaryotic and prokaryotic genomes (since V4.0)
- EuGène team:
  - Thomas Schiex DR - MIAT INRA Toulouse
  - Erika Sallet IE & Jérôme Gouzy IR - LIPM INRA/CNRS Toulouse
- Evolution of EuGene according to the needs emerging from research projects
- C++, artistic license

Modified from E. Sallet 2014

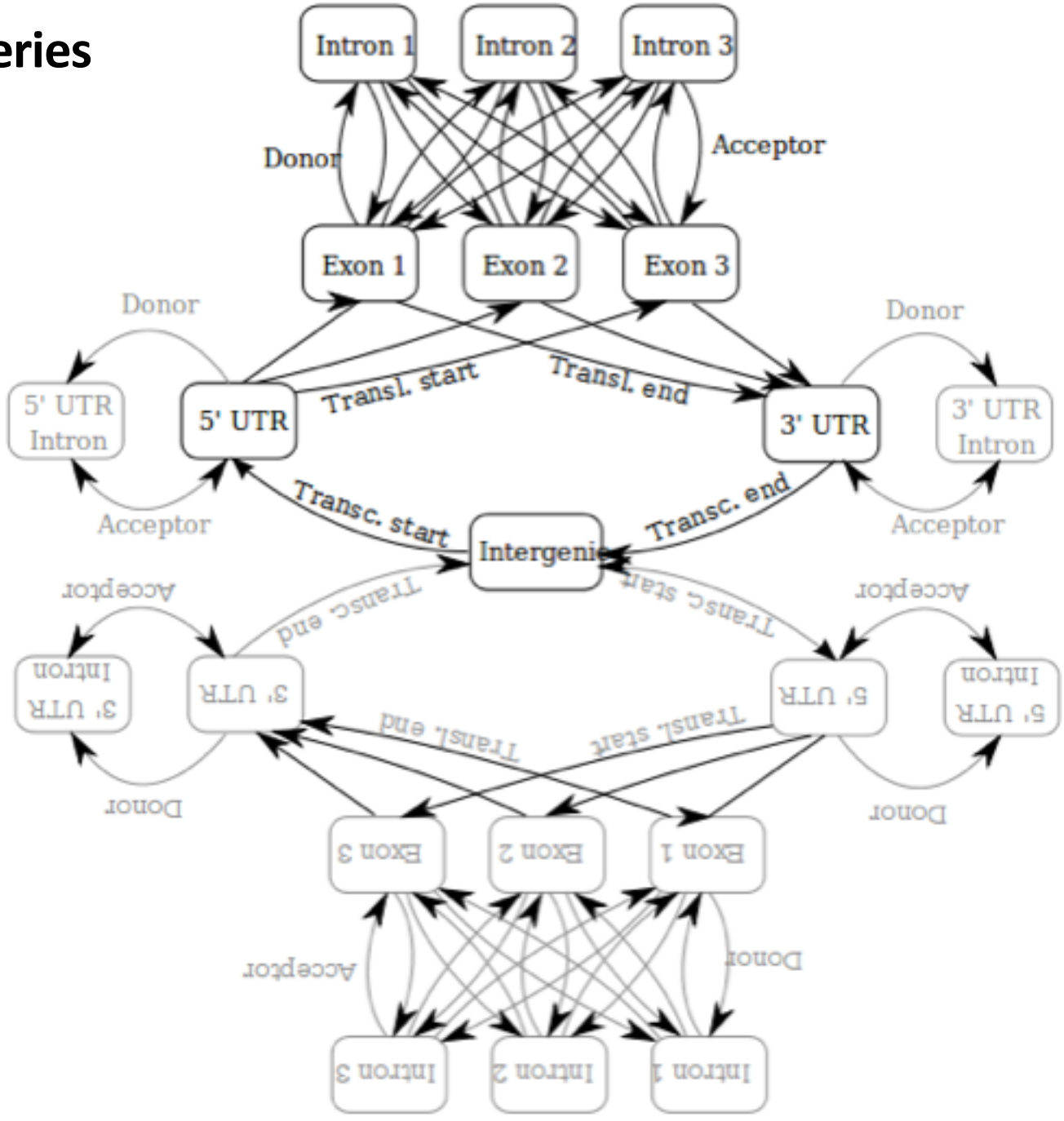
Eukaryotic gene

A series of regions separated by signals



Modified from E. Sallet 2014

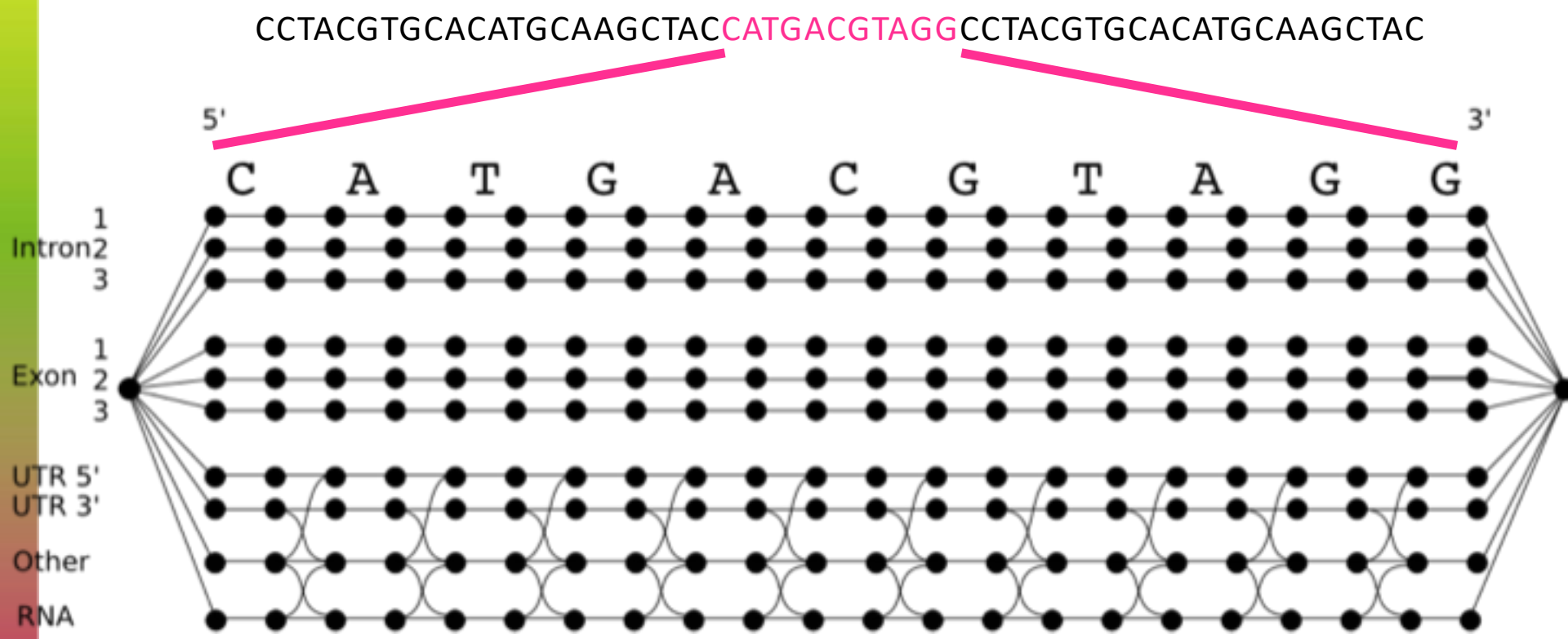
Eukaryote gene series



Foissac et al., 2008

How EuGène works

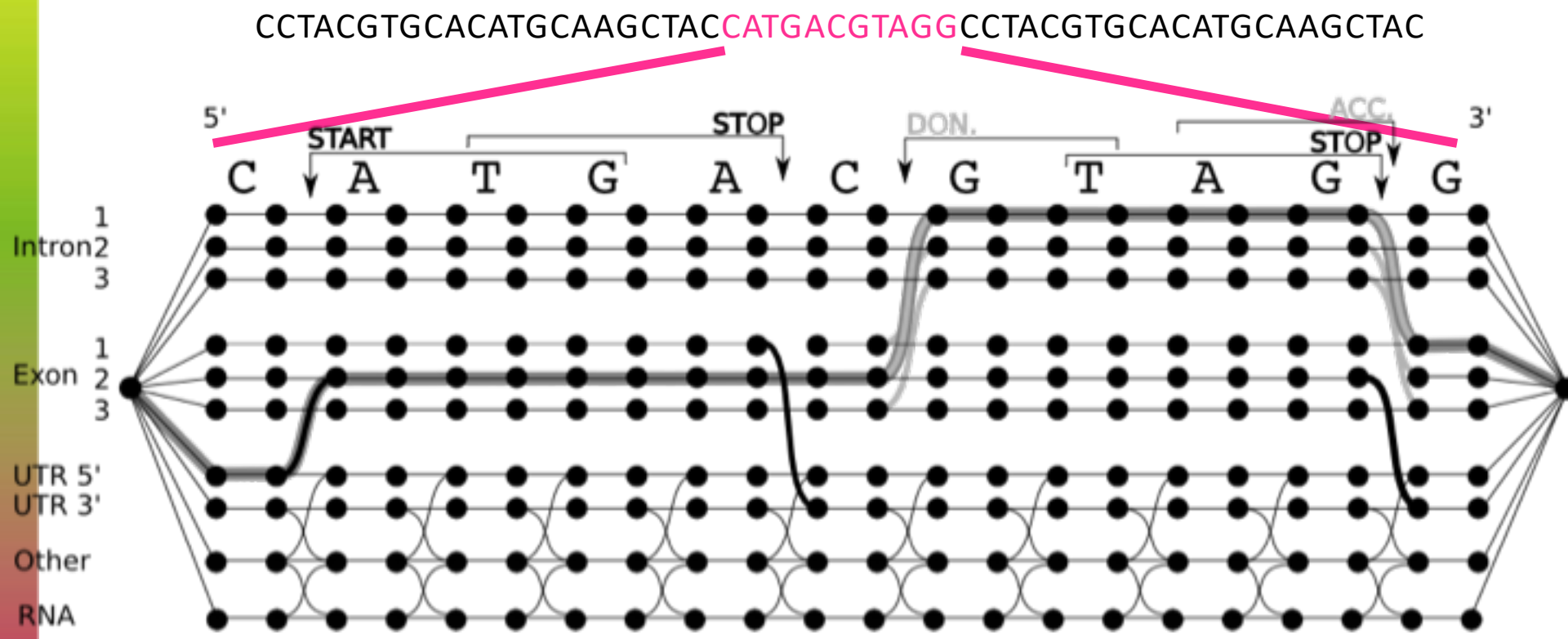
- Transforms a nucleic sequence into a graph.
- Edit / weight the graph according to the biological and statistical "evidences" available



Modified from E. Sallet 2014

How EuGène works

- Transforms a nucleic sequence into a graph.
- Edit / weight the graph according to the biological and statistical "evidences" available
- A graph path = a correct gene structure
- An **optimal** path of the graph = the prediction



Foissac et al., 2008

Integration of biological and statistical information

- One **plugin** per data type to be integrated
- Examples of commonly used plugins:

Plugin	Information	Format read by the plugin	Program commonly used to generate data
MarkovIMM	Statistical properties of regions	Binary matrix	egn_MarkovIMM
SMachine	Start codon & splicing sites	tsv	SpliceMachine
BLASTX	Protein similarities	GFF3	BLASTX
EST	Similarities with transcribed sequences (EST, RNA-Seq assembly)	GFF3	Gmap or genomeThreader
NSTRETCH	N Stretchs in the sequences to be annotated	-	-
…			

Modified from E. Sallet 2014

AnnotaStruct generic plugin

- Reads the Generic Feature Format GFF3

- Many recognized features
  - gene/exon/intron/CDS/transcript_region, ncrna....
  - Signals: tstart, start,...

- Allows, for example, to integrate the results of other predictors of:
  - Protein coding genes (e. g. Fgenesh, Augustus)
  - ncRNA (ex rfam-scan, tRNAscanSE)

- Creating a new plugin is relatively simple
  - API, inheritance

EuGène recipe examples

- Classical recipe
  - MarkovIMM, Smachine, BLASTX, EST
- Combiner
  - X annotations to be integrated -> X instances of AnnotaStruct
- Annotation transfer
  - Mapping of transcripts -> EST
  - Extraction of regions around start codons -> AnnotaStruct

Some features of EuGène

- EST / RNA-Seqs exploitation
  - Detection of non-canonical splicing sites
  - Detection of splicing variants (keep the most represented)
- Specific strand prediction possible
  - detection of overlapping genes
- Reading the codon table from a file
- GFF3 format at output
  - Interoperability

Modified from E. Sallet 2014

The complete annotation process before: 1) Training & optimisation with Toolkit-eugene

- ✓ Building an expertised gene set (either available or built from EST mapping)
- ✓ **Training** of EuGene and SpliceMachine
  - ➢ matrix + SpliceMachine program parameterized
- ✓ **Optimization** of data integration parameters
  - ➢ EuGène configuration file
- ✓ **Test**

The complete annotation process before: 2) Data preparation with Toolkit-eugene & EuGène annotation

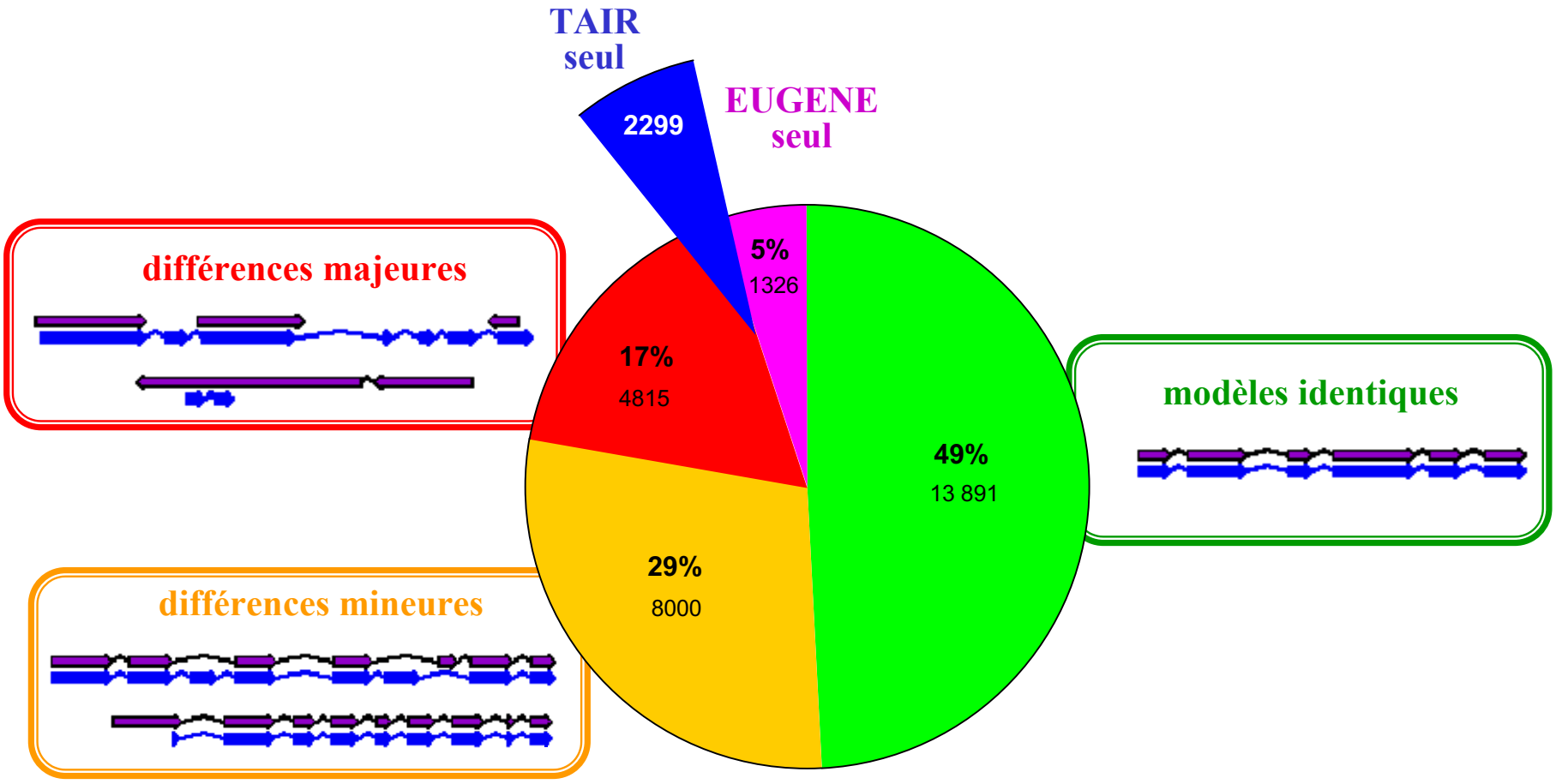
- ✓ **Computing & reformatting** of evidences
  - ❑ BLASTX, EST mapping, etc.
  - ➢ For each sequence to be annotated, X data files read by EuGène
- ✓ **Annotation**

- ➢ Relatively complex and long process
  - ❑ Perl scripts automated many steps but not yet enough
  - ❑ Relatively long SpliceMachine training
  - ❑ At the genome scale, the process of obtaining "evidences" takes much longer than the execution of EuGene

- ➢ The perspectives are the objectives of the **EGN-EP** part

Modified from E. Sallet 2014

Arabidopsis genome: EuGène v4.0 versus TAIR R6.0



S. Aubourg, J-P Renou et al

Rice genome: EuGène v3.2 optimisation: protocole & definition

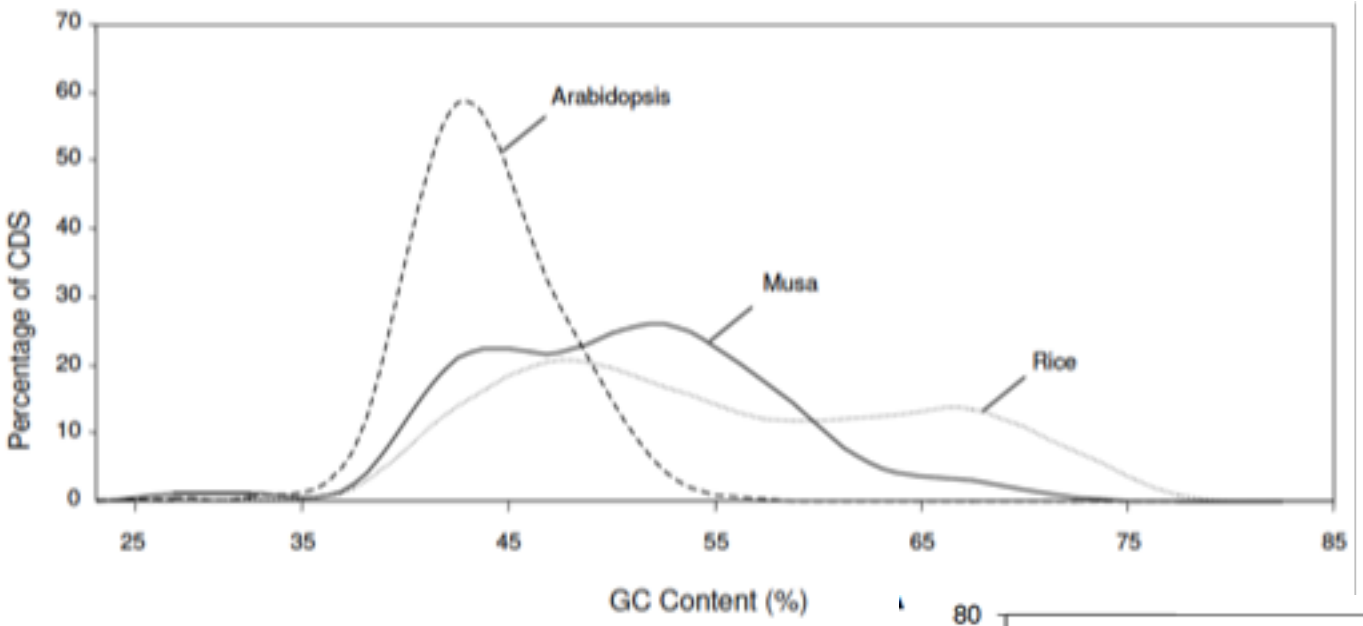
- ✓3 training sets of 100 rice genes (300 in total): A, B, C
- ✓4 sensors: SpliceMachine (SM), Fgenesh (FH), BLASTX, Sim4
- ✓Gene sensitivity (Sng) = nb correctly predicted genes / tot nb of genes in the curated dataset
  - ➢Sensitivity increase -> false positives (missing true genes) decrease
- ✓Gene specificity (Sne) = nb correctly predicted genes / tot nb of predicted genes
  - ➢Specificity increase -> false negatives (overpredicted false genes) decrease
- ✓Fitness = $\sqrt[4]{(Sne * Spe * Sng * Spg)}$ (Foissac et al., 2008)

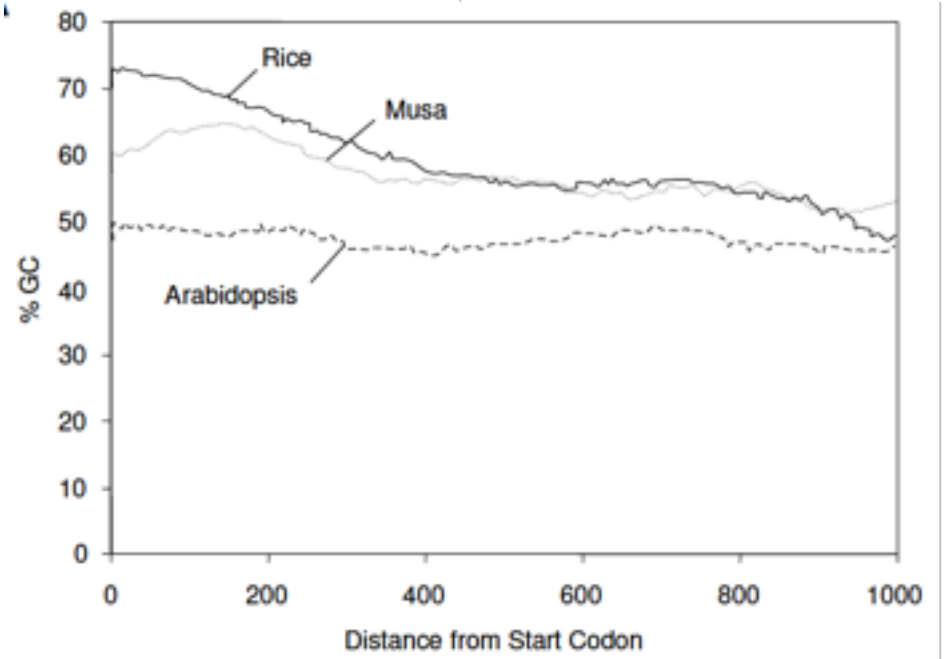
Rice genome: EuGène v3.2 optimisation: results

Sensor	Set	SNG	SPG	SNE	SPE	Mean		Fit
SM	Opt1.0B	23.53	21.28	56.36	72.68	43.46	44.46	37.84
	Test1.0C	27.38	26.14	56.18	72.14	45.46		41.27
	Opt1.0C	33.33	31.11	51.63	75.32	47.85	44.52	44.81
	Test1.0B	23.53	21.05	48.09	72.06	41.18		36.20
FH	Opt1.0B	51.76	41.90	84.75	82.30	65.18	61.41	62.37
	Test1.0C	38.10	32.00	81.34	79.11	57.64		52.92
	Opt1.0C	54.76	46.00	84.82	83.55	67.28	62.22	65.00
	Test1.0B	40.00	30.91	80.51	77.24	57.16		52.66
Blastx	Opt1.0B	55.29	46.08	86.65	83.64	67.92	65.68	65.55
	Test1.0C	46.43	39.00	86.12	82.19	63.44		59.83
	Opt1.0C	46.43	39.39	86.77	82.99	63.89	65.16	60.24
	Test1.0B	52.94	43.69	86.02	83.03	66.42		63.75
Sim4	**Opt1.0B**	**62.35**	**56.38**	**89.19**	**89.19**	**74.28**	**72.44**	**72.72**
	Test1.0C	**55.95**	**50.00**	**90.02**	**86.46**	**70.61**		**68.31**
	Opt1.0C	58.33	52.13	89.80	86.97	71.81	71.55	69.81
	Test1.0B	58.82	51.55	87.5	87.32	71.30		69.38

Opti	Set	SNG	SPG	SNE	SPE
Piégu	Test 1.0C	NA	NA	67.46	71.99
Bocs 2006		55.95	50.00	90.02	86.46

Rice genome: EuGène v3.2 optimisation: discussion





Lescot et al., 2008 – BMC Genomics

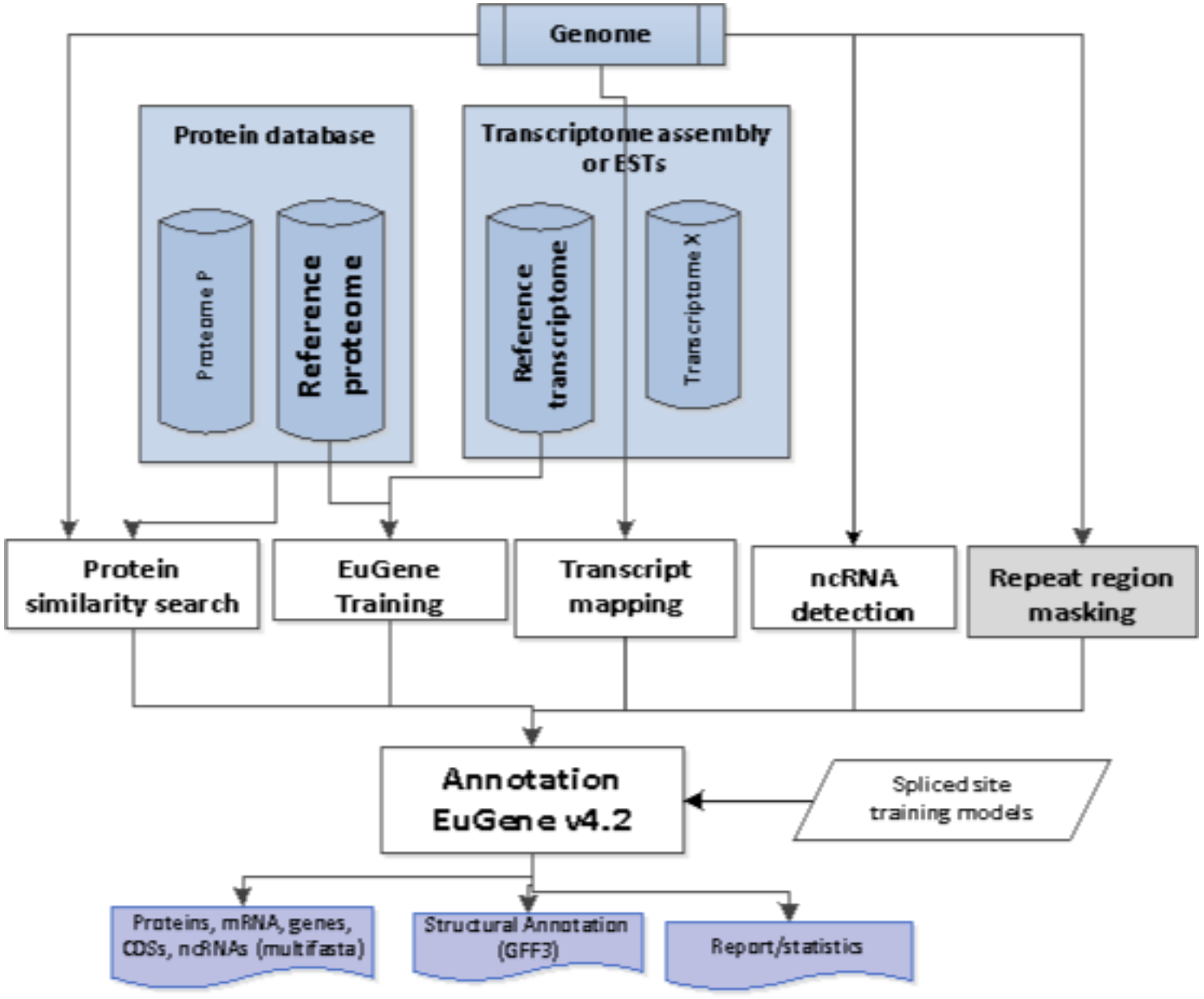
Rice genome: EuGène v3.2 optimisation: sensor priorities

	Piégu opti		Bocs opti
MarkovIMM	1	MarkovIMM	5
NStart	1	SMachine	1
NG2	1	AnnotaStruct	1
SPred	1		
GSplicer	1		
Est	2	Est	30
BlastX	2	BlastX	10
Repeat	8	Repeat	(5)

EuGene-EP automatic pipeline to annotate gene in plant genomes

- Structural annotation of genes
  - Protein coding
  - Non-coding (tRNA, rRNA, ncRNA)
- Fully automate annotation
  - Reduce manual settings as much as possible
- Optimize execution times
  - Optimize the protocols of certain steps
  - Parallelize certain tasks
  - Facilitate the error recovery
- Eliminate dependence on licensed software

EGN-EP main steps



RNAseq Data Integration Strategy

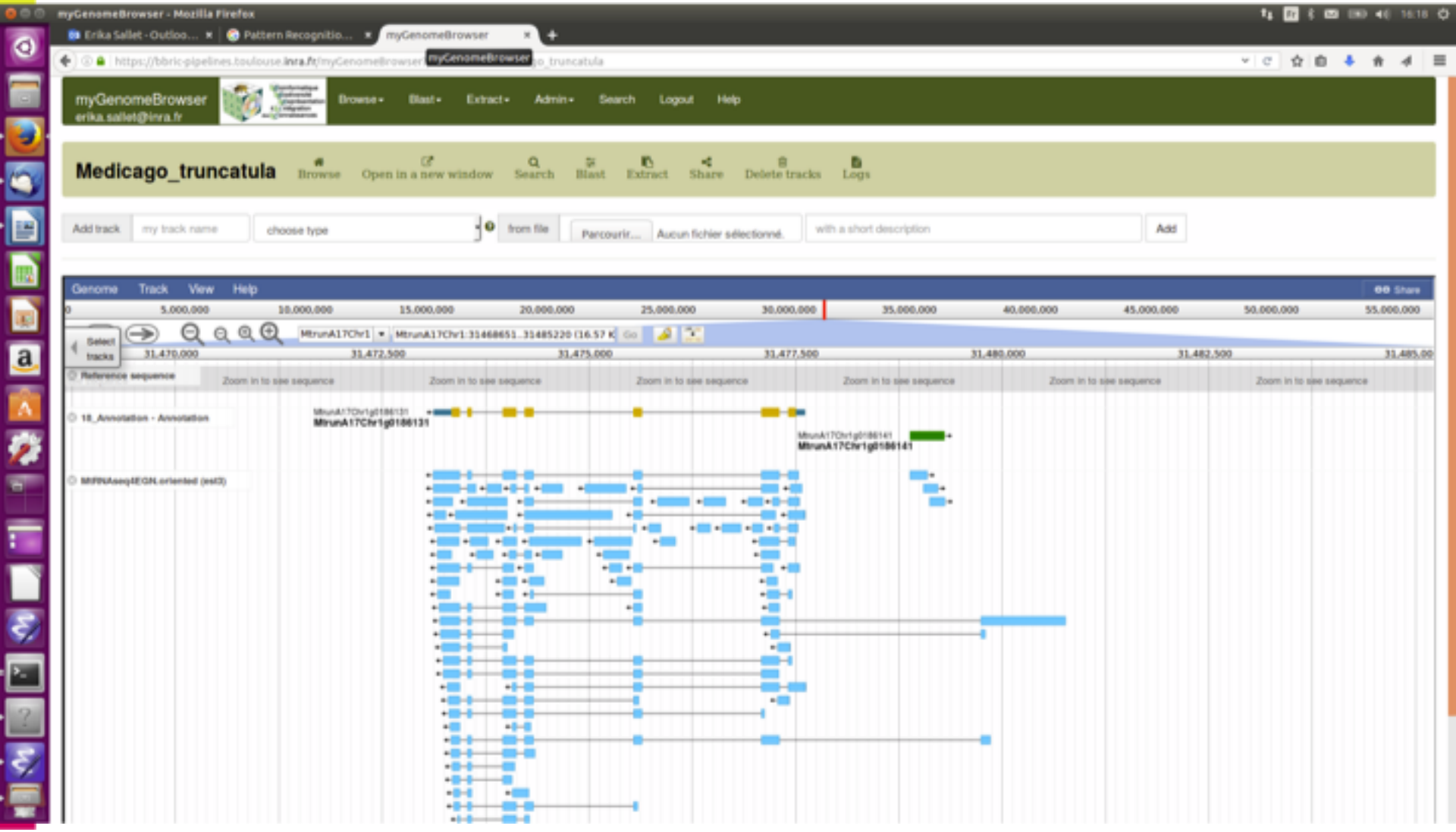
- Alignment with **gmap** of assembled transcripts
  http://research-pub.gene.com/gmap

- Much more weight is given to the prediction supported by the RNAseq data than to the *ab initio* prediction (i.e. only based on IMM models)

- More weight is given if the aligned transcripts are spliced

RNAseq Data Integration : management of local inconsistencies

- Option to filter the alignments of the transcripts when there are inconsistencies, due for example to the presence of splicing variants. The most represented introns are preferred.



Repetitive element masking

- EGN-EP does not aim to annotate repeated elements (other tools are dedicated to this task)

- We do not wish to annotate the genes associated with the transposable elements: we must work on a **masked genome**

- Masking principle
  1) A bank of repeated elements is built (= RepBase + species-specific TE proteins)
  2) Repeated elements are masked on the genome (detected by Red, LTRharvest, BLASTX against enriched RepBase) but **protected** areas are unmasked (= hit with a known transcript, protein bank or ncRNA)
  3) We annotate this masked genome

Fully automated steps

- EuGene's training
  - Differentiate coding from non-coding

- Detection of "non canonical" splicing sites
  - Search in mapping results, at the exon/intron border, of sites other than GT/AG
  - If a site is read in more than 1% of cases, it is authorized

- Training for the detection of splicing sites already done for plants
  - Models trained on one species work as well on another species as those of the species in question
  - Use of splicing sites of several species (Ha, Rosa, At, Mt) to build **weight array method** (WAM) models 'dicot plants'

Optimization of execution times

- Gmap is multithreaded and very fast

- BLASTX Optimizations
  - On pseudomolecule: parallelization via sliding windows
  - On a small scaffold: parallelisation by launching several BLASTX in parallel

  Rq: the two types of sequences can be mixed
  - For each window we launch **UBLAST** to select the proteins that can match and we do BLASTX on a subbank (if UBLAST crashes then normal BLASTX)

- When a calculation is done, a .success file is created. If the pipeline is restarted then the calculation is not restarted if the .success file exists.
  - Addition of transcript or protein banks: previous calculations not restarted
  - If you change parameters, then you must delete the .success.

Options

- "Rescue" mode possible
  - Genes, which have a CDS size < a threshold, are converted to ncRNA.

- Two "independent strand" annotations
  - Allows to detect overlapping genes on opposite strands, antisense ncRNA.

- Integration of all types of information
  - 1 file GFF3
  - 1 additional configuration file

  Example: ChiP-Seq data (start of transcription)

Basic use

- A configuration file to fill in:
  - protein databases and RNA-seq assemblies

```
# List the protein database numbers for blastX
blastx_db_list= 1 2 3 4
blastx_db_1_file=/db/swissprot_noTE
blastx_db_1_weight=0.3
blastx_db_1_pcs=50
blastx_db_1_activegaps=0
blastx_db_1_remove_repet=0
blastx_db_1_preserve=1
blastx_db_1_training=1
# List the trancriptome numbers
est_list=1 2
est_1_file=/db/PLVIT20100804Clusters
est_1_pcs=30
est_1_pci=97
est_1_remove_unspliced=0/1/2
est_1_preserve=1
est_1_training=1
```

  - the paths and parameters of the programs

Additional integration

- Any additional information

```
additional_list=1
# chipseq: on vote contre IG.
additional_1_file=%i/ADDITIONAL/MtrunA17r5.0-ANR.6387_H3K9Ac.gff3
additional_1_cfg_template=%i/ADDITIONAL/plugin_AnnotaStruct_H3K9Ac.cfg
```

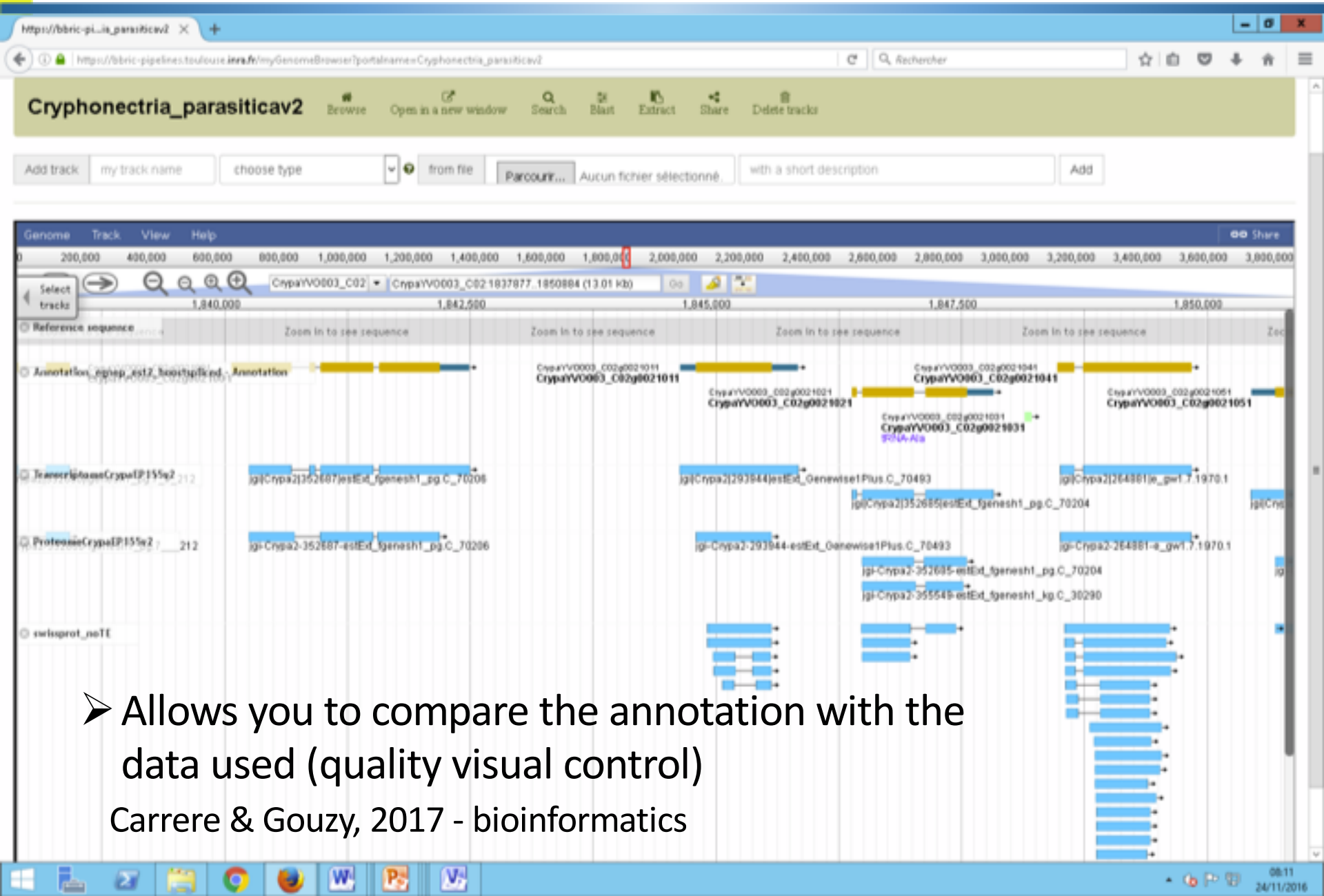
- One command line

$EGNEP/bin/int/egn-euk.pl --indir $PWD/indir
--outdir $PWD/outdir --cfg $EGNEP/cfg/egnep.cfg

Results

- Annotation in GFF3 format
- Fasta files of genes, mRNA, CDS, proteins, ncRNA
- Counting file
  - number of genes, average gene size, GC% of regions, etc
  - Information on the annotation steps
- Information on the annotation steps
  - % of aligned transcripts
  - Non-canonical sites detected
  - % of the sequence masked by repeated regions

MyGenomeBrowser: a tool to load output gff3 files



- Allows you to compare the annotation with the data used (quality visual control)

Carrere & Gouzy, 2017 - bioinformatics

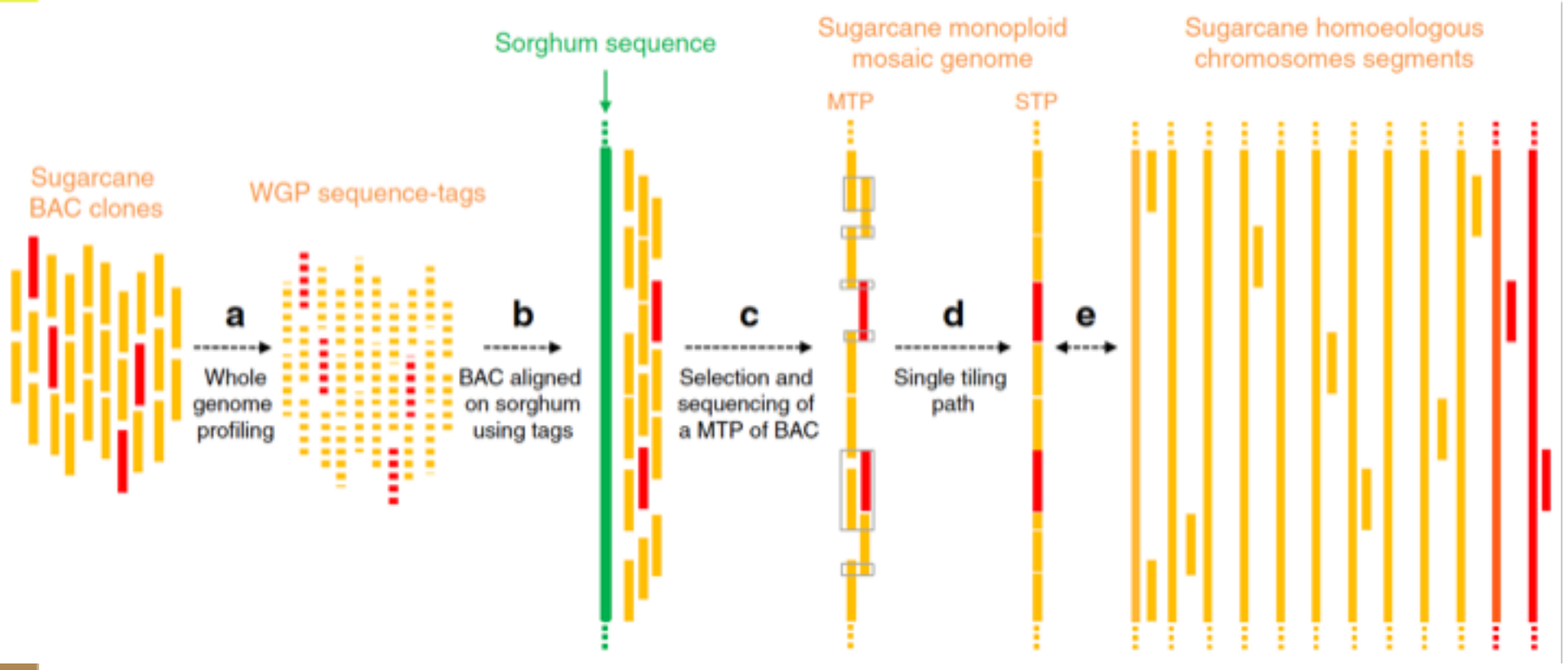
Sugarcane BACs: EuGène v3.2 annotation

- Use rice parameter files
- Expression data adapted to sugarcane related to sorghum

Plugin	Information	Format read by the plugin	Program commonly used to generate data
MarkovIMM	Statistical properties of regions	Binary matrix	egn_MarkovIMM
SMachine	Start codon & splicing sites	tsv	SpliceMachine
AnnotaStruct	Statistical properties of regions	GFF3	Fgenesh monocots
BLASTX	Similarities with Uniprot viridiplantae	GFF3	BLASTX
EST	Similarities with sugarcane EST	GFF3	GenomeThreader

Garsmeur et al., 2010 – New Phytol.

Sugarcane genome: sequencing strategy to the single tiling path



Garsmeur et al., 2018 – Nature Com.

Sugarcane genome: EGN-EP v1.4 & EuGène v4.2a annotation

- MarkovIMM trained on sugarcane
- Expression data = sugarcane RNA-Seq assembly

Plugin	Information	Format read by the plugin	Program commonly used to generate data
MarkovIMM	Statistical properties of regions	Binary matrix	egn_MarkovIMM
SWAM	Splicing sites (& Start codon)	tsv	SignalWAM
BLASTX	Similarities with Uniprot viridiplantae	GFF3	BLASTX
EST	Similarities with sugarcane transcripts from seven genotypes	GFF3	Gmap
AnnotaStruct	Several non coding gene prediction	GFF3	Rfam, Rnammer, tRNA-scan-SE
Repeat	Similarities with Repbase & RepeatDom & de novo repeat search	GFF3	Red, LTRHarvest

Garsmeur et al., 2018 – Nature Com.

Sugarcane genome: EGN-EP v1.4 & EuGène v4.2a result

Table 1 Selection and sequencing of BACs targeting the gene-rich part of the sugarcane monoploid genome

Selection and sequencing of a sugarcane minimum tiling path (MTP) of BACs			Sugarcane single tiling path (STP)			
Sorghum chromosome (Mb)	Nb of sugarcane BACs anchored	Nb of BACs sequenced (Mb)	Mosaic super scaffolds (Mb)	Genes		TE
				Nb	%	%
Sb01 (81)	1924	778 (94)	Sh01 (67)	4614	15	44
Sb02 (78)	1598	594 (68)	Sh02 (49)	3270	13	43
Sb03 (74)	1624	634 (74)	Sh03 (51)	3540	14	43
Sb04 (69)	1261	496 (56)	Sh04 (42)	2881	14	44
Sb05 (72)	827	289 (33)	Sh05 (22)	1337	11	41
Sb06 (61)	1060	404 (48)	Sh06 (33)	2189	13	45
Sb07 (66)	871	305 (36)	Sh07 (28)	1903	13	42
Sb08 (63)	623	265 (31)	Sh08 (23)	1381	12	43
Sb09 (59)	891	391 (46)	Sh09 (34)	2143	13	44
Sb10 (61)	1053	379 (45)	Sh10 (32)	2058	16	43
683 Mb	11,732	4535 (531 Mb)	382 Mb	25,316	13	43

Garsmeur et al., 2018 – Nature Com.

Sugarcane genome: EGN-EP v1.4 & EuGène v4.2a example

EuGène based Eukaryotic Pipeline conclusion

- Eugene is a gene structure combiner
  - Based on a directed acyclic graph and a scoring system.
  - Includes its own *ab initio* gene finder (egn_MarkovIMM) that it trains.

- In Eugene v4.2a
  - SignalWAM sensor is proposed as a replacement for Smachine in order to analyse splicing signals.
  - However Apache Spark Big Data architecture could allow the re-engineering of the SpliceMachine software (SVM Machine Learning)

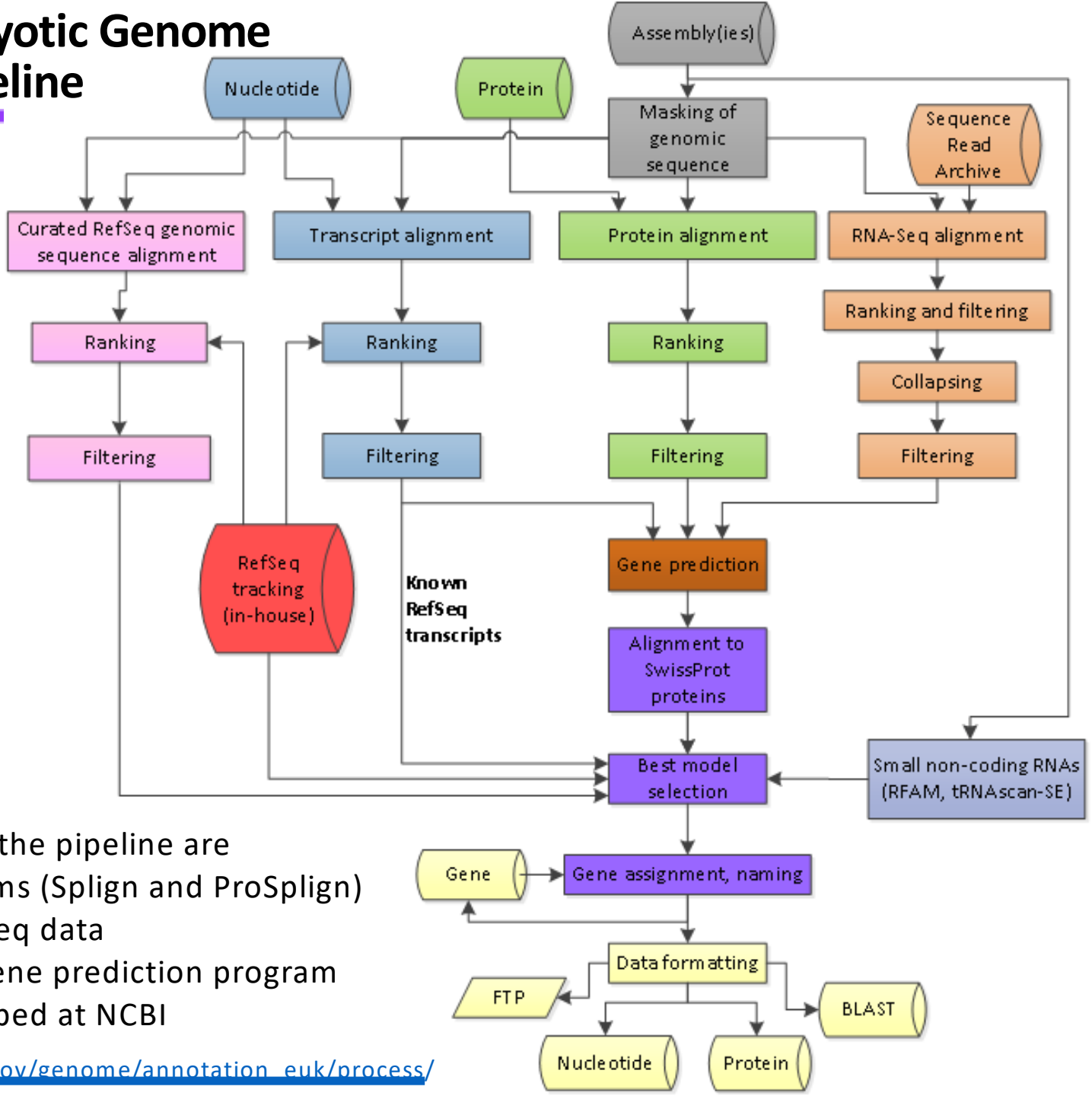
- Next in EGN-EP is to replace UBLAST by open source Diamond to accelerate BLAST.

EGN-EP / Maker comparison

	EGN-EP	MAKER
Transcriptomic data	EST, transcript contigs	EST, transcript contigs, RNA-Seq
Start & splicing site prediction model	Yes (SWAM)	Natively no
Alternative transcript prediction	No	**Yes** (*e.g.* Augustus integration, merge annotation)
Automatic training	Yes (*i.e* egn_MarkovIMM)	No
Automatic Optimisation	No	No
Combiner vs integrator	**Combiner**	Chooser (MK2) Combiner (MK3/EVM)
Customisable	Yes	Yes
Weights / priorities	Yes	No (MK2) Yes (MK3/EVM)

These tools are complementary

The NCBI Eukaryotic Genome Annotation Pipeline

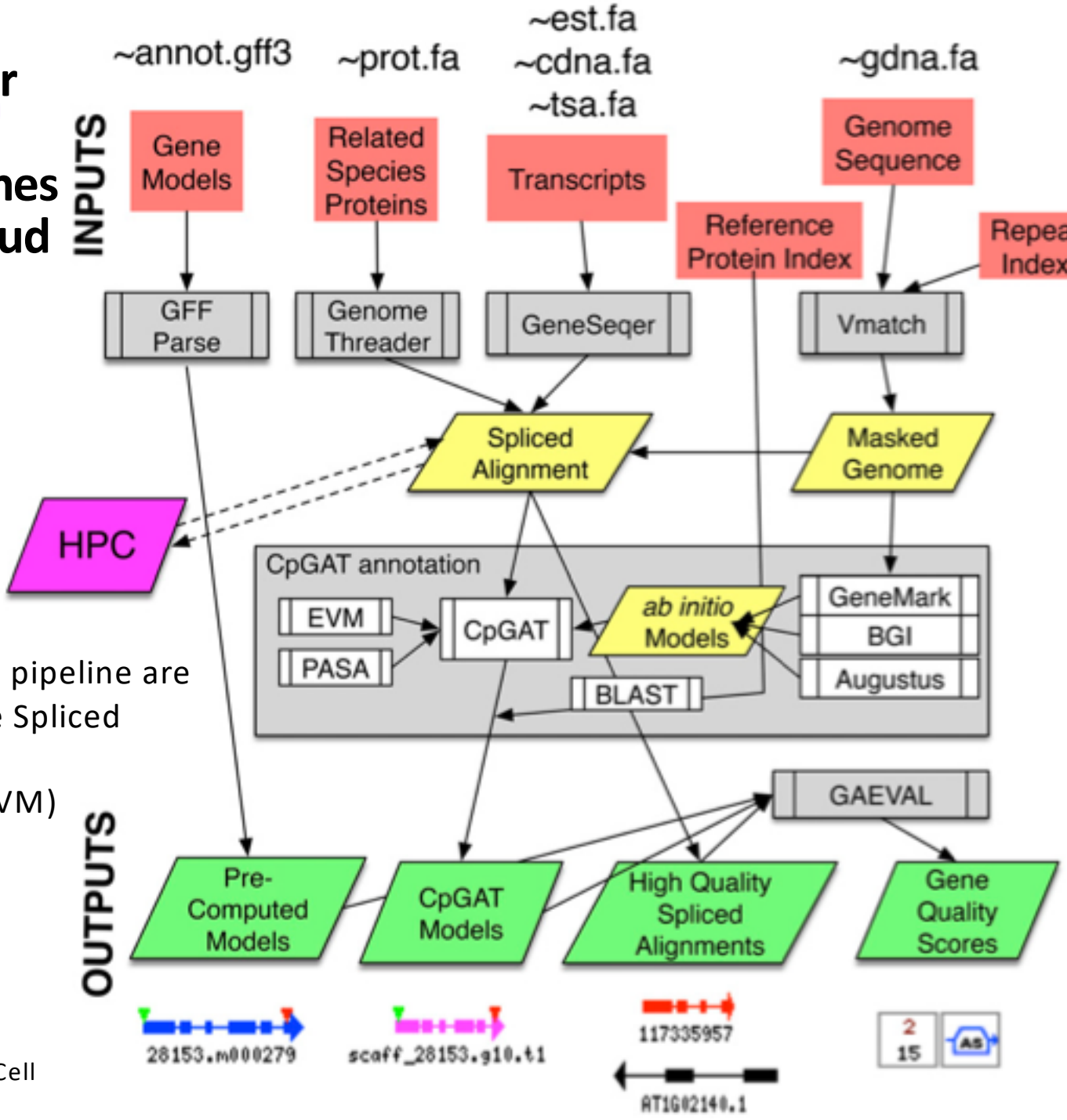


Core components of the pipeline are

- alignment programs (Splign and ProSplign) with curated RefSeq data
- an HMM-based gene prediction program (Gnomon) developed at NCBI

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/

xGDBvm: A Web GUI-Driven WK for Annotating Eukaryotic Genomes in the Cyverse cloud



Core components of the pipeline are

- Program to Assemble Spliced Alignments (PASA)
- EvidenceModeller (EVM)

Duvick et al., 2016 – Plant Cell