Traitement des séquences NGS chez les insectes ravageurs de cultures

Fabrice Legeai
fabrice.legeai@rennes.inra.fr

Journée de la plate-forme - 26 octobre 2010

Les insectes provoquent des dommages

Chaque jour 10-20% de la production mondiale agricole est détruite par les insectes herbivores

Aphids

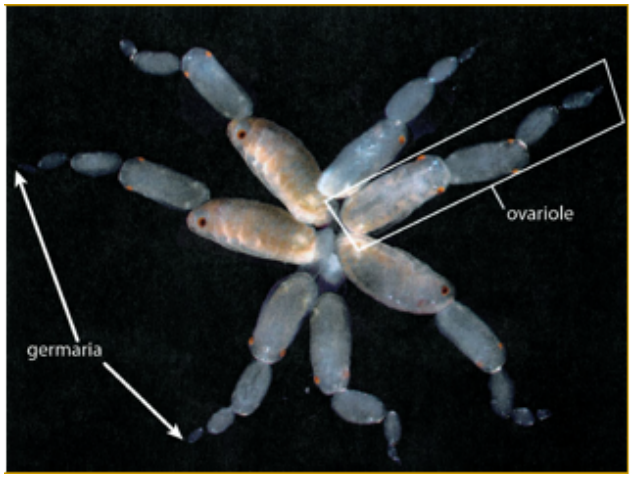
- brêche dans le phloème
- lésion
- transmission de virus
- prolifération des pathogènes
- miellat

Spodoptera

- megapest
- polyphage (coton, maize)
- immune
- resistant
- chemoreception évoluée

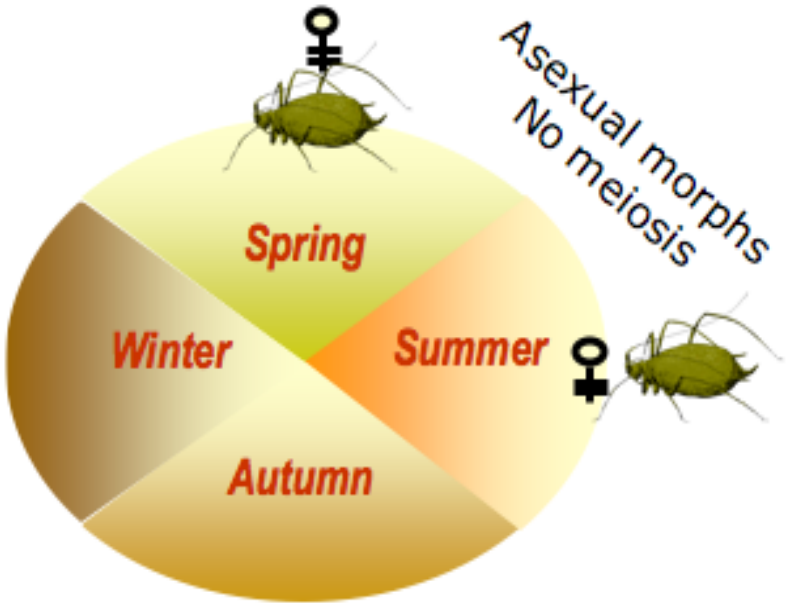
Pucerons du pois

Parthénogénèse vivipare et forte démographie



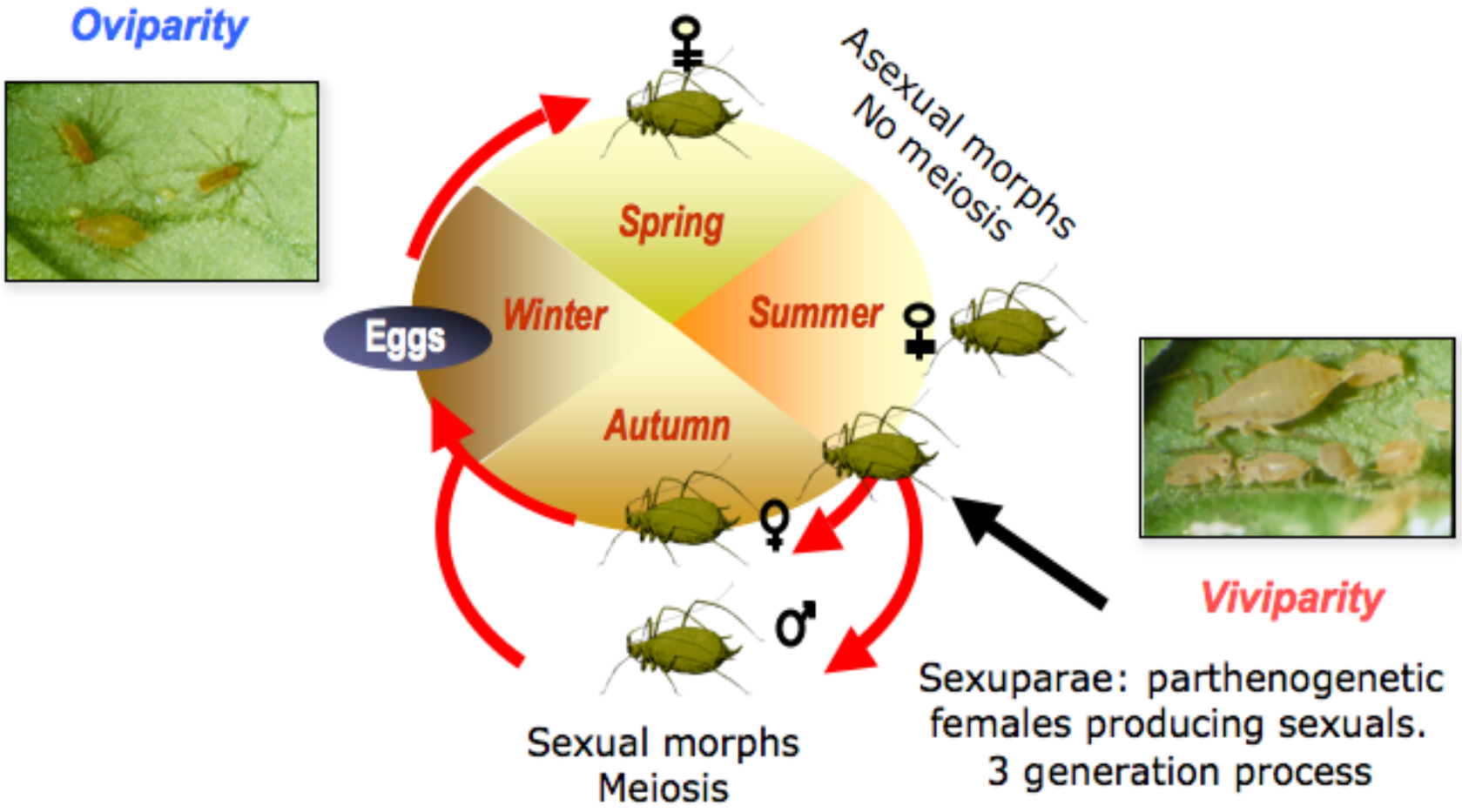
Pucerons du pois

Cycle annuel



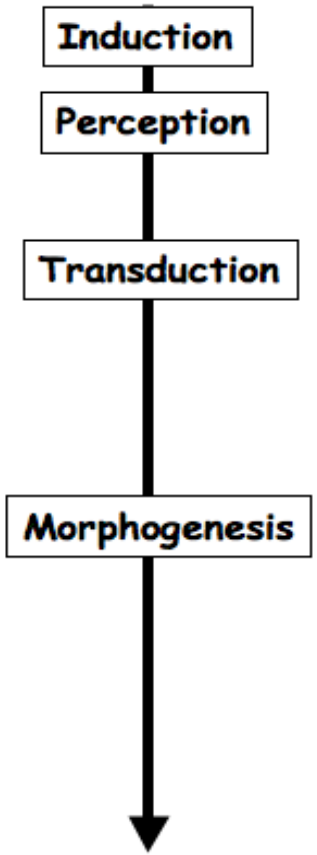
Pucerons du pois

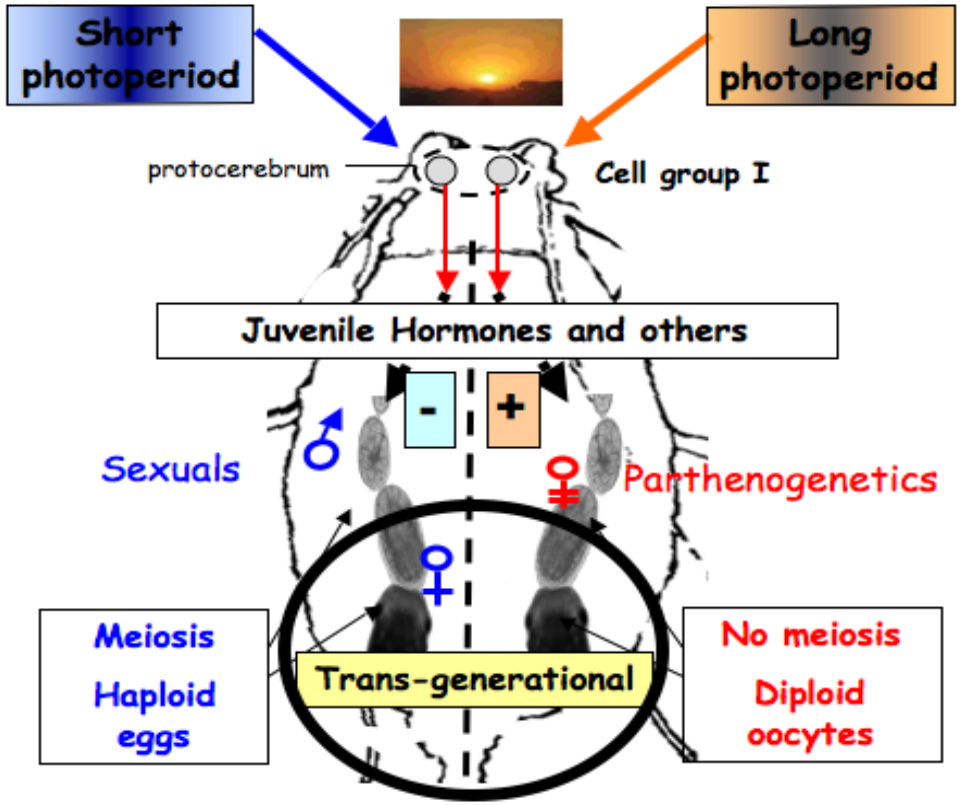
Cycle annuel



Pucerons

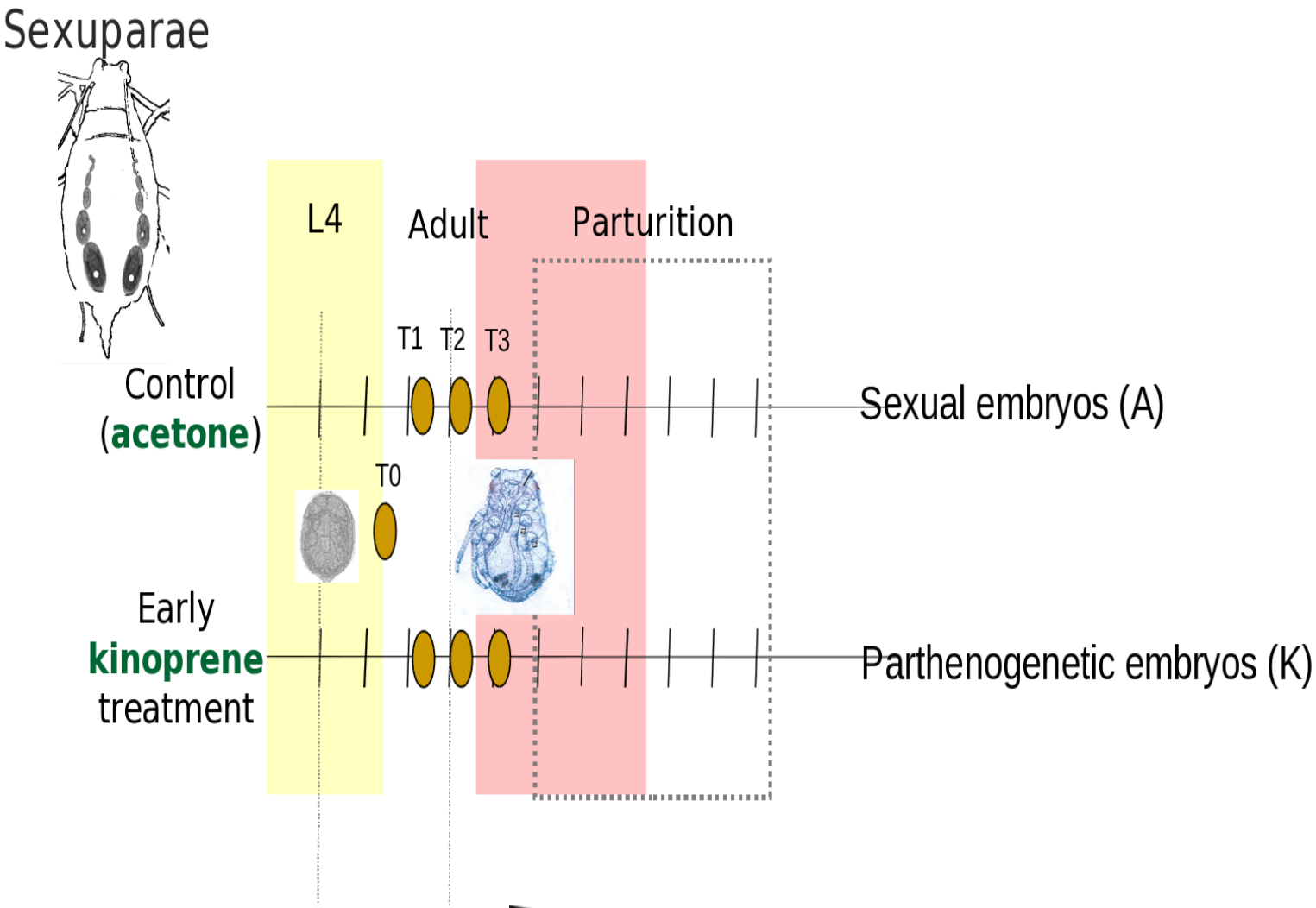
Régulation du polyphénisme de reproduction





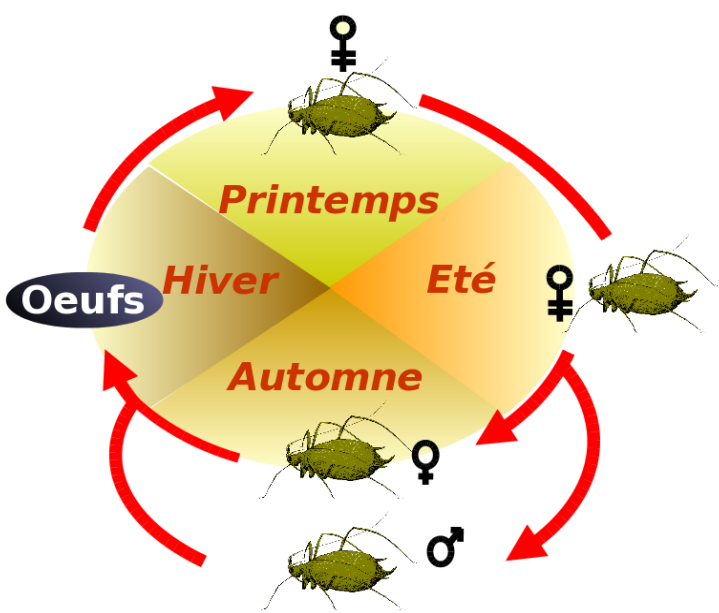
Pucerons

Régulation du polyphénisme de reproduction



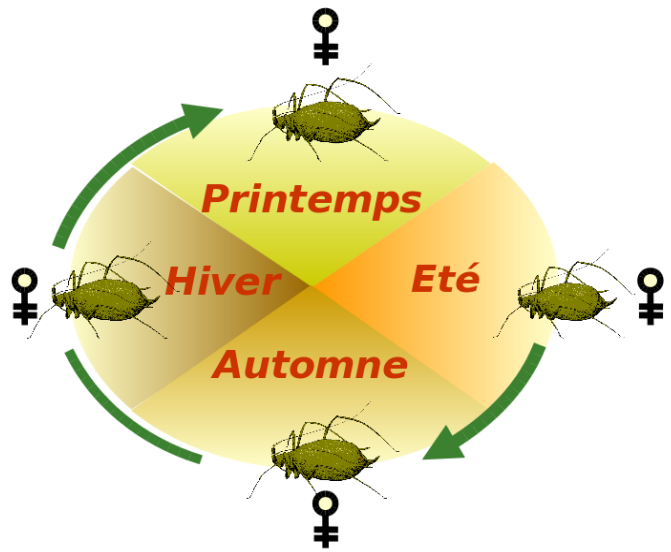
Pucerons

Régulation du polyphénisme de reproduction



Lignées "sexuées"
(N gén. clonales + 1 gén. sexuée)

≠



Lignées asexuées
(perte totale ou partielle du sexe)

Pucerons

Régulation du polyphénisme de reproduction

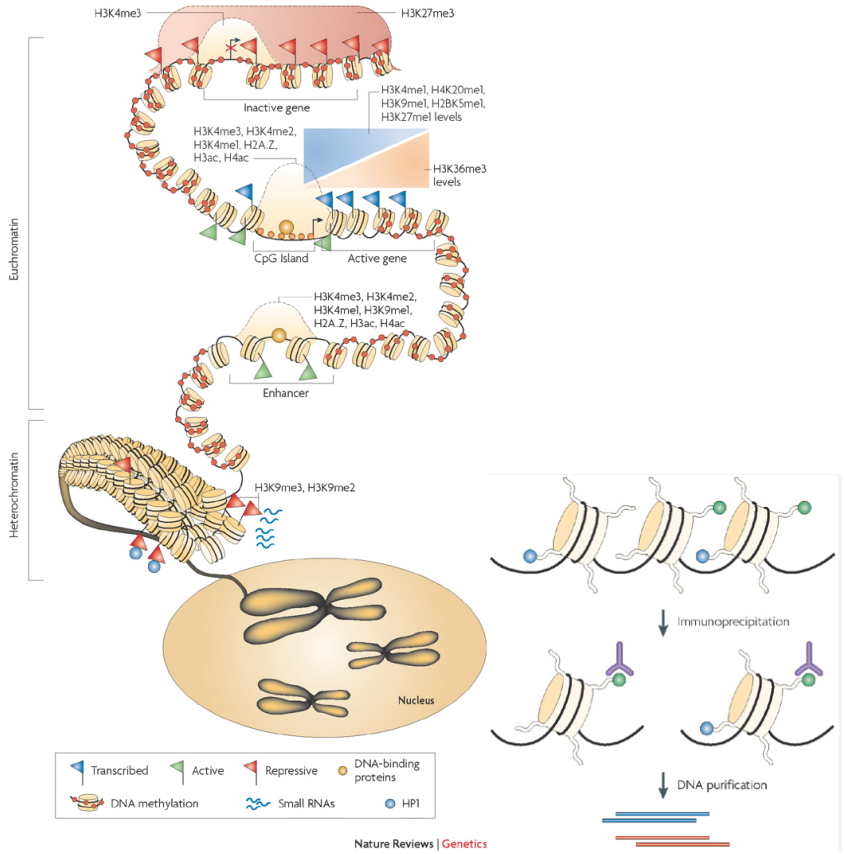
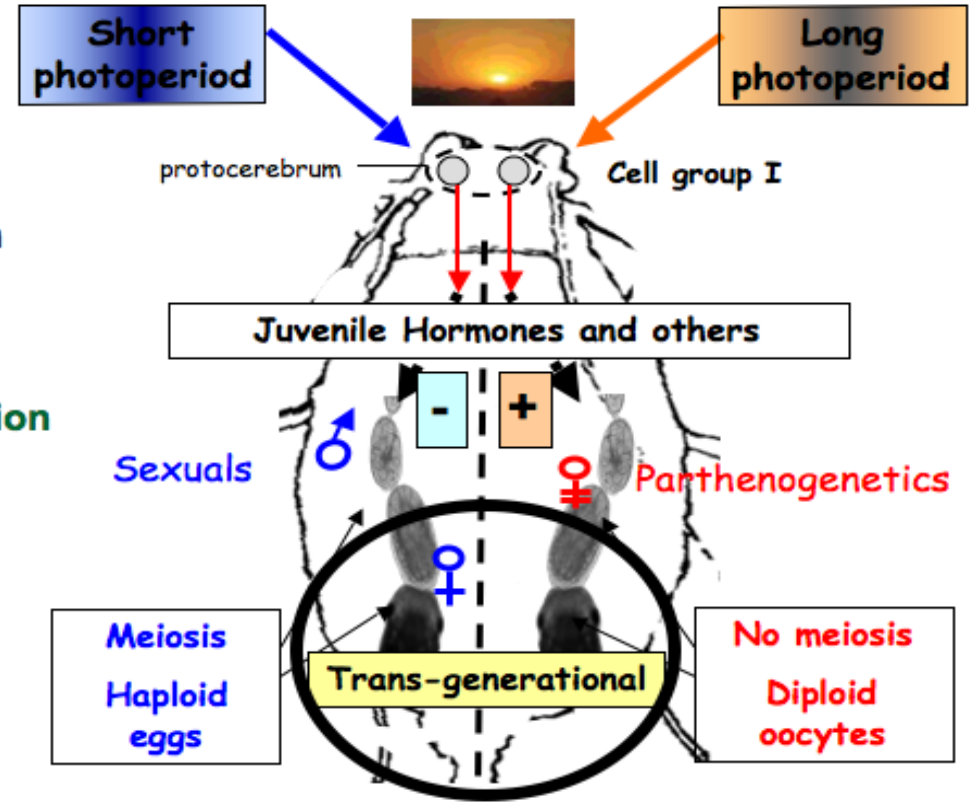


FIG. *1:* D.E. Schones and K. Zhao Nature Reviews Genetics 2008

Pucerons

Régulation du polyphénisme de reproduction

Epigenetic regulation
Gene
↓ Transcription
mRNA
Post-transcriptional regulation
↓ Translation
Proteins



Données NGS

RNA-Seq

1. 2 génotypes distincts
   - LSR1 (cycle sexué/parthénogénèse) : 11 632 206 paires ends
   - L121 (asexué) : 11 778 570 paires ends
2. 2 stades différents
   - sexupare L4 : +20 M
   - sexupare adulte : +20 M

Données NGS

ChIP-Seq

1. diméthylation de la lysine 4 de l'histone H3 (H3K4me2) : marque de chromatine centromérique chez la drosophile
   - H3K4me2 : 24 715 862 paires
   - contrôle : 33 067 689 paires
2. diméthylation de la lysine 9 de l'histone H3 (H3K9me2) : marque d'hétérochromatine chez la drosophile
   - en cours

Données NGS

Small RNAs

1. adulte parthénogénétique : 3 M
2. sexuperae L4 : 26M
3. sexuperae traité
   - 3 stades (T1, T2, T3) : 37M seq nettoyées
4. sexuperae non-traité :
   - 3 stades (T1, T2, T3) : 43M seq nettoyées

Génome du puceron du pois

Acyrthosiphon pisum

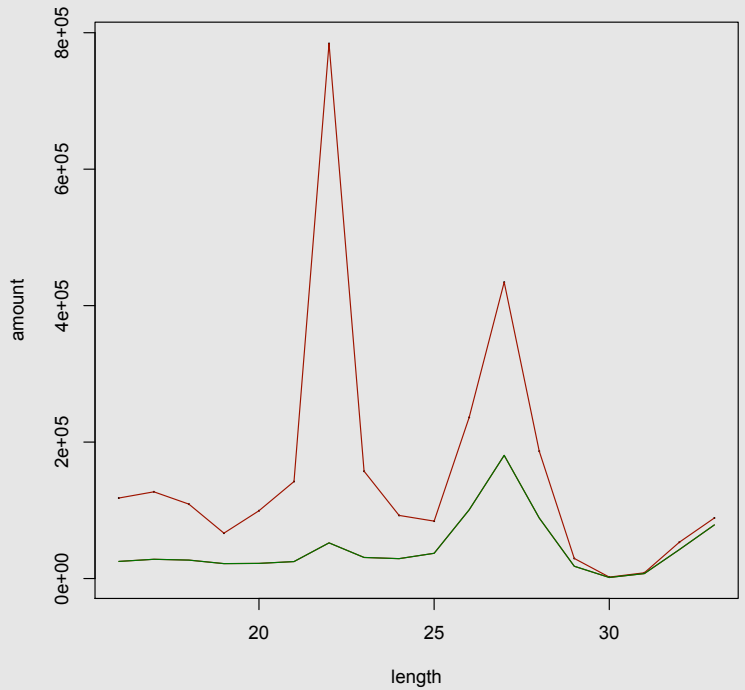
- 4 chromosomes (525 Mb)
- 464 Mb assembled in Dec 2007 (6.2X)
- 72 844 contigs (N50 : 10.7kb) - 22 801 scaffolds (N50 : 88.5 kb)
- 34821 genes (7% genome)
- 13911 different Transposable elements spread over 466727 loci (cov : 28.5%)

Deep sequencing prediction

From Sequences

3,000,000 Illumina-Solexa reads

851,979 short RNA sequences (mean : 25.2 bp, median : 26 bp, max : 33 bp, min : 1 bp)



to alignments on the genome

Deep sequencing prediction

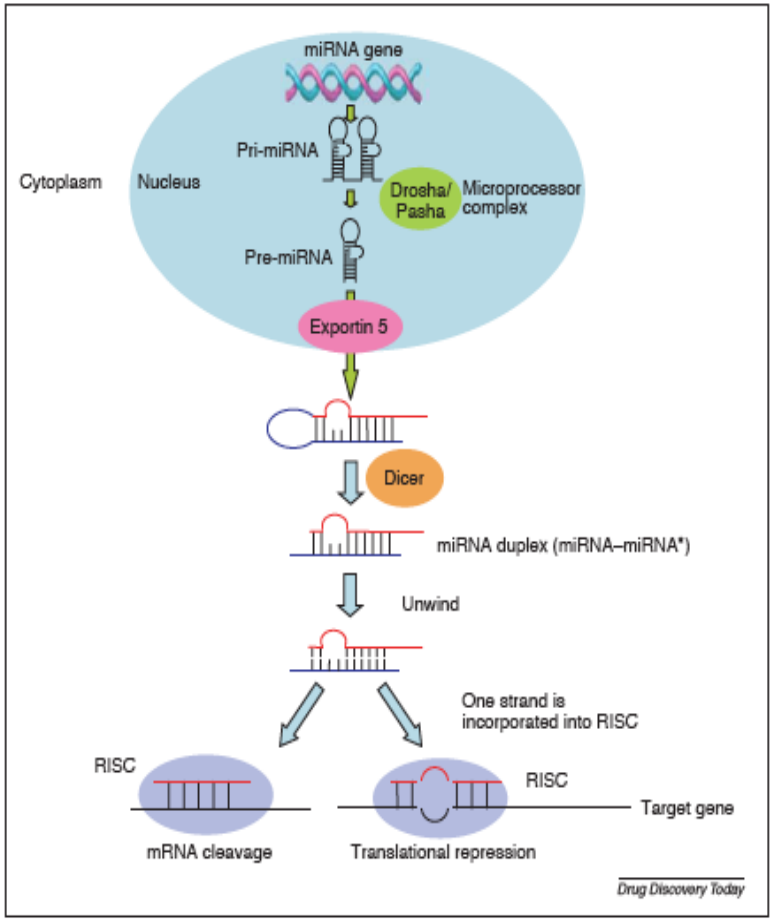
From Sequences

to alignments on the genome

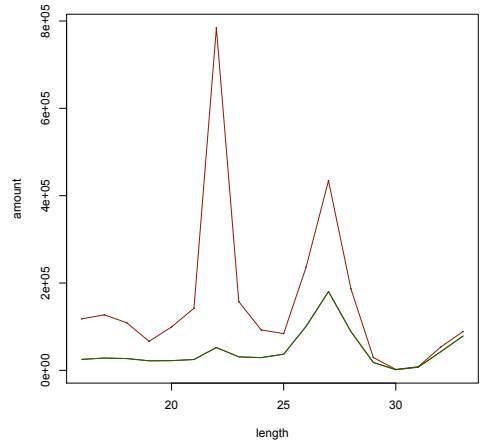
Using global alignment process (*Gassst* Rizk and Lavenier, Bioinformatics 2010)

Similarity percentage	mapped uniques	mapped reads	loci
90%	503,928	2,372,927	12,608,723
95%	335,844	1,798,056	4,068,078
100%	230,791	1,470,943	1,890,385

miRNA

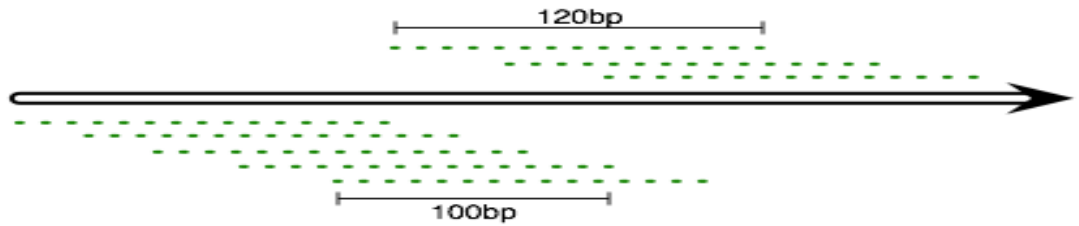


Brown et Sanseau DDT vol10, 2005



exhaustive extraction

The hairpin review



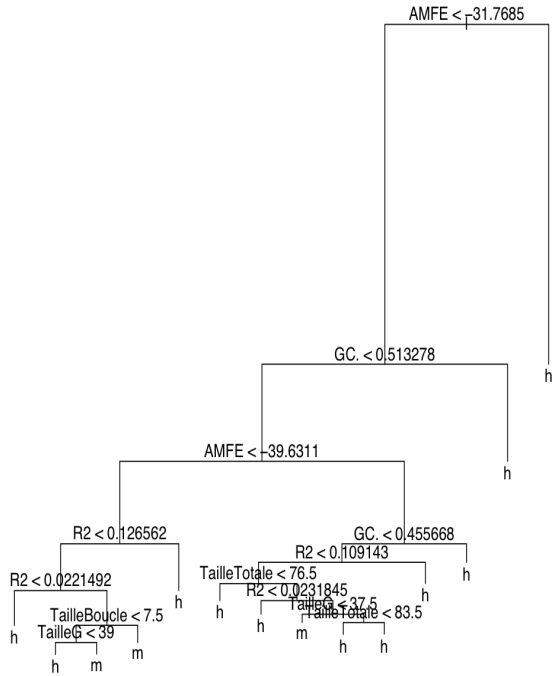

```
>SEQ:Contig1   POS:0   REV:0 TAI_CONTIG:496
UAGAACAGCAGUGCCACCGGUAGCAGGCGACCGCGGACGGGCGGCAGAUCGGUUCCUGUAUGUGCAGUAGUUAGGAAUGUGAAGUUUAGUAGAAGGGGAUAUACGGGUUUCGUUAAGUAA
......((..((((.(((((..((.((...)).)).)).))))..))))...((..((((((((((.(.........((((.....))))).........).)))))))))...)).....)).  (-31.27)
>SEQ:Contig1   POS:20   REV:0 TAI_CONTIG:496
GAGCAGGCGACCCGCGGACGGGCGGCAGAUCGGUUCCUGUAUGUGCAGUAGUUAGGAAUGUGAAGUUUAGUAGAAGGGGAUAUACGGGUUUCGUUAAGUAAGAUUAUGUCCGGUUUAAAGUA
((((.(.((.((((....)))))).)...((..((((((((((.(.........((((.....))))).........).)))))))))...))...................)))).....  (-26.17)
>SEQ:Contig1   POS:40   REV:0 TAI_CONTIG:496
GGCGGCAGAUCGGUUCCUGUAUGUGCAGUAGUUAGGAAUGUGAAGUUUAGUAGAAGGGGAUAUACGGGUUUCGUUAAGUAAGAUUAUGUCCGGUUUAAAGUAGGAGUGUGAGGGAAUGUAGU
..(((((((((...(((((((((.(.........((((.....))))).........).)))))))))..............)))).)))))..............................  (-19.36)
```

Selection
size > 60 bp
pairing > 70%

```
>SEQ:Contig1 POS:167 TT:78 TB:9 TG:37 TD:32 UNP:13 TAIC:496 REV:0
GUUUAUGUUUGAUACUGUUGGAGUUCCAAAAUAGAAUUUUUAAGUUAUUCAUUUUGAGGCGAGAGUAUGGACAGGAAU
((((-(((((-(((((-((---((--((((((-((((.........))))))))))..-))))).))))))))))-))))  (-15.90)
>SEQ:Contig1 POS:172 TT:68 TB:9 TG:32 TD:27 UNP:11 TAIC:496 REV:0
UGUUUGAUACUGUUGGAGUUCCAAAAUAGAAUUUUUAAGUUAUUCAUUUUGAGGCGAGAGUAUGGACA
.((((-(((((-((---((--((((((-((((.........))))))))))..-))))).))))))))).  (-14.10)
>SEQ:Contig2 POS:104 TT:64 TB:4 TG:32 TD:28 UNP:16 TAIC:1022 REV:0
UCUUCCGGCGAUCGGGUAACCUUGCGGGUCACUAAUGCGACCCAGUUCCGCCUGGUCCAGAAGA
(((((--(-(((((((((.....-(((((((-(....)-)))))-)))---)))))))))-)))))  (-24.30)
>SEQ:Contig2 POS:60 TT:66 TB:4 TG:32 TD:30 UNP:12 TAIC:1022 REV:1
```

exhaustive review
Hairpin filtering

Feature
minimum free energy /sequence length *100
unpaired base percentage
monomer repetition percentage
dimer repetition percentage
GC percentage
Maximum asymmetry observed in a single loop terminal loop region excluded
Total asymmetry observed over all loops terminal loop region excluded
Maximum asymmetry observed in the terminal loop region
Maximum size of internal loop
Total size of internal loops
left arm length
right arm length
total sequence length
score reported by Microprocessor
SVM (Snorre A. Helvik et al. 2006)



mirDeep

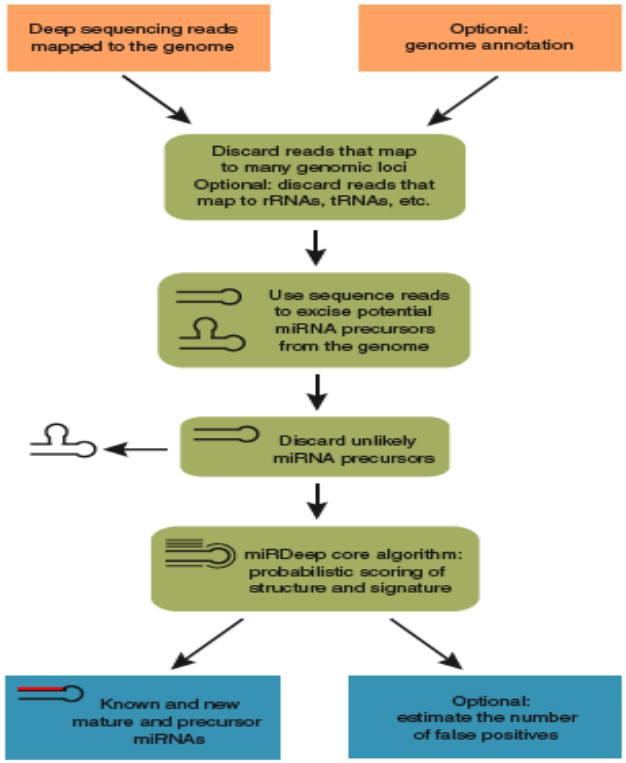
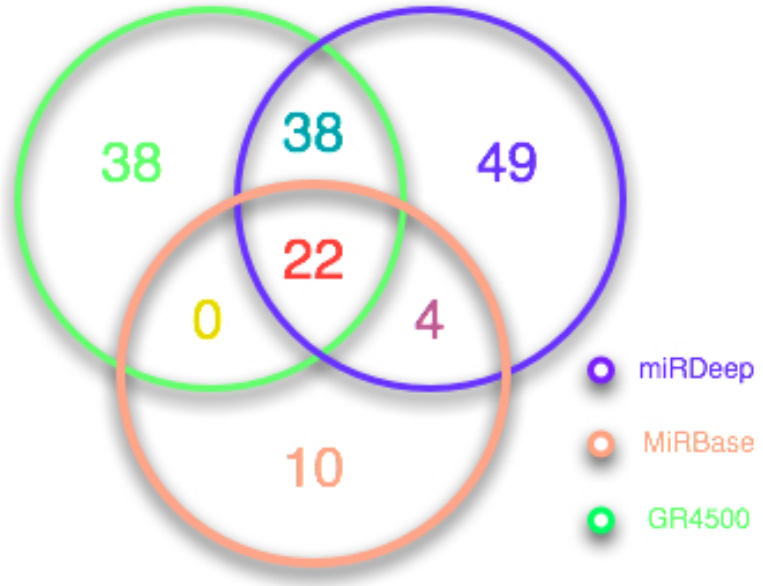


FIG. *1:* Friedländer et al., Nat Biotechnol. 2008

Results

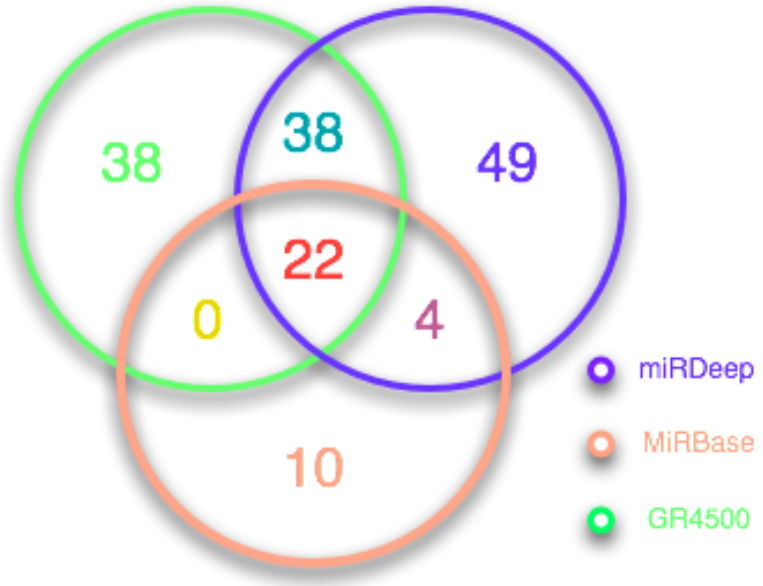
- 161 mature microRNA
- 189 precursors



- Only 53 (32.9%) of the 161 aphid mature miRNAs showed significant homology with known miRNAs.

Results

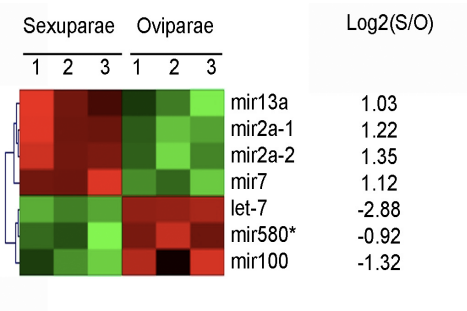
- 161 mature microRNA
- 189 precursors

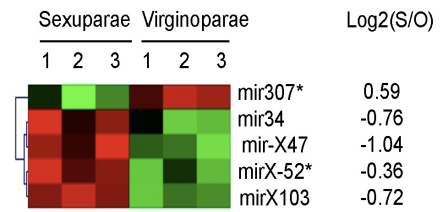


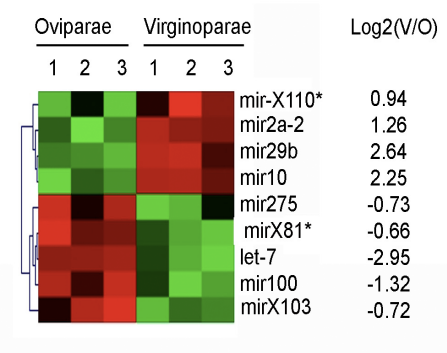
- Only 53 (32.9%) of the 161 aphid mature miRNAs showed significant homology with known miRNAs.

miRNA expression and the switch of reproductive mode

- 10 replicates of 149 miRNA and their mir*
- only four mature miRNAs and nine mir* gave no signal with any of the hybridizations.
- 17 miRNAs showed significant differences between the morphs







Legeai *et al. BMC Genomics* 2010, **11**:281
http://www.biomedcentral.com/1471-2164/11/281

BMC Genomics

RESEARCH ARTICLE Open Access

Bioinformatic prediction, deep sequencing of microRNAs and expression analysis during phenotypic plasticity in the pea aphid, *Acyrthosiphon pisum*

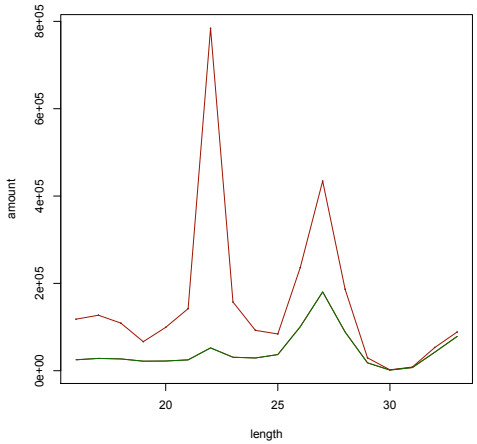
Fabrice Legeai[1,2], Guillaume Rizk[3], Thomas Walsh[4], Owain Edwards[4], Karl Gordon[5], Dominique Lavenier[6], Nathalie Leterme[1], Agnès Méreau[7], Jacques Nicolas[2], Denis Tagu[1] and Stéphanie Jaubert-Possamai*[1]

Validation avec un nouveau jeu de données

- transcription de 119 / 161 microARN
- transcription de 69 / 108 microARN spécifiques
- 46 microARN différentiellement exprimés entre les 2 conditions à au moins 1 des stades

piRNA features

- 26-30 nucleotides long
- no structural motifs
- no conservation among species
- mainly associated with repeats (rasiRNA)
- in euchromatin, often distributed in discrete loci from at least 20 piRNAs (flamenco, Maf in flies)
- strong and strict interaction with a transposon
- preference for a 5' term U



piRNA transposon silencing mechanism

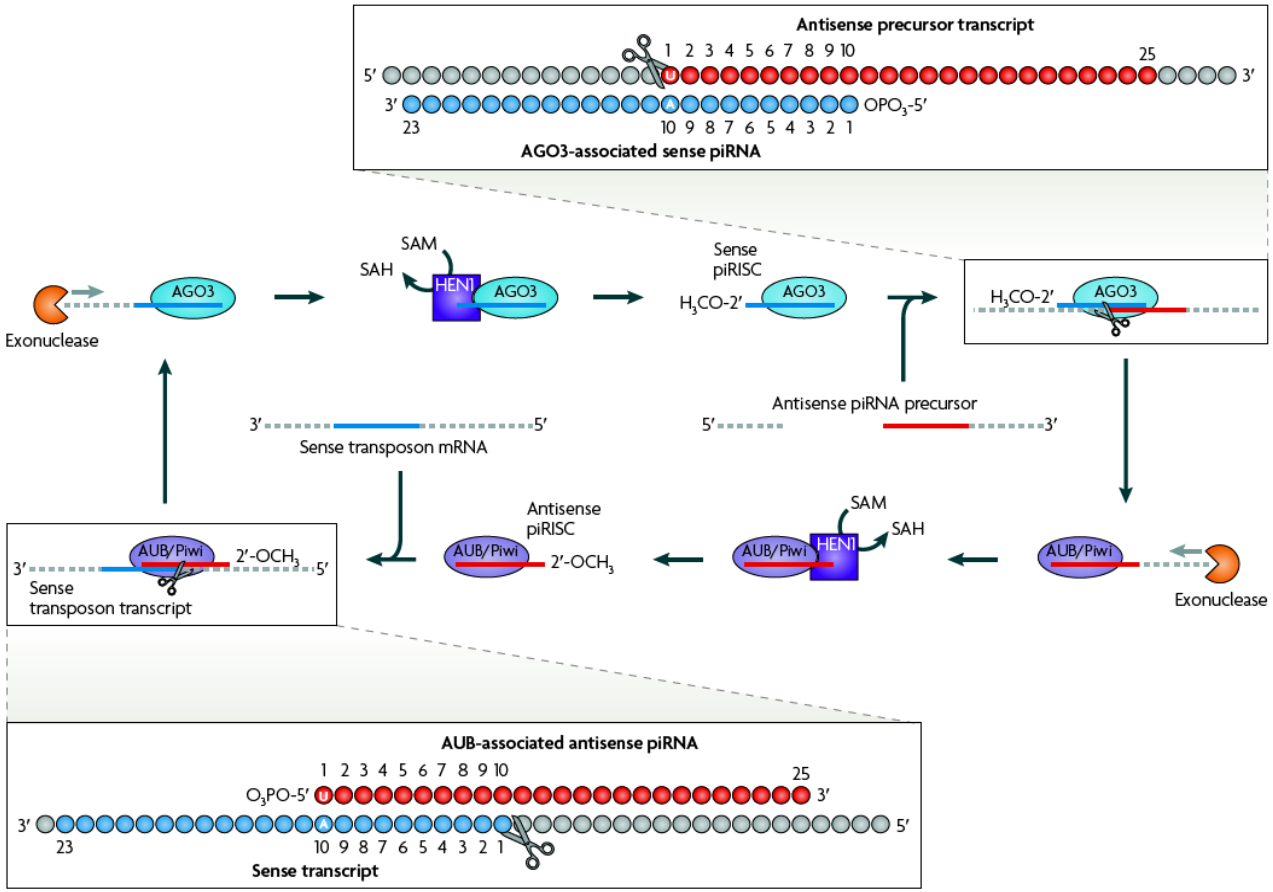


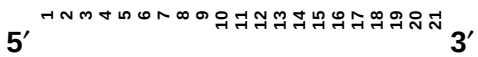
FIG. 2: Ghildayal M and Zamore D, Nat Rev Genet 2009

Nucleotide bias

size	gene sense	gene antisense	repeat sense	repeat antisense	intronic	intergenic
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						

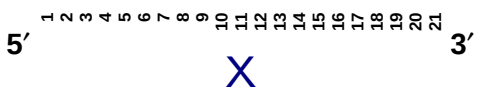
Nucleotide bias

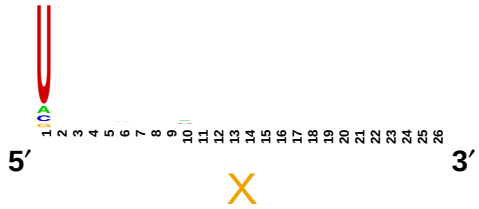
size	gene sense	gene antisense	repeat sense	repeat antisense	intronic	intergenic
20	X		X	X	X	
21	X	X	X	X	X	
22	X	X	X	X	X	
23	X		X	X	X	
24	X		X			
25						
26						
27						
28						
29						



Nucleotide bias

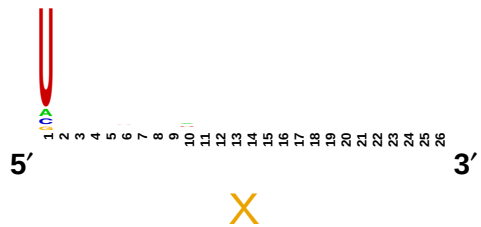
size	gene sense	gene antisense	repeat sense	repeat antisense	intronic	intergenic
20	X	X	X	X	X	X
21	X	X	X	X	X	X
22	X	X	X	X	X	X
23	X	X	X	X	X	X
24	X	X	X	X	X	X
25		X	X	X	X	X
26	X	X	X	X	X	X
27	X	X	X	X	X	X
28	X	X	X	X	X	X
29	X	X	X	X	X	X

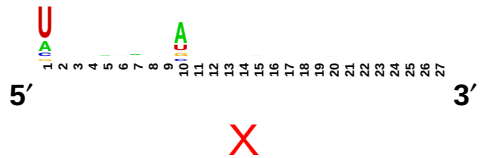




Nucleotide bias

size	gene sense	gene antisense	repeat sense	repeat antisense	intronic	intergenic
20	X	X	X	X	X	X
21	X	X	X	X	X	X
22	X	X	X	X	X	X
23	X	X	X	X	X	X
24	X	X	X	X	X	X
25		X	X	X	X	X
26	X	X	X	X	X	X
27	X	X	X	X	X	X
28	X	X	X	X	X	X
29	X	X	X	X	X	X





piRNA
Masterloci

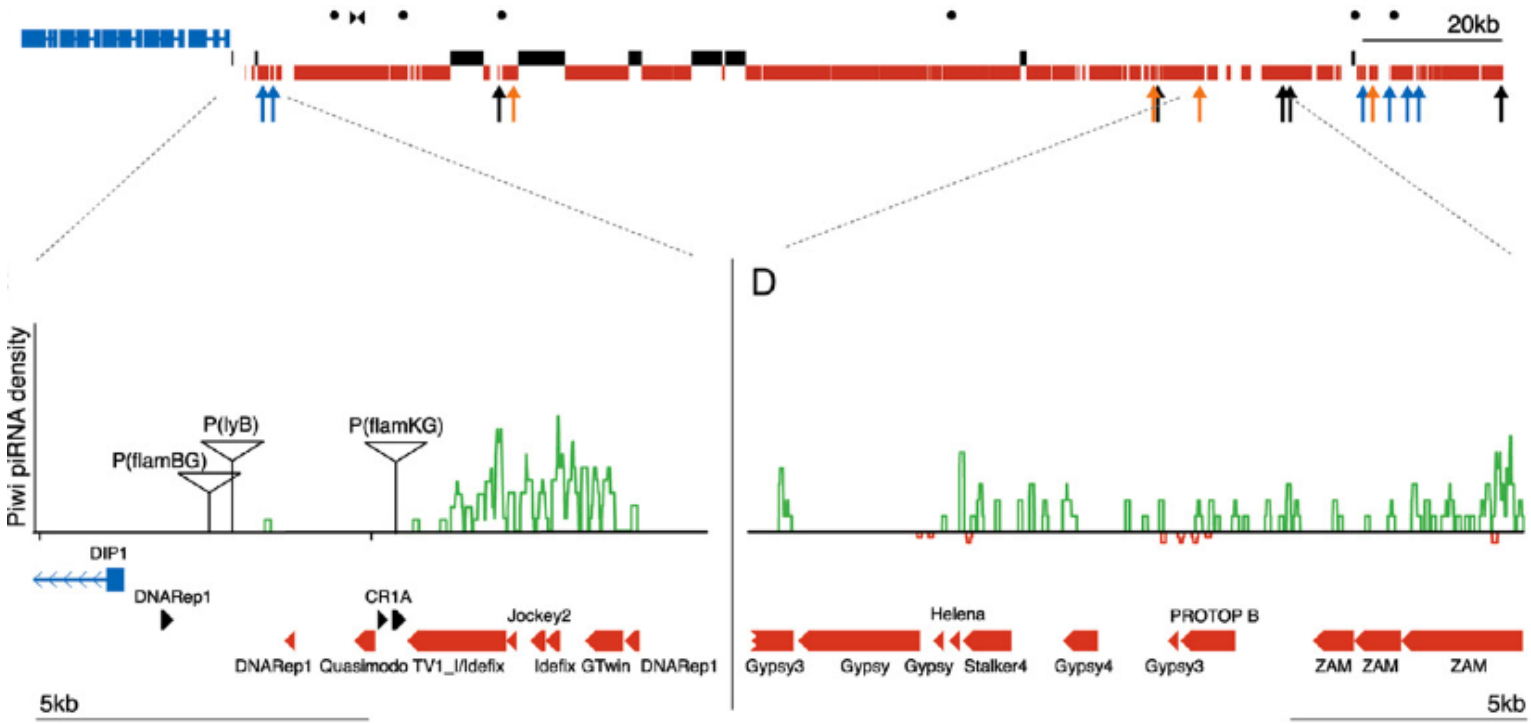


FIG. 3: Drosophila Flamenco locus, Brennecke et al Cell 2007

piRNA
Masterloci

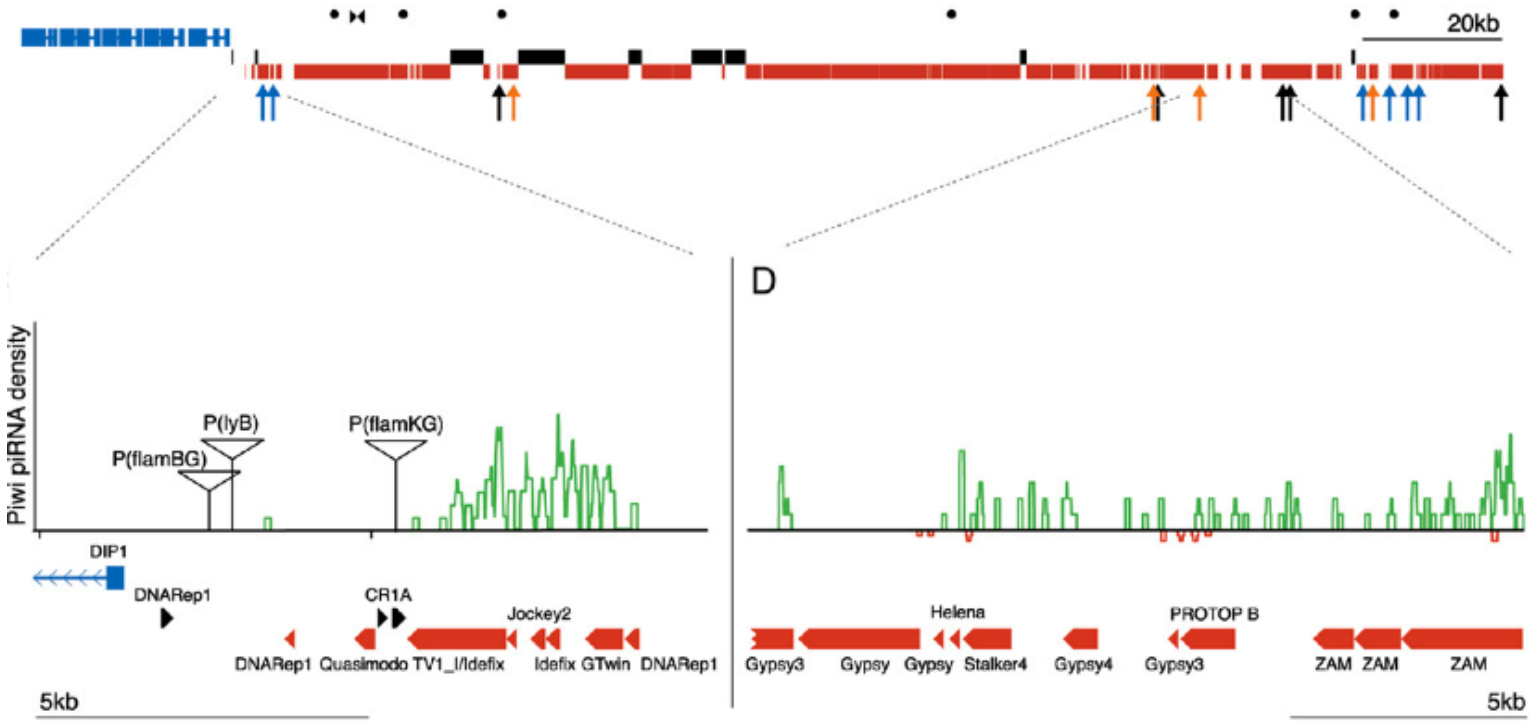


FIG. *3:* Drosophila Flamenco locus, Brennecke et al Cell 2007

=> Region selection where several small RNAs 26-28bp long were uniquely mapped (141644 / 230791)

piRNA
Masterloci

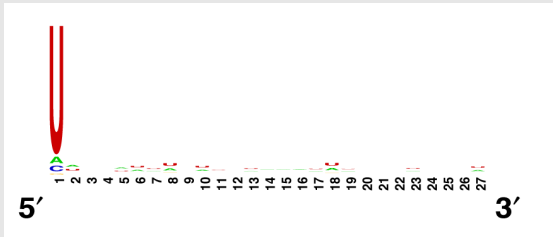
EQ120815

Region from 20,000 to 34,000

type	number (strand)	gene sense	gene antisense	transposon sense	transposon antisense
unique	572 (566+, 6-)	4	223	27	63
non unique	273 (252+,21-)	4	72	44	56

piRNA
Masterloci

EQ120815

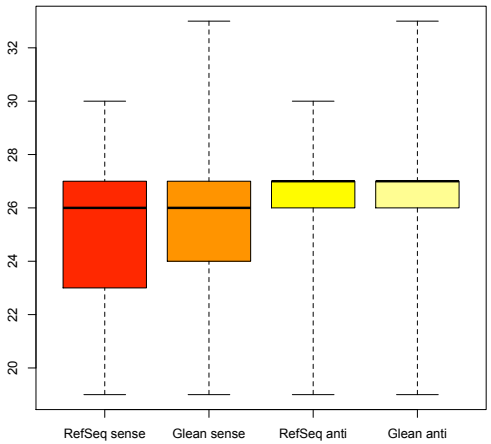


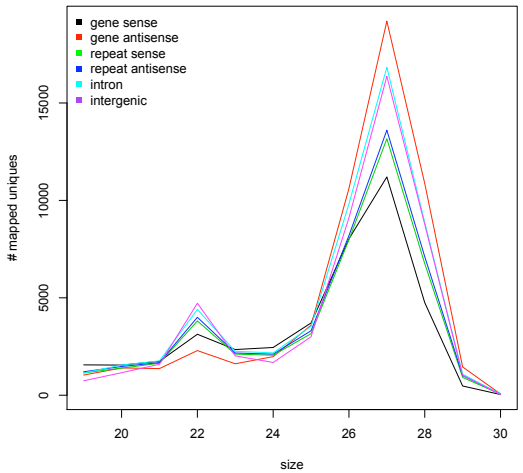
piRNA
Masterloci

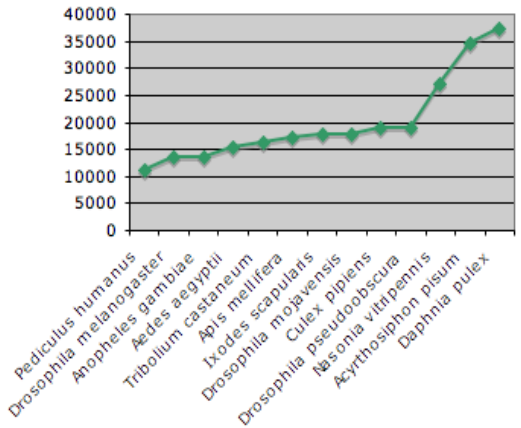
EQ120815

Gene	Type	Annotation	Alignment to Repbase
XM_001947766	RefSeq	piggyBac transposable element derived 1	Vandal6 piggyBag
Glean_32271	Glean	Hypothetical protein	none
Glean_32270	Glean	Hypothetical protein	none
Glean_32269	Glean	Hypothetical protein	none
Glean_32268	Glean	Hypothetical protein	none
Glean_32267	Glean	reverse transcriptase	Gipsy

Antisense genes

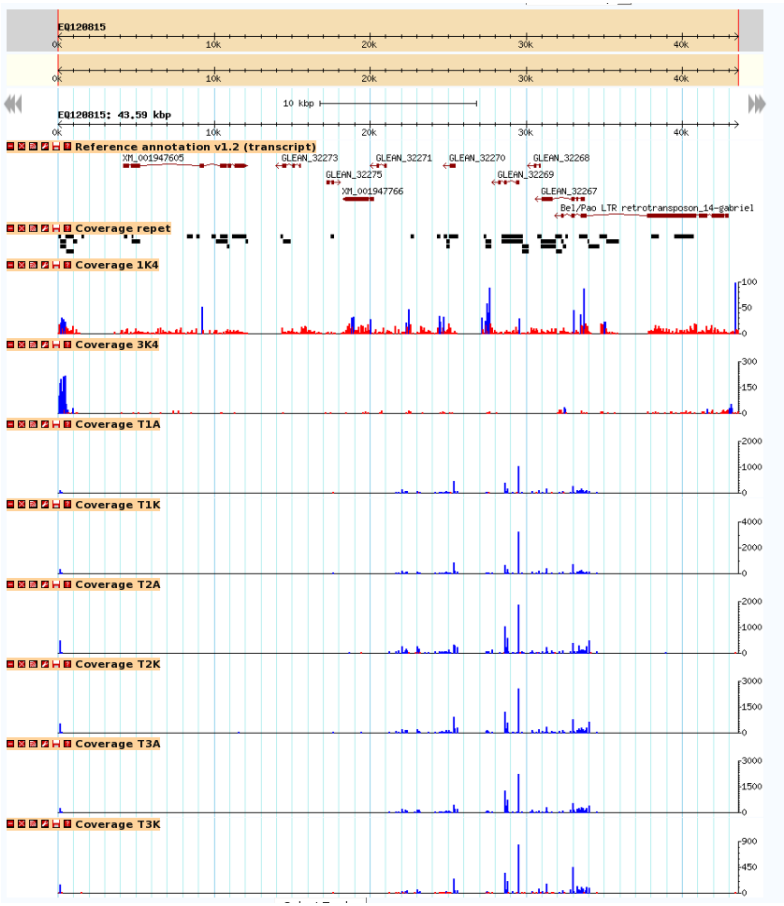




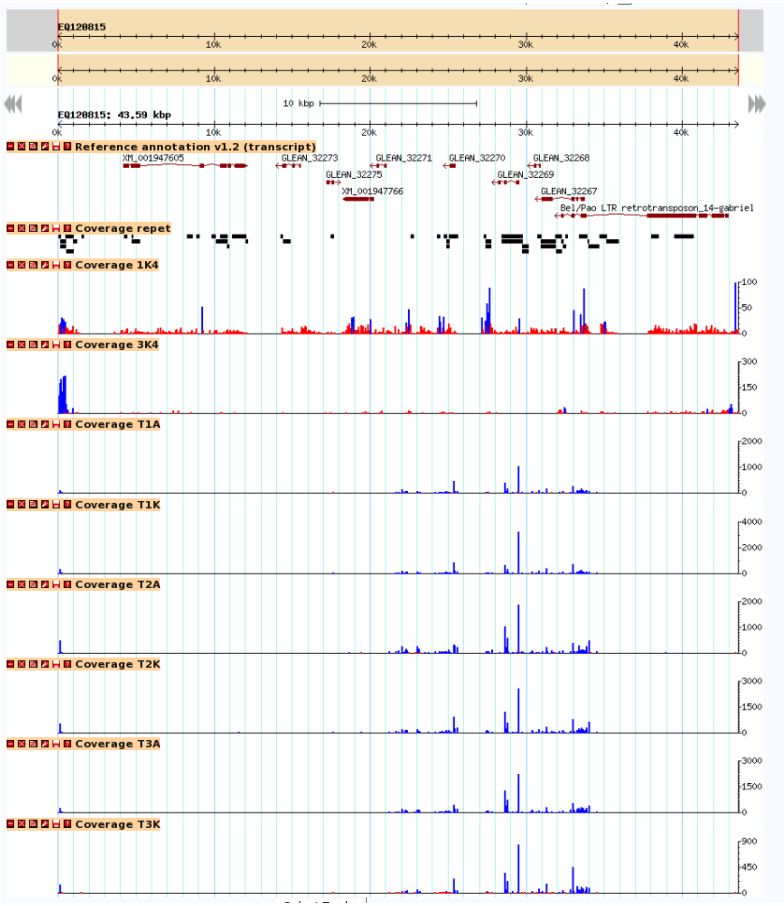


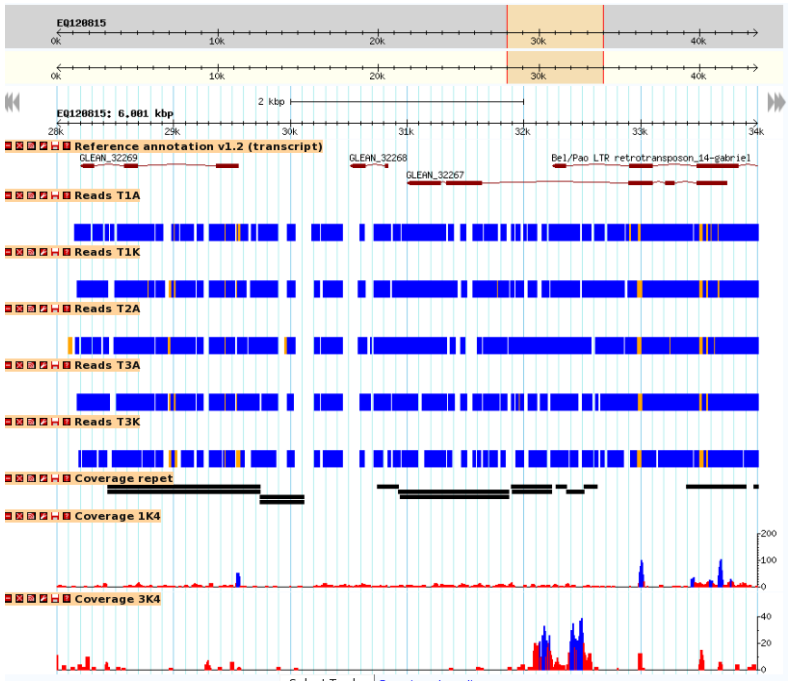
	Genes		Loci		Uniques	
	RefSeq	Glean	RefSeq	Glean	RefSeq	Glean
in the aphid genome	10466	24355				
with sense piRNA	4740	7452	16896	76358	15335	26526
with antisense piRNA	1061	8086	7148	202,625	6682	49846

Validation des clusters avec un nouveau jeu de données



Validation des clusters avec un nouveau jeu de données





Extension et intégration de données

Photoperiod

Epigenetic regulation

Gene

Transcription

mRNA

Post-transcriptional regulation

Translation

Proteins

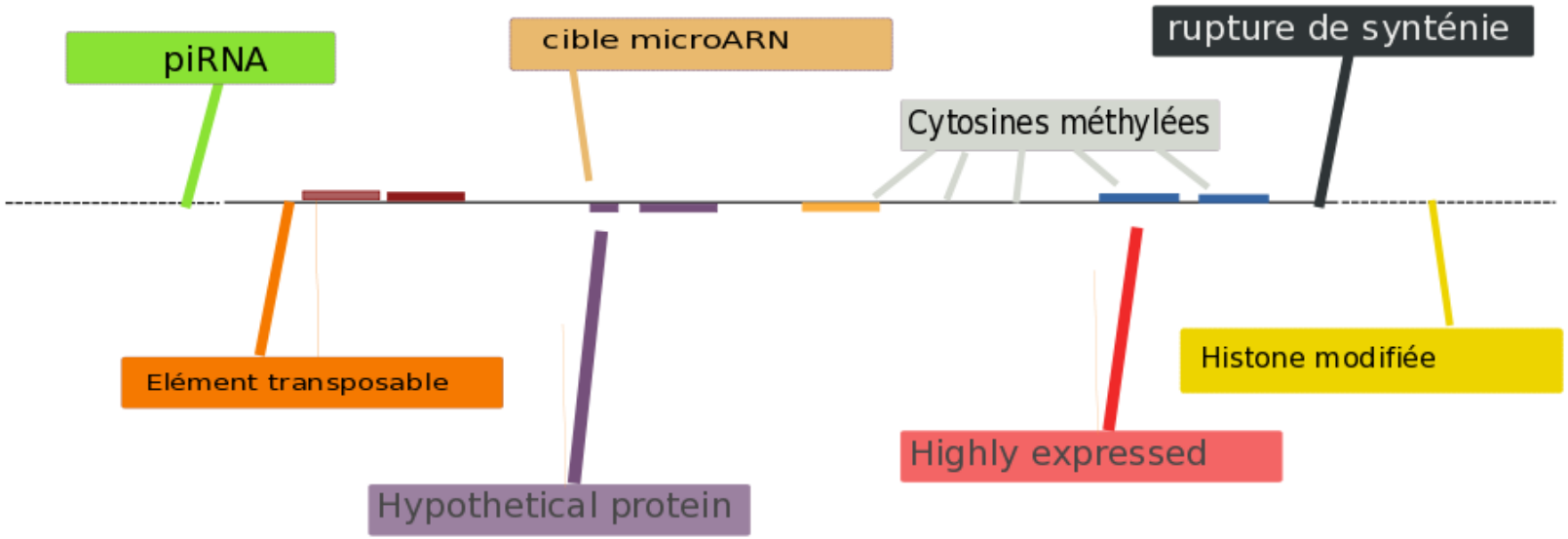
Methylome
Histone marks

Transcriptome
- microarray
- RNA-Seq
- alternative splicing

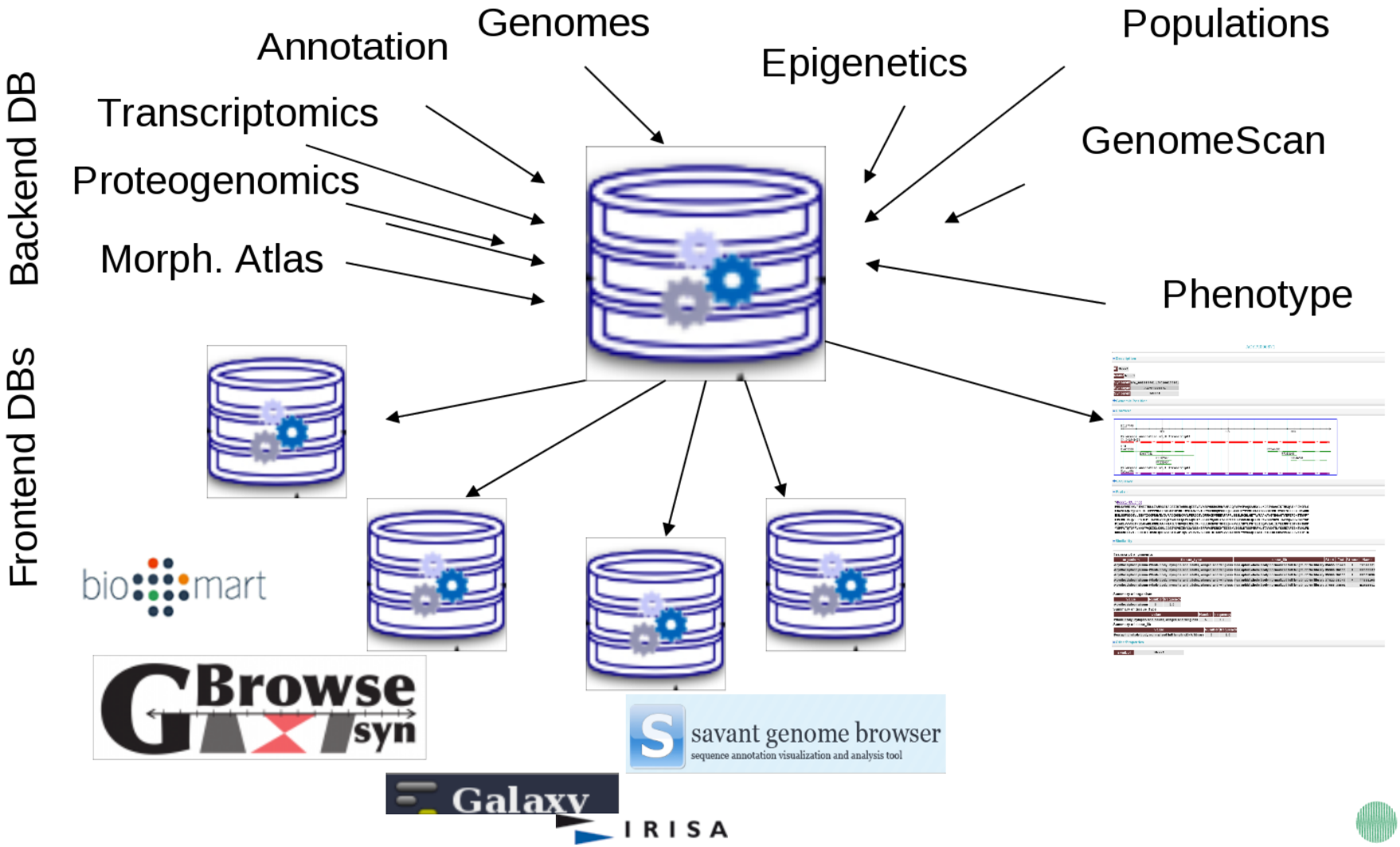
Small non coding RNAs
- miRNAs
- piRNA
- endosiRNAs

Proteomics

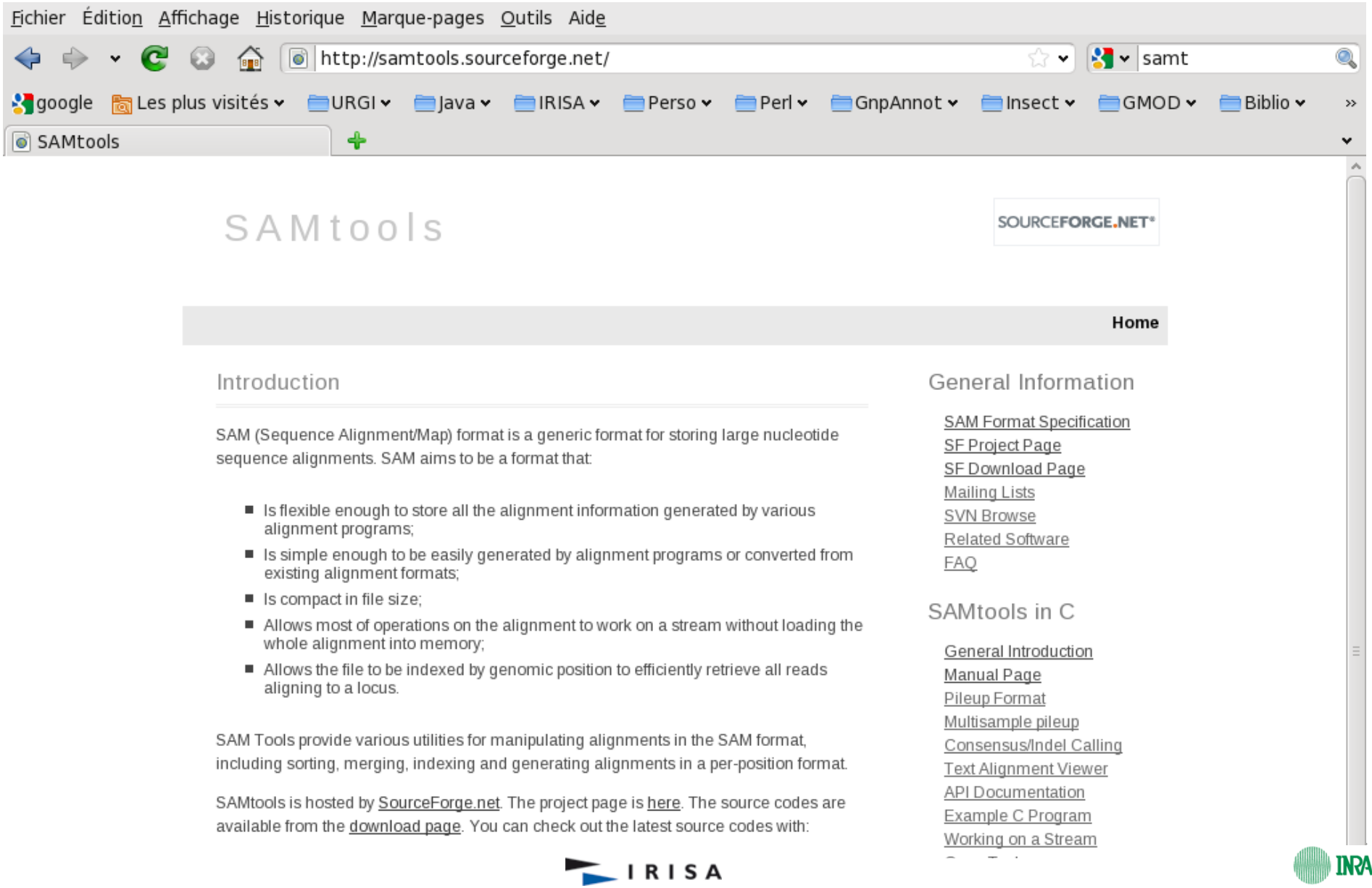
Extension et intégration de données



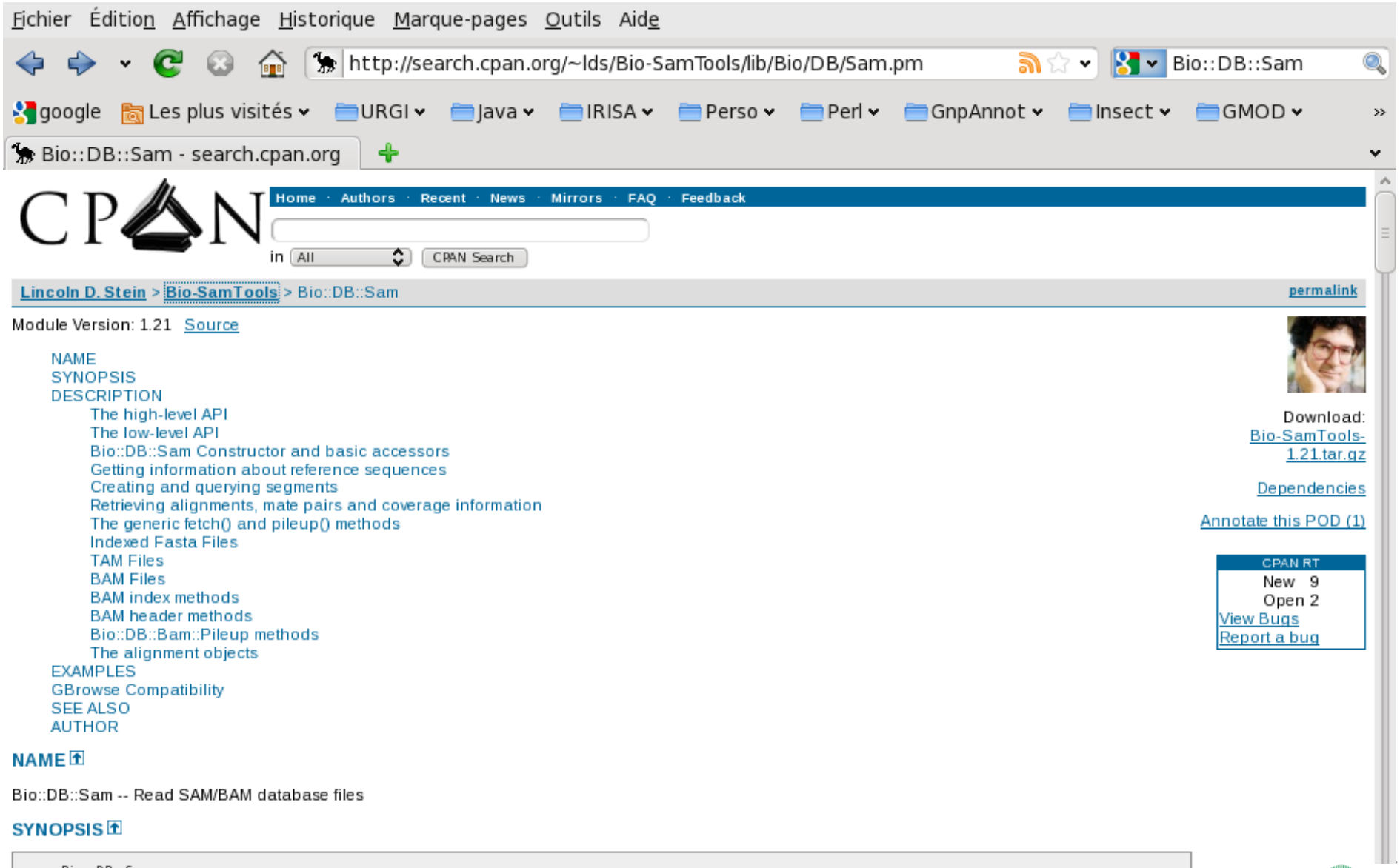
Extension et intégration de données



Extension et intégration de données



Extension et intégration de données



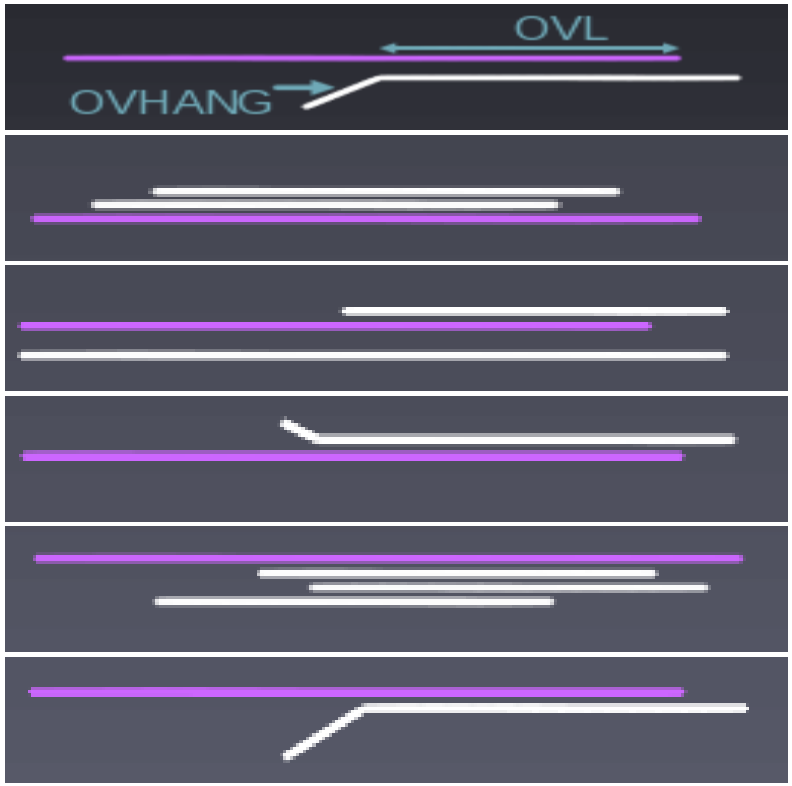
Autre collection de NGS

	Aphids	Spodoptera	
		frug.	litt.
RNA-Seq (454)		X	X
RNA-Seq (Illumina)	X	X	X
RNA-Seq (pair ends)	X		
ChIP-Seq	X	X	
Genome sequencing	X	X	
resequencing	X		

RNA-Seq 454 Assembly

TGICL++, a multi-step Transitive Clustering and Assembling

S. Carrere, J. Gouzy INRA, Toulouse



IRISA

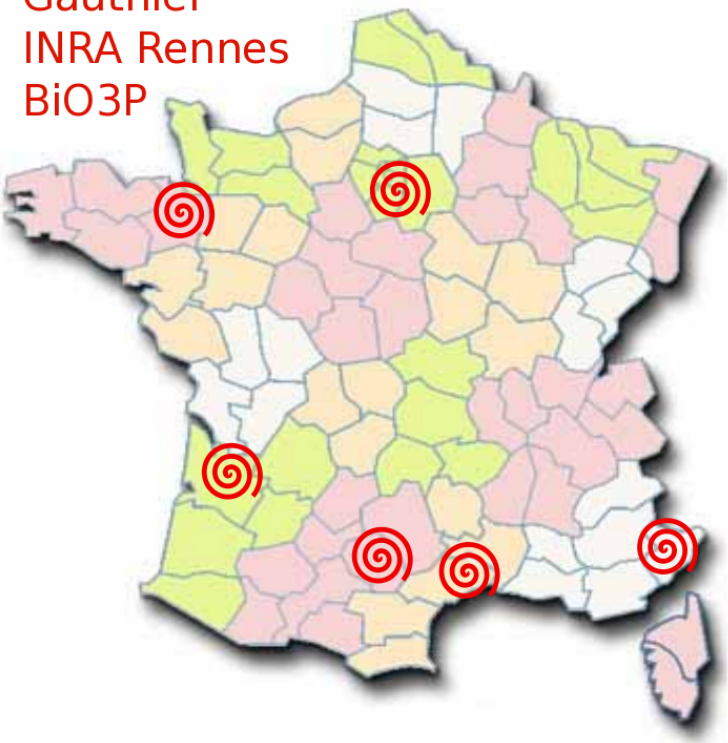
INRA

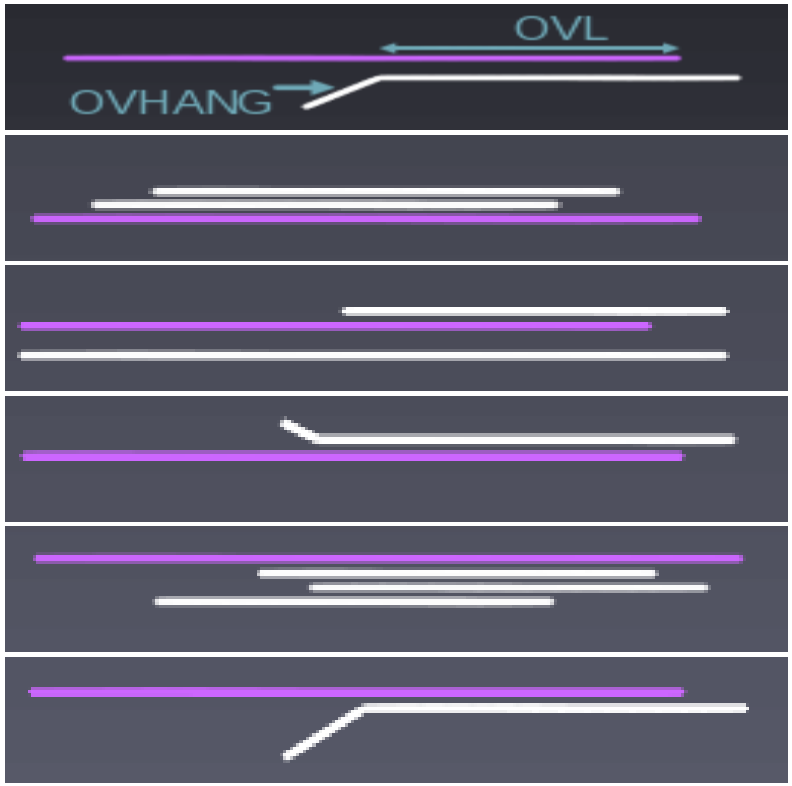
RNA-Seq 454 Assembly

TGICL++, a multi-step Transitive Clustering and Assembling

S. Carrere, J. Gouzy INRA, Toulouse

Jean-Pierre
Gauthier
INRA Rennes
BiO3P





IRISA

INRA

Acknowledgments

- Aurore Gallot, Marie Trap-Gentil, Stephanie Jaubert-Possamai, Nathalie Leterme, Julien Malaboeuf, Claude Rispe, Denis Tagu, Jean-Christophe Simon
- Owain Edwards, Karl Gordon and Tom Walsh
- Dominique Lavenier, Jacques Nicolas, Guillaume Rizk
- Emmanuelle Permal, Hadi Quesneville
- Olivier Collin, Aurelien Roult, Olivier Sallou,