

➤ Successes and pitfalls of the new genome annotation softwares

Few feedbacks on insect and plant genomes annotation

Maya Gawinoski, Audrey Chathuant, Nicolas Lapalu, David Gilardot, Frédérique Hilliou, Christophe Klopp, Christine Gaspin, and Fabrice Legeai



➤ Genome wide overview

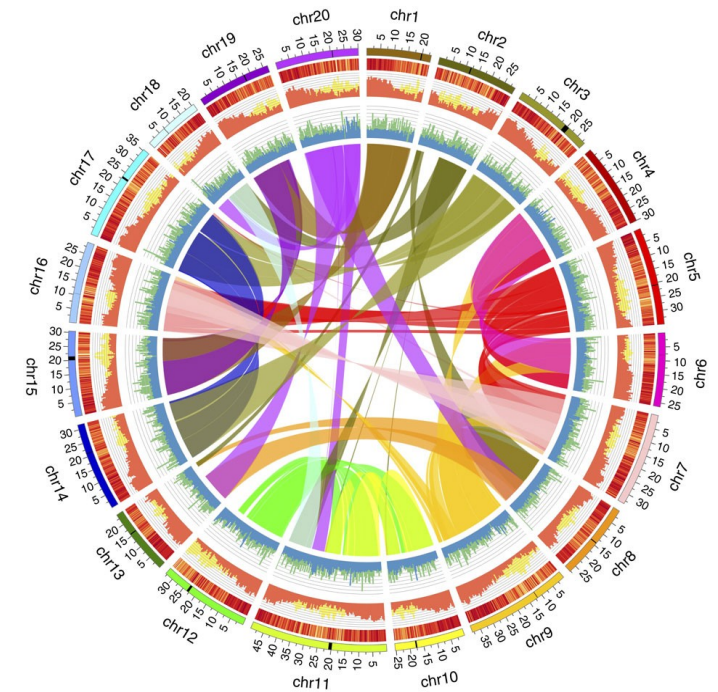
➤ Plant genome

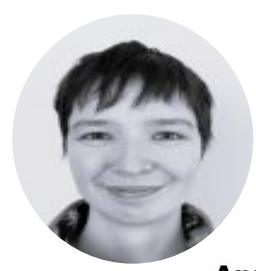
The *Camelina sativa* example



- An oilseed plant in the Brassicaceae family
- hexaploid
- 650Mbp, 20 chromosomes
- Reference annotation (RefSeq)
 - 82,569 protein coding genes
 - 107,481 mRNAs
 - 94,343 fully supported (87.8%)

Pipeline EGAP (NCBI 2016)

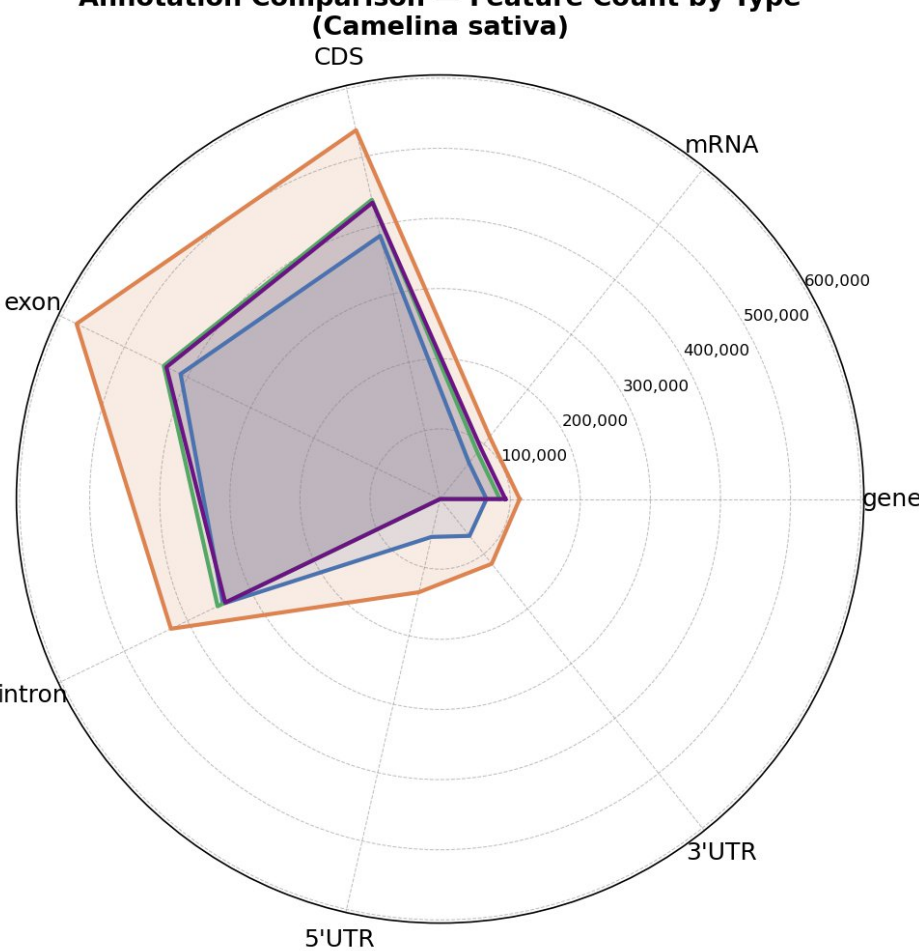




Annotation of the *Camelina sativa* genome

Are the number of genomic features similar ?

Annotation Comparison — Feature Count by Type
(*Camelina sativa*)



	EGAPx* v0.5.0	Helixer v0.3.6	Tiberius V2.0.5	Annevo V2.3.2
nb_genes	66,116 (-20%)	113,434 (+37%)	85,207 (+3%)	93,078 (+13%)
nb_transcripts	83,807	113,434	85,207	93,078
average_exons/transcript	6.9	5.1	5.1	4.7
nb_partial_CDS	4,573	2,635	893	1,367
average_CDS_length	1301	1052	1199	1,080
median_CDS_length	1007	849	1011	915
min_CDS_length	90	100	201	15
max_CDS_length	16,152	19,764	20,070	14,904
nb_monoexonic_genes	7,552 (11%)	22,717 (20%)	19,295 (23%)	21,608 (23%)

- with 4 RNASeq libraries including flower, leaves, stem and fruit.



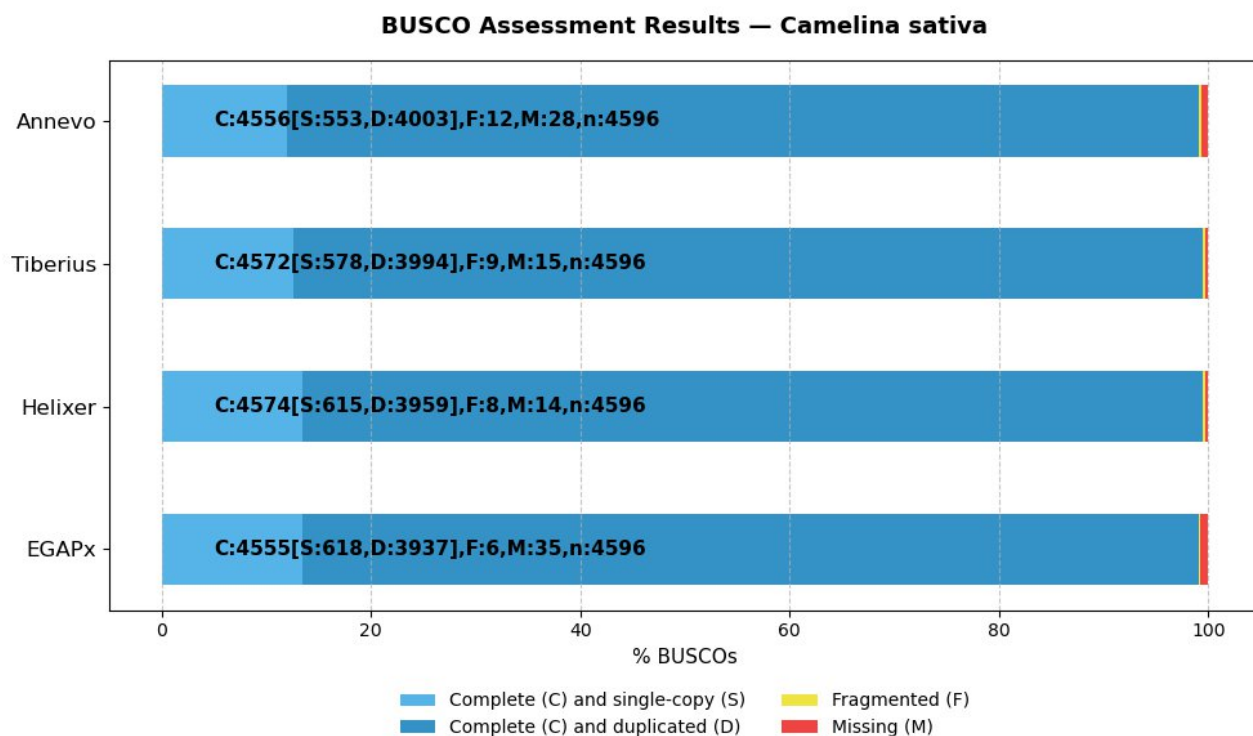
INRAE



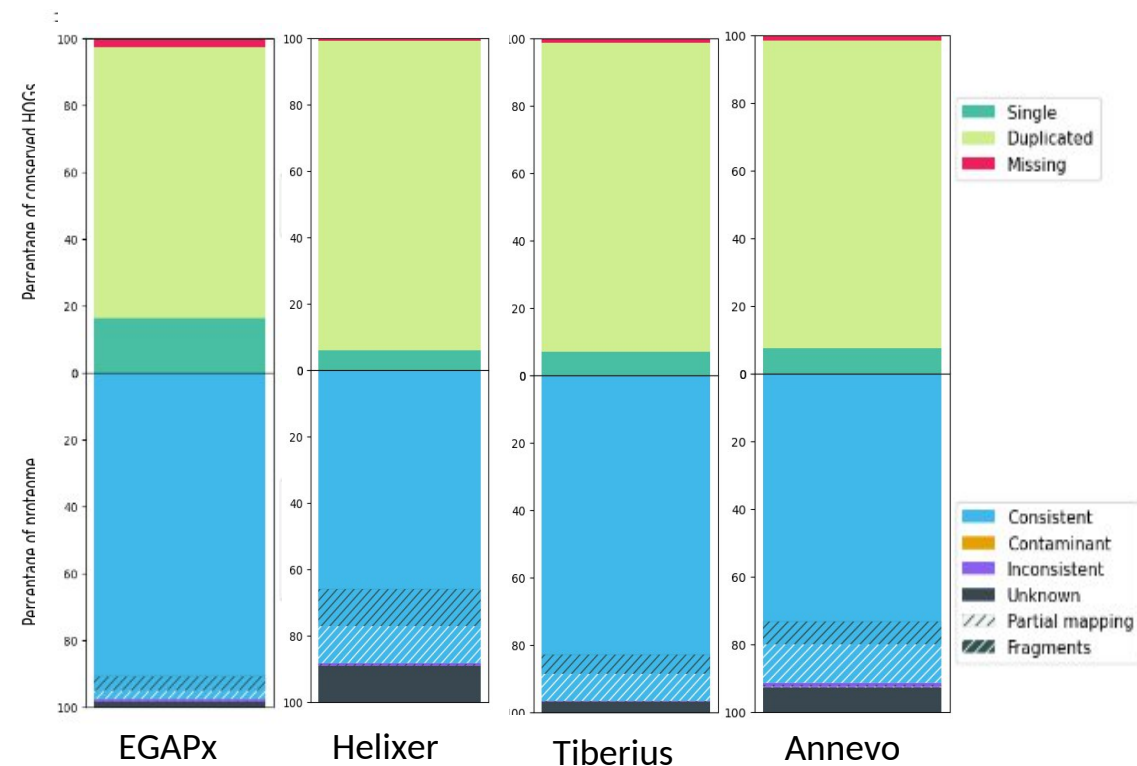
Annotation of the *Camelina sativa* genome

Completion of the annotations

BUSCO v6.0.0, with odb10_brassicales



OMark V0.4.1 / Viridiplantae HOG database





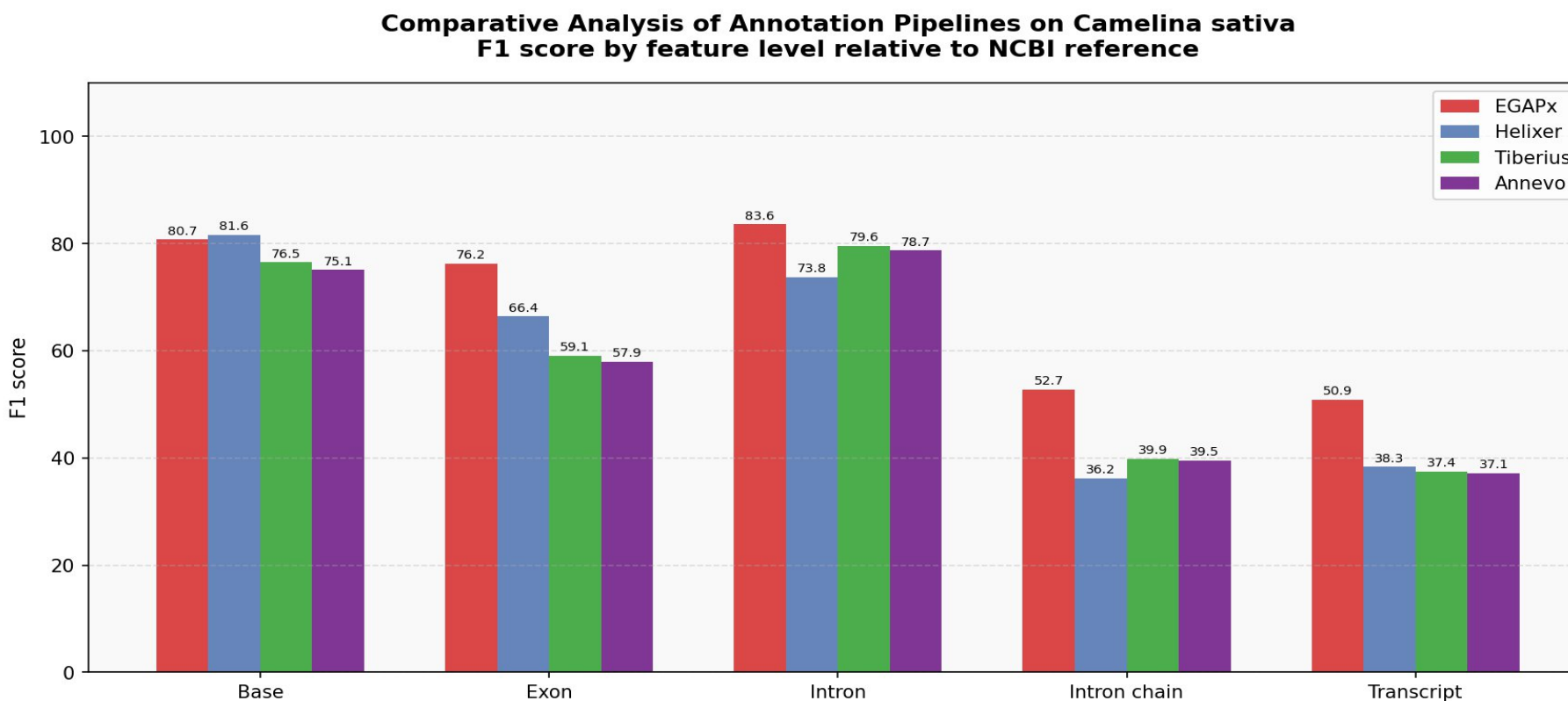
Annotation of the *Camelina sativa* genome

How similar are the predictions to the RefSeq annotation

Compared each annotation to a reference annotation with gffcompare

Applied F1 score formula to gffcompare results

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$



INRAE

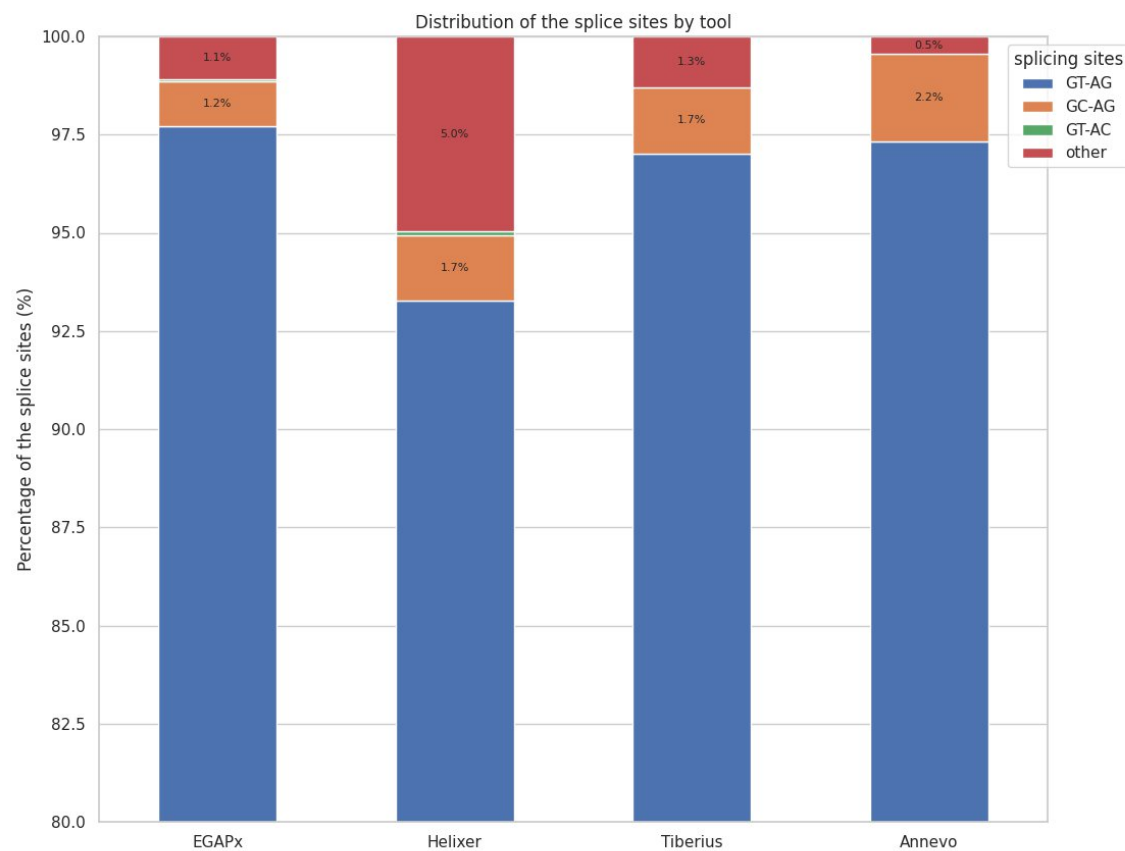
Successes and pitfalls of the new genome annotation softwares

JOBIM - 2 juillet 2026 - Chathuant, Legeai



Annotation of the *Camelina sativa* genome

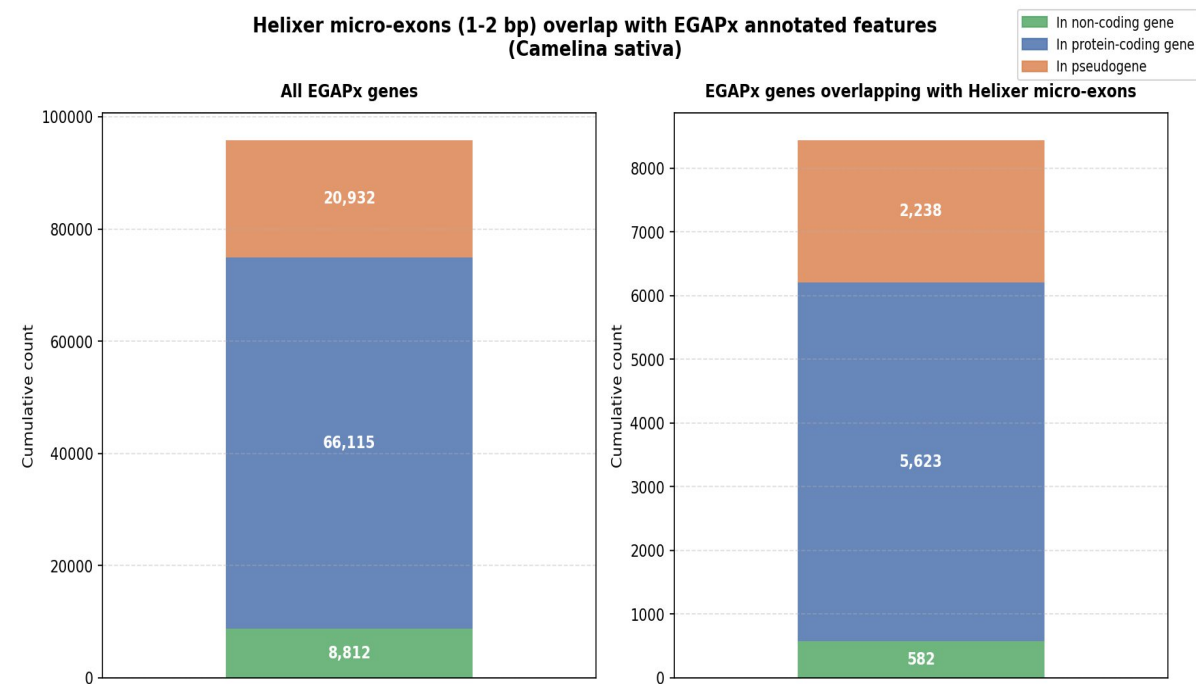
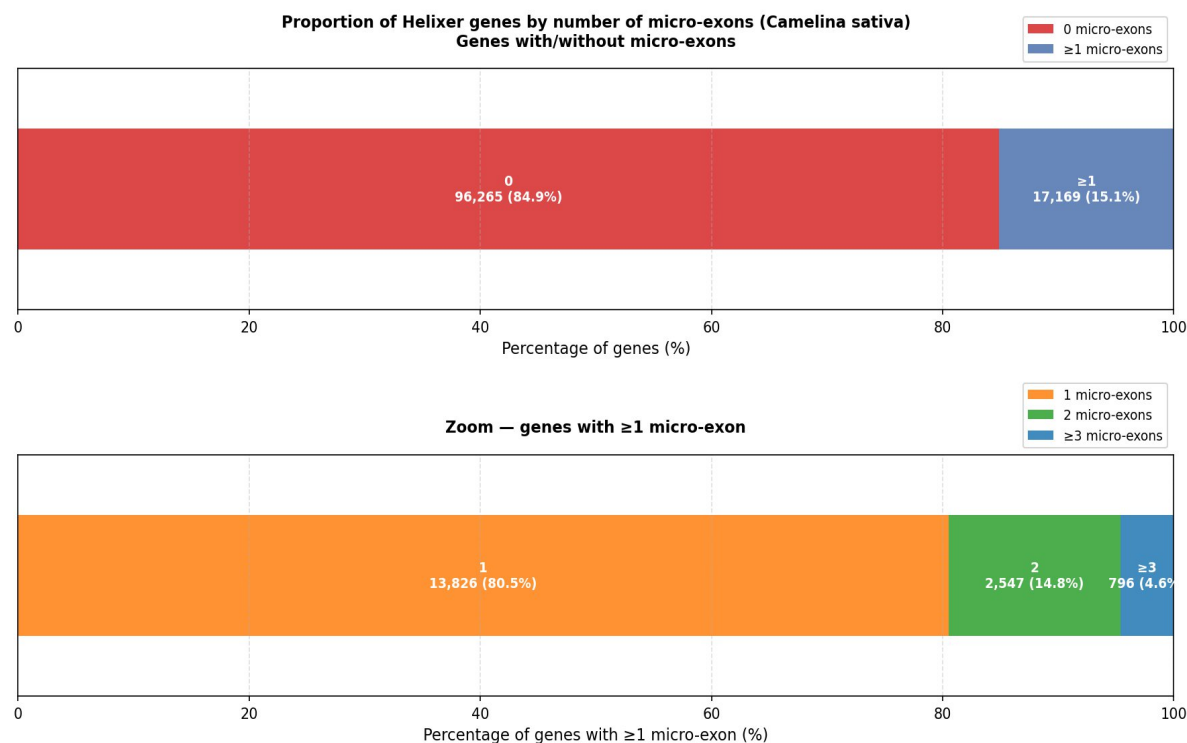
Splice-sites





Annotation of the *Camelina sativa* genome

Helixer's micro-exons (<3 bp)



INRAE

Successes and pitfalls of the new genome annotation softwares

JOBIM - 2 juillet 2026 - Chathuant, Legeai



Annotation of the *Camelina sativa* genome

What is the proportion of specific genes

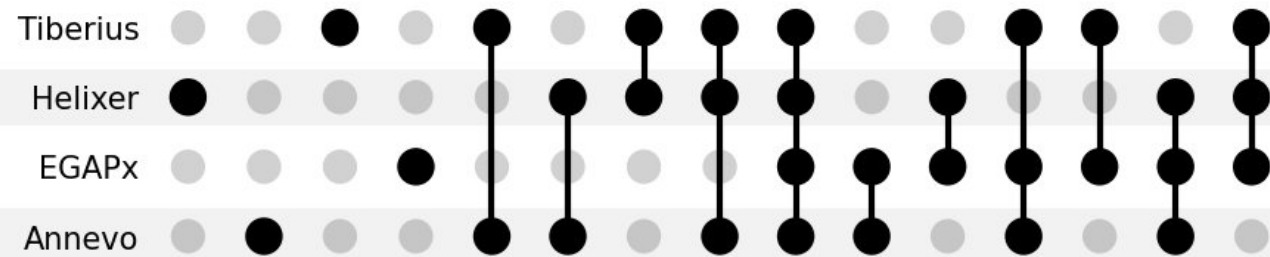
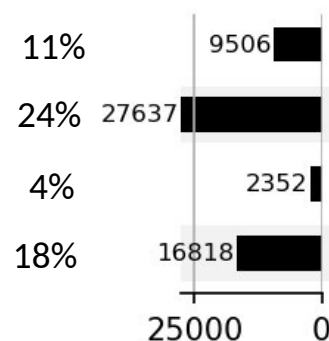
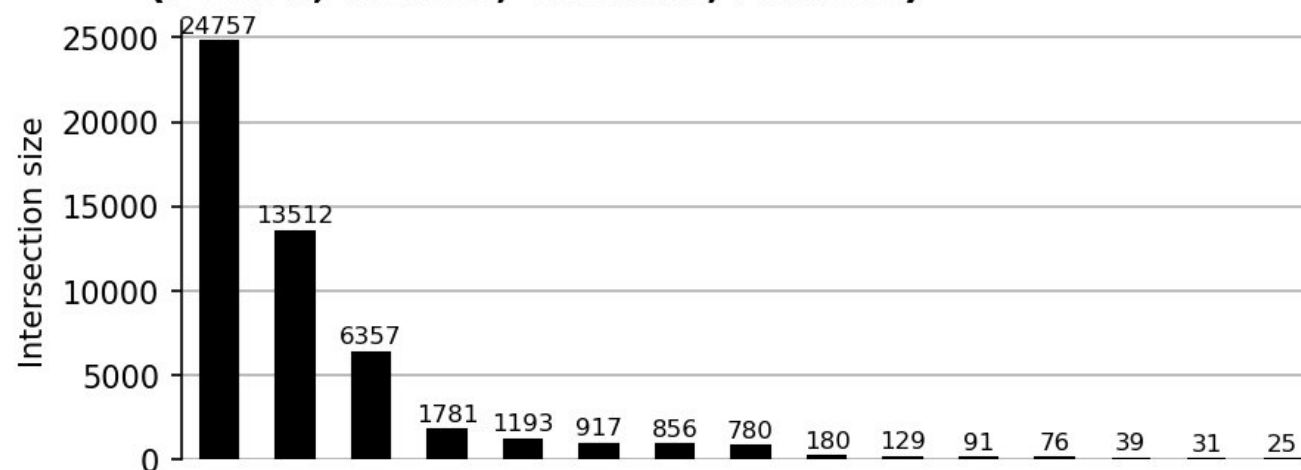
Ran gffcompare with the four annotations against the NCBI reference

Only kept the class code « u » for « unknown »

About EGAPx, only one isoform was kept for each gene using agat longest isoform tool

UTR were removed before running gffcompare

Novel transcript (code='u') - not in reference (EGAPx, Helixer, Tiberius, Annevo)



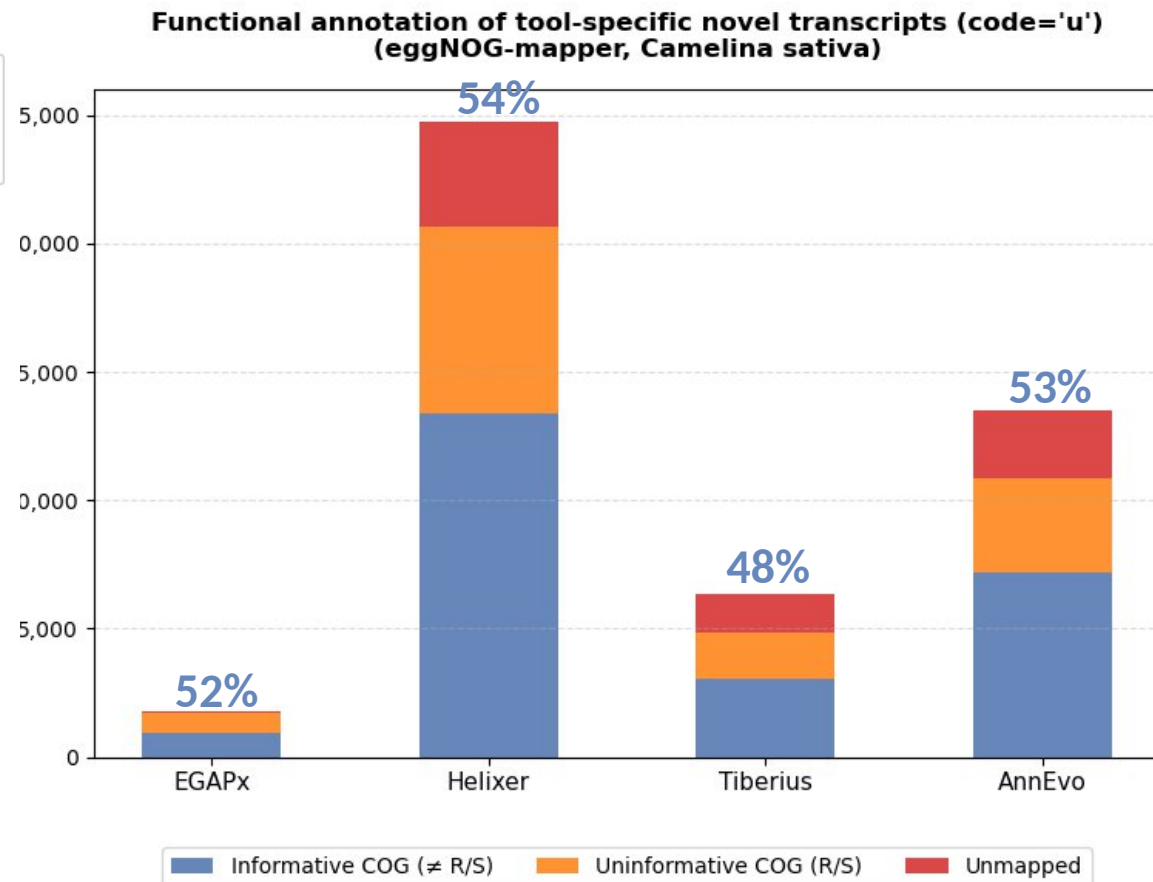
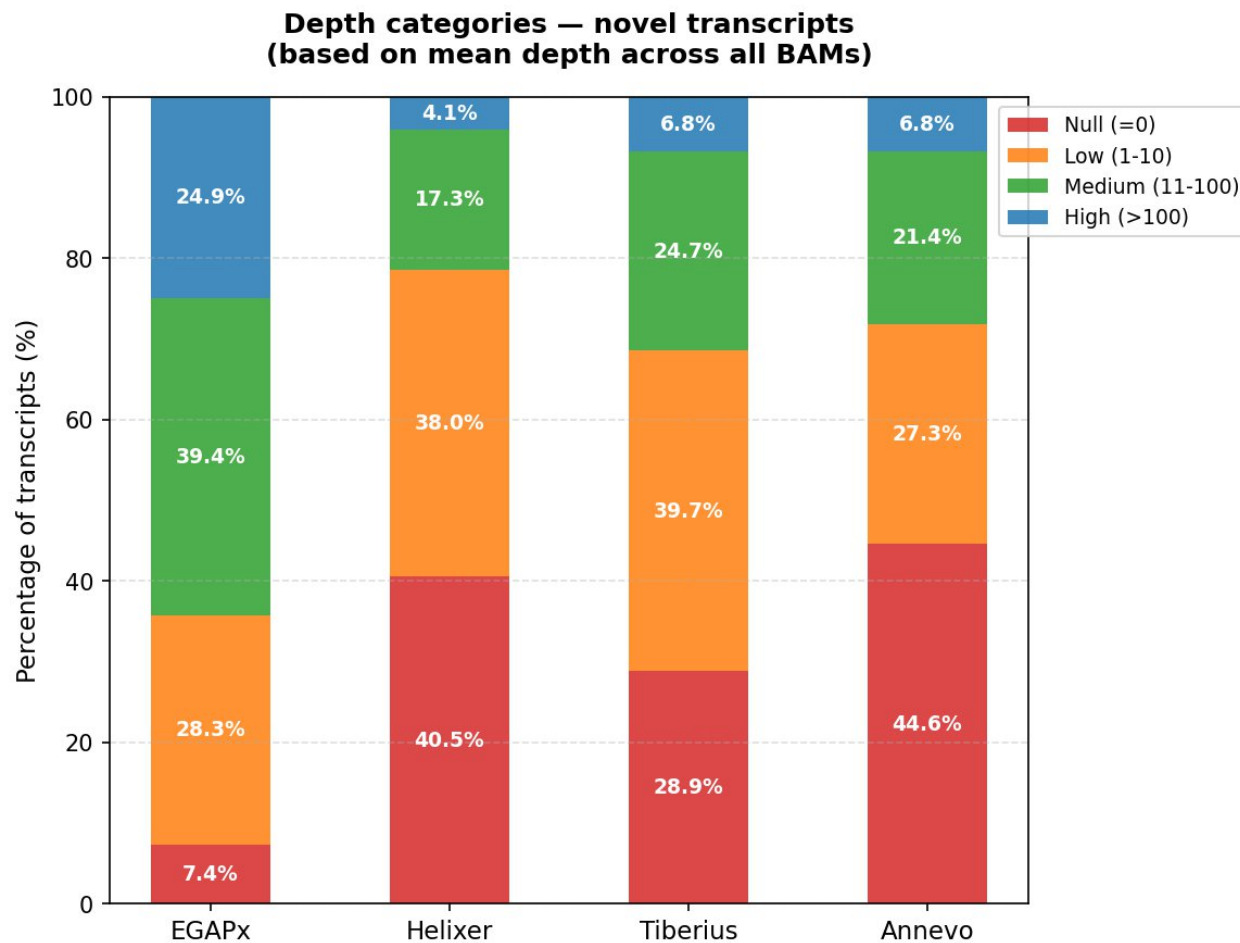
INRAE

Successes and pitfalls of the new genome annotation softwares
JOBIM - 2 juillet 2026 - Chathuant, Legeai



Annotation of the *Camelina sativa* genome

Are the novel transcripts biologically relevant ?

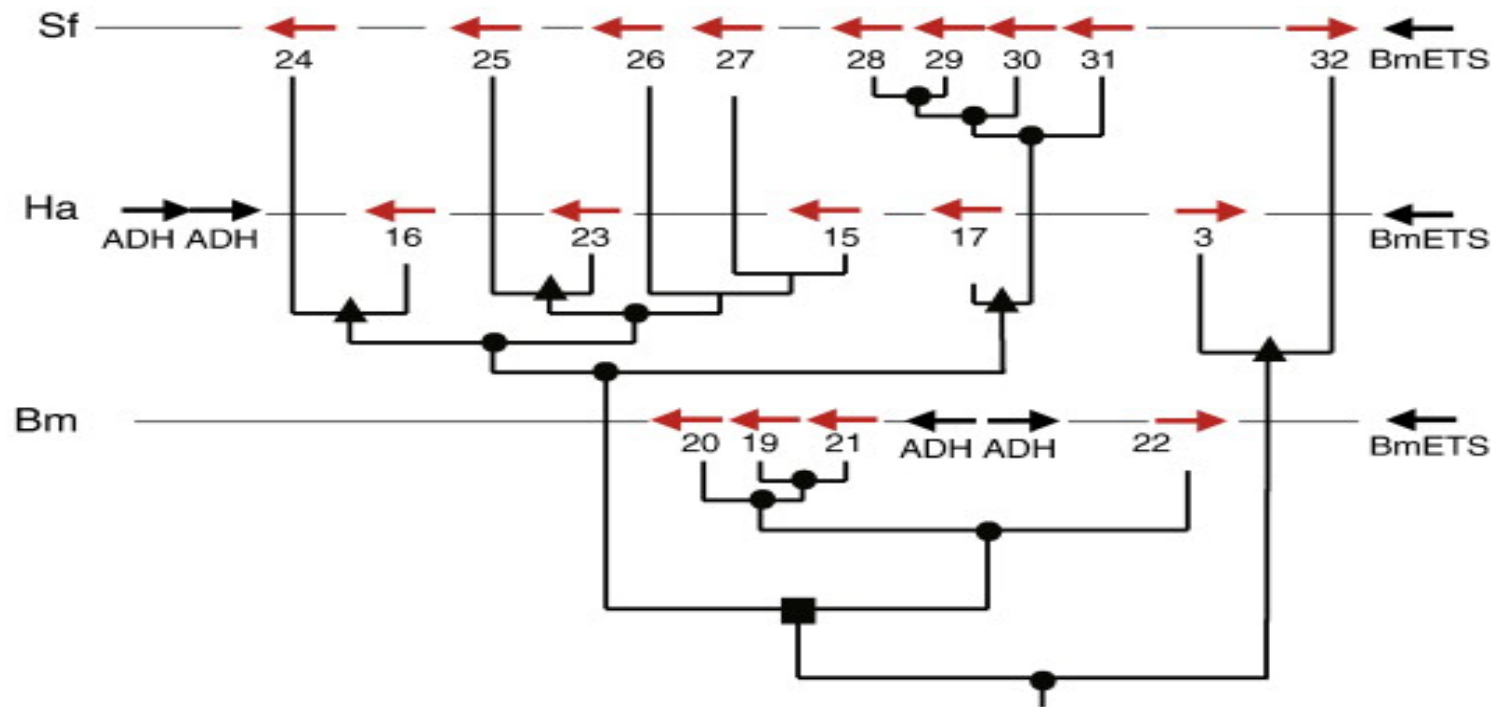




➤ Comparison with expertised gene structures models

➤ Insect Cytochrome P450

Insect's Cytochrome P450 are involved in the detoxification of plant defense chemistry, insecticides and environmental pollutants. They are evolving rapidly, but are very similar depending of their classification, clustered in the genome and repeated tandemly.



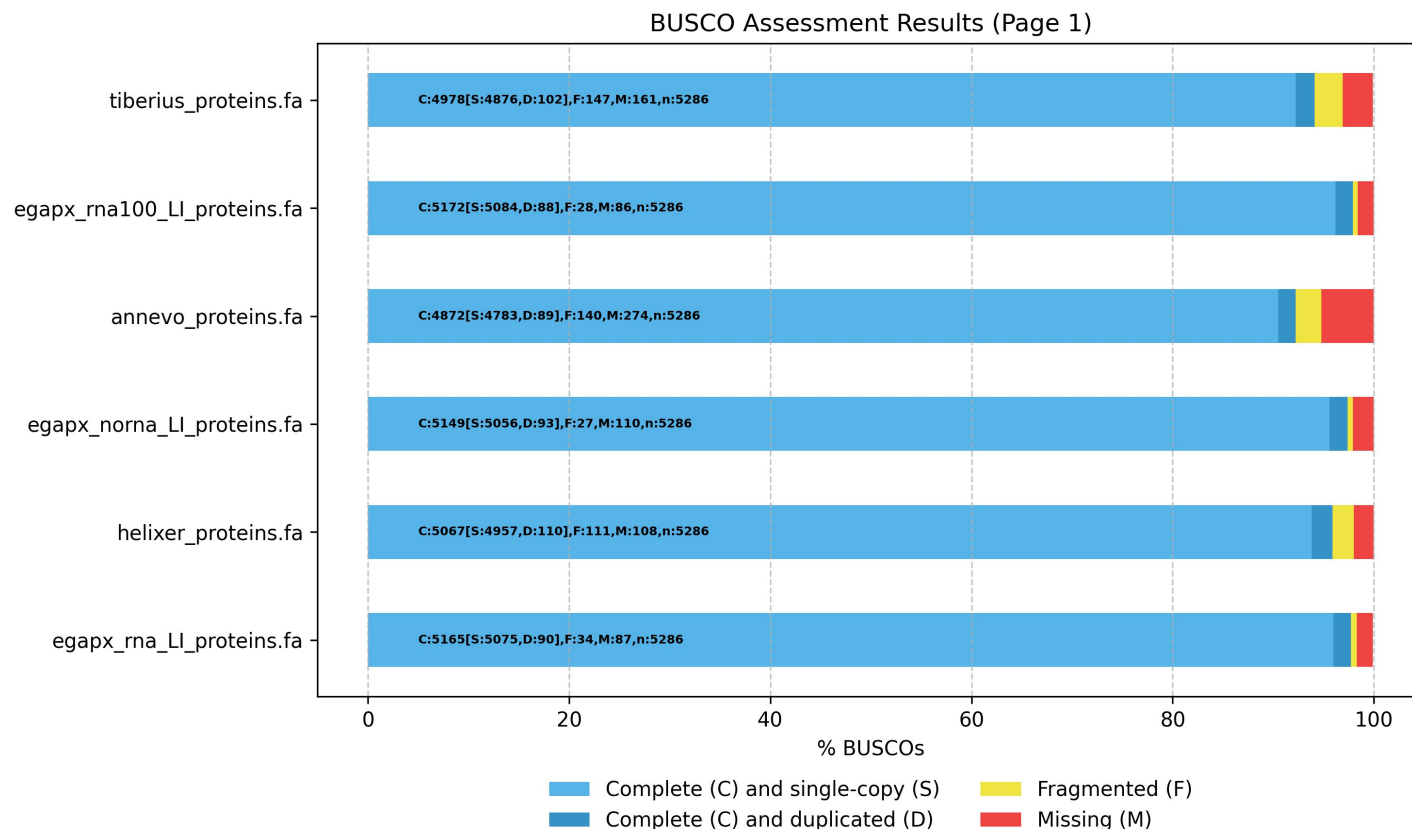
Feyereisen BBA, 2011

➤ Insect Cytochrome P450

Annotation of the *Spodoptera frugiperda* genome

Spodoptera frugiperda v7 genome (385Mb, 31 chrs)

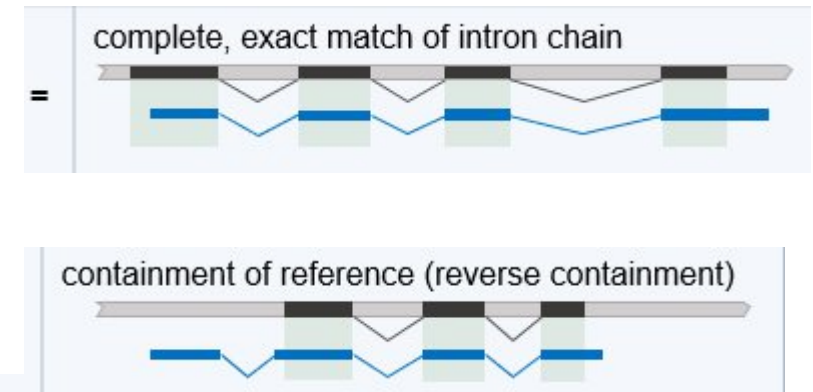
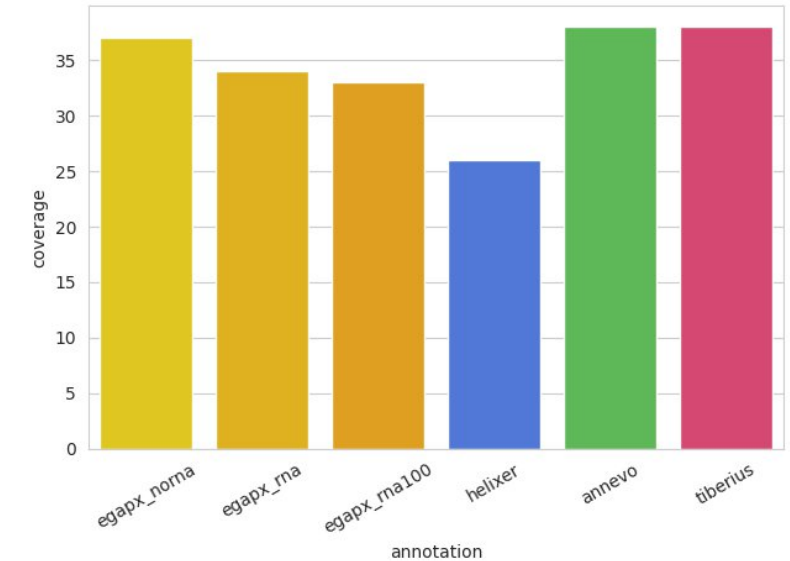
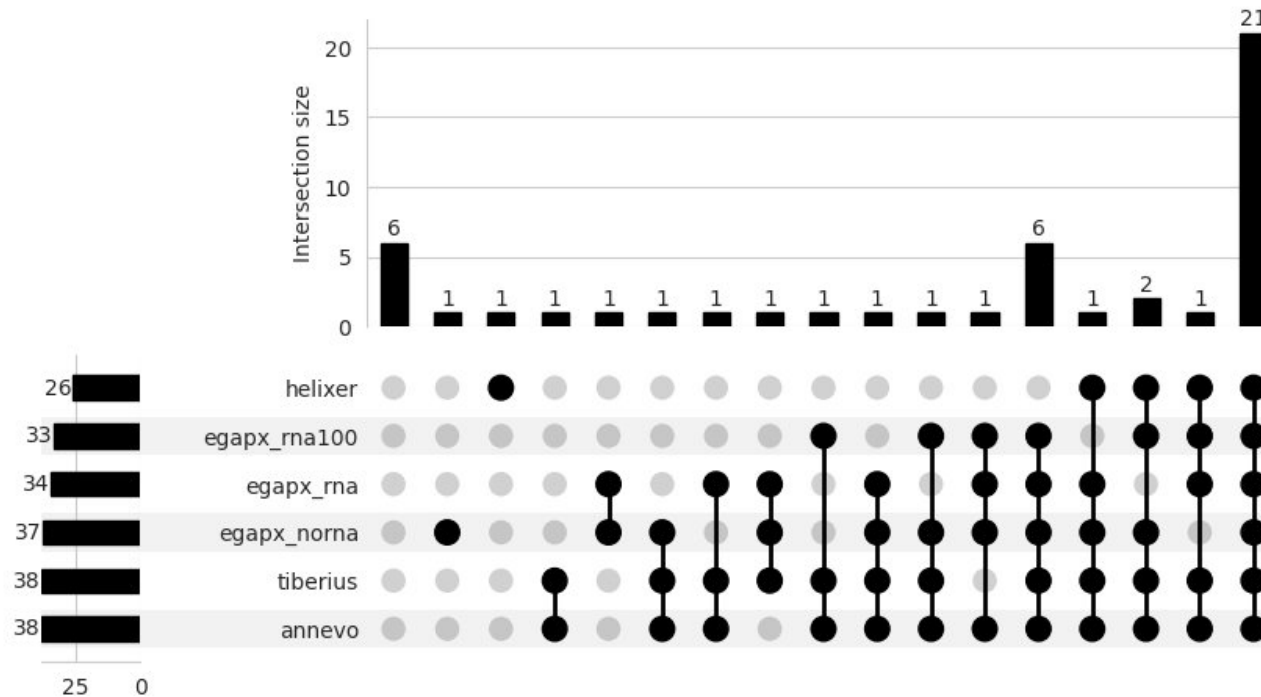
6 annotations (EGAPx without RNASeq, EGAPx + 20 RNASeq libs, EGAPx + 100 RNASeq libs)



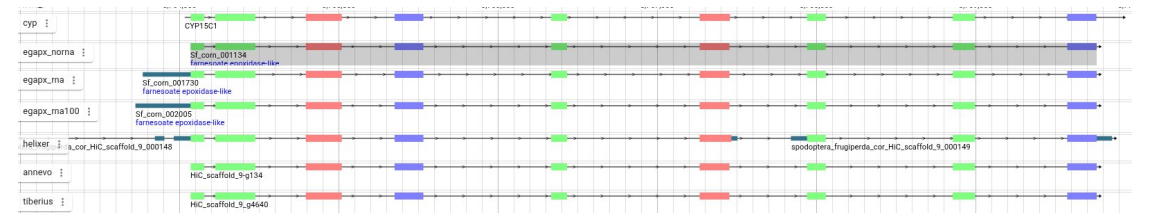
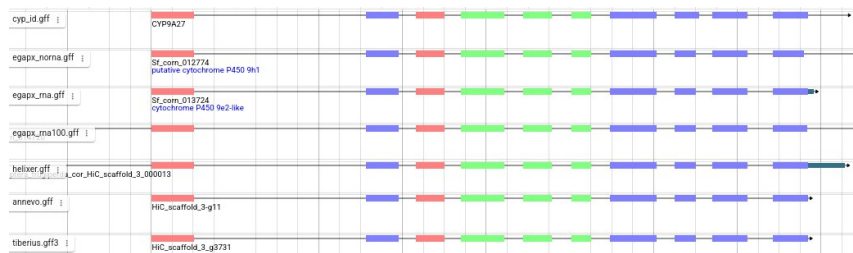
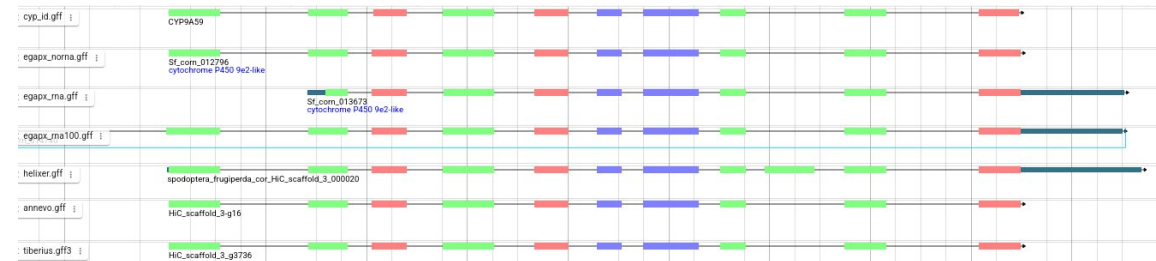
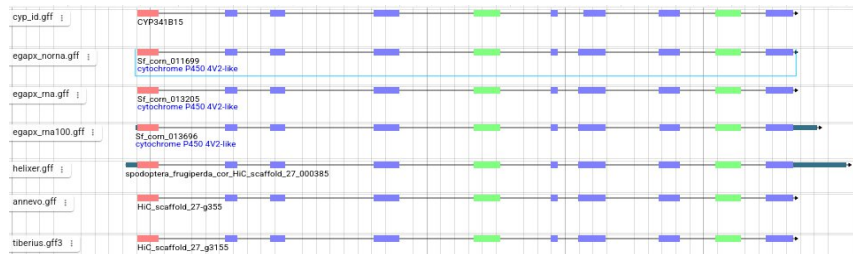
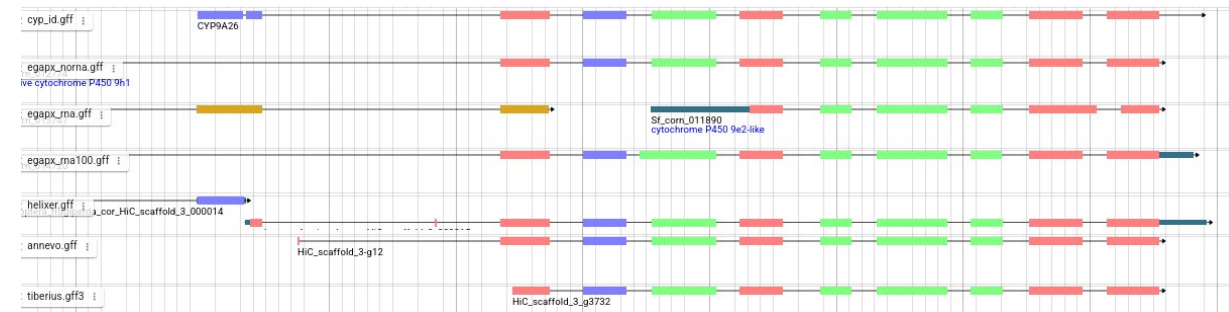
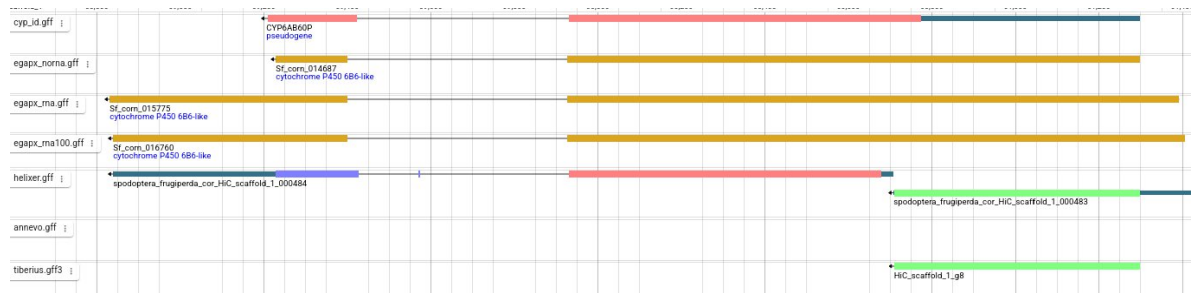
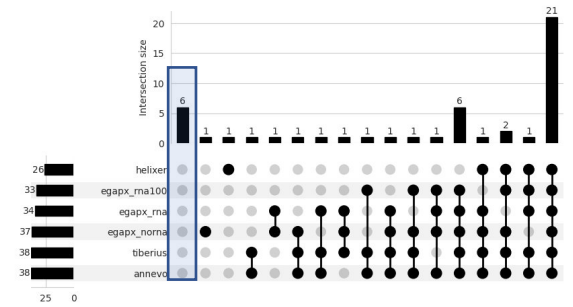
	# genes	# transcripts
EGAPx no RNASeq	14,547	15,698
EGAPx 20 RNASeq libs	15,593	18,457
EGAPx 100 RNASeq lib	16,126	20,642
Helixer	17,465	17,465
Tiberius	17,282	17,282
Annevo	16,224	16,224

➤ Insect Cytochrome P450

48 CYP identified and curated by expert



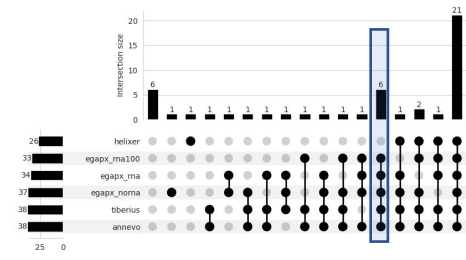
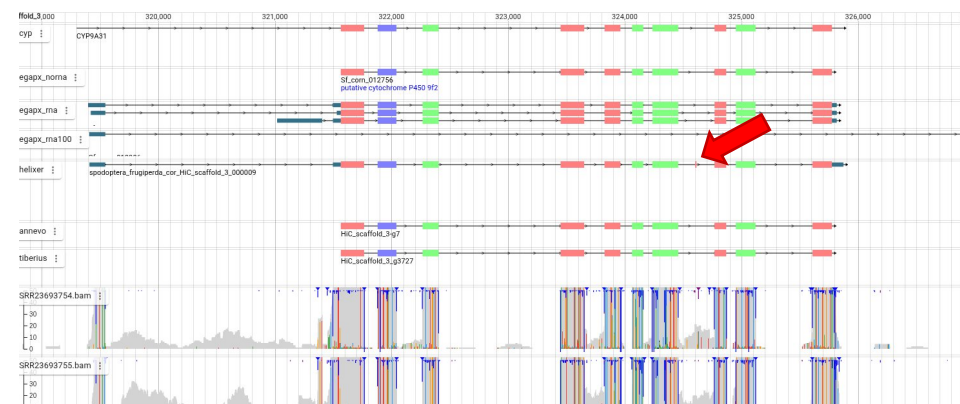
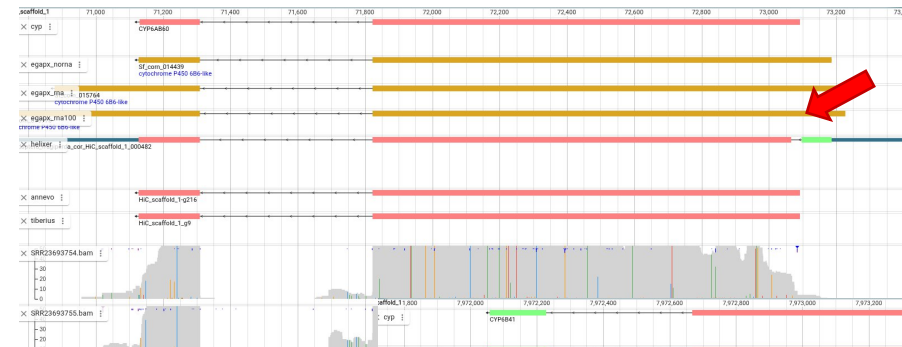
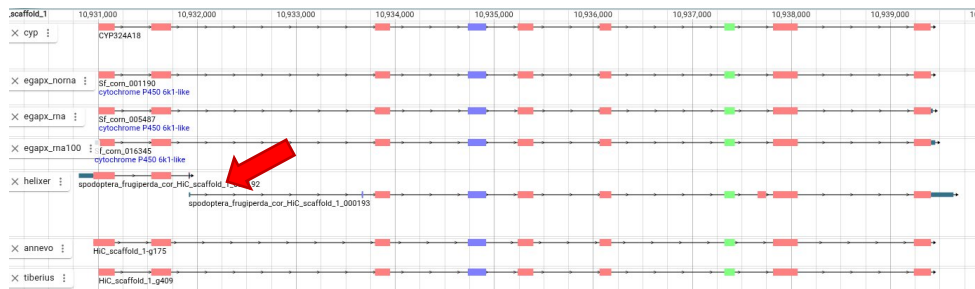
➤ Insect Cytochrome P450

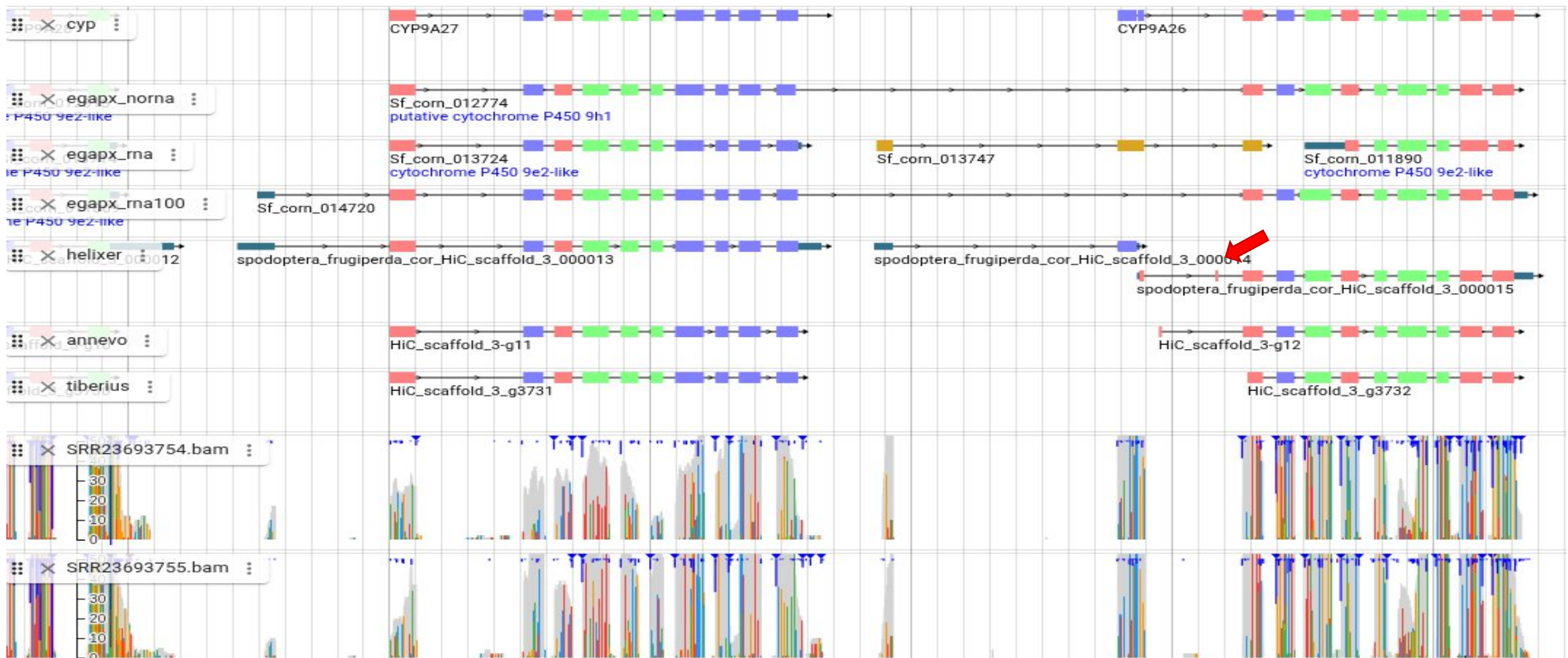


Successes and pitfalls of the new genome annotation softwares

JOBIM - 2 juillet 2026 - Chathuant, Legeai

➤ Insect Cytochrome P450





➤ Take home messages

- New ab-initio tools based on deep learning are much better than previous ab-initio predictor tools (e.g. Augustus), they run also much faster (usually less than 1 hour)
- EGAPx annotation is different from RefSeq, and the annotation quality is sensible to the given RNAseq libraries
- Helixer predicts much more genes than the other predictors, including thousands of genes with no biological support
- Helixer predicts numerous short and aberrant exons
- Annevo and Tiberius annotation outperform Helixer ones, and gave comparable results, with a slight advantage for Tiberius
- For many genes differences between the gene structures lie in the details (choice of the splicing site).

For challenging cases, manual curation is still needed or it is worth to use a combiner to reconcile annotation from the different tools

Can fine-tuning Helixer improve annotation?

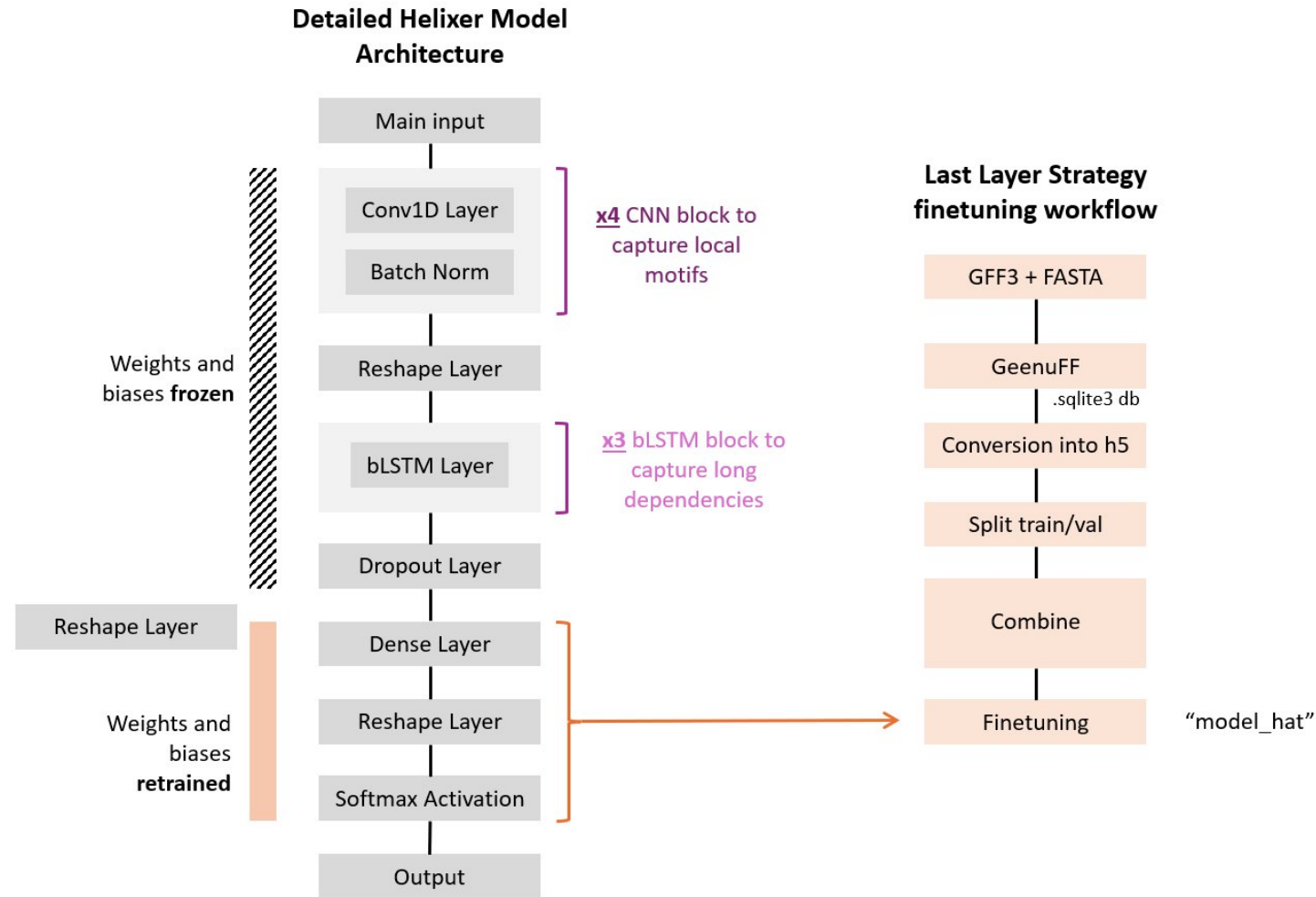




➤ Tuning the models

➤ Finetuning Helixer : Last Layer Strategy

Can we improve the results of Helixer by finetuning the model ?



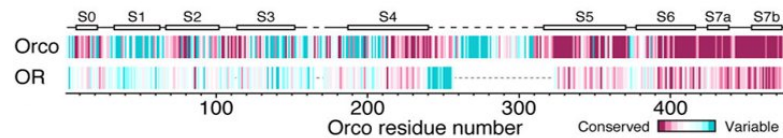
➤ Finetuning Helixer : Last Layer Strategy

To improve insect odorant receptor (OR) genes annotation

The case of insect odorant receptors (ORs)

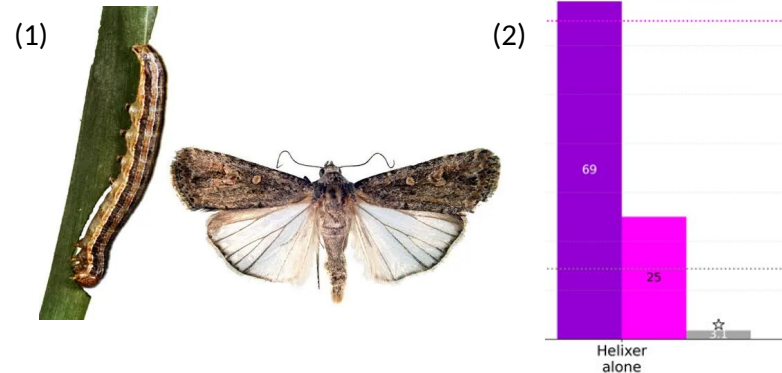
Challenging family :

- Highly divergent protein sequences especially the beginning



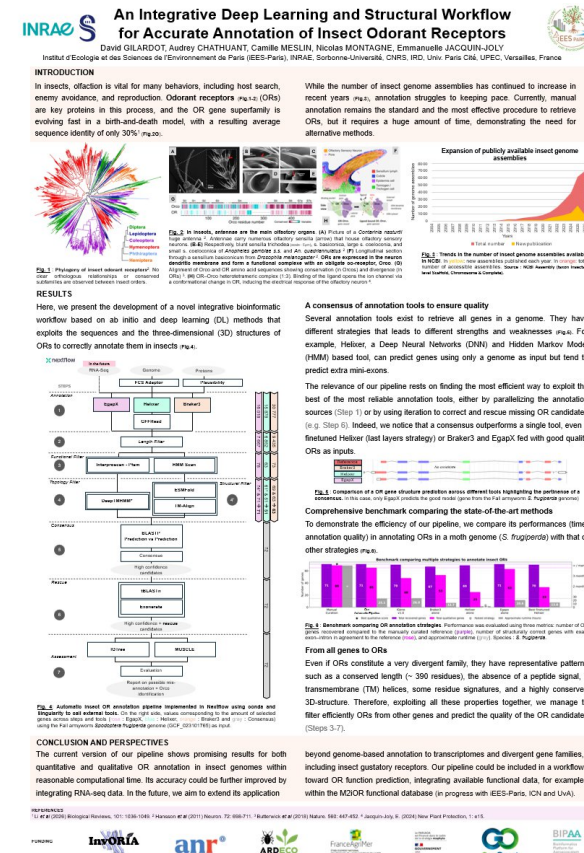
Butterwick et al (2018)

- Organization in cluster
- Multiple short exons / Long introns



Default Helixer on *S.frugiperda* ⁽²⁾ ORs : Assessment with **splice sites coordinates comparison** and visual check ⁽¹⁾

→ 25/71 correctly annotated *S.frugiperda* ORs



Poster B 79



INRAE

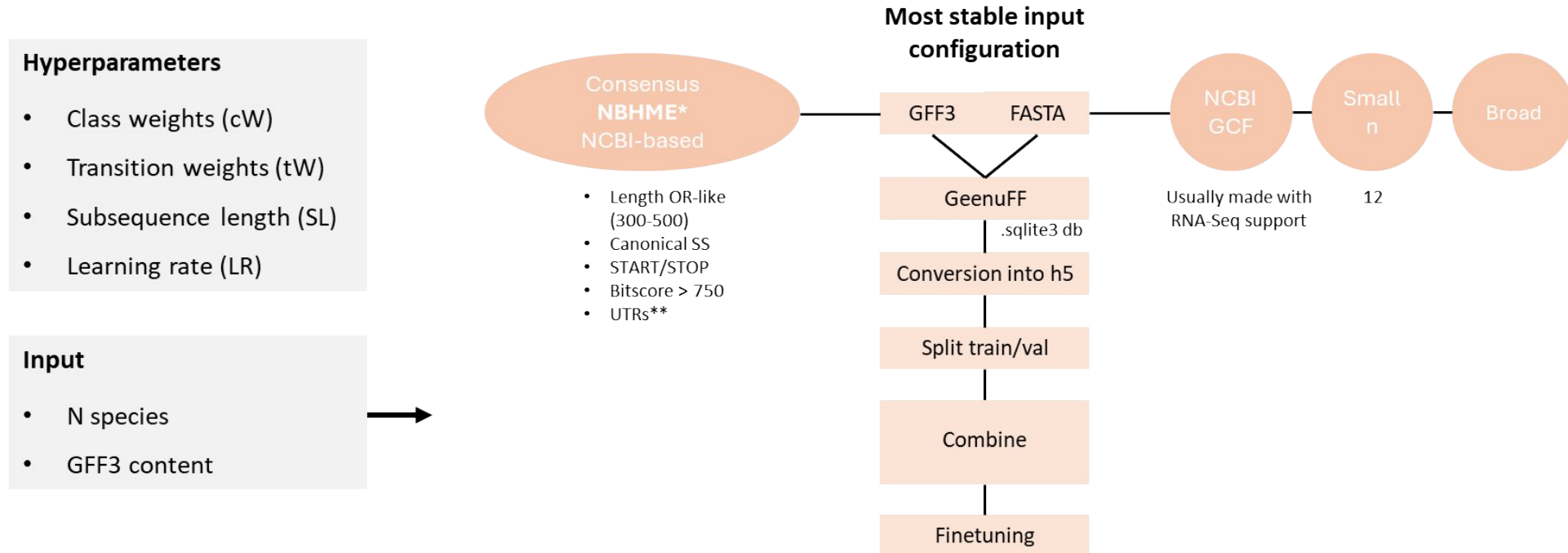
Successes and pitfalls of the new genome annotation softwares

JOBIM - 2 juillet 2026 - Chathuant, Legeai

➤ Finetuning Helixer : Last Layer Strategy

What is mutable during finetuning ?

The case of insect odorant receptors (ORs)



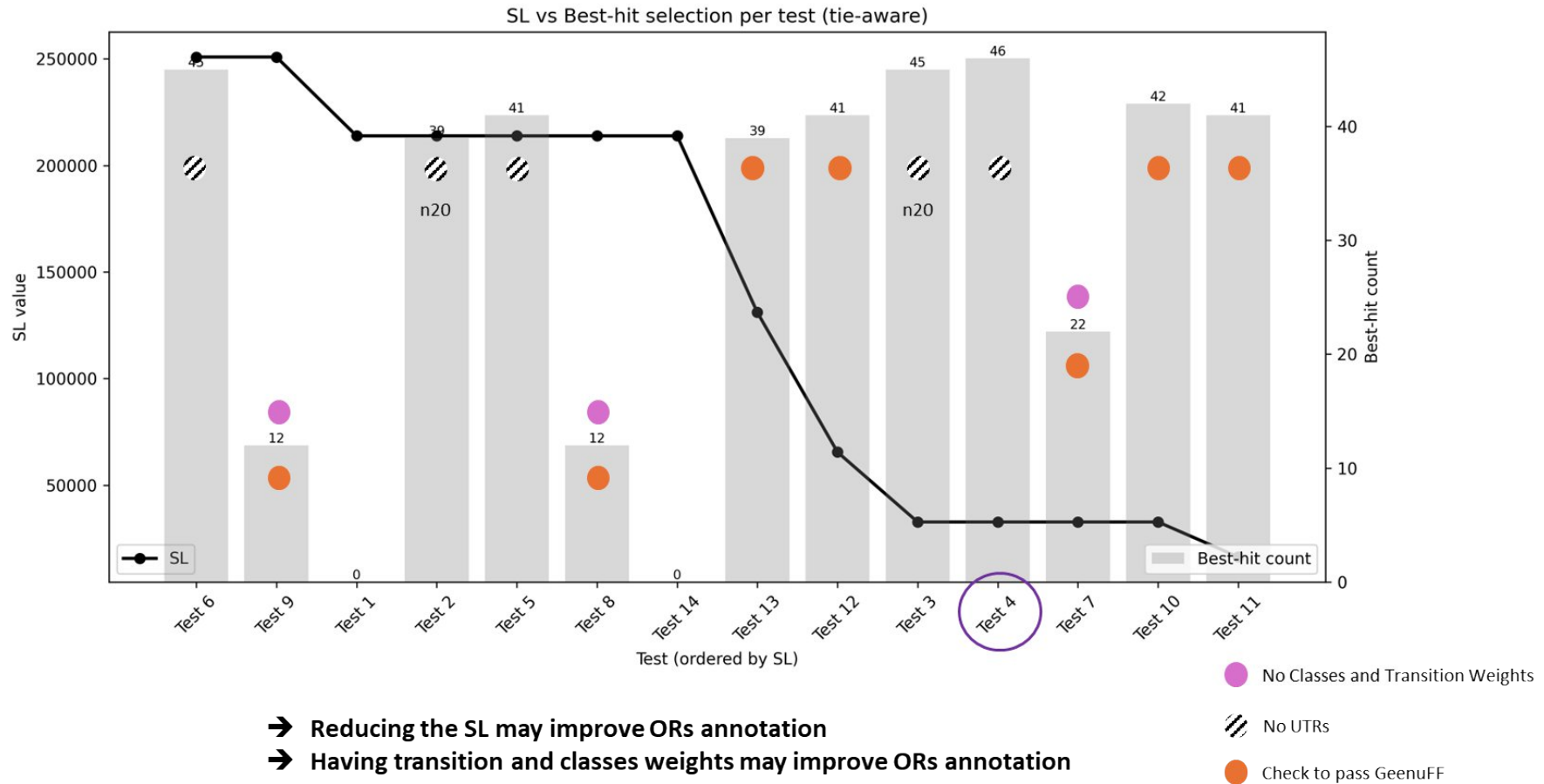
*NBHME : Respectively : NCBI, Braker3, Helixer, Miniprot and Exonerate

**Patch same-size UTRs from NCBI replaced genes to alternatives or made them up

➤ Finetuning Helixer : Last Layer Strategy

Results: Which hyperparameters configuration is the best for now ?

The case of insect odorant receptors (ORs)

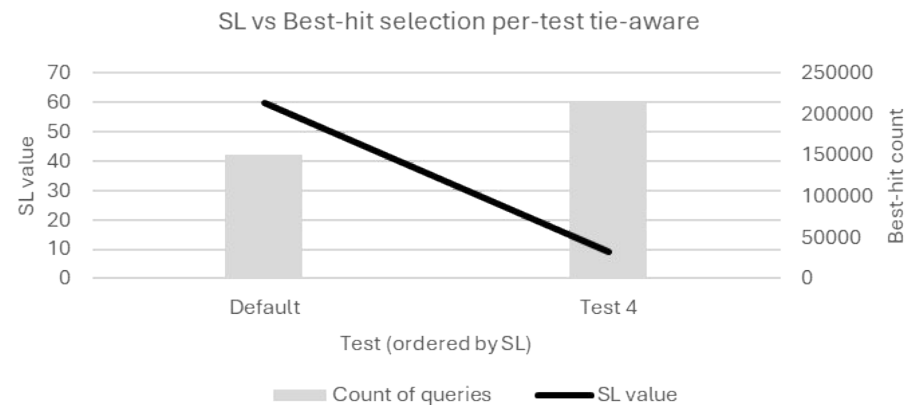


➤ Finetuning Helixer : Last Layer Strategy

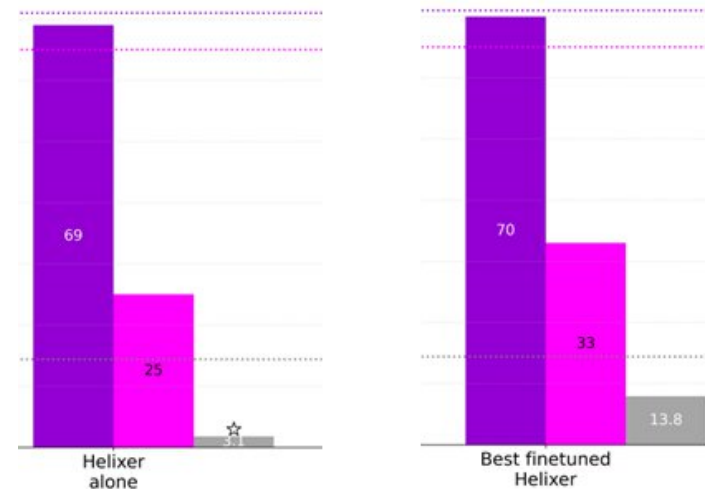
Results: Is finetuning really an added value comparing to the default ?

The case of insect odorant receptors (ORs)

Assessment by bitscore against known reference :
On a same query which one has the best prediction ?



Assessment with splice sites coordinates comparison
and visual check



About quality (ignoring UTRs)

→ ~ +8 ORs

About retrieval

→ +1 OR

→ Finetuned Helixer tend to be better than default Helixer

➤ Take home messages

- New ab-initio tools based on deep learning are much better than previous ab-initio predictor tools (e.g. Augustus), they run also much faster (less than 1 hour for a genome)
- EGAPx annotation is different from RefSeq, and the annotation quality is sensible to the given RNAseq libraries
- Helixer predicts much more genes than the other predictors, including thousands of genes with no biological support
- Helixer predicts numerous short and incorrect exons
- Annevo and Tiberius annotation outperform Helixer ones, and gave comparable results, with a slight advantage for Tiberius
- For many genes differences between the gene structures lie in the details (choice of the splicing site).

For challenging cases, manual curation is still needed or it is worth to use a combiner to reconcile annotation from the different tools

- **Fine-tuning Helixer is complex and does not significantly improve the results**
- **It must possible to tune or train specifically Tiberius model, but the process is complex and lacks documentation and requires lot of data and resources**
- **Since its version 2.0.5, the Tiberius pipeline is able to take into account transcriptomics data to improve the transcripts (new isoforms and UTR).**



➤ Acknowledgment

Maya Gawinowski

Audrey Chathuant

Matéo Boudet

David Gilardot

Frédérique Hilliou

Christophe Klopp

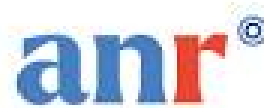
Christine Gaspin

Nicolas Lapalu et le CATI BARIC

Funding : PEPR BreIF, PARSADA ARDECO



BARIC



INRAE

Successes and pitfalls of the new genome annotation softwares

JOBIM - 2 juillet 2026 - Chathuant, Legeai