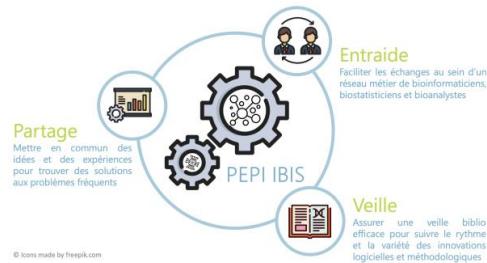


Arabidopsis genome annotations, differences between official and Helixer

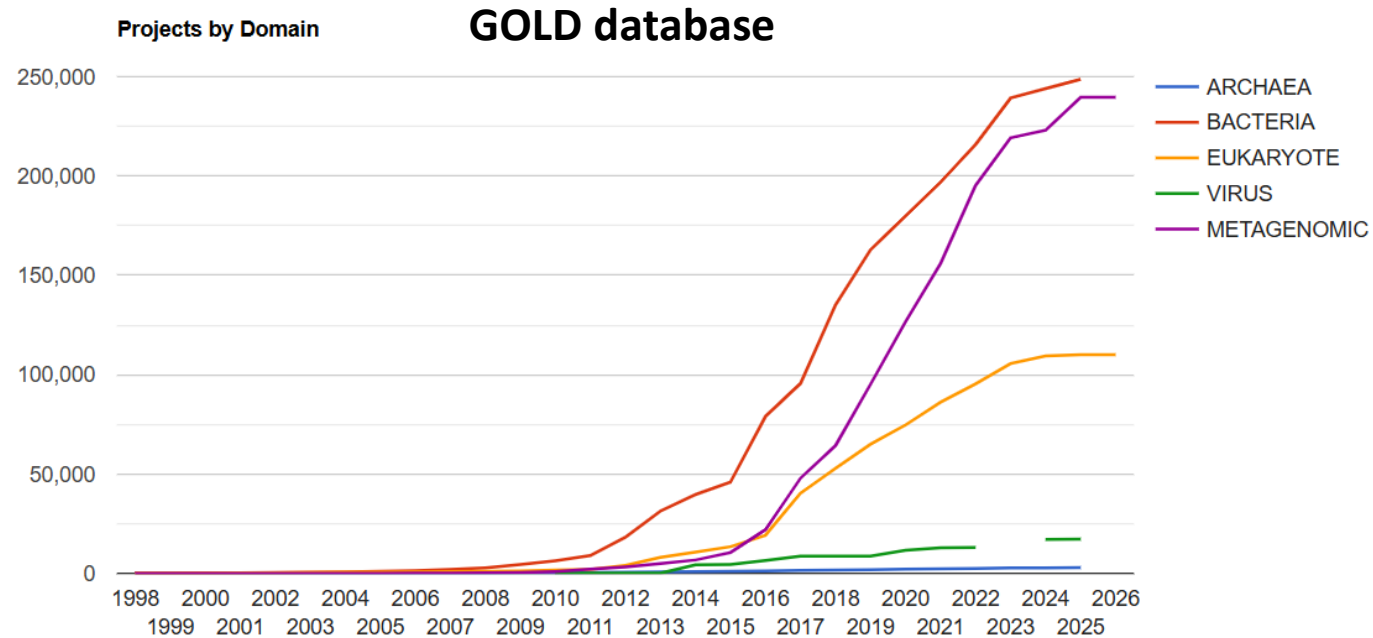


Véronique Brunaud – Compas Team
Institute of Plant Sciences Paris-Saclay
(IPS2)



Minisymposium Structural Annotation of Eukaryote genomes -2 juillet 2026

Whole Genome Sequencing



NCBI genome database

	Genome number	Reference	Annotated
Total	3.550.000	1%	78%
Bacteria	3.160.000	1%	80%
Eukaryote	64.500	40%	20%

→ 3452 references on Plants,
15% annotated

➡ Need new tools to annotate Eukaryote genomes

IA tools to do a first structural annotation

Helixer (Stiehler et al. 2021, Host et al. 2025) pre-training for vertebrates, invertebrates, plants, fungi

Tiberius (Gabriel et al. 2024) : first for mammals, updated in 2026 for pre-training for plants, insects, fungi

→ Using CNN + HMM

→ Very easy to use (one command!) just with fasta genome and possibility to add RNA-seq

→ Prediction of structural annotation for coding genes

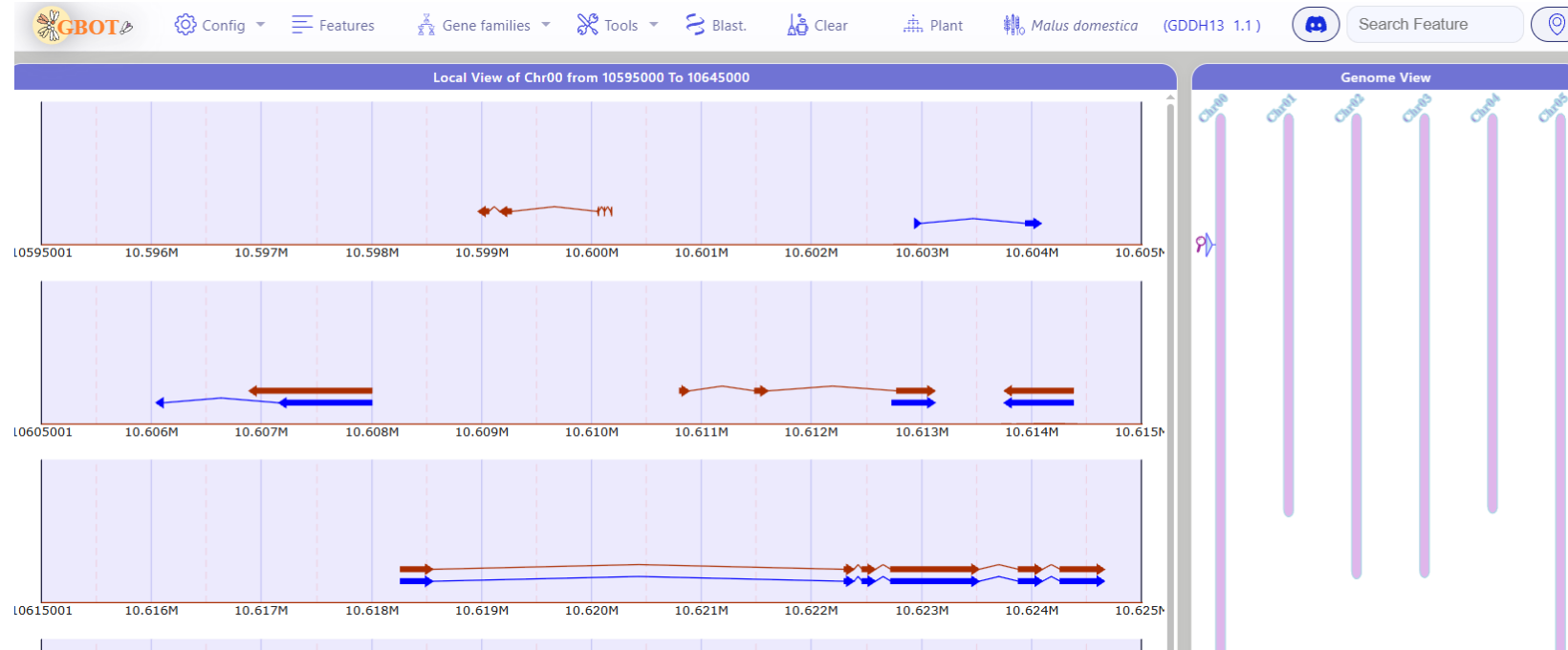
 We used Helixer because first model training with plant genomes

Helixer on Plant genomes

Akshay Khokan : student of M1 LaMME (Evry)

GBOT a genome browser and database developed by **Franck Samson**
8 plant genomes, official annotations and Helixer annotations

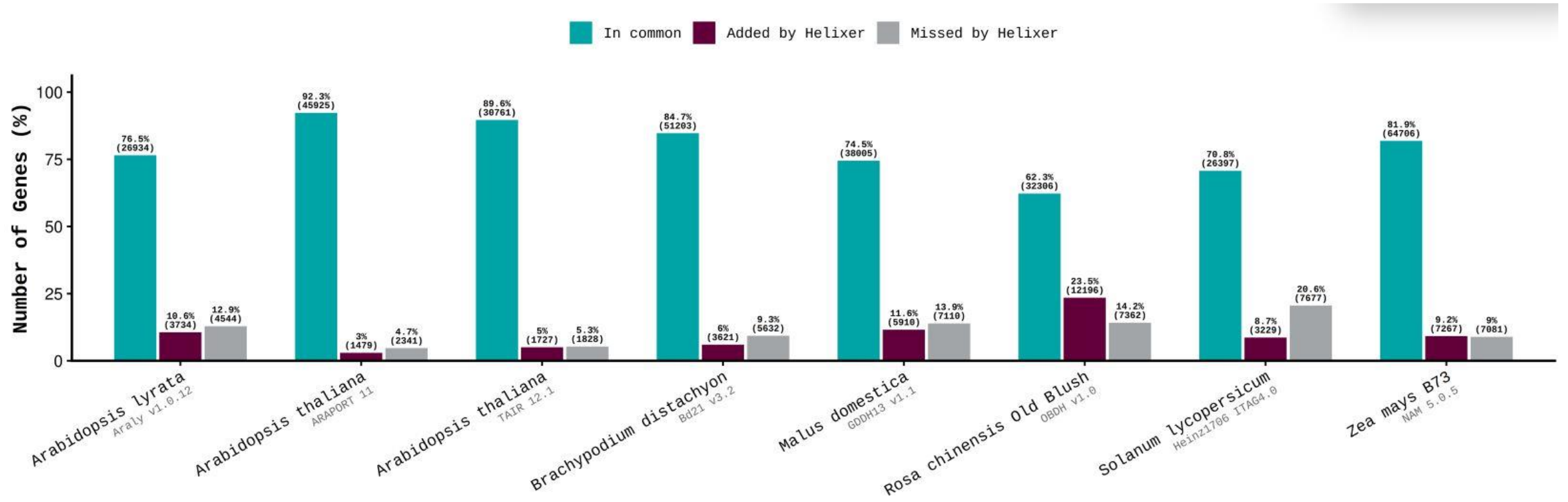
<http://stat.genopole.cnrs.fr/server/GBOT> -JOBIM 2024-2025 poster



Comparing structural gene annotation between official and Helixer, 3 categories

- 1- **Common** overlapping annotations, not necessarily identical structures
- 2- **Added** by Helixer, new gene annotation
- 3- **Missed** by Helixer in regard of official annotation

Helixer on 8 plant genomes



→ Results seem correspond to the version and quality of genome annotation

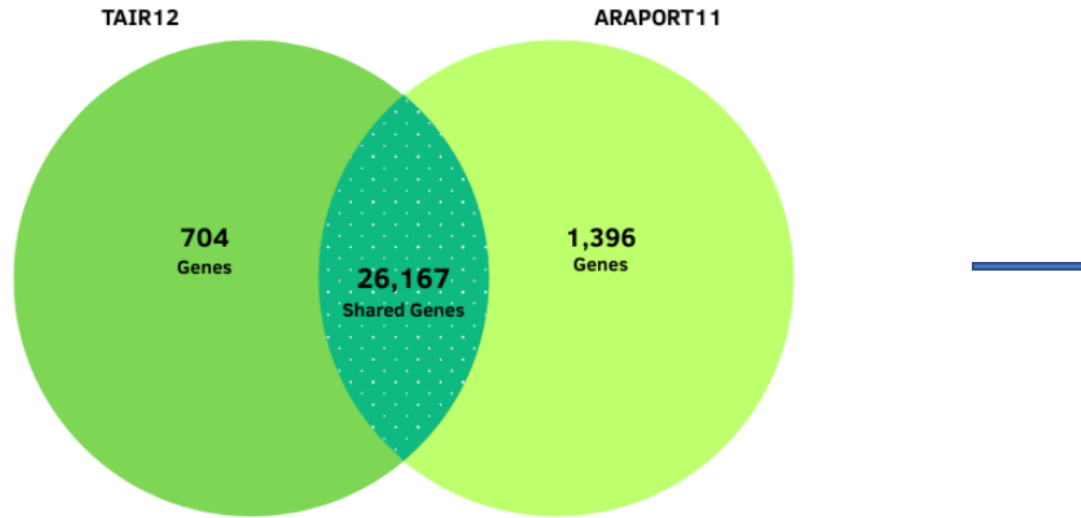
→ The best score of Helixer annotation is A. thaliana V11 and not V12 (last version 2026) but Helixer trained on V11

Zoom on *Arabidopsis thaliana* genome

- First plant genome to be sequenced in 2000 (Nature AGI consortium) : Col-0, small genome (120Mb, 5 nuclear chr + mito + chloro)
- The TAIR10 annotation (Lamesch *et al.* 2012) was informed by *ab initio* gene models, EST sequences from Sanger platforms, and two RNA-Seq datasets available at that time → This annotation is oriented to detect coding genes
- Araport11 annotation (cheng *et al.* 2016) used 113 RNA-Seq datasets partitioned into 11 groups according to their tissue or organ of origin → oriented new genes, new isoformes and non-coding genes.
- **TAIR12 new genome** (145Mb) and annotation (Reiser *et al.* 2026) integrates new high quality long-read assemblies, updated structural annotations, and expert manual review and (~100 scientists) → telomeric sequencing and curated annotation with effort on TE and ncRNA

Version	Size of genome (chr. 1–5)	Protein-coding genes	Transposons and pseudogenes	Alternatively spliced genes
TAIR12 (2025)	145.2 Mb	26,868	19,364	5,701
Araport11 (2016)	119.1 Mb	27,655	4,853	10,695
TAIR10 (2010)	119.1 Mb	27,411	4,827	5,885

Zoom on *Arabidopsis thaliana* genome



TAIR12: Genome annotations summary

TAIR12 VS ARAPORT11 : Protein Comparison

Shared Proteins	Count
With Same Sequence	21,294
With Different Sequence	4,873
Total	26,167

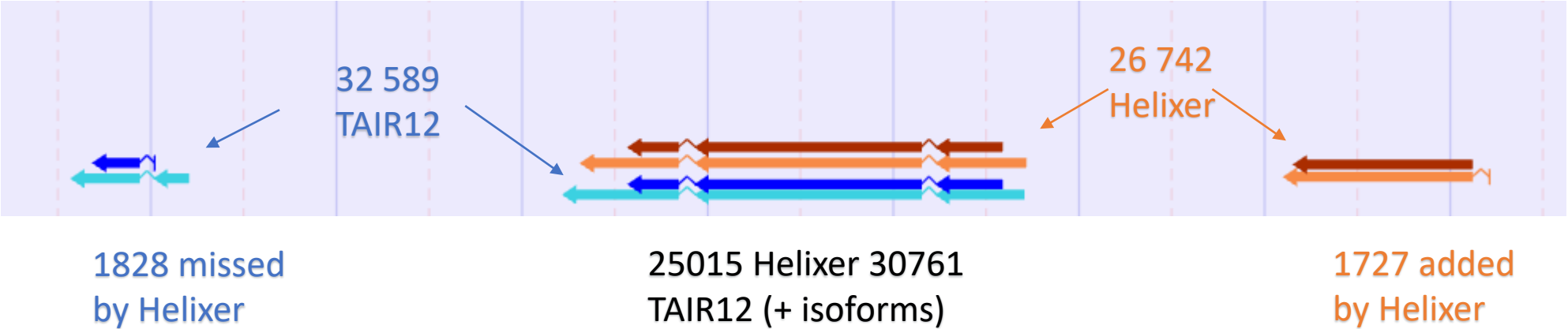
97% of common ID genes between V12 and V11

→ 81% identical proteins

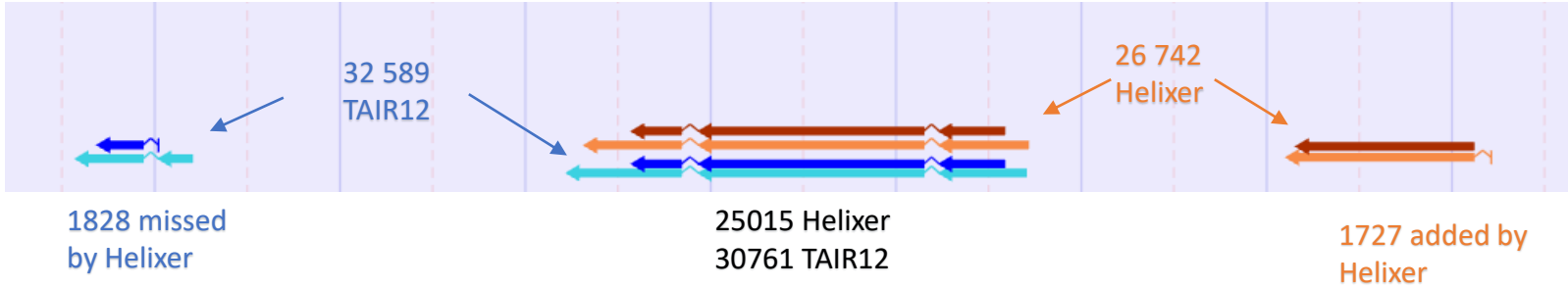
→ 14% small divergence (> 90% identity)

Zoom on TAIR12 Arabidopsis genome and prediction from Helixer

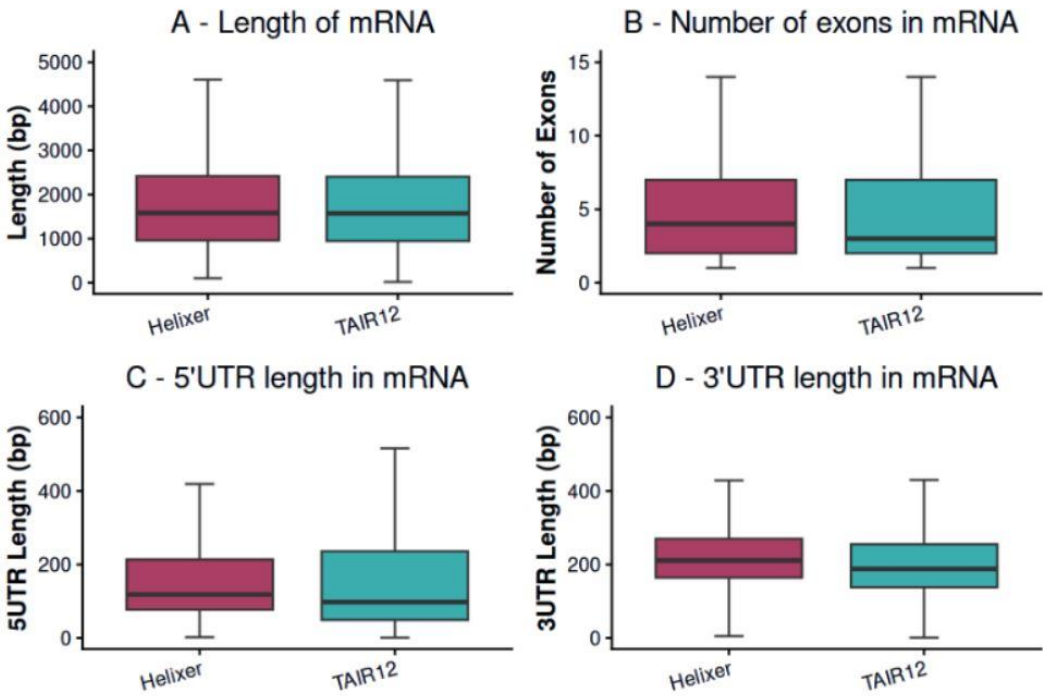
GBOT view



Global characteristics of genes TAIR12 vs Helixer



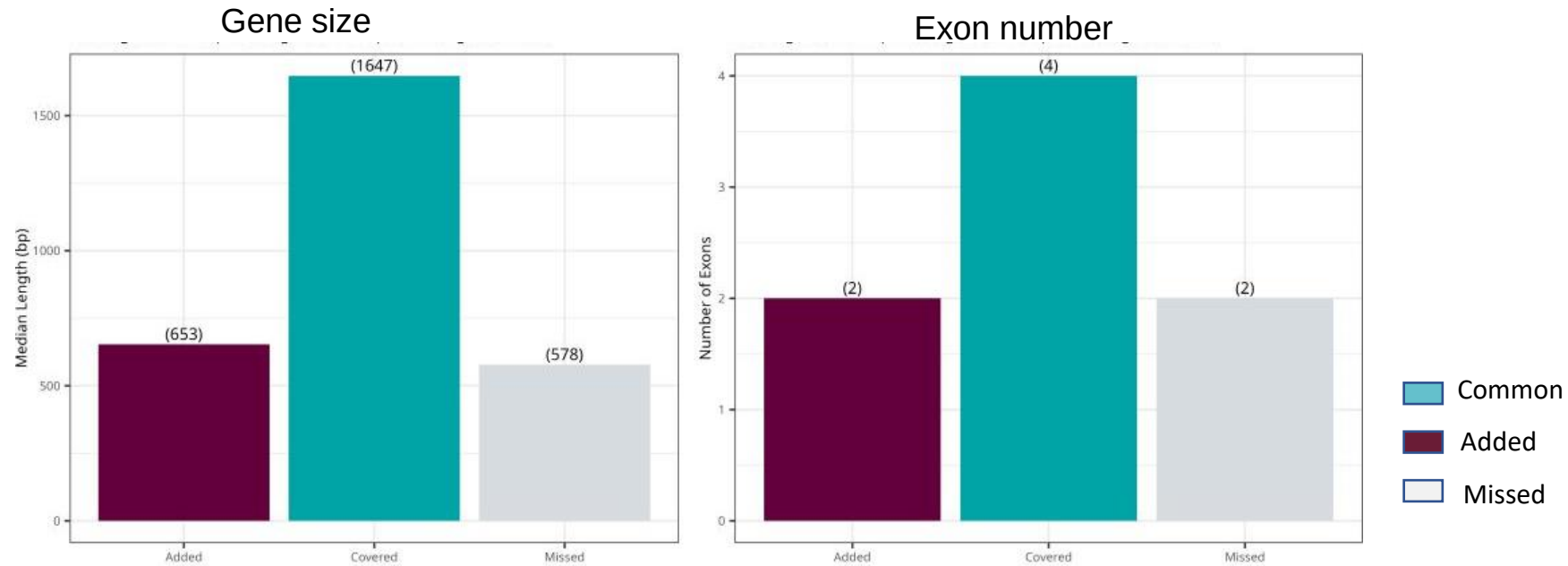
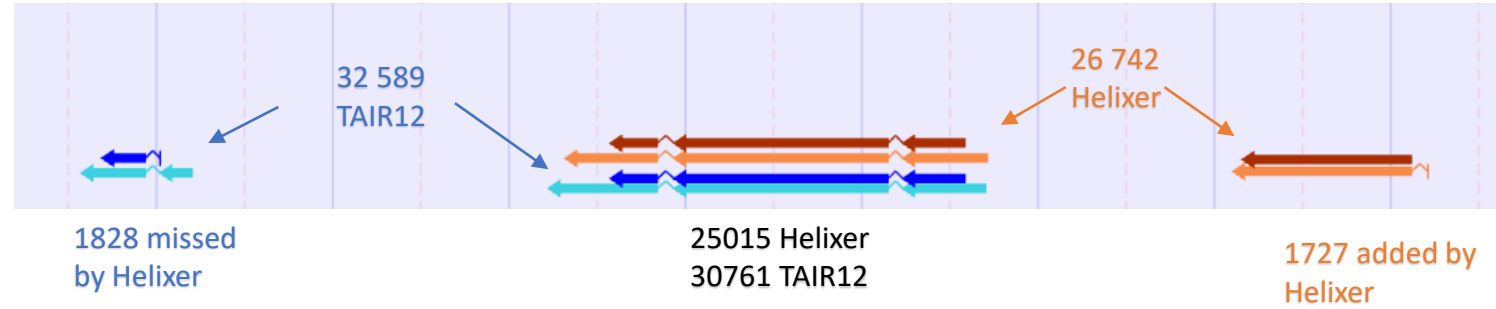
Differences between Helixer & TAIR12 mRNA predictions



Type	Length (bp)	Number of Exons	5'UTR (bp)	3'UTR (bp)
Helixer	1,582	4	118	211
TAIR12	1,573	3	98	188

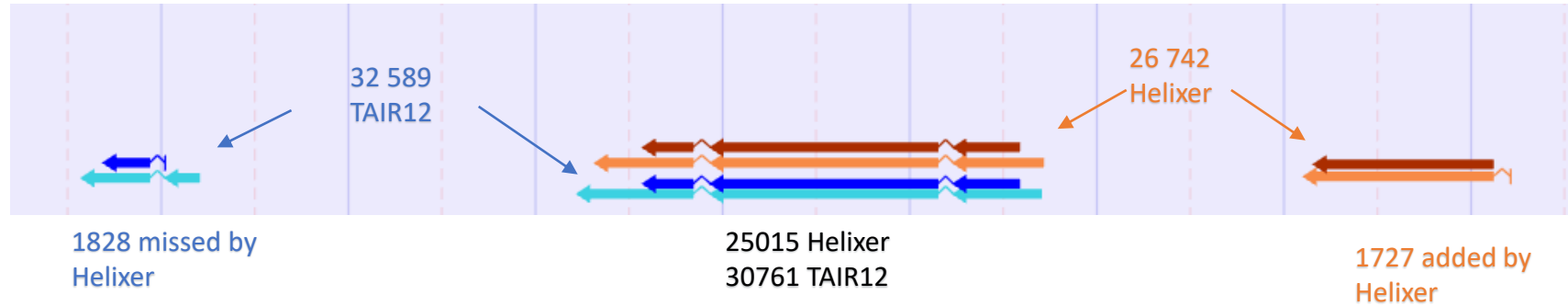
- Same gene size, UTRs 20 bases larger for Helixer
- Number of exons differs, one more for Helixer

Characteristics of genes predicted by Helixer



- Significant differences across gene classes in terms of gene size and number of exons
- New genes from Helixer are smaller than expected, missed also !

Characteristics of proteins predicted by Helixer



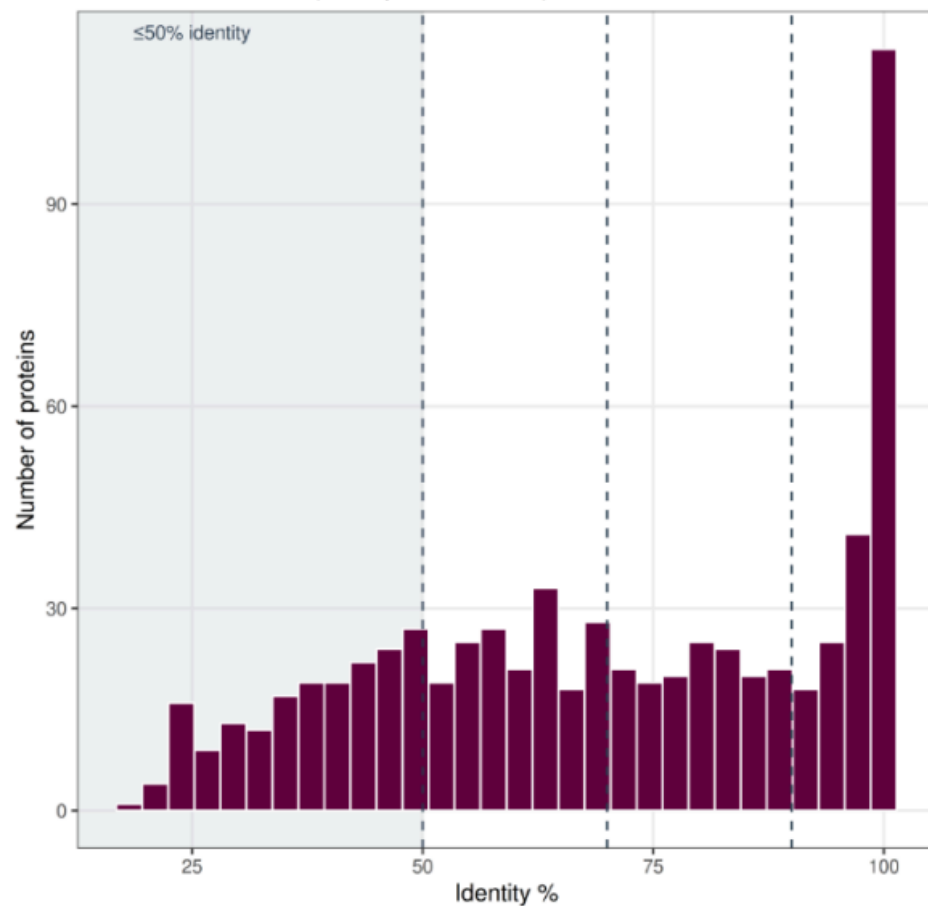
BUSCO results of Helixer proteome : Complete 99%, Single 97.6%

Common features

- 81% identical proteins
- 14% small divergence (> 90% identity)

Do the new proteins predicted by Helixer match other proteins ?

1- search 1727 new prot in Uniprot

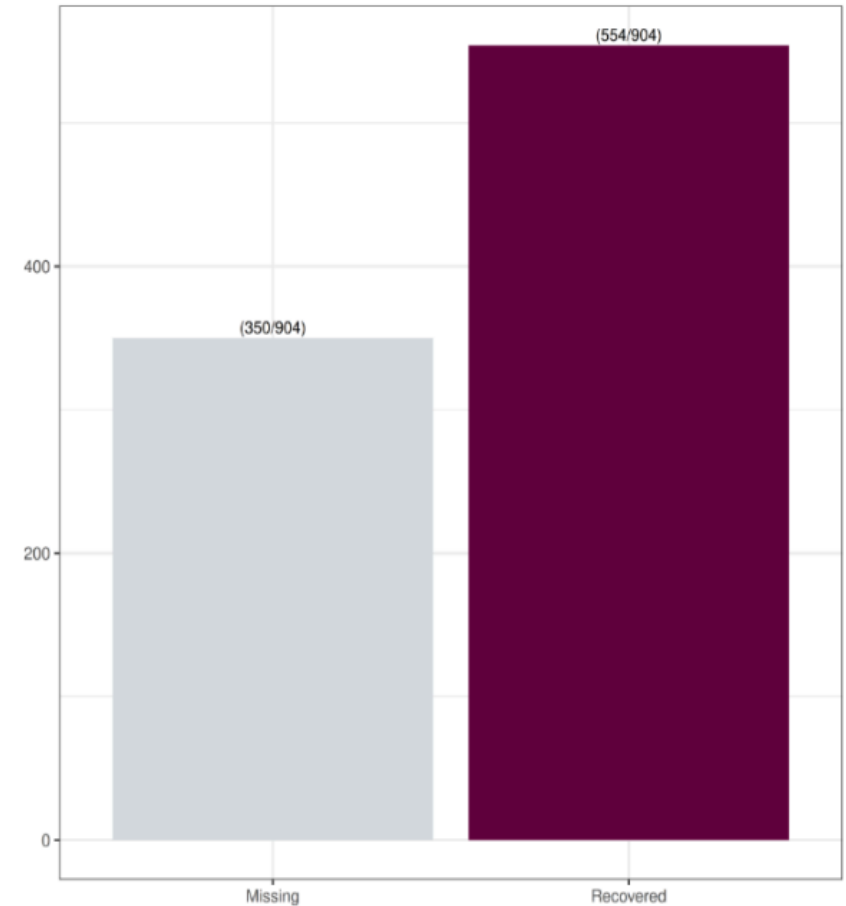


904 no Uniprot



Strict Filters
< 50% Identity
< 60% length cover

2- search for ortholog group in eggNOG



→ 80 % of new proteins predicted by Helixer are homologous to other proteins

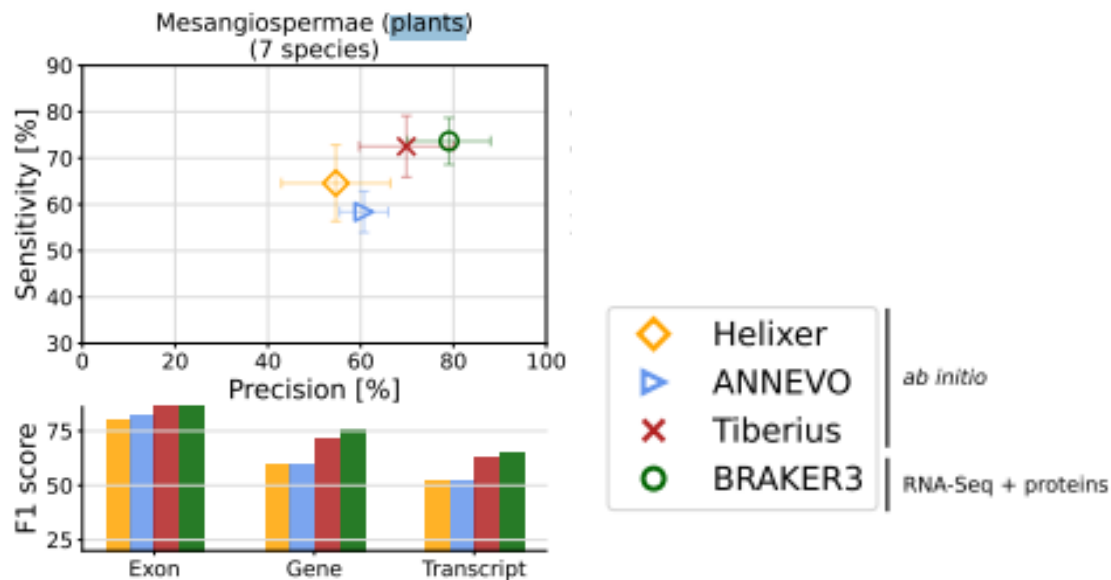
Conclusion

Helixer is not perfect but easy to use, rapid and simple for first annotation level

But only coding genes, no isoforms and no ncRNA predicted

We used Helixer but last version of

Tiberius applied to plant seems better, but only CDS are predicted (no mRNA)



Note on the implications of each new annotation

- The definition of the 5' and 3' ends of genes is not always accurate, case of ARAPORT11 compared to TAIR10
 - Poor definition of core or proximal promoters
- Transcriptome analyses depend on gene annotations
 - Change in list of differentially expressed genes (DEGs)

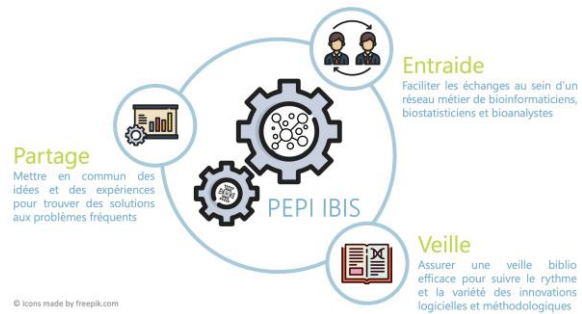
Example with one differential expressed data studied in ARAPORT11 and TAIR12

→ 97 % of common DEG

→ 265 specific to V11 and 510 specific to V12

Thank you

And thanks to Akshay KHOKAN, Franck
SAMSON, Carene RIZZON



Réseau **MERIT**
MatiER en
bioinformaTique