Improved gene annotation of the fungal wheat pathogen *Zymoseptoria tritici* based on combined Iso-Seq and RNA-Seq evidence

Nicolas Lapalu

Lucie Lamothe

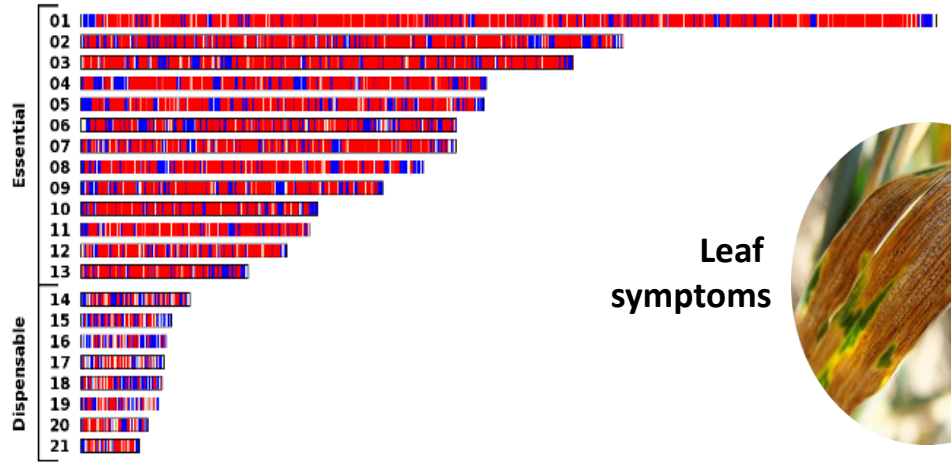
Nicolas Lapalu[1], Lucie Lamothe[1], Yohann Petit[1], Anne Genissel[1], Camille Delude[2], Alice Feurtey[3,4], Leen N. Abraham[3], Dan Smith[5], Robert King[5], Alison Renwick[6], Mélanie Appertet[2], Justine Sucher[2], Andrei S. Steindorff[7], Stephen B. Goodwin[9], Gert H.J. Kema[11], Igor V. Grigoriev[7,8], James Hane[6], Jason Rudd[5], Eva Stukenbrock[10], Daniel Croll[3], Gabriel Scalliet[2], Marc-Henri Lebrun[1]

[1]Université Paris-Saclay, INRAE, UR1290 BIOGER, Palaiseau, France
[2]Syngenta Crop Protection AG, Stein, Switzerland
[3]University of Neuchâtel, Neuchâtel, Switzerland
[4]ETH Zurich, Zurich, Switzerland
[5]Dept of Protecting Crops and the Environment, Rothamsted Research, UK
[6]Centre for Crop and Disease Management, Curtin University, Perth, Australia
[7]U.S. Department of Energy Joint Genome Institute, Berkeley, USA
[8]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, USA
[9]USDA-Agricultural Research Service, West Lafayette, USA
[10]Environmental Genomics, Max Planck Institute for Evolutionary Biology, Germany
[11]Wageningen University and Research, The Netherlands

Lapalu et al. 2023. bioRxiv. doi: https://doi.org/10.1101/2023.04.26.537486

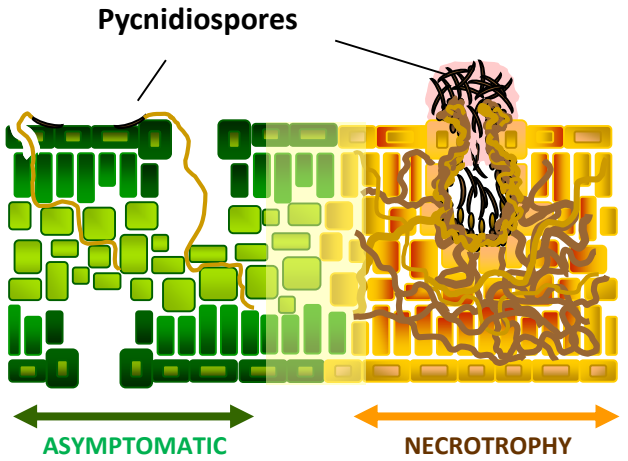
Zymoseptoria tritici, a fungal pathogen of wheat

Fully sequenced genome 21 chromosomes,
39.7 Mb, JGI: 10.952 genes, 17-20 % transposons



Haploid hemi-biotrophic ascomycete
Life cycle: sexual and asexual reproduction

Leaf symptoms



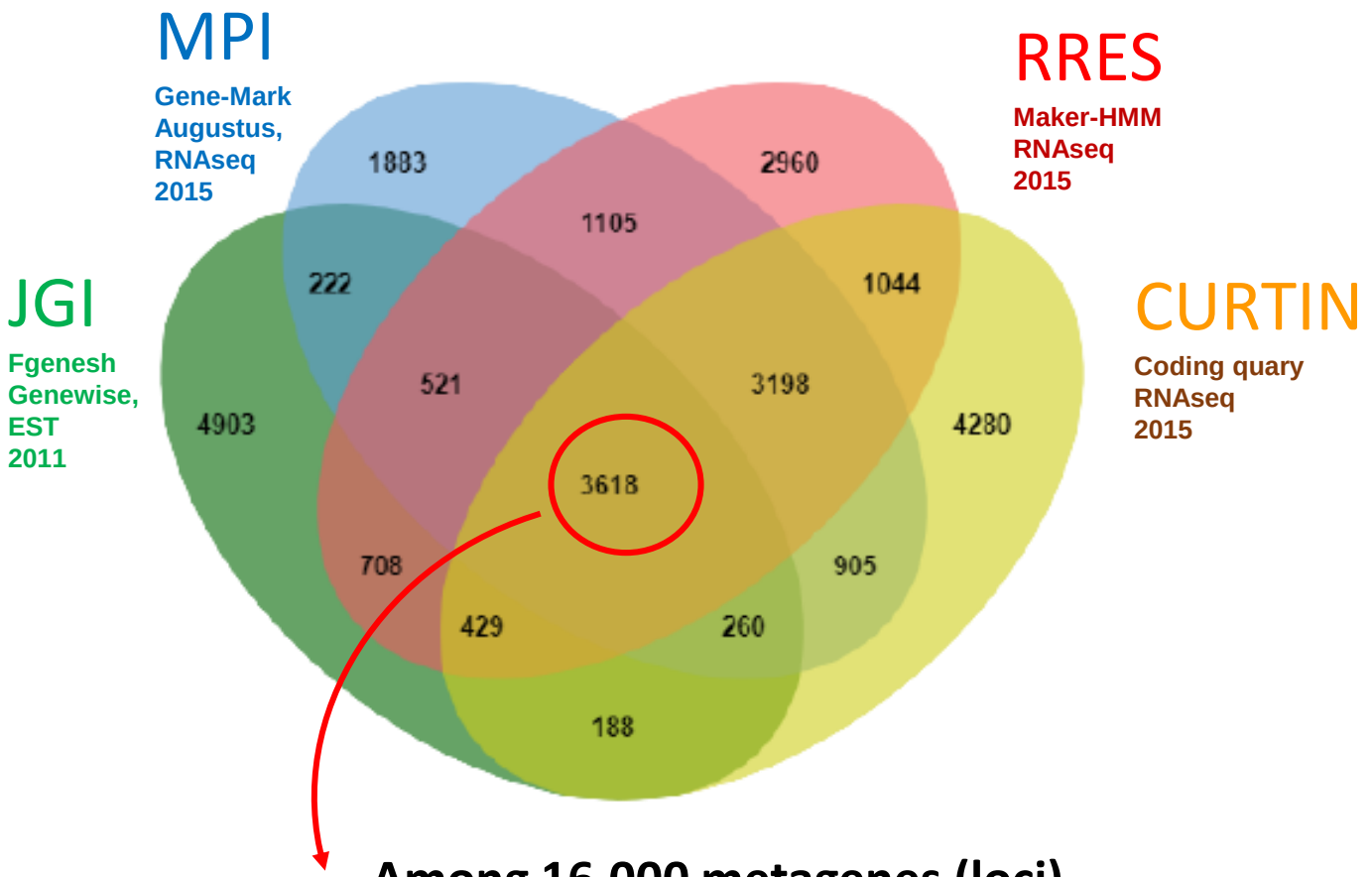
Genetics crosses, genome sequences, genome annotations
Transformation, reverse genetics using Ku70 deficient strains, cellular biology, expression vectors
Large infection RNAseq datasets,
GWAS, populations genomics, experimental evolution

Improvement of *Z. tritici* gene annotation

Why do we need to improve *Z. trititi* gene annotation ?

Why do we need to improve *Z. tritici* gene annotation ?

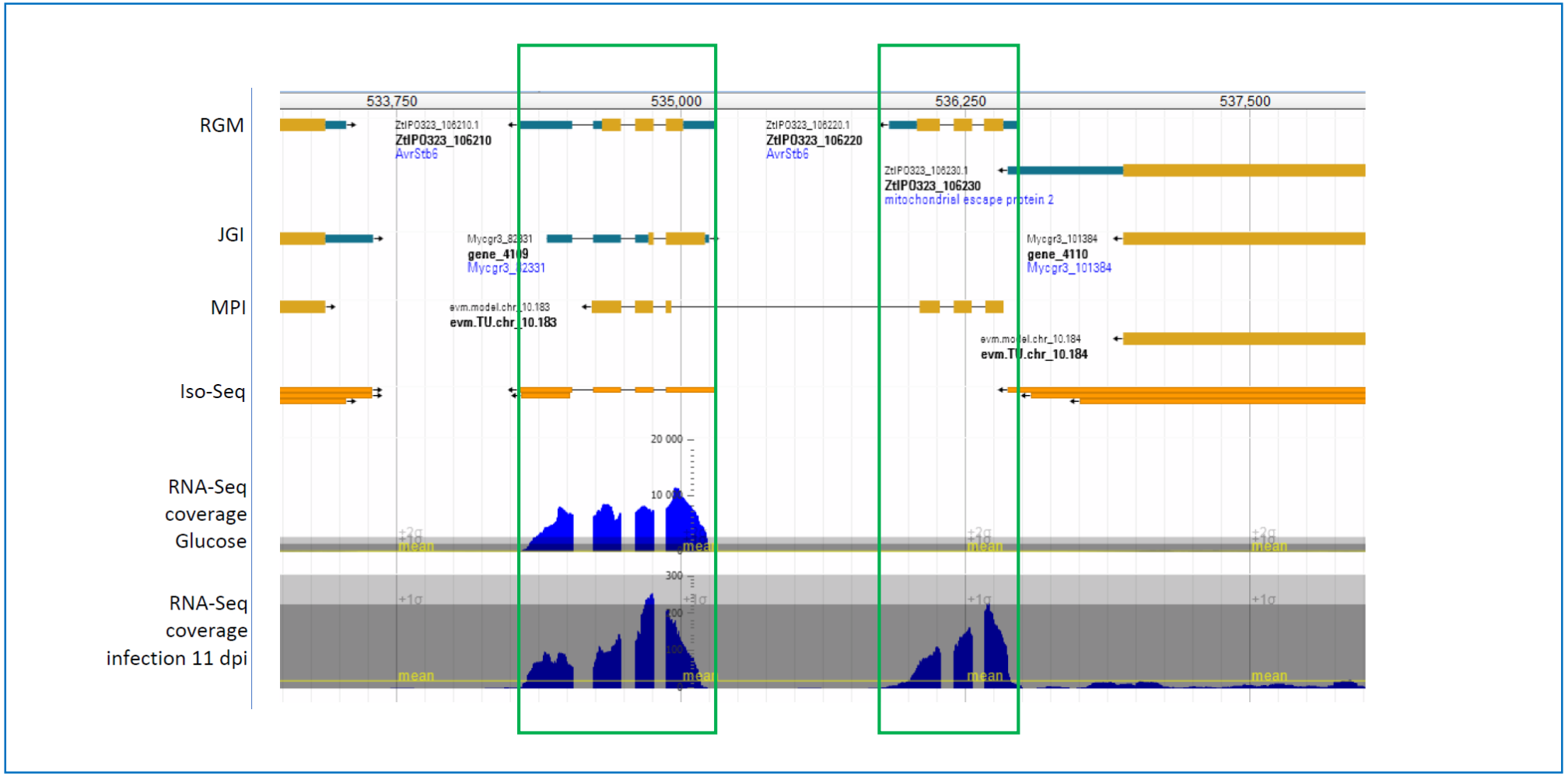
Comparison of existing IPO323 *Z. tritici* gene annotations



Among 16.000 metagenes (loci),
only 3618 (22 %) displayed identical gene structures

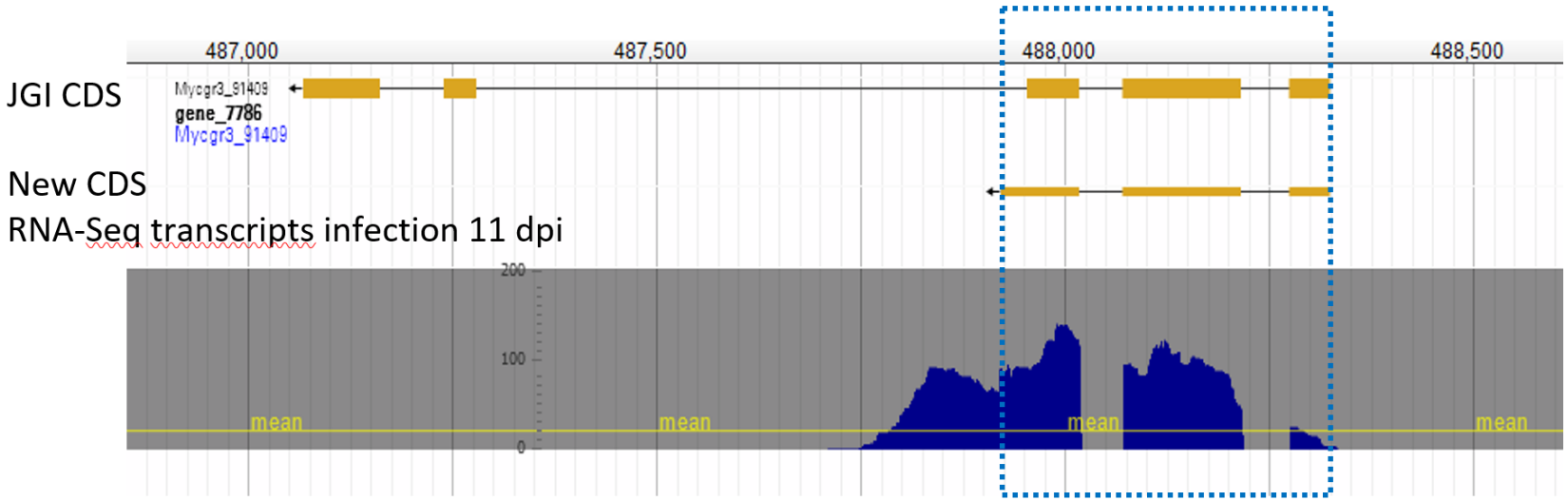
Why do we need to improve *Z. tritici* gene annotation ?

Avr-Stb6 encoding genes were not predicted by any previous annotations



Why do we need to improve *Z. trititi* gene annotation ?

Zt-NIP1 encoding gene was not predicted accurately by previous JGI annotation (NCBI)



Zt-Mycgr3-91409

MAPIFTYAVAALAFAQSAYAVVYAARCKFGNPLVQNNRITRAVCDLTNEHTTKDGSWHYVEVDNECKYLAGDNPRDQPGWAVFVKYCPEQNAAADKSRARGEQTCGFVRTPVDDVSRRAESATTPVCEGD

NEW Zt-Mycgr3-91409-2

MAPIFTYAVAALAFAQSAYAVVYAARCKFGNPLVQNNRITRAVCDLTNEHTTKDGSWHYVEVDNECKYLAGDNPRDQPGWAVFVKYCTYYKGVPDA

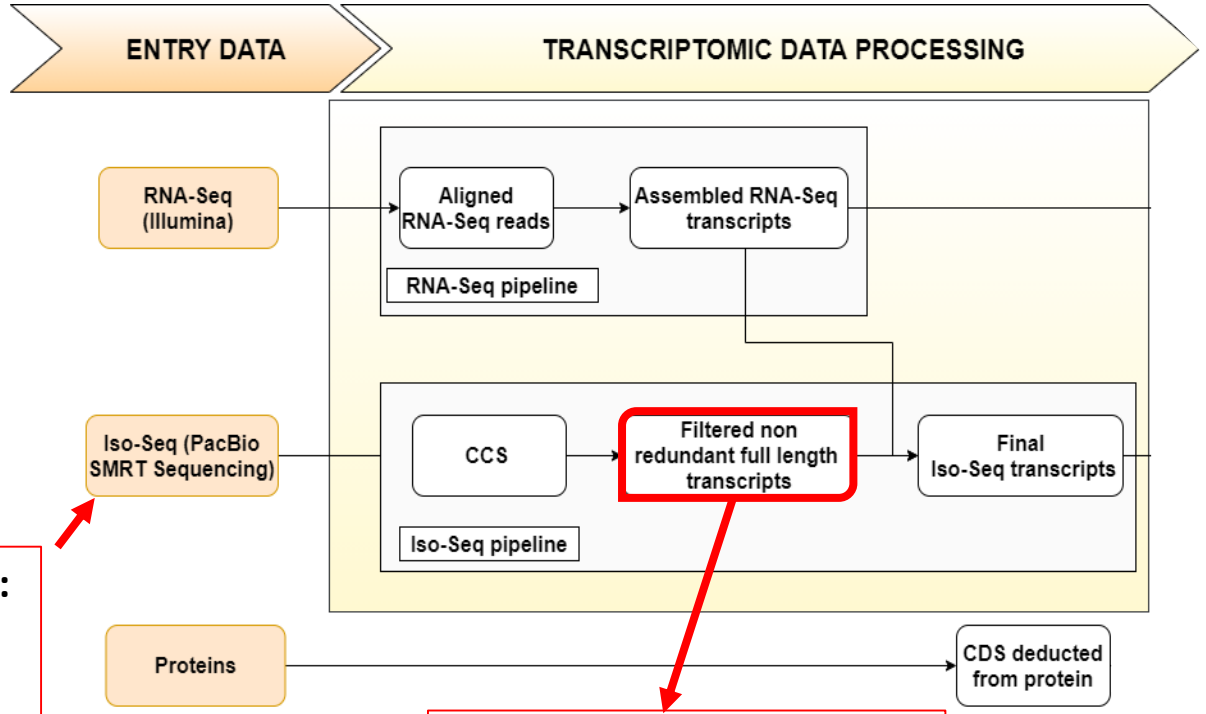
Errors in database searches

Improvement of *Z. tritici* gene annotation

Search for a modulable strategy
to improve *Z. trititi* gene annotation

Improvement of *Z. tritici* gene annotation

Workflow for gene model prediction and selection

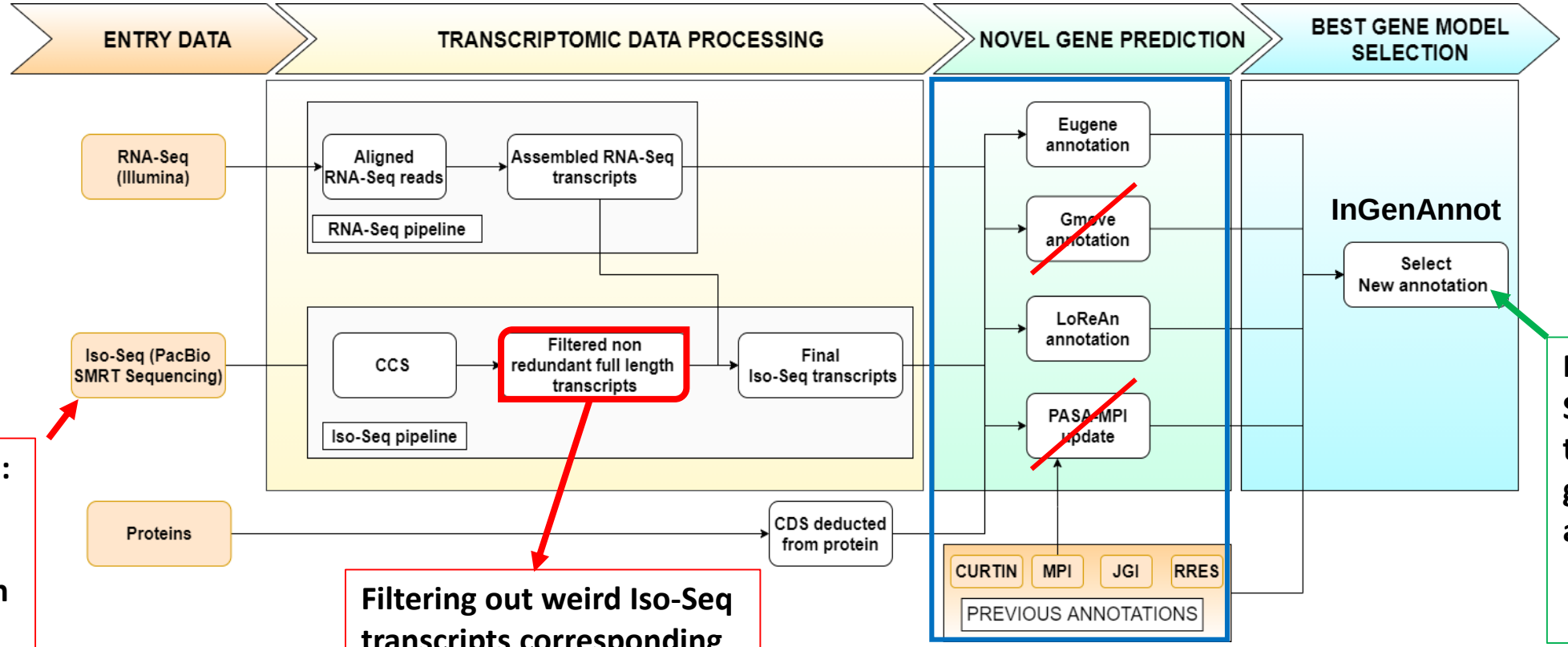


New data: Iso-Seq (PacBio full length cDNAs)

Filtering out weird Iso-Seq transcripts corresponding to rare transcripts

Improvement of *Z. tritici* gene annotation

Workflow for gene model prediction and selection



New data: Iso-Seq (PacBio full length cDNAs)

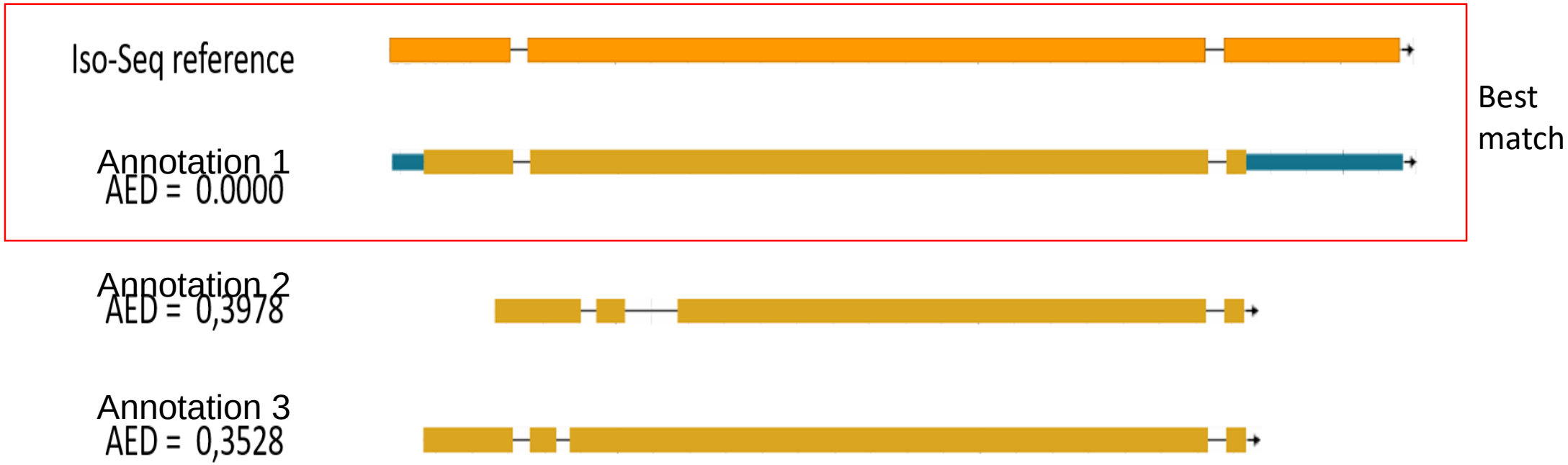
Filtering out weird Iso-Seq transcripts corresponding to rare transcripts

Six independant gene model predictions

New Tool: Selection of the best gene model according to :
- **transcripts**
- **proteins**

Improvement of *Z. tritici* gene annotation

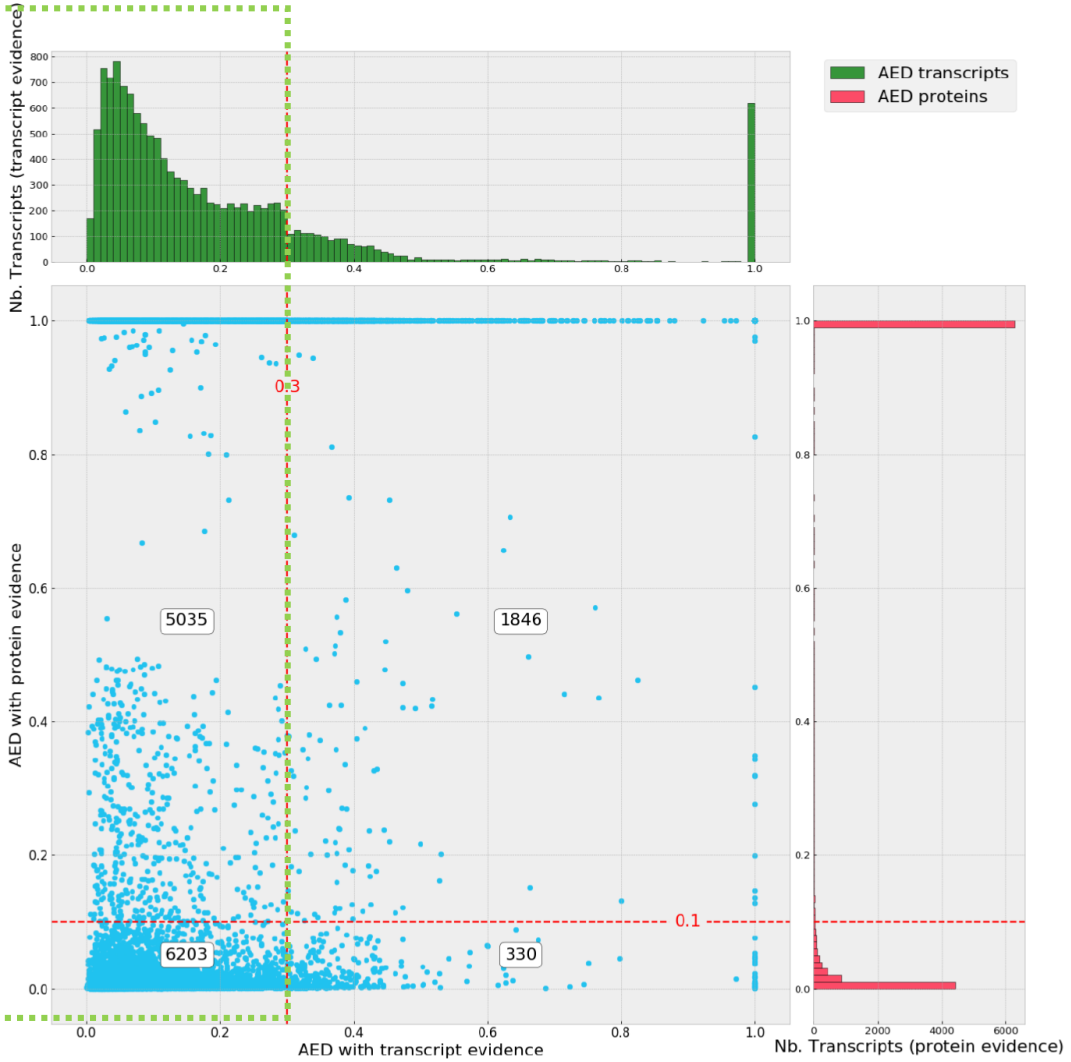
Selection of the best gene models according to AED scores
(AED = value integrating evidences from transcripts and proteins)



Improvement of *Z. tritici* gene annotation

New *Z. tritici* annotation with 13,414 gene models fitting to transcript (0.3 < AED) and/or protein (0.1 < AED) evidence

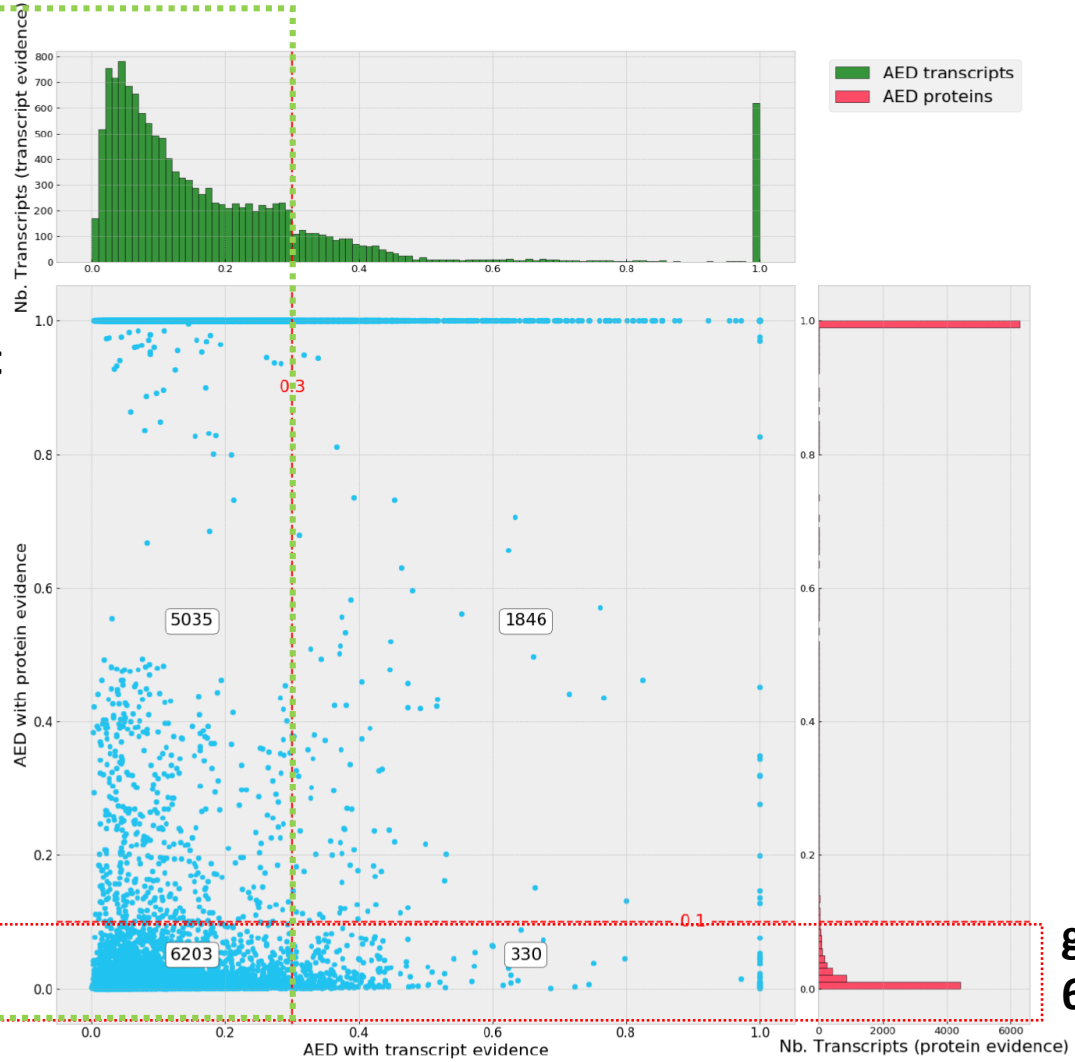
good transcript evidence: 11238 gene models



Improvement of *Z. tritici* gene annotation

New *Z. tritici* annotation with 13,414 gene models fitting to transcript (0.3 < AED) and/or protein (0.1 < AED) evidence

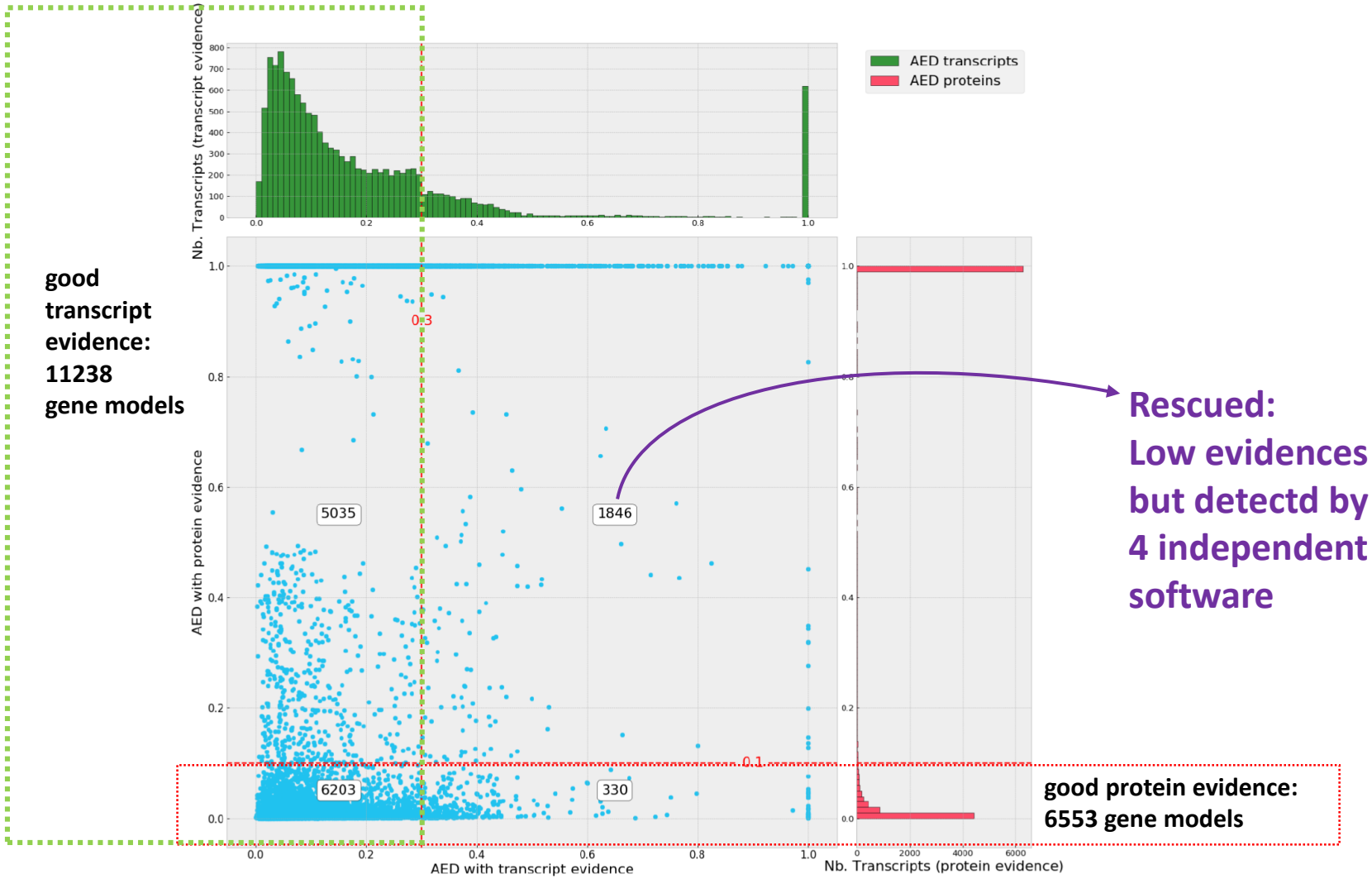
good transcript evidence: 11238 gene models



good protein evidence: 6553 gene models

Improvement of *Z. tritici* gene annotation

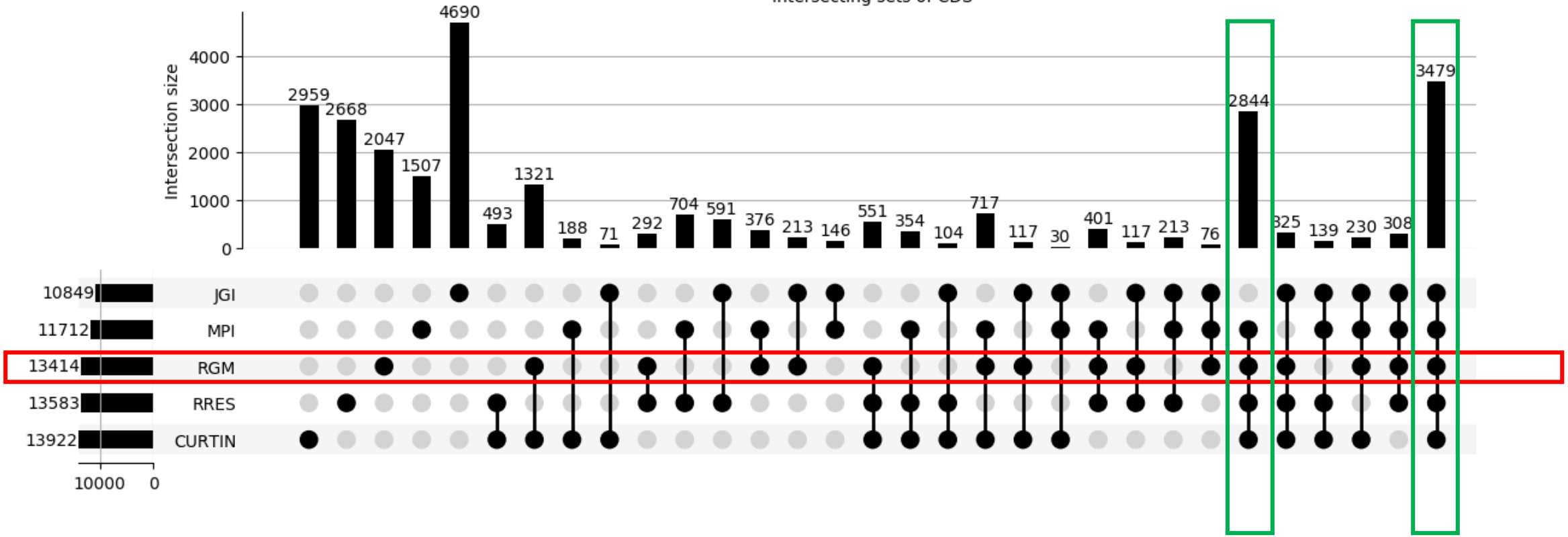
New *Z. tritici* annotation with 13,414 gene models fitting to transcript (0.3 < AED) and/or protein (0.1 < AED) evidence or rescued



Improvement of *Z. tritici* gene annotation

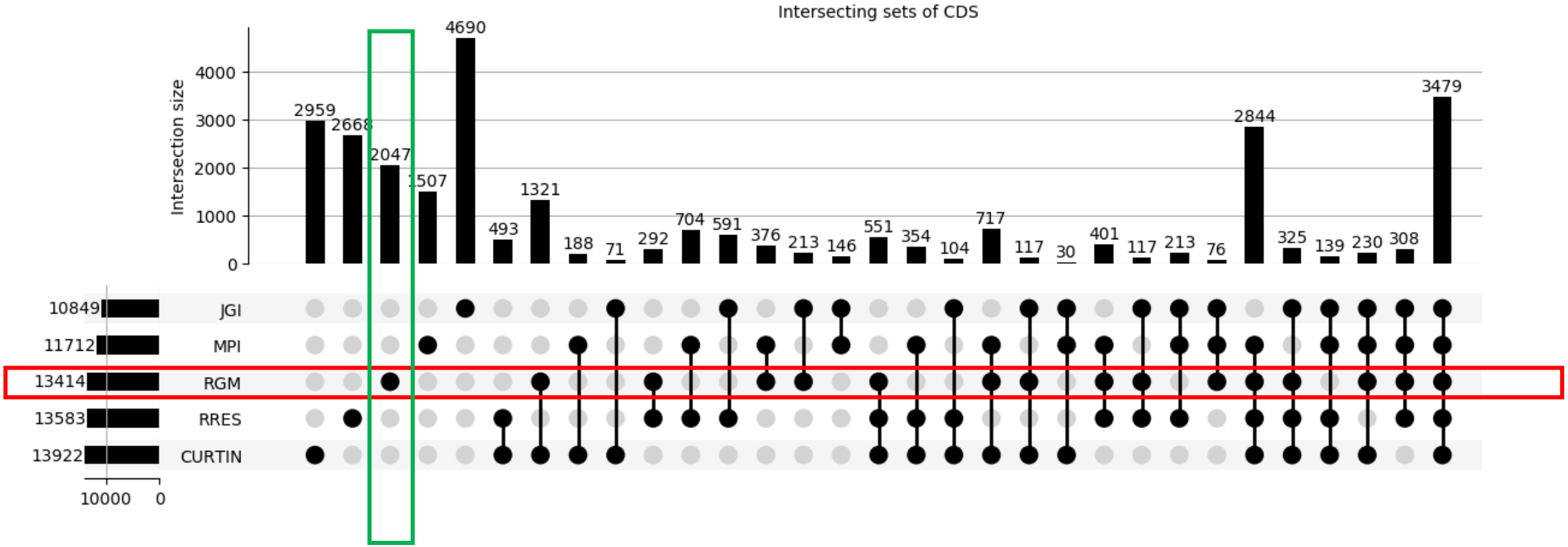
Comparison to previous annotations: % of identical predictions

6323 genes (47%) correctly identified in MPI-RRES-Curtin



Improvement of *Z. tritici* gene annotation

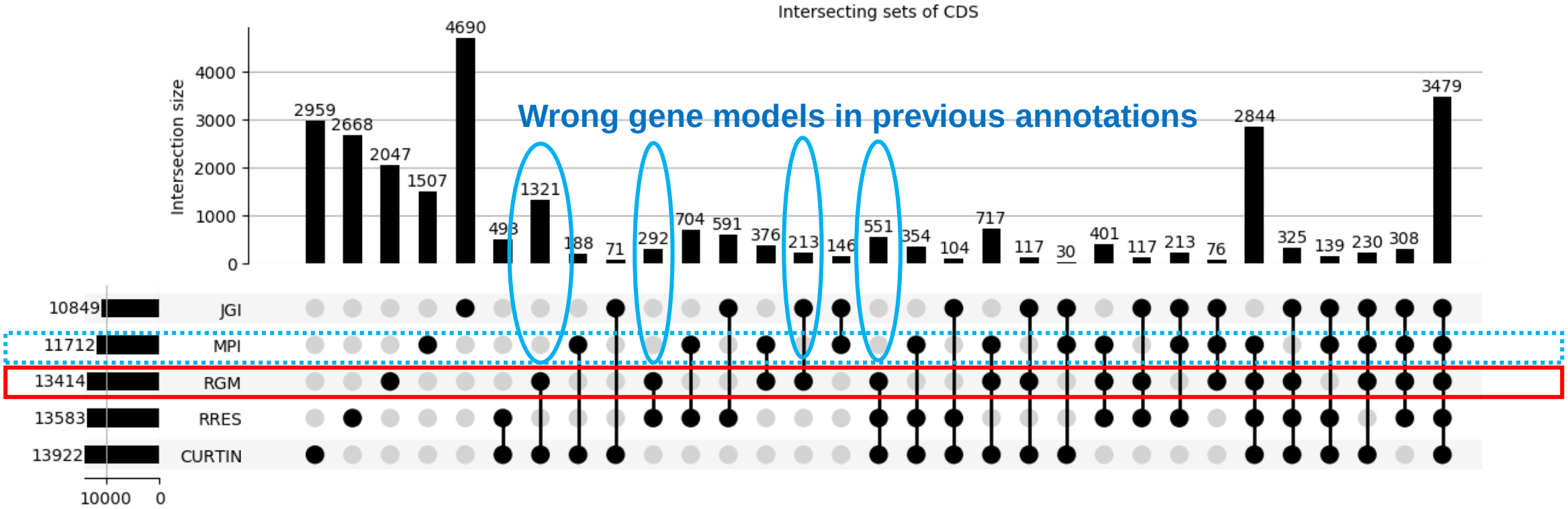
Comparison to previous annotations: new genes ?



2047 new genes

Improvement of *Z. tritici* gene annotation

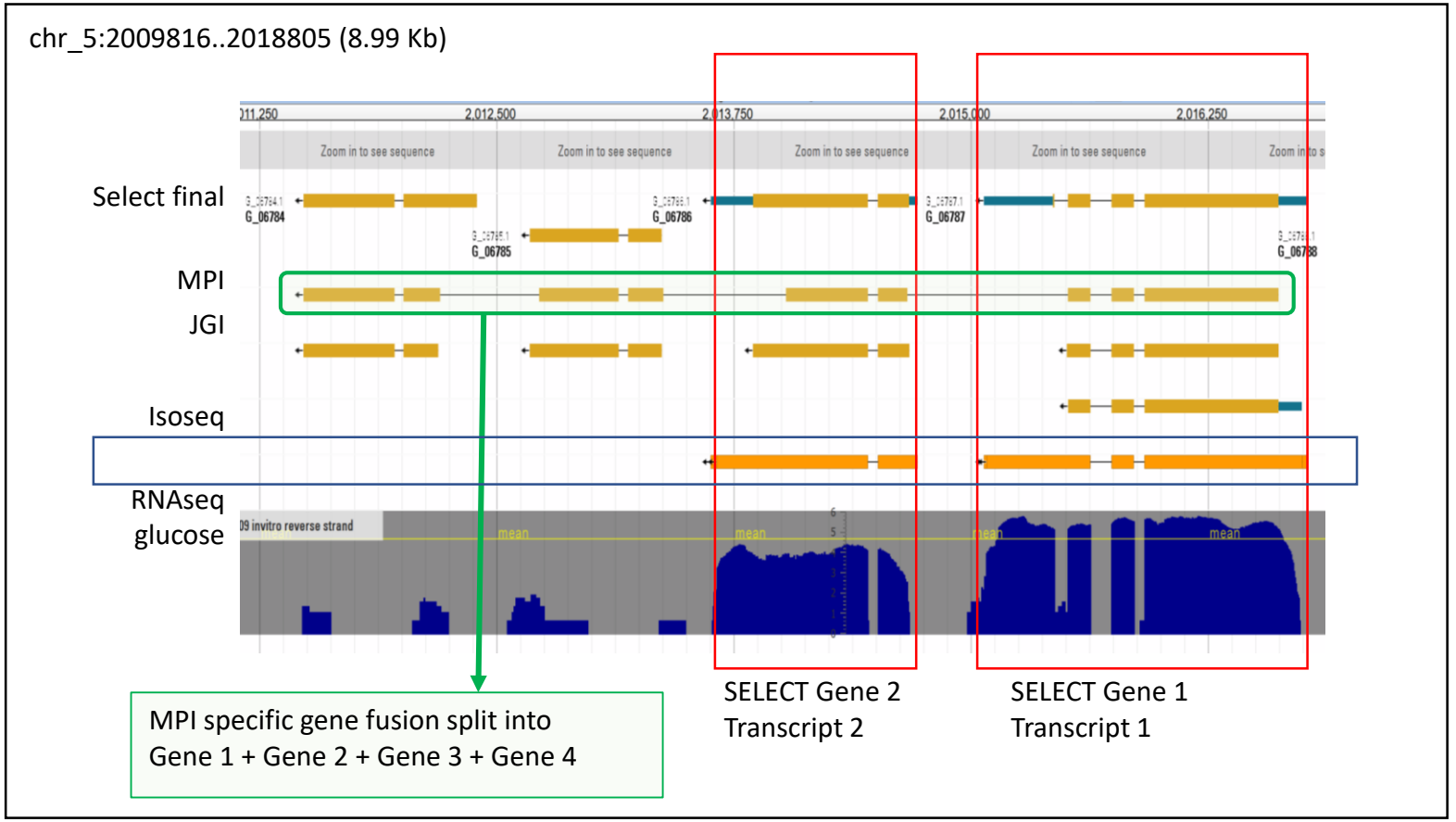
Comparison to previous annotations: systematic errors ?



Improvement of *Z. tritici* gene annotation

Systematic errors in previous gene models: gene fusions

Exemple of a MPI gene model splitted into 4 new genes



Improvement of *Z. tritici* gene annotation

Contribution of each software to the best gene models

	Annotation	Identical	Specific to	% total
Available annotations	JGI	4865	157	36
	MPI	8431	91	63
	RRES	8317	175	62
	CURTIN	9584	506	71
New annotations	Eugene	10224	1603	76
	LoReAn	7769	199	58

No single software is able to predict all best gene models: need for many to select the best

Improvement of *Z. tritici* gene annotation: Conclusions

- **New datasets to improve CDS and gene (UTR) predictions:**
  Iso-Seq full length transcripts

- **New tool to select the best gene model according**
  to transcript and protein evidence using different annotations pipelines
  (Ingenannot)

  Improved *Z. tritici* IPO323 gene annotation with 13,414 gene models
  Annotation of 5' and 3' UTRs for 9,856 genes (73 %)

- **Identification of transcript isoforms**
  13 % of *Z. tritici* expressed genes have transcript isoforms
  (alternative transcriptional start, stop, exon skipping, intron retention)
  mostly intron retention

Improvement of *Z. tritici* gene annotation

What's next

- Add more ab initio software to InGeAnnot
Saturation in the prediction of novel genes ?

- Manual annotation still needed for some gene models
Web site for manual annotation

- Compare our pipeline (InGeAnnot) to other software
Braker3-Tsebra, Deep Learning

- Improve other fungal genomes with InGeannot and PacBio Isoseq data