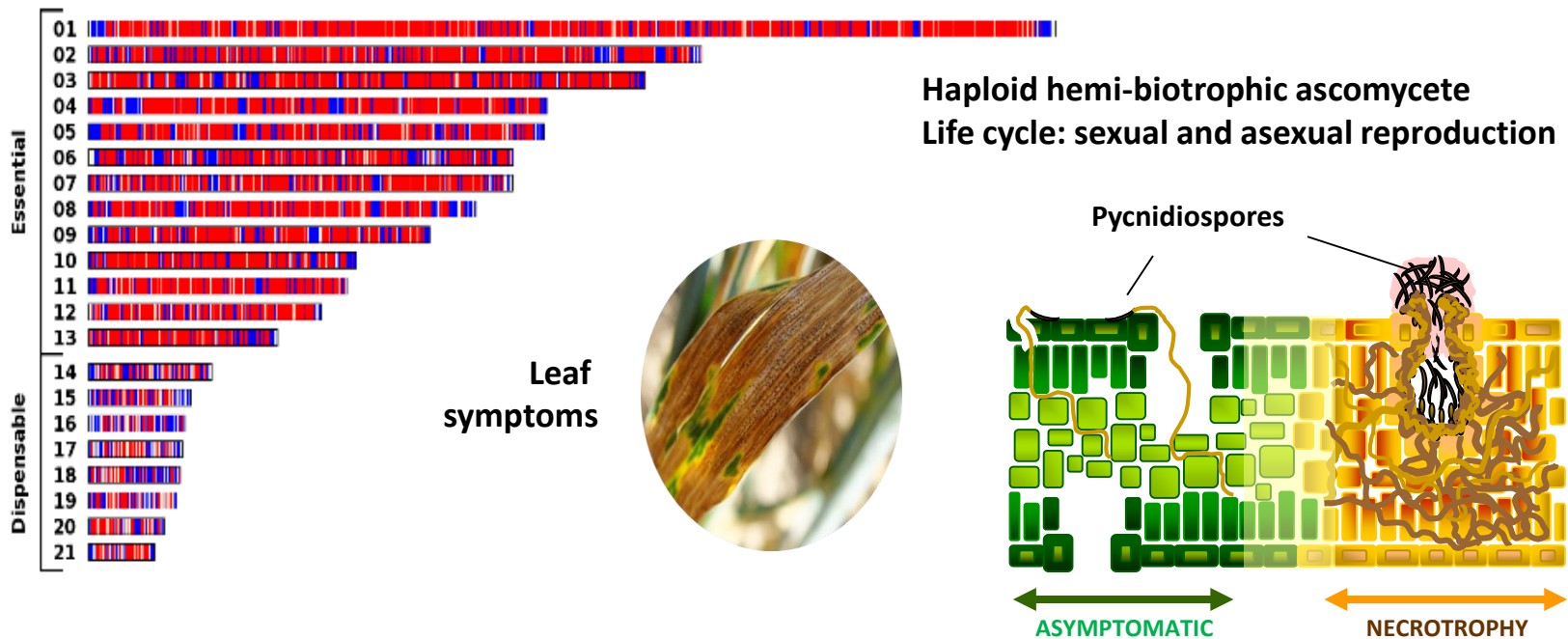


# ➤ Evaluation of coding gene annotations: tools and measurements

Nicolas Lapalu –PEPI IBIS  
14-15th September 2023

# ➤ Biological context: *Zymoseptoria tritici*, a wheat pathogen

Fully sequenced genome\* 21 chromosomes,  
39.7 Mb, 18 % transposons



\* Goodwin, et al 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet.



## ➤ Genome annotation releases of *Z.tritici* IPO323

Several annotations were obtained, each with a specific method (ab initio prediction software, evidence used, pipeline)

- 1) JGI: Release 1, GeneWise and FGenesh ab initio software, EST and protein evidence

Published in 2011\*, **10849** genes

- 2) MPI: Release 2, EVM, PASA and GeneWise ab initio software, RNA-Seq and protein evidence, JGI predictions kept

Published in 2015\*, **11712** genes

- 3) - CURTIN, **13922** genes, CodingQuarry ab initio software using RNA-Seq data, RNA-Seq and protein evidence

- RRES, **13583** genes , Maker2 ab initio software, RNA-Seq and protein evidence

## Why another annotation ?

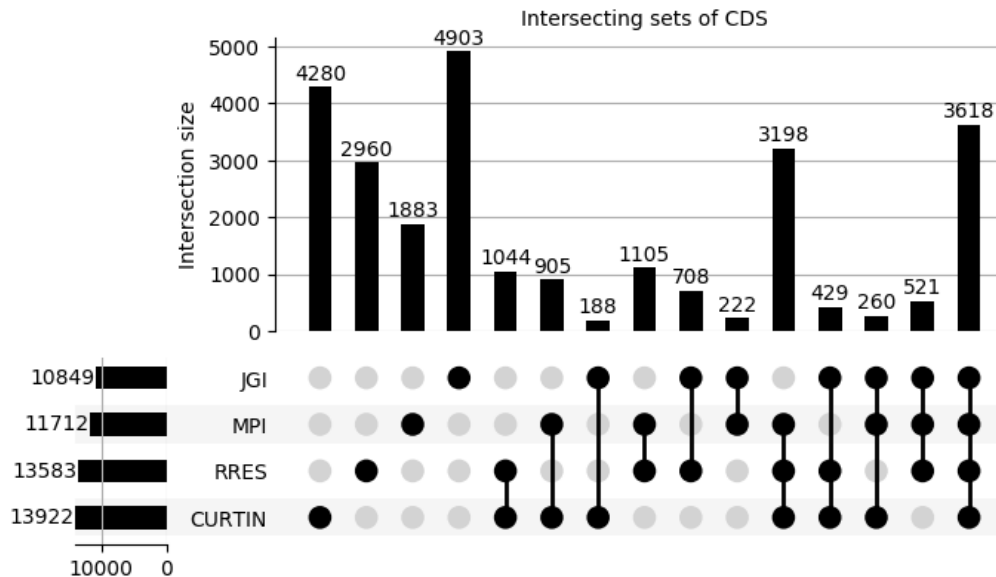
\* Goodwin, et al 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet.* 7

\*\*Grandaubert et al 2015. RNA-seq-Based gene annotation and comparative genomics of four fungal grass pathogens in the genus *Zymoseptoria* identify novel orphan genes and species-specific invasions of transposable elements. *G3*



# ➤ Genome annotation releases of *Z. tritici* IPO323

## CoDing Sequences (CDS) congruence

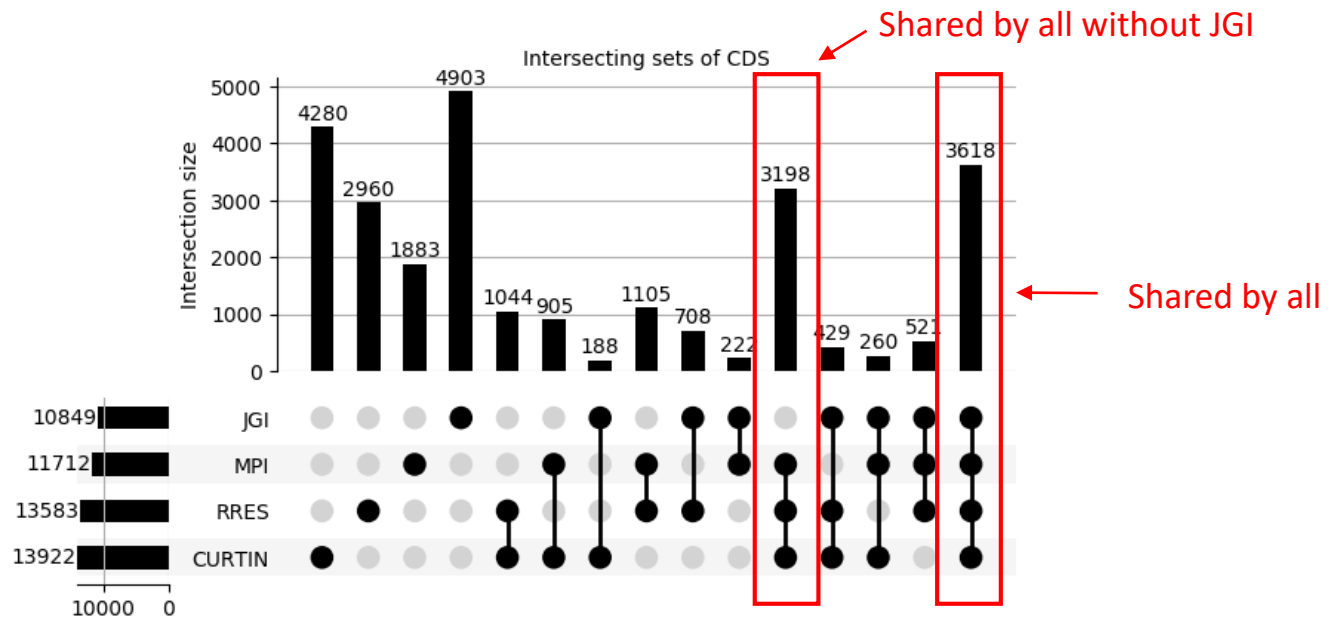


	JGI	MPI	RRES	CURTIN
<b>pairwise identical CDS</b>	JGI 10849	MPI 4621	RRES 5276	CURTIN 4495
<b>dissimilar CDS (same locus)</b>	4752	1871	2367	3844
<b>specific CDS (specific locus)</b>	151	12	593	436



# ➤ Genome annotation releases of *Z. tritici* IPO323

## CoDing Sequences (CDS) congruence

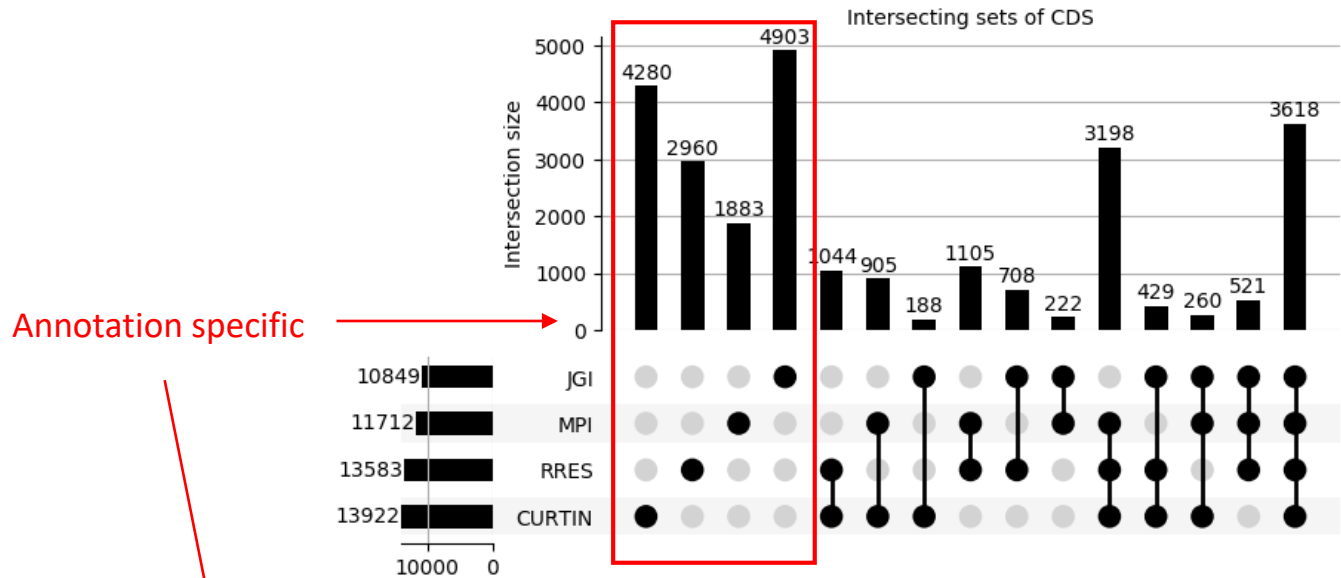


	JGI	MPI	RRES	CURTIN	
<b>pairwise identical CDS</b>	JGI	10849			
	MPI	4621	11712		
	RRES	5276	8442	13583	
	CURTIN	4495	7981	8289	13922
<b>dissimilar CDS (same locus)</b>		4752	1871	2367	3844
<b>specific CDS (specific locus)</b>		151	12	593	436



# ➤ Genome annotation releases of *Z. tritici* IPO323

## CoDing Sequences (CDS) congruence



	JGI	MPI	RRES	CURTIN
<b>pairwise identical CDS</b>	JGI 10849	MPI 11712	RRES 13583	CURTIN 13922
	JGI	10849		
	MPI	4621	11712	
	RRES	5276	8442	13583
	CURTIN	4495	7981	8289
<b>dissimilar CDS (same locus)</b>		4752	1871	2367
<b>specific CDS (specific locus)</b>		151	12	593



# ➤ Genome annotation releases of *Z. tritici* IPO323

## Overall statistics

	JGI	MPI	RRES	CURTIN	
nb_CDS	10849	11712	13583	13922	
average_CDS_length, bp	1307	<b>1465</b>	1293	1287	← Longer CDS
median_CDS_length, bp	1071	1203	1044	1041	
min_CDS_length, bp	150	150	96	93	
max_CDS_length, bp	13842	18297	18423	14523	
nb_exons	28313	29728	30772	30564	
average_exons_per_CDS	2.6	2.5	2.2	2.2	
average_exon_length, bp	531	577	570	586	
min_exon_length	2	1	1	1	
max_exon_length	12888	12975	18423	9987	
nb_transcript_mono_exon	3153	3746	<b>5233</b>	<b>5594</b>	← More mono-exons
nb_introns	17464	18016	17189	16642	
average_introns_per_transcript	1.6	1.5	1.2	1.2	
average_intron_length	<b>133</b>	93	109	92	← Longer intron
min_intron_length	11	23	4	10	
max_intron_length	<b>42135</b>	7292	<b>59574</b>	5000	← Very long introns

# ➤ Genome annotation releases of *Z. tritici* IPO323

## Overall statistics

	JGI	MPI	RRES	CURTIN	
nb_CDS	10849	11712	13583	13922	
average_CDS_length, bp	1307	<b>1465</b>	1293	1287	← Longer CDS
median_CDS_length, bp	1071	1203	1044	1041	
min_CDS_length, bp	150	150	96	93	
max_CDS_length, bp	13842	18297	18423	14523	
nb_exons	28313	29728	30772	30564	
average_exons_per_CDS	2.6	2.5	2.2	2.2	
average_exon_length, bp	531	577	570	586	
min_exon_length	2	1	1	1	
max_exon_length	12888	12975	18423	9987	
nb_transcript_mono_exon	3153	3746	<b>5233</b>	<b>5594</b>	← More mono-exons
nb_introns	17464	18016	17189	16642	
average_introns_per_transcript	1.6	1.5	1.2	1.2	
average_intron_length	<b>133</b>	93	109	92	← Longer intron
min_intron_length	11	23	4	10	
max_intron_length	<b>42135</b>	7292	<b>59574</b>	5000	← Very long introns

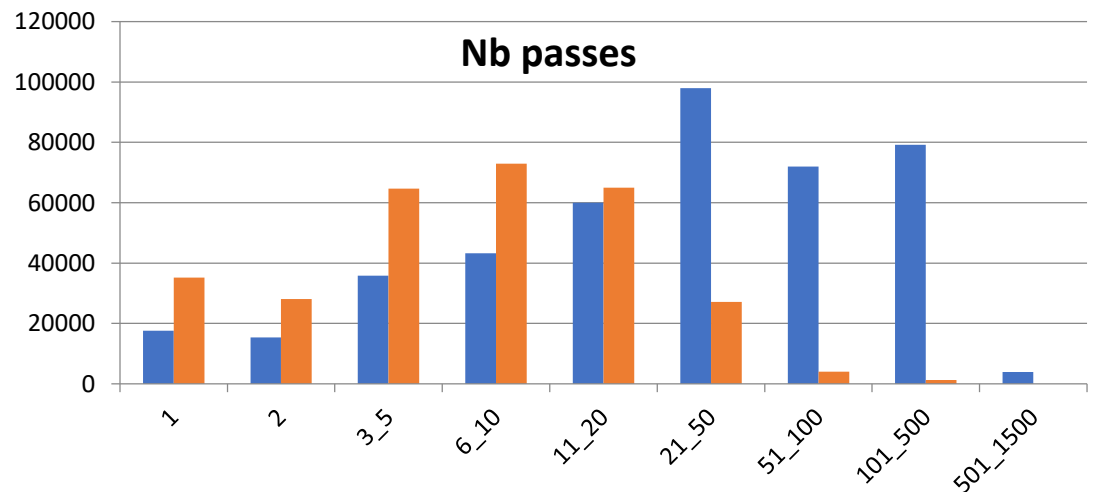
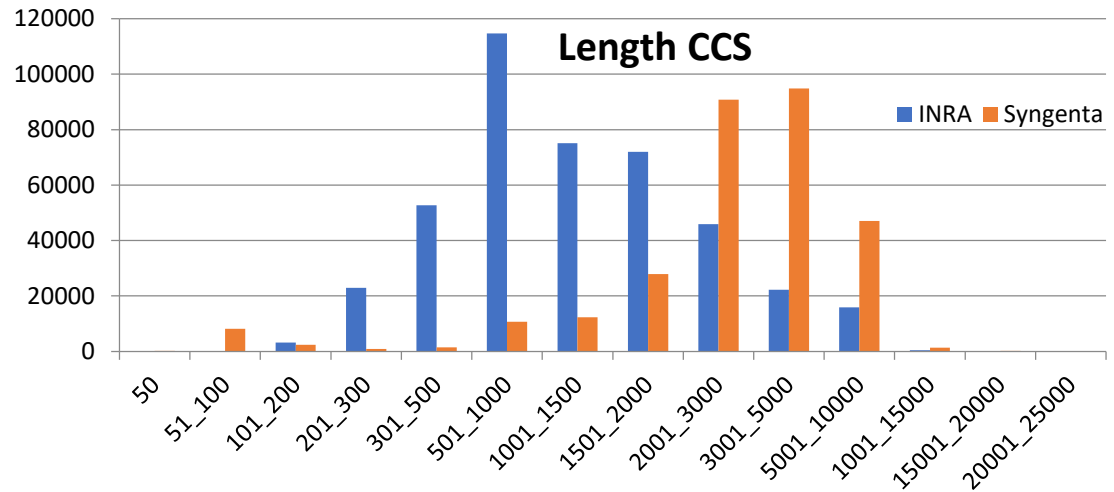
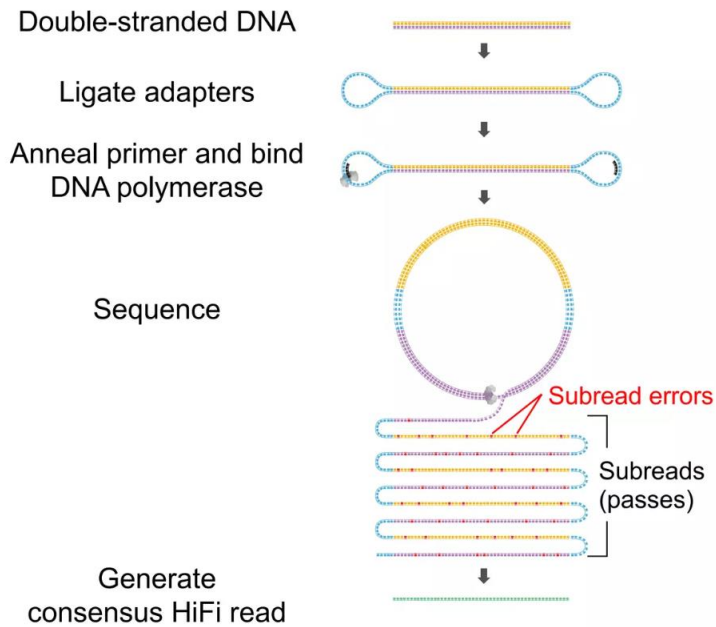
**Goal: providing a new release with the best from each annotation !**





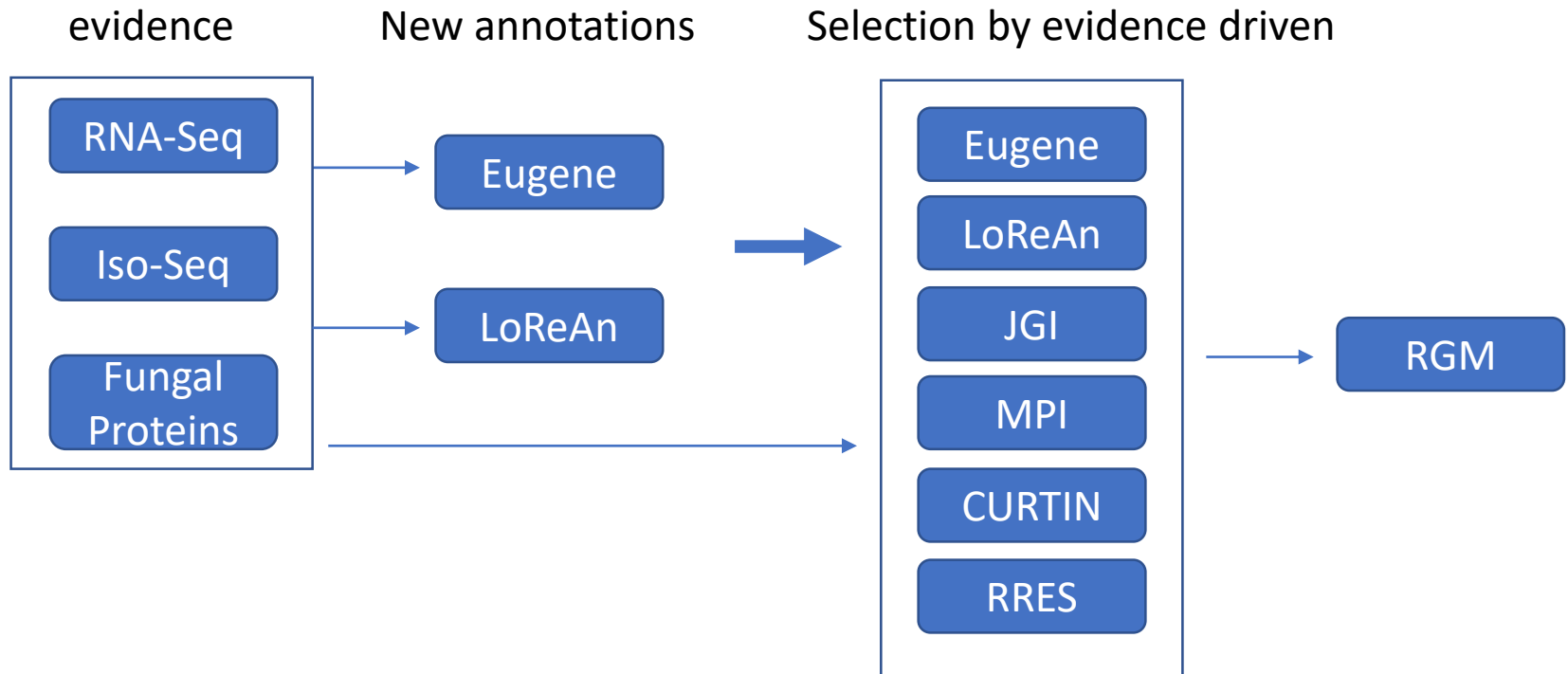
# ➤ Iso-Seq long reads: new source of evidence

Add new source of evidence to fix common errors and facilitate the selection of the best model



## ➤ Reannotated Gene Models (RGM)

Generate new gene predictions using Iso-Seq and RNA-Seq data (Eugene, LoReAn) and select the best gene models using all annotations and Iso-Seq, RNA-Seq, protein evidence



\*Sallet, E. et al 2019. EuGene: An automated integrative gene finder for eukaryotes and prokaryotes. *Methods in Molecular Biology*

\*\* Cook et al. 2019. Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing. *Plant Physiology*

## ➤ Reannotated Gene Models (RGM)

Integration of different sources / usefull tools:

- EvidenceModeler (EVM) → done with MPI annotation, need to specify a weight for each source.

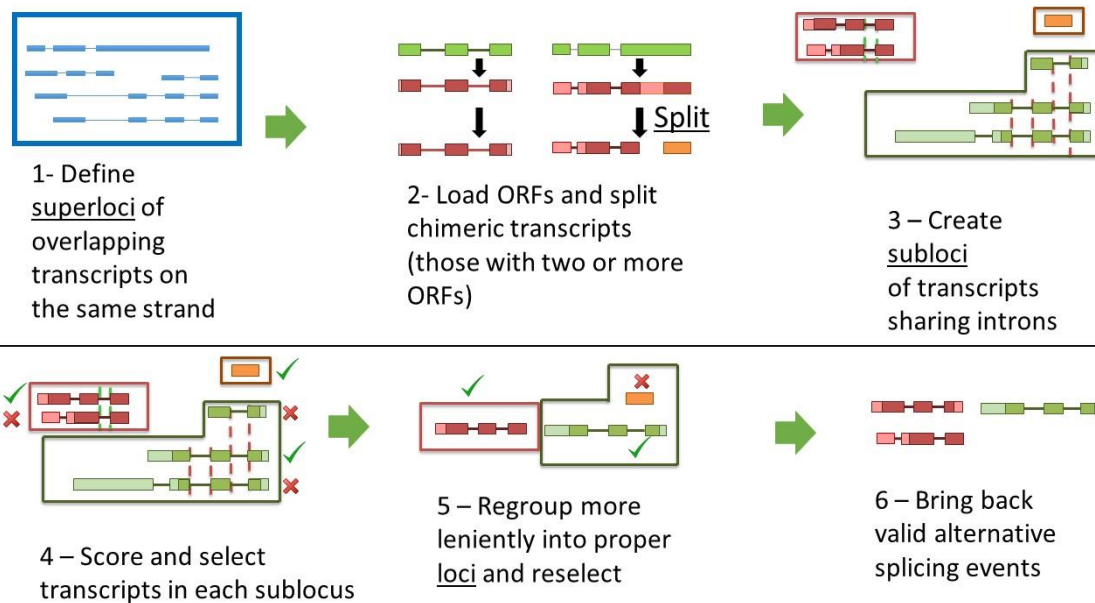
ABINITIO_PREDICTION	augustus	1
ABINITIO_PREDICTION	twinscan	1
ABINITIO_PREDICTION	glimmerHMM	1
PROTEIN	spliced_protein_alignments	1
PROTEIN	genewise_protein_alignments	2
TRANSCRIPT	spliced_transcript_alignments	1
TRANSCRIPT	PASA_transcript_assemblies	10
OTHER_PREDICTION	PASA_transdecoder	5

$$\begin{aligned} \text{Score}(a, b) = & \sum_{a \leq i \leq b} \text{ScoringVector}[i] \\ & + \sum_{\substack{\text{evidence\_end5}'=a \\ \text{evidence\_end3}'=b}} \text{featureLength} * \text{weight}(\text{evidence}) \end{aligned}$$

## ➤ Reannotated Gene Models (RGM)

Integration of different sources / usefull tools:

- Mikado 2 (2019) -> integration multiple sources, compare structure. Main goal, improve RNA-Seq transcripts with fusion detection. Complete and complexe configuration (external score available)

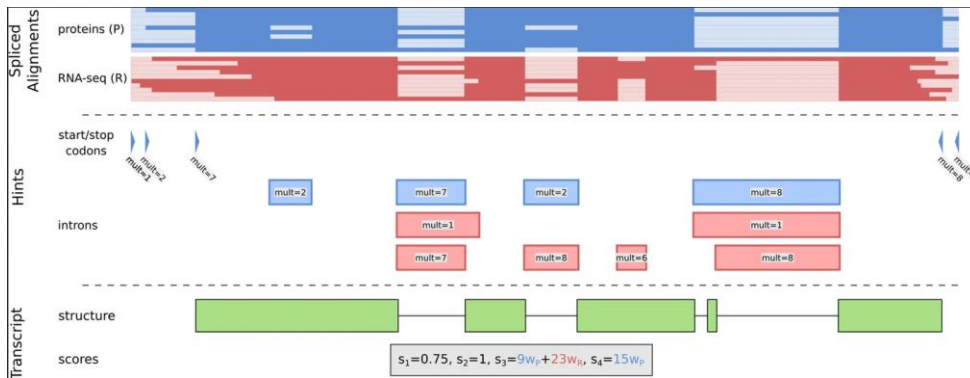


[https://mikado.readthedocs.io/en/stable/Scoring\\_files/#scoring-files](https://mikado.readthedocs.io/en/stable/Scoring_files/#scoring-files)

# ➤ Reannotated Gene Models (RGM)

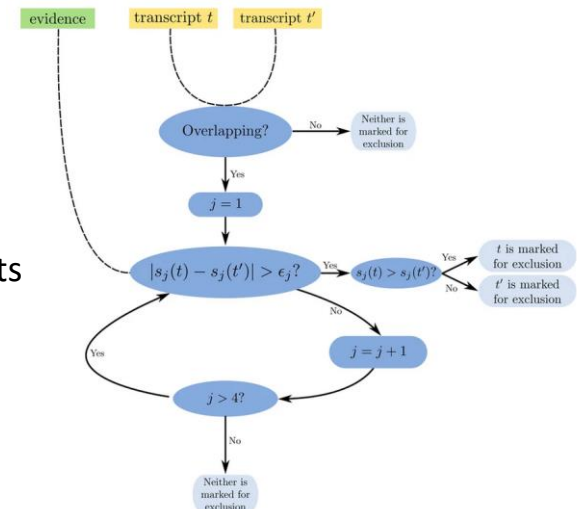
Integration of different sources / usefull tools:

- TSEBRA (2021) -> based on BRAKER outputs, improve selection of genes comparing BRAKER1/BRAKER2 outputs. => new preprint BRAKER3 (2023), fully automated pipeline ; BRAKER (AUGUSTUS, GeneMark.hmm)



transcript scores quantify the support of the transcript structure

Comparison rule for two transcripts



# ➤ How to evaluate annotation improvement ?

With reference !

TSEBRA, MIKADO:

```
Command line:
/usr/local/bin/mikado compare -r reference.gff3 -p mikado.loci.gff3 -o compare -l compare.log
7 reference RNAs in 5 genes
15 predicted RNAs in 8 genes
----- | Sn | Pr | F1 |
      Base level: 85.74 64.73 73.77
      Exon level (stringent): 63.83 42.86 51.28
      Exon level (lenient): 80.00 52.94 63.72
      Intron level: 89.47 59.65 71.58
      Intron chain level: 33.33 14.29 20.00
      Transcript level (stringent): 0.00 0.00 0.00
      Transcript level (>=95% base F1): 28.57 13.33 18.18
      Transcript level (>=80% base F1): 42.86 20.00 27.27
      Gene level (100% base F1): 0.00 0.00 0.00
      Gene level (>=95% base F1): 40.00 25.00 30.77
      Gene level (>=80% base F1): 60.00 37.50 46.15
```

```
# Matching: in prediction; matched: in reference.
```

```
      Matching intron chains: 2
      Matched intron chains: 2
      Matching monoexonic transcripts: 1
      Matched monoexonic transcripts: 1
      Total matching transcripts: 3
      Total matched transcripts: 3
```

```
      Missed exons (stringent): 17/47 (36.17%)
      Novel exons (stringent): 40/70 (57.14%)
      Missed exons (lenient): 9/45 (20.00%)
      Novel exons (lenient): 32/68 (47.06%)
      Missed introns: 4/38 (10.53%)
      Novel introns: 23/57 (40.35%)
```

```
      Missed transcripts: 0/7 (0.00%)
      Novel transcripts: 6/15 (40.00%)
      Missed genes: 0/5 (0.00%)
      Novel genes: 2/8 (25.00%)
```

- Exons, ok
- Transcripts, UTRs, Isoforms ?
- UTRs, CDS ?



## ➤ How to evaluate annotation improvement ?

Without reference ? Find metrics/measurement usefull to compare annotations

- Eugene: gff tags (Field 9) est\_cons=74.8;est\_incons=0.0, CDS (est\_cons = adequation with splicing site, est\_incons = presence of different splicing site/non-splicing site)
- Augustus: gff score, CDS , 0->1, posterior probability, ex 0.8 = sampling 80/100
- TSEBRA: transcript score
- Maker: AED = Annotation Edit Distance

AED:0.41 eAED:0.41 QI:0|0.5|0.4|1|1|1|5|0|278

- Length of the 5 UTR
- Fraction of splice sites confirmed by an EST alignment
- Fraction of exons that overlap an EST alignment
- Fraction of exons that overlap EST or Protein alignments
- Fraction of splice sites confirmed by a SNAP prediction
- Fraction of exons that overlap a SNAP prediction
- Number of exons in the mRNA
- Length of the 3 UTR
- Length of the protein sequence produced by the mRNA



## ➤ How to evaluate annotation improvement ?

Without reference ? Find metrics/measurement usefull to compare annotations

D. S. Standage and V. P. Brendel, "ParsEval: Parallel comparison and analysis of gene structure annotations," *BMC Bioinformatics*, vol. 13, no. 1, p. 187, Aug. 2012, doi: 10.1186/1471-2105-13-187.

Comparison annotations vs ref, or annotation versions. Comparison of structures, no measurement of biological relevance





## ➤ How to evaluate annotation improvement ?

Without reference ? Find metrics/measurement usefull to compare annotations

D. S. Standage and V. P. Brendel, "ParsEval: Parallel comparison and analysis of gene structure annotations," *BMC Bioinformatics*, vol. 13, no. 1, p. 187, Aug. 2012, doi: 10.1186/1471-2105-13-187.

Comparison annotations vs ref, or annotation versions. Comparison of structures, no measurement of biological relevance

D. S. Standage, "AEGeAn: an integrated toolkit for analysis and evaluation of annotated genomes," 2015.  
<http://standage.github.io/AEGeAn>. => GAEVAL

**Compute coverage score**: percentage of nucleotides in exons that have coverage from one or more transcript alignments.

**Compute integrity score**:

- $A$ : the percentage of introns confirmed by an alignment gap; for single-exon gene predictions lacking introns,  $A$  represents the ratio of the observed CDS length to the expected CDS length (with a maximum of 1.0)
- $B$ : the exon coverage
- $\Gamma$ : the ratio of the observed 5' UTR length to the expected 5' UTR length (with a maximum of 1.0)
- $E$ : the ratio of the observed 3' UTR length to the expected 3' UTR length (with a maximum of 1.0)

A weight is applied to each of these 4 values, and the final integrity score  $\Phi$  is computed as follows.

$$\Phi = \alpha A + \beta B + \gamma \Gamma + \epsilon E$$

The sum of the weights must be 1.0, and their default values are as follows.

- $\alpha = 0.6$
- $\beta = 0.3$
- $\gamma = 0.05$
- $\epsilon = 0.05$

Expected lengths for UTRs and CDSs should be determined empirically. The original GAEVAL tool calculated these values as the length achieved by 95% of the evaluated features.



## ➤ How to evaluate annotation improvement ?

Without reference ? Find metrics/measurement usefull to compare annotations

D. S. Standage and V. P. Brendel, "ParsEval: Parallel comparison and analysis of gene structure annotations," *BMC Bioinformatics*, vol. 13, no. 1, p. 187, Aug. 2012, doi: 10.1186/1471-2105-13-187.

Comparison annotations vs ref, or annotation versions. Comparison of structures, no measurement of biological relevance

D. S. Standage, "AEGeAn: an integrated toolkit for analysis and evaluation of annotated genomes," 2015.  
<http://standage.github.io/AEGeAn>. => GAEVAL

**Compute coverage score**: percentage of nucleotides in exons that have coverage from one or more transcript alignments.

**Compute integrity score**:

- $A$ : the percentage of introns confirmed by an alignment gap; for single-exon gene predictions lacking introns,  $A$  represents the ratio of the observed CDS length to the expected CDS length (with a maximum of 1.0)
- $B$ : the exon coverage
- $\Gamma$ : the ratio of the observed 5' UTR length to the expected 5' UTR length (with a maximum of 1.0)
- $E$ : the ratio of the observed 3' UTR length to the expected 3' UTR length (with a maximum of 1.0)

A weight is applied to each of these 4 values, and the final integrity score  $\Phi$  is computed as follows.

$$\Phi = \alpha A + \beta B + \gamma \Gamma + \epsilon E$$

The sum of the weights must be 1.0, and their default values are as follows.

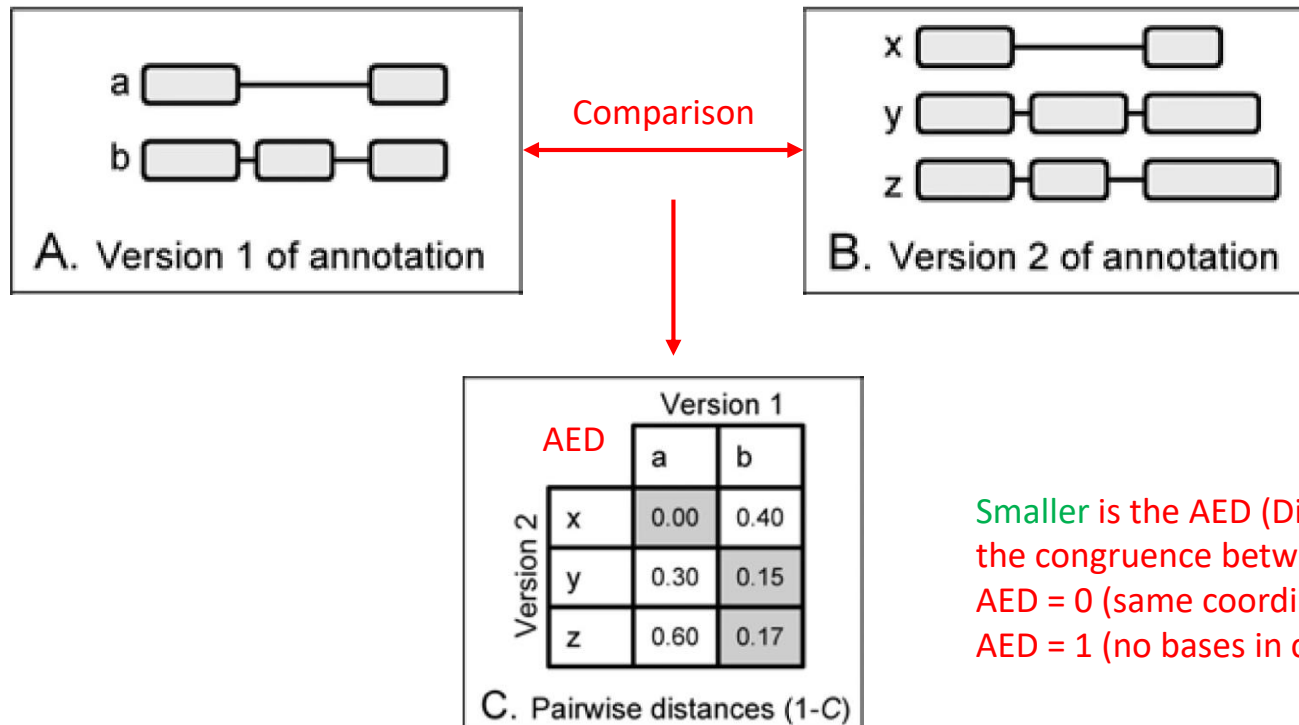
- $\alpha = 0.6$
- $\beta = 0.3$
- $\gamma = 0.05$
- $\epsilon = 0.05$

Expected lengths for UTRs and CDSs should be determined empirically. The original GAEVAL tool calculated these values as the length achieved by 95% of the evaluated features.



## ➤ Annotation Edit Distance (AED)

AED proposed by MAKER\* developers to evaluate the match between different gene predictions

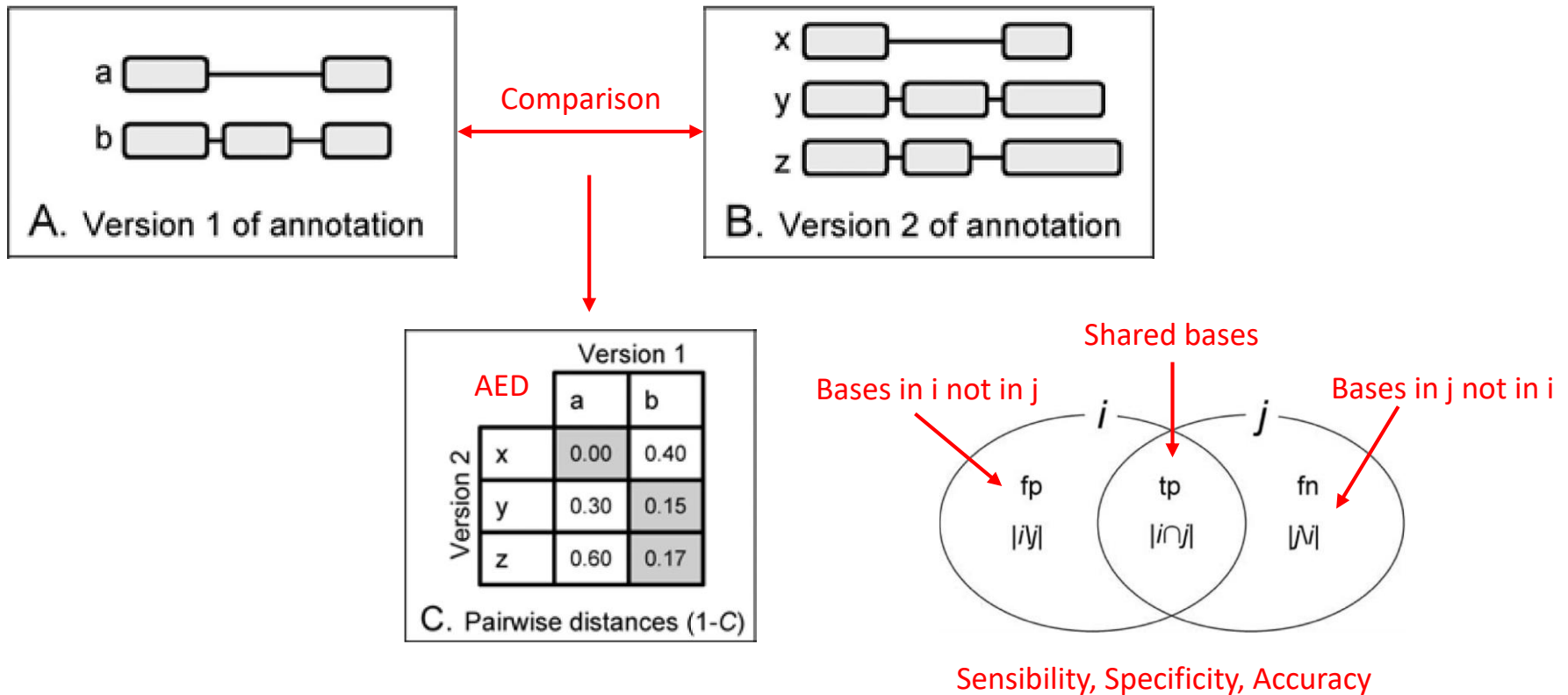


Smaller is the AED (Distance), stronger is the congruence between annotations.  
AED = 0 (same coordinates) = identical  
AED = 1 (no bases in common)

\*Eilbeck, K., Moore, B., Holt, C., and Yandell, M. 2009. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics. 10:67

## ➤ Annotation Edit Distance (AED)

AED proposed by MAKER\* developers to evaluate the match between different gene predictions

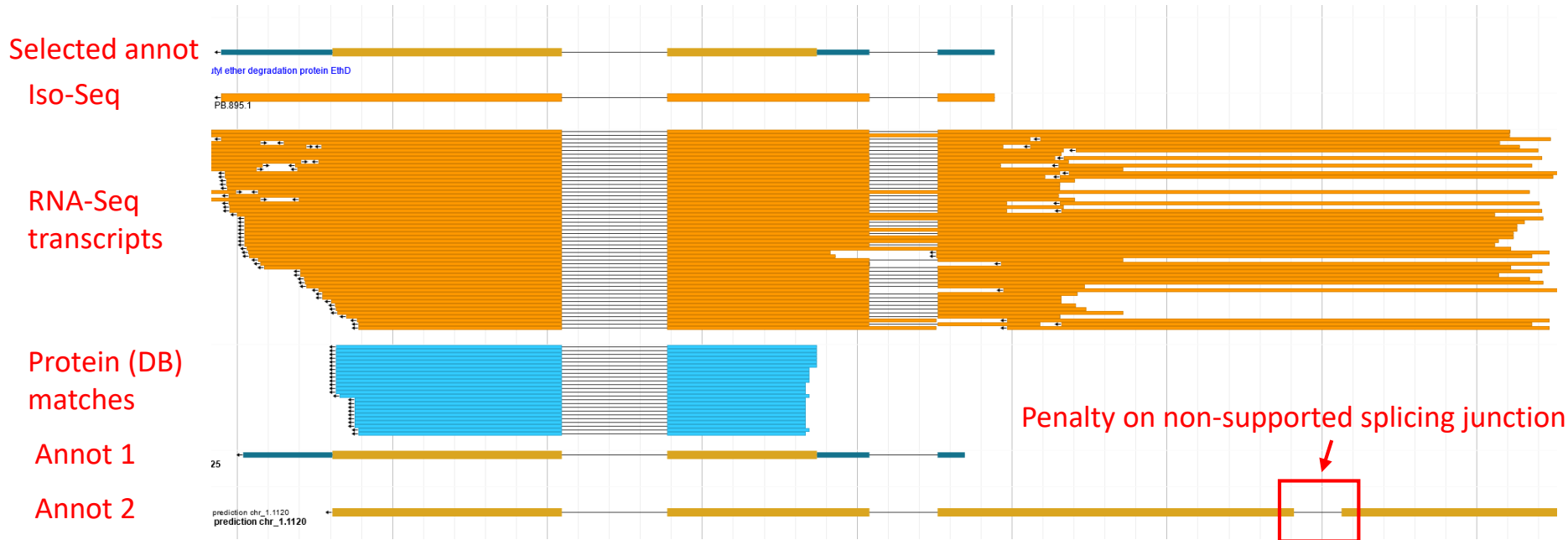


\*Eilbeck, K., Moore, B., Holt, C., and Yandell, M. 2009. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics. 10:67

# ➤ AED in InGenAnnot (Inspection of Gene Annotation)

Adapted AED to evaluate distance between gene models and evidence

Selection of best gene models according to AED scores



	Annot 1	Annot 2
AED_RNA-Seq	0.00738 (penalty=NO)	0.3007 (penalty=YES)
AED_protein	0.0049	0.3164
AED_long-read	0.1839 (penalty=NO)	0.6042 (penalty=YES)

Ranking

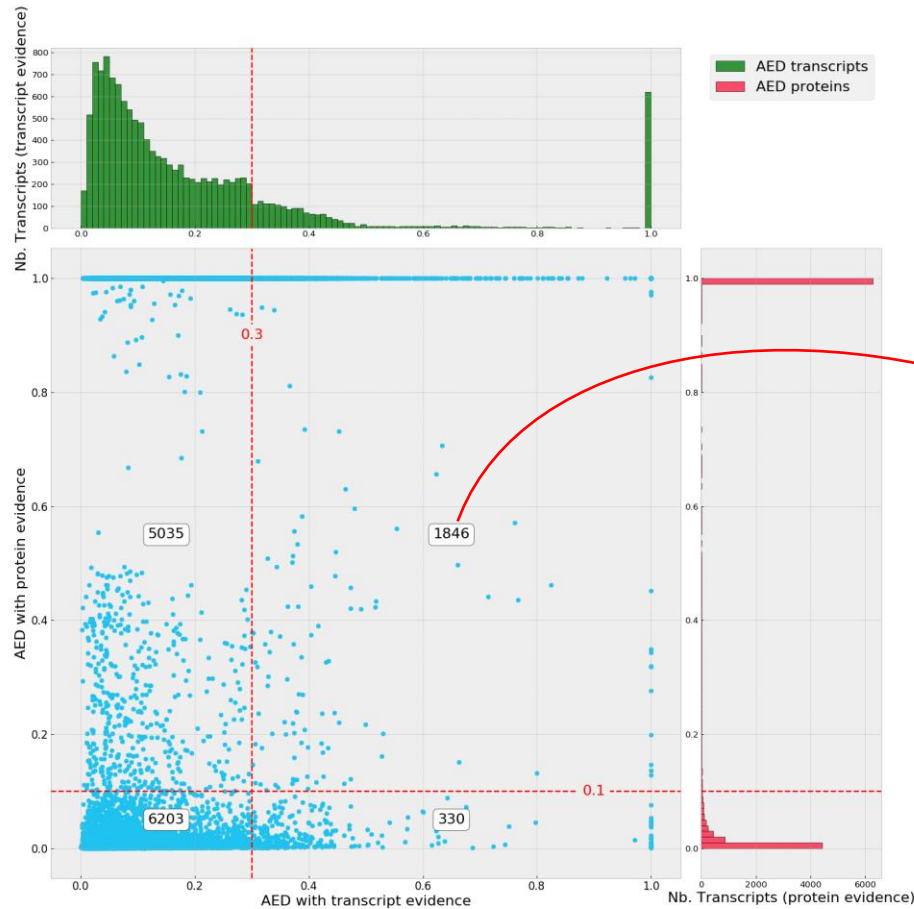


Annot1, Annot2



# ➤ New annotation release driven by evidence

RGM release (Reannotation Gene Models) => 13414 gene models



Selection thresholds:

AED transcripts  $\leq 0.3$

AED proteins  $\leq 0.1$

11568 genes

1846 rescued genes (locus supported by 4 annotations)

⇒ 574 genes no AED (fully ab-initio), 224 co-localized in a specific region of chromosome 7 (repressive histone marks\* as mini-chromosomes)

\*Schotanus et al. 2015. Histone modifications rather than the novel regional centromeres of *Zyoseptoria tritici* distinguish core and accessory chromosomes. *Epigenetics and Chromatin*



## ➤ Genome annotation release comparison

RGM displays coherent statistics

	JGI	MPI	RRES	CURTIN	RGM
nb_CDS	10849	11712	13583	13922	13414
average_CDS_length, bp	1307	<b>1465</b>	1293	1287	1287
median_CDS_length, bp	1071	1203	1044	1041	1041
min_CDS_length, bp	150	150	96	93	102
max_CDS_length, bp	13842	18297	18423	14523	16506
nb_exons	28313	29728	30772	30564	30946
average_exons_per_CDS	2.6	2.5	2.2	2.2	2.3
average_exon_length, bp	531	577	570	586	<b>782</b>
min_exon_length	2	1	1	1	1
max_exon_length	12888	12975	18423	9987	16680
nb_transcript_mono_exon	3153	3746	<b>5233</b>	<b>5594</b>	4850
nb_introns	17464	18016	17189	16642	17532
average_introns_per_transcript	1.6	1.5	1.2	1.2	1.3
average_intron_length	<b>133</b>	93	109	92	73
min_intron_length	11	23	4	10	5
max_intron_length	<b>42135</b>	7292	<b>59574</b>	5000	3166

← UTRs

## ➤ Genome annotation release comparison

### BUSCO

<b>BUSCO category</b>	<b>JGI</b>	<b>MPI</b>	<b>CURTIN</b>	<b>RRES</b>	<b>RGM</b>
<b>Complete BUSCOs (C)</b>	1633	1679	1681	1693	1696
<b>Complete BUSCOs (C) %</b>	95.7%	98.4%	98.5%	99.2%	99.4%
<b>Complete and single-copy BUSCOs (S)</b>	1632	1678	1615	1692	1695
<b>Complete and duplicated BUSCOs (D)</b>	1	1	66	1	1
<b>Fragmented BUSCOs (F)</b>	25	3	8	5	2
<b>Missing BUSCOs (M)</b>	48	24	17	8	8
<b>Total BUSCO groups</b>			1706		

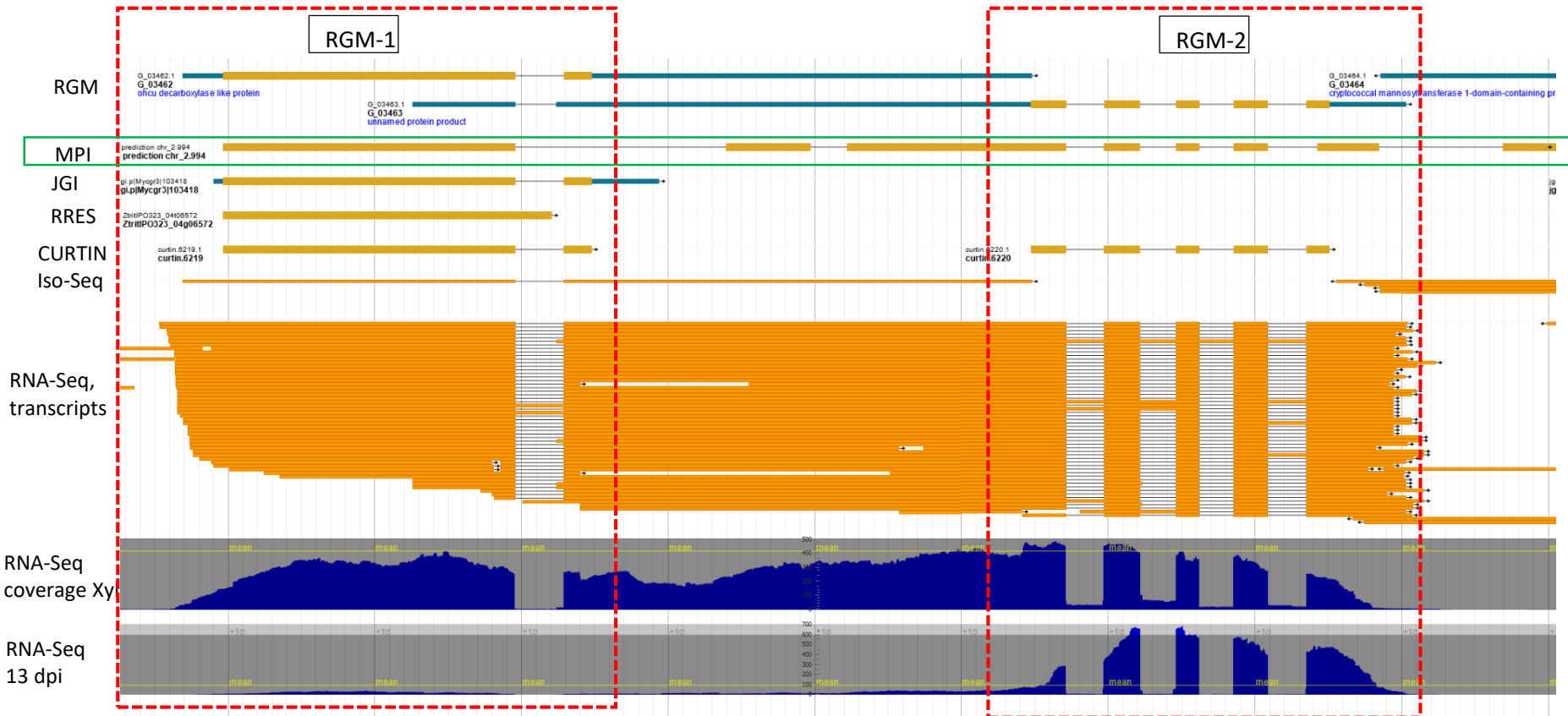
**BUSCO: not enough to evaluate improvement !**





# ➤ Gene model Improvements

Fixing fused genes specific of MPI annotation

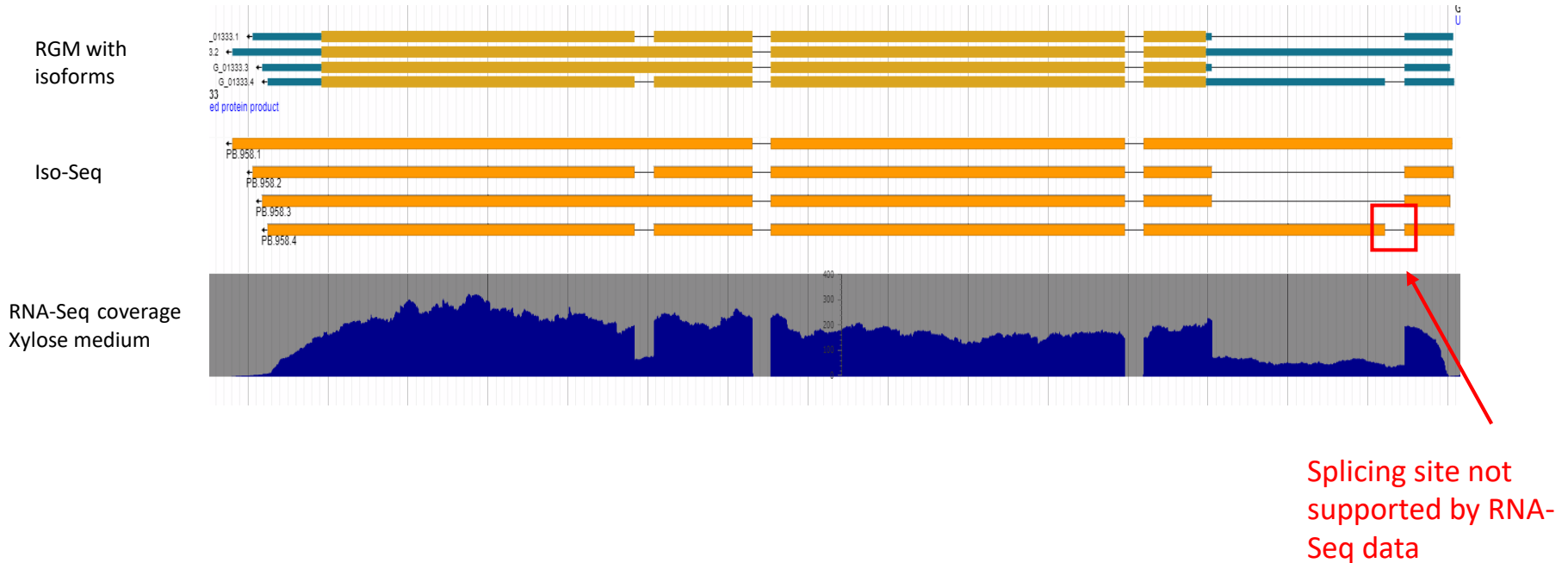


Iso-Seq (RGM-1) and RNA-Seq (infection 13 dpi, RGM-2) identified two distinct genes



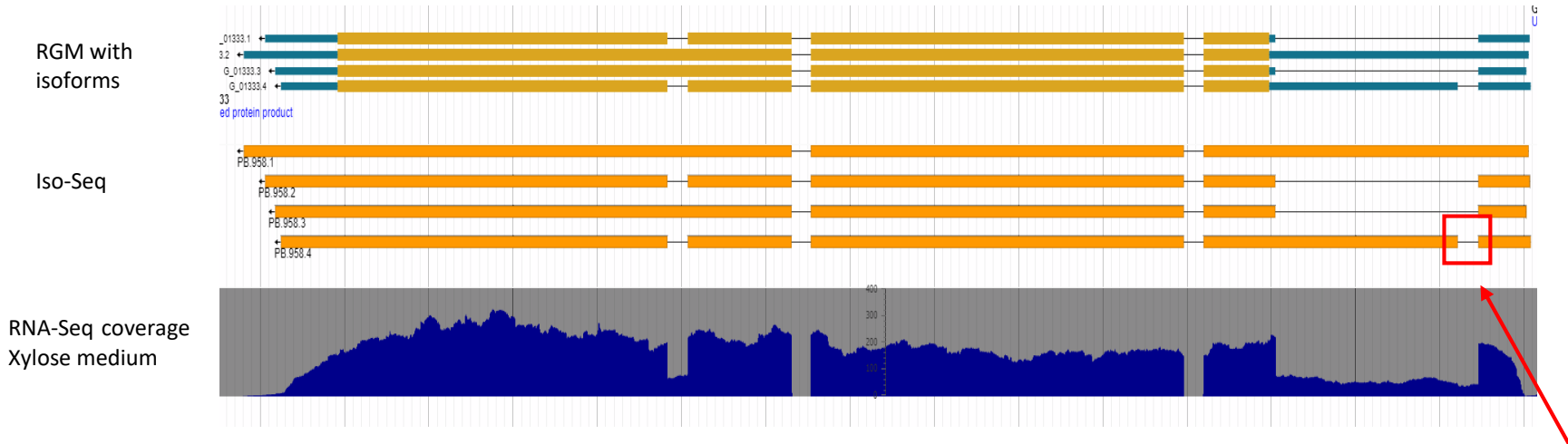
## ➤ Alternative splicing: Iso-Seq + RNA-Seq

Iso-Seq detection of transcript isoforms validated by RNA-Seq (expression level)



# ➤ Alternative splicing: Iso-Seq + RNA-Seq

Iso-Seq detection of transcript isoforms validated by RNA-Seq (expression level)



categories	counts
genic <sup>1</sup>	664
Intron retention (IR)	1571
novel_in_catalog (NIC) <sup>2</sup>	7
novel_not_in_catalog (NNC) <sup>3</sup>	474

➔ 1342 genes with at least one isoform supported by Iso-Seq

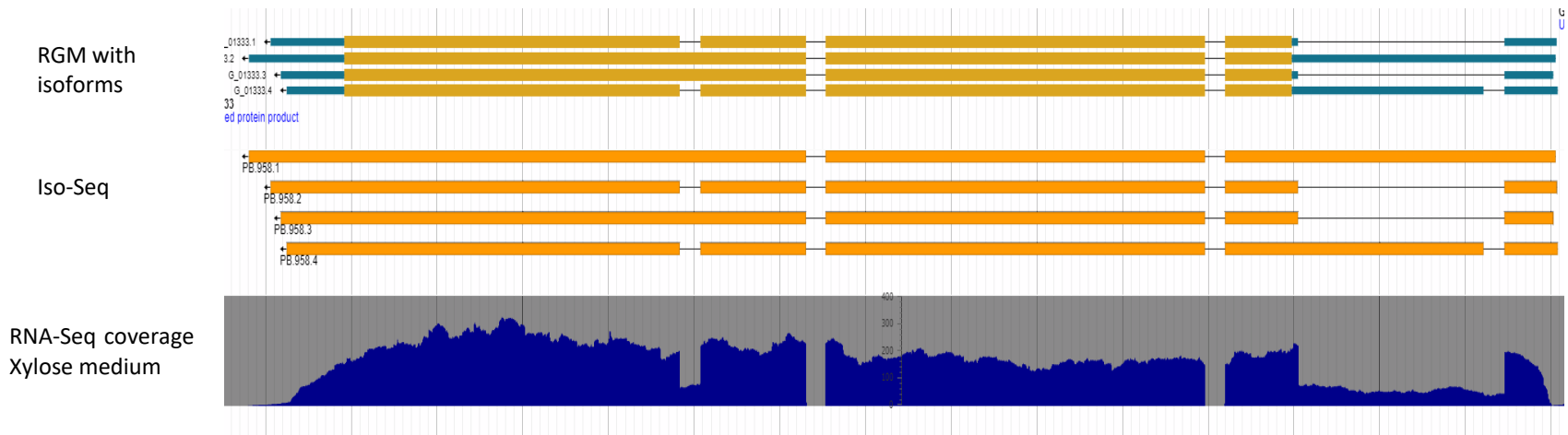
Splicing site not supported by RNA-Seq data

1: partial overlaps of intron and exons not compliant with intron/exons coordinates  
 2: use combination\_of\_known\_splice sites  
 3: at\_least\_one\_novel\_splice site detected



## ➤ Alternative splicing: Iso-Seq + RNA-Seq

Iso-Seq detection of transcript isoforms validated by RNA-Seq (expression level)



categories	counts
genic <sup>1</sup>	664
Intron retention (IR)	1571
novel_in_catalog (NIC) <sup>2</sup>	7
novel_not_in_catalog (NNC) <sup>3</sup>	474

➔ Differential Isoform Usage (DIU) ➔

Few DIU statistically validated between in vitro and infection. Only few reads during infection: bias  
No clear signals

<sup>1</sup>: partial overlaps of intron and exons not compliant with intron/exons coordinates

<sup>2</sup>: use combination\_of\_known\_splice sites

<sup>3</sup>: at\_least\_one\_novel\_splice site detected



## Tool suite to select, compare and filter gene annotation

<https://bioger.pages.mia.inra.fr/ingenannot>

<https://pypi.org/project/ingenannot/>

**Select best gene models from predictions**

The first goal of ingenannot is to help you in gene selection and curation when you ran several gene predictors. Many tools are available to predict gene structure with variable sensibility / specificity. In the same way as EvidenceModeler<sup>1</sup>, ingenannot propose a tool to select the best gene model fitting transcriptomic or protein evidence. The selection is based on the best Annotation Evidence Distance (AED) described in this paper<sup>2</sup>. Non supported splicing-junction could be penalized to maximize the suitability to evidence. If you only want evidence supported gene models, you can set thresholds of required AED and if you want to rescue fully ab-initio gene models, you can define a minimal number of source to keep a gene in a locus without evidence.

**Workflow:**

```

    graph TD
      A[Assemble transcripts] --> B[Prepare / Validate data]
      C[Select long read isoforms] --> B
      D[Align proteins] --> B
      B --> E[Filter TE genes]
      E --> F[AED annotation]
      F --> G[Select]
      G --> H[Rescue effectors]
      G --> I[Compare the selection with all sources]
      G --> J[Add UTRs]
      H --> K[Add other isoforms]
      I --> K
      J --> K
  
```

**Steps:**

- 1) Generate / Assemble new transcripts from RNA-Seq / long reads

**ingenannot 0.0.4**

pip install ingenannot

Demière version : 9 Nov. 2022

InGenAnnot: Inspection of Gene Annotation

**Description du projet**

InGenAnnot is a set of utilities to inspect and generate statistics for one or several sets of gene annotations. It allows structure comparison and can help you to prioritize your efforts in manual curation. InGenAnnot uses among other things, the Sequence Ontology gene-splicing classification SO [1] that aims to classify alternative transcripts in seven categories or the Annotation Edit Distance AED [2] proposed as a metric for evidence support.

As several approaches and tools exist to annotate genes in newly assembled genomes, it could be useful to compare predictors and extract best evidence supported.

InGenAnnot can handle multiple gffs from different sources. In case of several annotations, gene boundaries are often divergent (especially if you tried to predict UTR regions), that implies to clusterize genes, to propose new loci sharing a list of transcripts. We define these new loci as 'meta-gene' and propose several options to clusterize them. We tried to summarize the pro and cons of classification feature type in the following table.

	pros	cons
==c-lu-type: gene	detect problem of missens predictions	overlaps of UTR merge different genes, not suitable for compact genomes
==c-lu-type: cds	detect problem of missens predictions	could not correct splitted CDS
==c-lu-type: gene ==c-lu-stranded	resolve conflict between genes and possible non-coding RNA on the opposite strand	will not detect severe problem due to divergent prediction on opposite strand, overlaps of UTR merge different genes
==c-lu-type: cds ==c-lu-stranded	resolve conflict between genes and possible non-coding RNA on the opposite strand	will not detect severe problem due to divergent prediction on opposite strand

In most cases, we recommended to use ==c-lu-type: cds with ==c-lu-stranded to avoid gene merge. A post-process is implemented to remove overlapping CDS, keeping gene models with the best AED scores.

**Selection of best gene structures, evidence-driven with Annotation Edit Distance (AED)**

Annotation Edit Distance AED [2] was proposed as metric for gene annotation prediction and was implemented in MAKER [3] to filter out predicted models based on their AED. Here we propose some options which modify the computation of this distance and take into account the different sources of evidence. All gene predictor tools are not able to predict UTRs, despite the RNA-Seq data and Long-read based transcripts. So to avoid penalizing gene models limited to CDS, we implemented an overflow penalty parameter to maximize the score of model fitting best with transcript evidence despite missing UTRs. In addition, we compute separately the AED with transcript and proteomic evidence. Some genes are only supported with a transcript evidence (new/specific genes), a protein evidence (gene not expressed in our data), or in both type of evidences. Then to select the best model, we classified genes according to their AED for tr and pr separately. In case where the first gene is the same in the both ranking, we select this last.

SO classification, Isoform selection, UTR, rescue effectors ...



# ➤ InGenAnnot

Tool suite to select, compare and filter gene annotation

<https://bioger.pages.mia.inra.fr/ingenannot>

<https://pypi.org/project/ingenannot/>

The screenshot displays the InGenAnnot web interface. On the left is a dark sidebar with the InGenAnnot logo and a search bar. Below the search bar are sections for 'Installing InGenAnnot', 'USE CASES' (listing tasks like adding isoforms and UTRs), and 'TOOLS' (listing various command-line tools like 'aed', 'clusterize', etc.). The main content area on the right has a header 'Gff\_file in GFF/GTF format.' followed by a section titled 'outputs'. This section lists four expected output files: 'isoforms.ranking.gff', 'isoforms.top.gff', 'isoforms.alternatives.gff', and 'isoforms.unclassif.gff'. Below the list, there is a paragraph explaining the 'isoform\_ranking' groups and a reference to an example visualization. The visualization itself shows a genomic track with exons and introns, overlaid with numerous colored bars representing different isoforms, and a bottom section showing specific isoform models labeled P17021 through P17024.

Isoform ranking: select best isoform to predict gene models



INRAE

Journées PEPI IBIS 2023

Nicolas Lapalu

# InGenAnnot

Tool suite to select, compare and filter gene annotation

<https://bioger.pages.mia.inra.fr/ingenannot>

<https://pypi.org/project/ingenannot/>

InGenAnnot

Search docs

Installing InGenAnnot

USE CASES

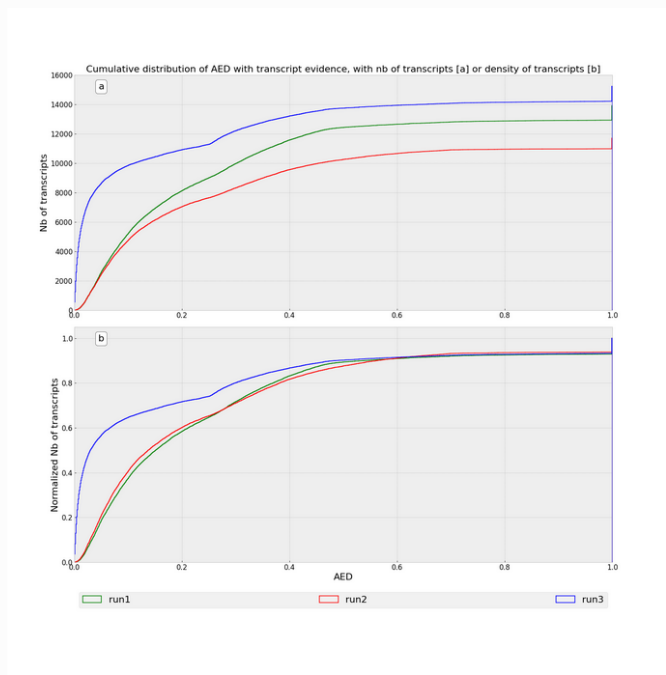
- Add new gene isoforms to your annotations
- Add UTRs to gene models
- Comparison of different annotation datasets
- Evaluate gene annotations
- Find new potential Small Secreted Proteins (SSP)
- Select best gene models from predictions

TOOLS

- aed
- aed\_compare
- usage
- inputs
- outputs
- aed\_strand\_annotation\_filter
- add\_sqanti3\_isoforms
- clusterize
- compare
- curation
- effector\_predictor
- exonerate\_to\_gff
- filter
- isoform\_ranking
- rename
- rename\_effector

comparison of aed score between datasets. You can compute some metrics as ... with `-statistics`

Output example of plot showing cumulative transcript aed scores:



Output of statistics:

#	mean (tr)	median (tr)	stdev (tr)	mean (pr)	median (pr)	stdev (pr)	median
run1	0.244	0.152	0.258	0.478	0.176	0.480	0.176
run2	0.239	0.138	0.255	0.444	0.089	0.476	0.154

AED comparison between several annotation sets. Cumulative AED like in MAKER2 publication: [10.1186/1471-2105-12-491](https://doi.org/10.1186/1471-2105-12-491)



# ➤ InGenAnnot

Tool suite to select, compare and filter gene annotation

<https://bioger.pages.mia.inra.fr/ingenannot>

<https://pypi.org/project/ingenannot/>

The screenshot shows the InGenAnnot web interface. On the left is a dark sidebar with a search bar and a list of navigation items under 'USE CASES' and 'TOOLS'. The main content area is titled 'inputs' and 'outputs'. Under 'inputs', it describes a 'File of Files (FoF)' format. Under 'outputs', it lists 'Statistics for each category:' and 'Categories defined by the SO such:'. Three categories are shown, each with a table and a corresponding gene model diagram.

Class	definition
N:0:0	No transcript-pairs share any exon sequence

Class	definition
N:N:0	Some transcript-pairs share sequence, but none have common exon boundaries

Class	definition
N:0:N	Some transcript-pairs share no sequence, others have common exon boundaries

SO classification : gene overlaps, aberrant mRNA isoforms



INRAE

Journées PEPI IBIS 2023

Nicolas Lapalu



# ➤ InGenAnnot

Tool suite to select, compare and filter gene annotation

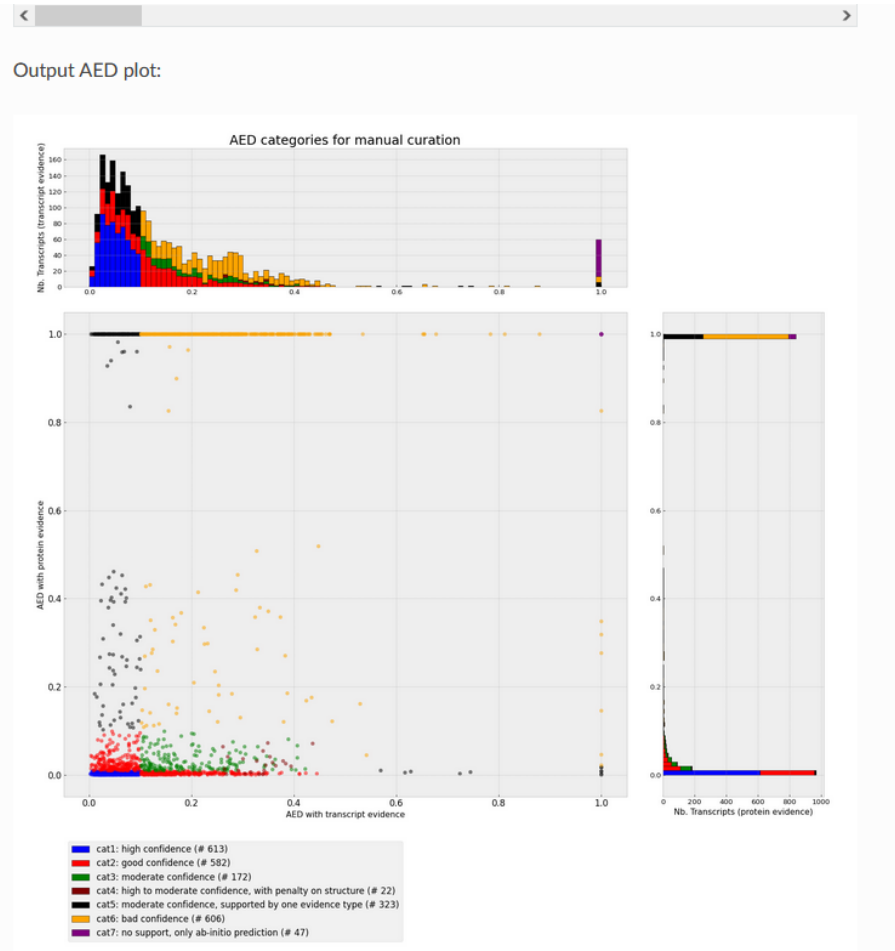
<https://bioger.pages.mia.inra.fr/ingenannot>

<https://pypi.org/project/ingenannot/>

- Add new gene isoforms to your annotations
  - Add UTRs to gene models
  - Comparison of different annotation datasets
  - Evaluate gene annotations
  - Find new potential Small Secreted Proteins (SSP)
  - Select best gene models from predictions
- TOOLS**
- aed
  - aed\_compare
  - aed\_strand\_annotation\_filter
  - add\_sqanti3\_isoforms
  - clusterize
  - compare

## ☰ curation

- usage
  - inputs
  - outputs
- effector\_predictor
  - exonerate\_to\_gff
  - filter
  - isoform\_ranking
  - rename
  - rescue\_effectors



Manual curation:  
prioritization

## > Conclusions & Perspectives

Iso-Seq sequencing method is a complementary method to RNA-Seq:

- Robust isoform detection
- UTR inference
- Resolution of overlapping genes coordinates (dense genome)
- Full lncRNA sequencing

Drawbacks:

- highlights rare transcripts / transcription machinery errors -> need quantitative control (RNA-Seq)

*Zymoseptoria tritici* IPO323 new release: RGM (preprint + InGenAnnot)

<https://doi.org/10.1101/2023.04.26.537486>

InGenAnnot: new suite of tools to deal with gene annotations.

Benchmark analysis in progress versus EvidenceModeler, TSEBRA, gaeval (AEGeAn) , FINDER.

Snakemake workflow in progress: comparison of annotations or evaluation of an annotation release.



## ➤ Conclusions & Perspectives

Snakemake workflow in progress: comparison of annotations or evaluation of an annotation release.

The screenshot displays a web interface for a Snakemake workflow. On the left, a sidebar menu includes 'Input data', 'Gene features', 'AED metrics', and 'AED curation thresholds'. The main content area is divided into two sections: 'Input data' and 'Gene features'. The 'Input data' section shows a table with columns 'source' and 'file', listing 'run1' and 'run2' with their respective annotation files. The 'Gene features' section shows a table with columns 'feature', 'run1', and 'run2', listing various gene features and their values for the two runs.

source	file
run1	run1_annotations.gff
run2	run2_annotations.gff






feature	run1	run2
average_CDS_length	1360.24	1351.41
average_exon_length	779.27	774.45
average_exons_per_transcript	2.41	2.41
average_five_prime_utr_length	285.29	286.67
average_gene_length	1980.00	1969.70
average_intron_length	74.78	74.68
average_introns_per_transcript	1.41	1.41
average_three_prime_utr_length	359.42	354.70
average_transcript_length	1980.00	1969.70

VARUS: Stanke, M., Bruhn, W., Becker, F. *et al.* VARUS: sampling complementary RNA reads from the sequence read archive. *BMC Bioinformatics* **20**, 558 (2019). <https://doi.org/10.1186/s12859-019-3182-x>




## ➤ Conclusions & Perspectives

<https://bsapubs.onlinelibrary.wiley.com/doi/10.1002/aps3.11533>

APPLICATION ARTICLE |  Open Access |    

### Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes

Vidya S. Vuruputoor, Daniel Monyak, Karl C. Fetter, Cynthia Webster, Akriti Bhattarai, Bikash Shrestha, Sumaira Zaman, Jeremy Bennett, Susan L. McEvoy, Madison Caballero, Jill L. Wegrzyn 

First published: 08 August 2023 | <https://doi.org/10.1002/aps3.11533> | Citations: 2

Services SFX pour INRAE



**INRAE**

Journées PEPI IBIS 2023

Nicolas Lapalu

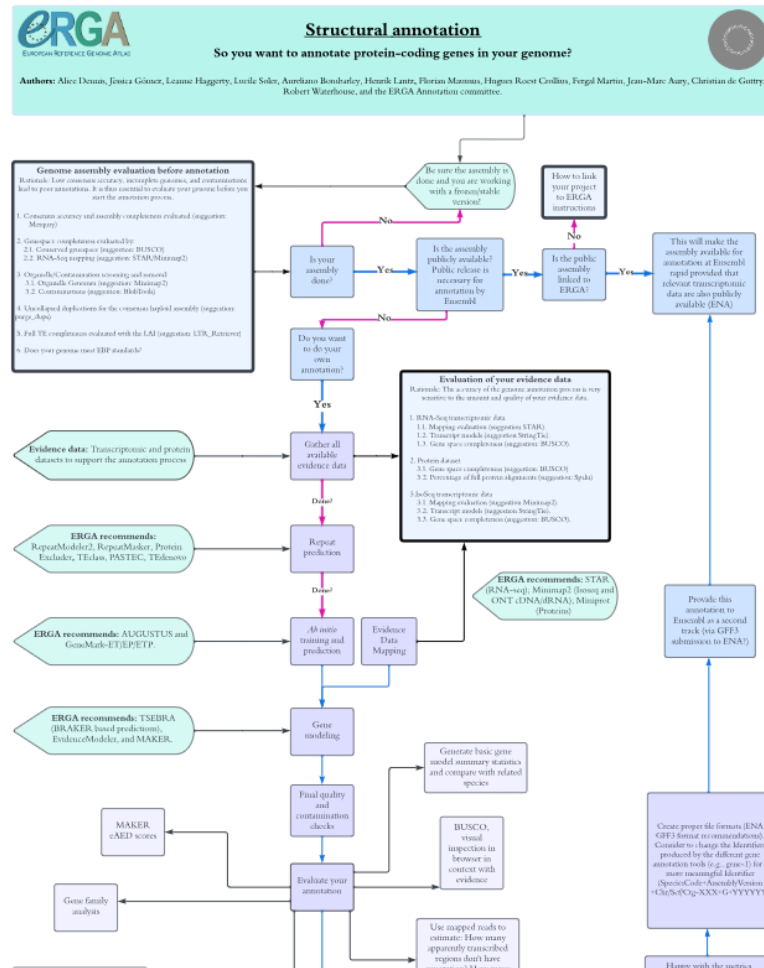
# ➤ Conclusions & Perspectives

<https://www.erga-biodiversity.eu/structural-annotation>



## Structural annotation - So you want to annotate protein-coding genes in your genome?

Version 1.0 - August 2023



INRAE

Journées PEPI IBIS 2023

Nicolas Lapalu

## > Acknowledgements

**Gabriel Scalliet  
and Syngenta  
bioinformatics**



**Lucie Lamothe**



**Yohann Petit**



**Marc-Henri Lebrun**



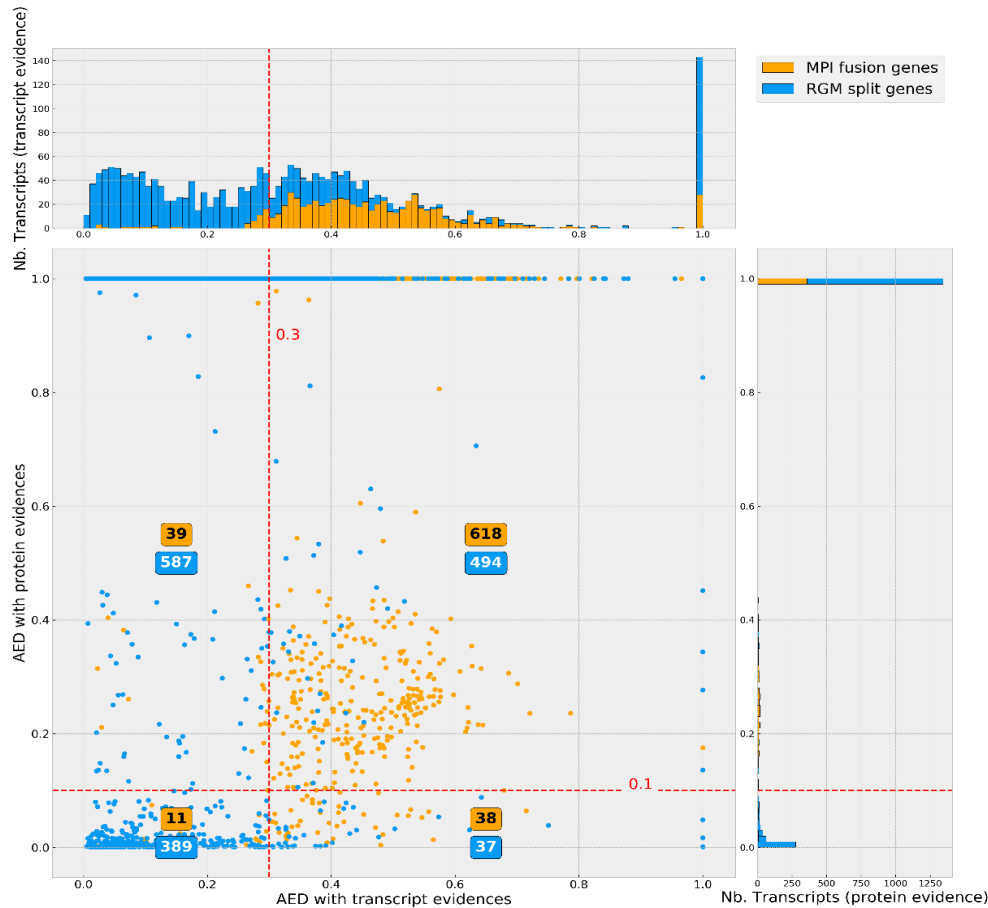
**INRAE**

Journées PEPI IBIS 2023

Nicolas Lapalu

# ➤ Gene model Improvements

Fixed many fused genes in MPI annotation

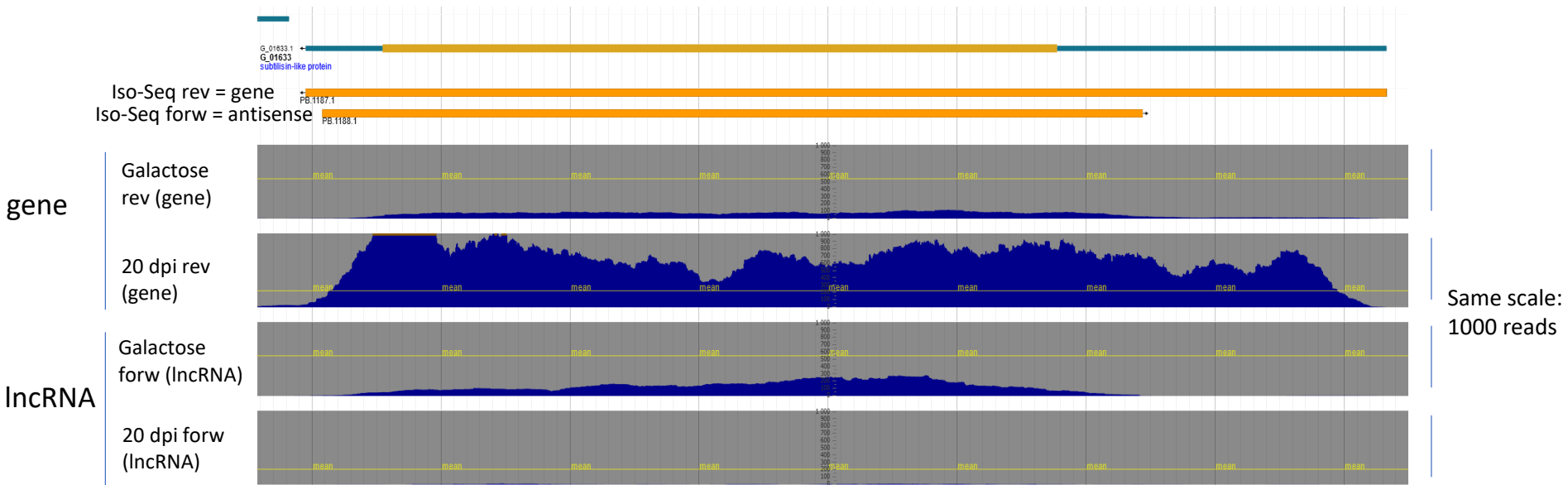


MPI 706 genes  
vs  
RGM 1507 genes



# ➤ Identification of Long Non-Coding RNAs using Iso-Seq

Identification of 51 reliable Long Non-Coding RNA (lncRNA) validated by RNA-Seq (mainly antisense transcripts)



Negative correlation between the expression of the subtilisin-like coding gene and its antisense:

Infection: subtilisin (up) antisense (down)

In vitro: subtilisin (down) antisense (up)

Hypothesis: Negative control of subtilisin-like coding gene expression by the antisense lncRNA.

Subtilisins are secreted proteases playing an important role in plant infection\*, \*\*.

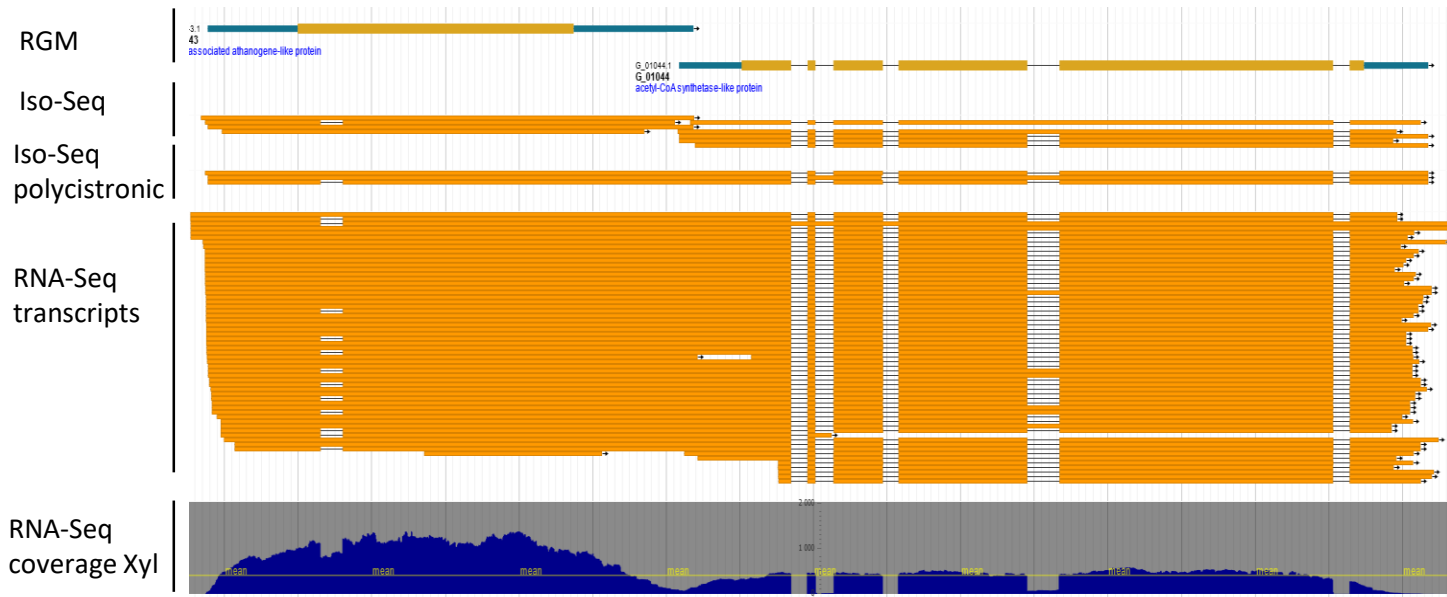
\* Li, J. et al. 2010. New insights into the evolution of subtilisin-like serine protease genes in Pezizomycotina. BMC Evol. Biol. 10

\*\* Figueiredo, et al, 2014. Subtilisin-like proteases in plant-pathogen recognition and immune priming: A perspective. Front. Plant Sci. 5



## ➤ Identification of Polycistronic mRNAs using Iso-Seq

Identification of polycistronic mRNAs validated by the occurrence of independent long Iso-Seq molecules and RNA-Seq



2.625 potential polycistronic mRNA (224 validated by Independent Iso-Seq molecules)

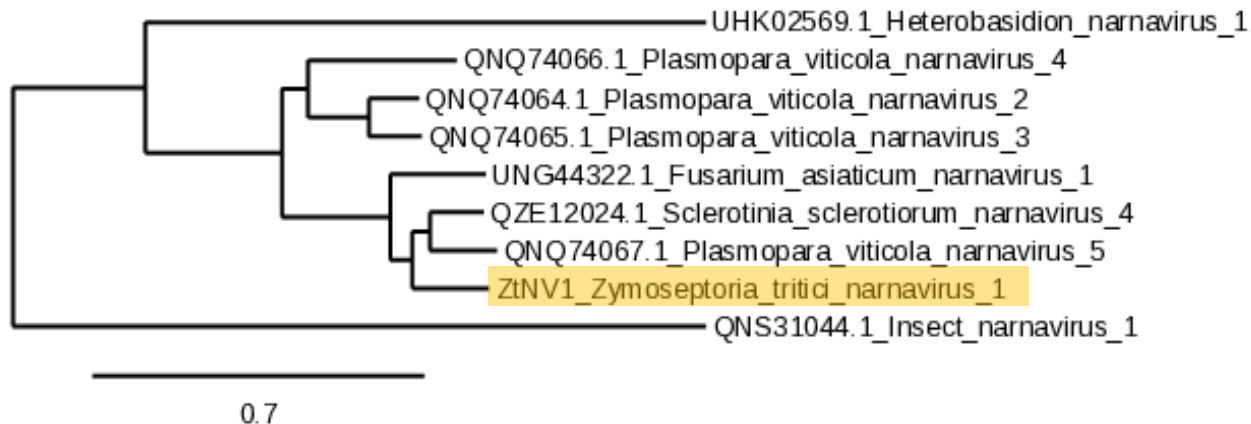
Polycistronic mRNAs observed in *Agaromyces*\* and *F.graminearum*\*\* . *Agaromyces* polycitronic mRNAs found for secondary metabolite gene clusters => stop codon, prevent for intermediate metabolite accumulation ?

\* P. Lu *et al.*, "Landscape, complexity and regulation of a filamentous fungal" *bioRxiv*, Nov. 2021

\*\* S. P. Gordon *et al.*, "Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing," *PLoS One*, 2015

## ➤ Identification of RNA Mycovirus using Iso-Seq

Detection of a new RNA mycovirus from narnavirus : ZtNV1



ZtNV1: 3091 nt (986 aa)

Iso-Seq detects rare Long RNAs -> Internal priming.

Not detected by RNA-Seq transcript assembly\*, ratio 1/70000 compared to ZtFV1 virus

\* Gilbert, K. B., Holcomb, E. E., Allscheid, R. L., and Carrington, J. C. 2019. Hiding in plain sight: New virus genomes discovered via a systematic analysis of fungal public transcriptomes. PLoS One. 14