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

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, F-91123, Palaiseau, France
[2]Syngenta Crop Protection AG, CH-4332 Stein, Switzerland
[3]University of Neuchâtel, CH-2000 Neuchâtel, Switzerland
[4]ETH Zurich, CH-8092 Zurich, Switzerland
[5]Dept of Protecting Crops and the Environment, Rothamsted Research, Harpenden, Herts AL52JQ, UK
[6]Centre for Crop and Disease Management, Curtin University, WA 6845, Perth, Australia
[7]U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[8]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94720, USA
[9]USDA-Agricultural Research Service, West Lafayette, IN 47907-2054, USA
[10]Environmental Genomics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany and Christian-Albrechts University of Kiel, 24118, Germany
[11]Wageningen University and Research, Laboratory of Phytopathology, Wageningen, 6700 AA, The Netherlands

Corresponding author: Nicolas Lapalu

Abstract

Despite large omics datasets, the prediction of eukaryotic genes is still challenging. We have developed a new method to improve the prediction of eukaryotic genes and demonstrate its utility using the genome of the fungal wheat pathogen *Zymoseptoria tritici*. From 10,933 to 13,260 genes were predicted by four previous annotations, but only one third were identical. A novel bioinformatics suite, InGenAnnot, was developed to improve *Z. tritici* gene annotation using Iso-Seq full-length transcript sequences. The best gene models were selected among different *ab initio* gene predictions, according to transcript and protein evidence. Overall, 13,414 re-annotated gene models (RGMs) were predicted, improving previous annotations. Iso-Seq transcripts outlined 5' and 3' UTRs for 73% of the RGMs, and alternative transcripts mainly due to intron retention. Our results showed that the combination of different *ab initio* gene predictions and evidence-driven curation improved gene annotation of a eukaryotic genome. It also provided new insights into the transcriptional landscape of this fungus.

Keywords: Septoria tritici blotch, gene prediction, genome annotation, transcripts, isoforms

Predicting genes in eukaryotic genomes is a challenging process (Salzberg 2019). The quality of a genome annotation depends on supporting evidence for coding regions, splice junctions and the algorithms used for predictions (Ejigu and Jung 2020). Known drawbacks for gene annotation are the complexity of eukaryotic gene structure, including difficulties in intron or start codon prediction and the quality of genome assembly. These drawbacks are particularly significant for fungal genomes. Indeed, the high gene density observed in fungal genomes leads to overlaps between adjacent transcripts (Donaldson et al. 2017; Hansen et al. 1998; Gerads and Ernst 1998) leading to wrong gene predictions such as gene fusions (Testa et al. 2015). In addition, fungi have short introns (70-100 bp, (Kupfer et al. 2004)), and frequently fragmented genome assemblies. These particularities have led to the development of fungal-specific annotation pipelines (Scalzitti et al. 2020; Birney et al. 2004; Brůna et al. 2021; Sallet et al. 2019; Holt and Yandell 2011; Stanke et al. 2006; Lukashin 1998; Min et al. 2017; Reid et al. 2014). Long-read sequencing is now providing fungal genome assemblies with almost no fragmentation. Experimental transcript evidence also has been improved by using large datasets of assembled Illumina single-stranded RNA-Seq short reads. Recently, Iso-Seq long-read sequencing is providing full-length transcript sequences (Raghavan et al. 2022). Iso-Seq can provide evidence for alternative intron splicing events and sometimes for alternative transcription start and termination sites. Still, RNA-Seq reads are required, since Iso-Seq is not quantitative (Beiki et al. 2019). Combining these two types of transcript sequencing improves the reliability of full-length transcript sequences (Amarasinghe et al. 2020). Other omics methods such as transcription start site sequencing (TSS-Seq) or cap-analysis gene expression sequencing (CAGE-Seq) are available to define transcript start sites (Casco et al. 2022; Chiba et al. 2022), but are still rarely used in fungi.

We have chosen the genome of the fungus *Zymoseptoria tritici* as a case study to develop novel methods for gene annotation of fungal genomes. *Z. tritici* is an ascomycete ((Quaedvlieg et al. 2011) causing a major foliar disease of bread and durum wheat (Petit-Houdenot et al. 2021). The *Z. tritici* genome was first sequenced in 2011 using the reference isolate IPO323 (Goodwin et al. 2011). This complete genome sequence from telomere to telomere has a size of 39.7 megabases (Mb) and is composed of 13 core chromosomes (CCs) and 8 accessory chromosomes (ACs). Twenty two additional fully assembled (long-reads) genome sequences of *Z. tritici* isolates are available (Badet et al. 2020; (Feurtey et al. 2020) (Moller et al. 2021), as well as genome sequences from four related *Zymoseptoria species* (*Z. ardibilae*, *Z. brevis*, *Z. passerinii*, *Z. pseudotritici*) (Feurtey et al. 2020). Around 14-22 % of *Z. tritici* genomes are composed of transposable elements (TEs) (Dhillon et al. 2014); (Grandaubert et al. 2015); (Badet et al. 2020); (Lorrain et al. 2021); (Oggenfuss et al. 2021). The interest of *Z. tritici* for fungal gene annotation comes from the occurrence of four independent annotations of the IPO323 *Z. tritici* genome. Large discrepancies in gene numbers and structures were observed across these four independent annotations. In addition, genes that are thought to be important for plant infection were not predicted by any of these annotation pipelines. For example, the avirulence gene *Avr-Stb6* was predicted using infection-related RNA-Seq data, but not by the existing annotations (Zhong et al. 2017). Clearly, the complete coding potential of this genome has not been identified despite four thorough annotations using different pipelines and large RNA-Seq datasets.

Using *Z. tritici* as a case study, we established a novel strategy to annotate genes in a compact eukaryotic genome using a large dataset of Iso-Seq full-length cDNA sequences (An et al. 2018; Zhang et al. 2019), and a novel bioinformatics suite, InGenAnnot, to select the best genes models among those predicted by different *ab initio* gene prediction software. This selection relies on a customized Annotation Edit Distance (AED) metric (Eilbeck et al. 2009). InGenAnnot computes an AED for each evidence with penalties for unsupported intron splicing sites (Figure S1). Using InGenAnnot, we identified 13,414 curated genes in the *Z. tritici* genome. Iso-Seq also identified alternative transcripts and long, non-coding RNA (lncRNA), improving our understanding of the *Z. tritici* transcriptional landscape.

Materials and Methods

Available *Z. tritici* IPO323 gene annotations

Currently, four annotations of the *Z. tritici* IPO323 genome are available (Table S1). The first , with 10,933 gene models, was developed in 2011 by the Joint Genome Institute (JGI) with *ab initio* gene prediction software FGENESH and Genewise (Birney et al. 2004) using EST (expressed sequence tag) and proteome evidence (Goodwin et al. 2011). The second annotation was performed in 2015 by the Max Planck Institute (MPI, Germany), resulting in 11,839 gene models (Grandaubert et al. 2015) identified with the Fungal Genome Annotation pipeline (Haas et al. 2011). This pipeline uses *ab initio* gene prediction software GeneMark-ES, GeneMark-HMM (Lukashin 1998) and Augustus (Stanke et al. 2006) combined by EVidenceModeler (Haas et al. 2008) with RNA-Seq evidence and keeping as much as possible of the first annotation provided by JGI. The third annotation was generated in 2015 by the Rothamsted Research Experimental Station (Chen et al. 2023) with 13,862 gene models (RRES, UK) obtained with the *ab initio* gene prediction software MAKER-HMM (Holt and Yandell 2011) and RNA-Seq evidence. The fourth annotation was performed in 2015 by the Centre for Crop & Disease Management, Curtin University (CURTIN, Australia) with 13,260 gene models obtained with *ab initio* gene prediction software CodingQuarry (Testa et al. 2015) and RNA-Seq evidence. All gene files corresponding to the annotations provided by JGI, MPI, RRES and CURTIN are accessible at (https://doi.org/10.57745/CVIRIB) and displayed at a dedicated INRAE genome browser (https://bioinfo.bioger.inrae.fr/portal/genome-portal/12) or at the *Zymoseptoria tritici* IPO323 JGI genome browser (https://mycocosm.jgi.doe.gov/Zymtr1/Zymtr1.home.html).

Fungal isolate, RNA extraction, PacBio Iso-Seq and Illumina RNA-Seq libraries

The reference isolate of *Z. tritici* IPO323 (Goodwin et al. 2011) was stored at -80°C as a yeast-like cell suspension (10^7 cells/mL in 30% glycerol), and grown at 18°C in the dark on solid (Yeast extract Peptone Dextrose (YPD) agar) or liquid (Potato Dextrose Broth (PDB)) media. For RNA production, different media were used (Table S3). Additional single-stranded RNA-Seq data were obtained from public databases (Table S3). Novel and public RNA-Seq data were cleaned and mapped to the *Z.tritici* IPO323 genome (see Table S3 for methods). Processed Iso-Seq data also were mapped to the *Z. tritici* IPO323 genome (see Table S3 for methods).

Gene prediction and selection of the best gene models

The two *ab initio* gene prediction software Eugene v1.6.1 (Sallet et al. 2019) and LoReAn v2.0 (Cook et al. 2019), handling long-read transcript sequences as evidence, were used to annotate the *Z. tritici* IPO323 genome sequence. Eugene was trained with filtered Iso-Seq transcripts (Table S3) and a dataset of proteins from four genomes of species phylogenetically related to *Z. tritici*: *Cercospora beticola* (GCF_002742065.1_CB0940_V2); *Ramularia collo-cygni* (GCF_900074925.1_version_1); *Zasmidium cellare* (GCF_010093935.1_Zasce1); and *Sphaerulina musiva* (GCF_000320565.1_Septoria_musiva_SO2202_v1.0), using the fungal matrix (WAM fungi matrix). After the training step, gene structures were predicted with assembled transcripts from RNA-Seq and a dataset of Dothideomycetes proteins obtained from Uniprot without *Zymoseptoria* sequences to avoid inference with previous *Z. tritici* annotations. Filtered Iso-Seq transcripts were used as strongly weighted evidence in model prediction with the parameter "`est_priority=2`". LoReAn was launched in the fungal mode with the Augustus retraining mode using the same Dothideomycetes Uniprot protein dataset and Iso-Seq transcript dataset as used for Eugene. RNA-Seq data were merged as a mapping file (BAM) obtained with the pipeline used to assemble transcripts and detect splicing sites (see Table S3 for methods). The new (Eugene, LoReAn) and previous (JGI, MPI, RRES, CURTIN) gene models were analyzed with *ingenannot filter* to filter out TE-encoding genes.

Filtered gene models were analysed with *ingenannot aed* to provide Annotation Edit Distance (AED) (Eilbeck et al. 2009) scores for each gene model compared to either the Uniprot fungal protein dataset without *Zymoseptoria* species (AED protein) or the filtered Iso-Seq and RNA-Seq transcripts (AED transcript). The original AED score proposed by Maker (Eilbeck et al. 2009), is a combination of

Sensitivity and Specificity computing to compare two gene models using the number of bases overlapping both annotations or specific to each of them. InGenAnnot computes a customized AED for each source of evidence with several options such as restriction to coding sequence (CDS) or penalty on unsupported intron splicing sites (Figure S1). AED were calculated with "`--aed_tr_cds_only`" to avoid bias between datasets with or without UTR annotations and with "`--penalty_overflow 0.25`" to penalize gene models with unsupported intron splicing sites. The best gene models were selected with InGenAnnot *select* based on an AED value below 0.3 for transcript evidence (AED transcript) or below 0.1 for protein evidence (AED protein). AED values of 0.5 or below are considered as indicative of good annotations, while values of 0.3 or below are classified as high-quality annotations (Hunt et al. 2020; Holt and Yandell 2011). As we benefit from extensive RNA-Seq and Iso-Seq datasets, we set a threshold of 0.3 for selecting the best gene models. The threshold for "AED protein" was set to 0.1, as it is challenging to evaluate accurately gene structure using only protein sequence alignments. In this context, the AED protein score was used as evidence for the presence—or absence—of a similar existing protein in other fungi, in particular for gene models without sufficient transcript support. Gene models with AED scores higher than the threshold values, but predicted by at least 4 independent *ab initio* gene prediction software, were also retained. Gene models with AED scores higher than the threshold values, but predicted by at least 4 independent *ab initio* gene prediction software, were retained. However, all the gene models without an ATG or stop codon were removed. The relatively high number of annotation sources (6) and the selection of loci detected by 4 independent gene predictors, allow us to use stringent AED thresholds, leading to well supported gene structures (see Figure S2 for a full description the bioinformatics workflow).

Potential new gene models encoding effectors were predicted with *ingenannot rescue_effector,* and added to the final dataset. Transcripts that did not co-localize with a gene model were tested in 3 frames to analyse the predicted peptides with the same criteria as those used to detect small secreted proteins (SSP) as described below. The final set of gene models was identified as RGMXXXX for Reannotated Gene Models from RGM00001 to RGM13414.

UTRs were inferred in two passes with the *ingenannot utr_refine*. First, all previously annotated UTRs and inferred new coordinates from a filtered set of Iso-Seq transcripts were withdrawn. Second, UTRs were inferred using a filtered set of RNA-seq assembled transcripts, considering only transcripts with no UTRs from the first step. Both sets were established with the *ingenannot isoform_ranking* for filtering and ranking UTR isoforms based on RNA-Seq evidence.

Gene models from each annotation were compared according to their AED scores using *ingenannot aed_compare*. Specific/shared gene models were identified using *ingenannot compare.* BUSCO (Manni et al. 2021) analyses with ascomycota_odb10 were performed to evaluate the completeness of datasets (See Table S5 for details, and comments).

Functional annotation of RGMs

RGMs protein sequences were analysed with Interproscan 5.0 (Jones et al. 2014) and Blastp (Camacho et al. 2009) (e-value <1e-5) against the NCBI nr databank to perform a Gene Ontology annotation (Gene Ontology Consortium 2004) with Blast2GO (Götz et al. 2008). Secreted proteins and effectors were annotated as described in (Gay et al. 2021), using a combination of TMHMM (v.2.0) (Möller et al. 2001), SignalP (v4.1) (Nielsen 2017) and TargetP (v1.1b) (Armenteros et al. 2019) with the following criteria: no more than one transmembrane domain and either a signal peptide or an extracellular localization prediction. The SSP repertoire was predicted by applying a size cutoff of 300 amino acids and keeping only proteins predicted as effectors by EffectorP (v2.0).

Analysis of Iso-Seq transcript isoforms, antisense and lncRNAs

The annotation of transcript isoforms was performed with sqanti3 (Tardaguila et al. 2018) using Iso-Seq transcripts (see Table S8 for methods). Iso-Seq transcripts annotated as antisense and

intergenic with sqanti3 were selected as long, non-coding (lnc) RNAs and further filtered (see Table S9 for Methods; File S1).

Detection of polycistronic Iso-Seq transcripts

Read-through Iso-Seq transcripts that were previously filtered out were merged to obtain the global counts of co-transcribed genes. These read-through transcripts were filtered out using RGMs and their Iso-Seq transcripts as evidence. Only polycistronic mRNAs that were supported by independent long-read single transcripts for each gene were conserved and considered as reliable. Detection of overlaps between transcripts and annotations was performed with intersect from BEDTools (Quinlan and Hall 2010).

Annotation Edit Distance as a metric for comparing gene models predicted by different tools

InGenAnnot RGMs were compared to gene models obtained with either funannotate (funannotate n.d.), Helixer (Stiehler et al. 2021) or BRAKER3 (Gabriel et al. 2024). Gene models obtained with these three tools were scored with AED using the same evidence as for RGM, and their AED scores were plotted for both transcript and protein evidence (Figures S11, S12 and S13). Gene models from each annotation were compared according to their AED scores using *ingenannot aed_compare*. Specific/shared gene models were identified using *ingenannot compare* (Figure S14).

Results

Comparisons of existing *Z. tritici* IPO323 genome annotations

The gene models from the four previous annotations of the *Z.* tritici IP0323 genome (MPI, JGI, RRES, and CURTIN) were filtered out for TE-encoding genes. These gene models were clustered into 13,225 metagenes, defined as the "gene locus" of ParsEval (Standage and Brendel 2012). These metagenes corresponded to 26,224 distinct CDS. The comparison of the different gene models occurring at each "locus" highlighted three categories of metagenes: a) identical gene models (same CDS); b) dissimilar gene models (same metagene, but different CDS); and c) specific gene models (predicted by a single gene predictor). Only 3,618 identical gene models were shared across the four annotations (27%, Figure 1). The MPI, RRES, and CURTIN annotations share more identical gene models, reaching the value of 6,816 (51%, Figure 1). The JGI and CURTIN annotations displayed the highest numbers of dissimilar gene models (4,752 and 3,844, respectively) compared to RRES and MPI (2,367 and 1,871, respectively; Figure 1). On the other side, the RRES, CURTIN and JGI annotations displayed higher numbers of specific gene models (593, 436, and 151, respectively; Figure 1), compared to the MPI annotation (12). Overall, this comparison showed that most metagenes displayed gene models predicted by at least two independent annotations (91%). Despite the low numbers of identical gene models across all four annotations, basic genomic statistics were similar (Table S1). Indeed, the JGI, MPI and CURTIN annotations displayed a similar distribution of gene models across chromosomes. However, the RRES annotation had more gene models on accessory chromosomes (Table S2). In addition, the average sizes of gene models differed between MPI (1465 bp) and the other annotations (~1300 bp). This difference could result from the occurrence of wrong gene models corresponding to the fusion of two or more adjacent gene models predicted as such by other annotations. Indeed, 533 and 801 gene fusions were detected in the MPI annotation compared to the RRES and CURTIN annotations, respectively. Overall, the low number of identical gene models among these four annotations (27%) likely resulted from drawbacks of each pipeline. To circumvent these problems, we generated a novel annotation of the IPO323 genome using a large set of transcript sequences, coming from either publicly available transcript sequences obtained by short-read, single-stranded RNA-Seq sequencing, or new transcript sequences obtained from long-read PacBio sequencing (Iso-Seq) and short-read, single-stranded RNA-Seq sequencing (Table S3).

Iso-Seq based annotation of the IPO323 genome and gene model selection

mRNAs obtained from a large set of *in vitro* mycelial growth conditions (Table S3) were used for the construction of either single-stranded Iso-Seq cDNA libraries or single-stranded Illumina cDNA libraries. After mapping and filtering, 22,659 Iso-Seq transcripts were identified, including alternative transcripts differing in their intron splicing or TSS/TTS (TSS: transcriptional starting site, TTS: transcriptional termination site). Alternative Iso-Seq transcripts either unsupported by RNA-Seq or in low relative abundance according to RNA-Seq (10%<), were filtered out. This filtering provided 21,052 transcripts corresponding to 8,927 loci. Most loci displayed only one isoform (50%), while other loci had either 2 to 5 isoforms (42%), or at least 6 isoforms (8%). Transcripts from each single-stranded RNA-Seq library were assembled separately and those with weak expression levels (TPM<1) were removed (Table S3). A total of 498,010 single-stranded RNA-Seq transcripts were obtained as evidence. Currently a few gene prediction tools such as Eugene (Sallet et al. 2019), and LoReAn (Cook et al. 2019) use Iso-Seq transcripts as evidence. Eugene identified 15,810 gene models in the *Z. tritici* genome in a two-pass mode and strand-specific prediction allowing overlapping gene models on opposite strands. This number was reduced to 15,245 gene models after filtering out genes corresponding to TEs. LoReAn predicted 11,537 gene models without overlapping predictions on the opposite strand, which were reduced to 11,497 after filtering out genes corresponding to TEs. Selection of the best gene model was performed with InGenAnnot using Eugene, LoReAn and previous annotations (JGI, MPI, RRES, CURTIN). All these gene models were clustered into 17,147 metagenes.

InGenAnnot computed an Annotation Edit Distance (AED) (Eilbeck et al. 2009) for each comparison (two gene models or one gene model and its evidence), taking into account the number of overlapping bases (Eilbeck et al. 2009). AED computation was limited to the CDS, and a penalty score of 0.25 was introduced if intron splicing sites differed between a transcript evidence and its gene model. In addition, different AED scores were computed for transcript and protein evidence. Gene models with AED values below 0.3 for transcript and/or 0.1 for protein evidence were selected (Figure 2). Gene models failing to pass the AED threshold, but predicted by at least four independent annotations, were retained to avoid the loss of gene models with low support from transcript or protein evidence (upper right square in Figure 2 corresponding to 1,846 gene models). These rescued genes models were mostly not conserved across fungi and had low transcriptional support (Figure 2). For CDS overlapping on opposite strands, only the gene model with the best AED score was selected. Finally, 97 additional effector-encoding genes were predicted with the *rescue_effector* tool of InGenAnnot. Overall, we predicted 13,414 re-annotated Gene Models (RGMs; File S1, Table S4). In addition, UTRs were inferred from Iso-Seq transcripts for 7,713 genes and for 9,856 genes when combined with filtered RNA-Seq assembled transcripts. The average lengths of 5' UTRs were 315 bp, while they were 389 bp for 3'UTRs (Table S4), similar to what was reported for the fungus *P. anserina* (5' UTR 275 bp, 3' UTR 303 bp) (Lelandais et al. 2022). A small proportion of genes displayed long 5' UTRs (1,000 to 7,000 bp, 6%), and/or long 3' UTRs (1,000 to 8,600 bp, 8.6%).

Comparison of the reannotated IPO323 gene models with available genome annotations

The 13,414 IPO323 RGMs were compared to previous gene models (JGI, MPI, RRES, CURTIN), using BUSCO and the ascomycota_odb as reference. High BUSCO scores were obtained with the RGM, RRES, MPI and CURTIN annotations (98.4-99.4%; Table S5), while the score obtained with the JGI annotation was lower (95.7 %), likely due to a high number of fragmented and missing BUSCO genes (Table S5). The comparison between annotations was then performed using AED scores (Figure 2, S3 and S4). Of the 13,414 RGMs, 11,568 gene models (86%) displayed AED values below the thresholds of 0.3 for transcript and 0.1 for protein evidence (Figure 2). CURTIN had a high level of support (10,716 gene models, Figure S3), followed by RRES (9518 gene models) and MPI (8936 gene models), while JGI was the least supported (7,730 gene models). Among the 1,846 RGMs failing to pass the AED threshold, but rescued as predicted by at least four annotations, 574 have no AED score. This implied that they were only predicted by *ab-initio* software (no evidence in Table S6). Half of these fully *ab-initio* RGMs were located on the 3' arm of chromosome 7 between positions 1,900,000 and 2,500,000 (Table S6). Almost none of these RGMs was expressed, including during infection. This region was described as enriched in repressive histone H3K27me3 and H3K9me3 marks as observed for accessory chromosomes (Schotanus et al. 2015). These genes were not expressed in the kmt1 and kmt6 mutant backgrounds that lacked these histone modifications. This observation suggested that they were either unexpressed pseudogenes, or that their expression was under a negative control independent of the H3K27me3 and H3K9me3 marks. In addition, none of these genes was conserved across fungi, suggesting either a recent origin or an artefact from annotation pipelines. The other fully *ab-initio* RGMs were enriched in genes localized on accessory chromosomes (Table S6).

Among the 13,414 RGMs, 7,888 were identical to at least one gene model from another annotation (Figure 3), while 3,479 RGMs were identical to all the gene models from the four previous annotations (Figure 3). Since 3,618 gene models were identical among the four previous annotations (see above), 139 of these genes were not identical to RGMs. Most of these 139 RGMs had a novel start codon that did not change the coding phase of the first open reading frame, but led to a shorter or longer version of the same protein. Ribosome profiling could solve this problem by identifying the real start codon (Ingolia 2014). 2,047 RGMs were either different from (1,376 modified RGMs, Table S6) or not predicted by previous annotations (671, specific RGMs, Table S6). Most of the 1,376 modified RGMs had either alternative ATGs or intron splicing sites supported by transcript evidence. The 671 specific RGMs were distributed evenly on all chromosomes (Table S6). 117 of these specific RGMs displayed more than 40% sequence identity to proteins from other fungi. Blastn and tblastn searches showed

that 654 (97%) of these specific RGMs were detected in the genome of other *Z. tritici* strains (File S1). This result showed that most of the specific RGMs are not IPO323 specific, but are shared across isolates.

One major improvement of this novel annotation was the identification of split RGMs corresponding to genes wrongly fused in previous annotations, by detecting overlaps between gene models. Fused genes were detected in high numbers in the MPI and JGI annotations (1,507 and 1,258, respectively, Table S7), and in a lower number in the RRES annotation (701), while they were almost absent from the CURTIN annotation (176). The average AED score of split RGMs was better (median AED score: 0.17) than that of fused genes (median AED score: 0.34). In addition, most MPI fused genes (87%) were not supported by transcript evidence, since their AED scores were higher than the cutoff value (>0.3, Figure S5). Even though most transcript AED scores of split RGMs (65 %) were lower than the cutoff value (0.3<, Figure S5), a significant number of split RGMs (494, 35%) had a low support from both transcript and protein evidence (Figure S5). These split RGMs were rescued since they were also identified in other annotations than MPI. The transcript evidence of two randomly chosen MPI fused genes and their corresponding split RGMs are shown in Figures S6 and S7. Both MPI fused genes had no Iso-Seq transcript support, while Iso-Seq transcripts supported the split RGMs. Assembled RNA-Seq transcripts supporting split RGMs were also observed for RGM-1 and RGM-2 from Figure S6. However, large, assembled RNA-Seq transcripts supporting fused MPI genes were observed (Figure 5). We hypothesised that these long transcripts were artefacts of the assembly of RNA-Seq reads from individual genes with overlapping transcripts. The final evidence supporting these split RGMs was obtained by identifying specific expression conditions (13 days post-inoculation, wheat infection, Figure S5) in which RGM-2 was strongly expressed, but not RGM-1.

Functional annotation of reannotated IPO323 gene models

Functional annotation of RGM proteins was performed using Blast2Go and InterProScan. 5,593 RGMs exhibited a GO term or an IPR and 2,838 were annotated with at least one Enzyme Code (EC). Several tools (Morais do Amaral et al. 2012; Grandaubert et al. 2015) were used to identify 1,895 RGMs encoding putative secreted proteins, including effectors (File S1, Table). A previous analysis predicted 970 secreted proteins using the JGI annotation [43] which were all identified as RGMs. However, they increased to 1,046, due to the split of fused genes from the JGI. The RGM secretome included 234 small secreted proteins (SSP) according to our criteria (peptide signal, size < 300 aa, EffectorP detection). Among these SSPs, 54 were detected by the effector rescue software of InGenAnnot including 43 SSPs that were not identified by new *ab initio* gene prediction software we used, nor by previous annotation pipelines. The effector rescue software searched for genes encoding SSPs according to our criteria (see before) among CDS inferred from transcripts not associated with a gene model. This strategy allowed the rescue of genes encoding SSPs that were difficult to predict by *ab initio* gene prediction software. Four of these 43 novel SSPs displayed a significant upregulation during infection. Four of these 43 novel SSPs (ZtIPO323_001210, ZtIPO323_072700, ZtIPO323_105940 and ZtIPO323_123970) displayed a significant upregulation during infection compared to *in vitro* culture, suggesting a possible role in infection. In addition, genes encoding effectors that were missing in previous annotations, such as *Avr-Stb6* located at the end of chromosome 5 (Zhong et al. 2017), were predicted as RGMs (Figure S8b). Two additional *Avr-Stb6* paralogs located on chromosome 10 were also predicted as RGM specific SSPs (Figure S8a).

Identification of alternative transcripts

The initial set of 21,052 Iso-Seq transcripts was filtered to exclude UTR length isoforms, yielding 11,690 Iso-Seq transcripts corresponding to coding and non-coding loci. Sqanti3 allocated 10,938 Iso-Seq transcripts to 8,199 RGMs (Table 1). 7,872 of these RGMs had the same structure as their Iso-Seq transcripts (full_splice_match). The other 327 RGMs classified as “ISM” or “genic” by Sqanti3 displayed a structure differing from Iso-Seq transcripts. In most cases, these Iso-Seq transcripts were partially covering RGMs, suggesting truncated cDNAs. These RGMs were supported either by other evidence (RNA-Seq, protein) or were rescued (*ab initio* only). 2,716 Iso-Seq transcripts identified as alternative

splice variants (25% of coding transcripts) were classified by Squanti3 into: combination of known splicing sites (NIC); new splicing sites (NNC); intron retention (IR); and genic (Table 1). Most alternative transcripts corresponded to intron retention events (IR, 75%). Transcripts with a premature termination codon (PTC) recognized by the non-sense-mediated decay (NMD) pathway were filtered out (Zhang and Sachs 2015), leaving 2,372 alternative transcripts corresponding to 1,742 RGMs. The numbers of RGMs with 2, 3, 4 and at least 5 isoforms were 1,342, 274, 77 and 49, respectively (Table S8). A total of 337 alternative transcripts corresponded to a novel combination of coding exons, 271 to a novel combination of UTR exons, and 16 to a novel combination of both (NIC, NNC and Genic events, Table 1). For example, RGM ZtIPO323_030030, encoding a putative SSP (SSP10) (Mirzadi Gohari et al. 2015), had an alternative splicing site generating a new exon that encoded a shorter SSP that was reduced by 34% at its C terminus (Figure 4a). The 1,753 remaining transcript isoforms with intron-retention events were likely un-spliced transcripts that were not detected by our NMD screen. Some alternative transcripts such as RGM ZtIPO323_013330 were detected in high abundance using RNA-Seq (Figure 4b). Its main transcript (Iso-Seq 2), corresponding to the selected RGM, had 4 splicing sites, one being in the 5' UTR. Two alternative Iso-Seq transcripts (Iso-Seq 1 and 2) with one or two intron-retention events were also supported by RNA-Seq. The last Iso-Seq transcript displayed an alternative splicing site for the fourth intron that was not supported by RNA-Seq. We identified a drawback of using Iso-Seq for annotation, as some alternative transcript isoforms were used as evidence for selecting the RGM as shown for ZtIPO323_030030 (Figure 4a) or ZtIPO323_013090 (Figure S9). These examples illustrated the difficulty in choosing between gene models with complex alternative splicing events leading to transcript isoforms with similar expression levels (Figure 4a). However, these events were not detected frequently.

RNA-Seq data were used to compute differential isoform usage (DIU) for coding genes using tappAS [29]. Only 22 RGMs had a significant DIU (p-value < 0.01) between Galactose/Sucrose and Mannose/Xylose culture conditions (File S1). A total of 163 RGMs displayed a significant DIU between infection and culture conditions (File S1), including 23 secreted proteins. Some of these RGMs were highly up or down regulated during infection such as ZtIPO323_042160 (unknown), ZtIPO323_042360 (unknown), ZtIPO323_043800 (PHD/RING finger protein), and two secreted proteins (ZtIPO323_016670, ZtIPO323_043500) that were significantly upregulated during late infection (13, 21 dpi). ZtIPO323_016670 encoded a carbohydrate esterase from family 8 involved in cell wall modifications and ZtIPO323_043500 encoded a SSP. Manual inspection of the RNA-Seq data associated with these DIU RGMs confirmed their differential expression, but not a different usage of isoforms. Indeed, the isoforms detected during infection corresponded to a low number of reads compared to *in vitro* culture conditions, leading to a bias in DIU analyses.

Identification of long, non-coding RNAs

Sqanti3 allocated 752 Iso-Seq transcripts to non-coding loci (Table 1, 395 antisense and 357 intergenic). A single study of fungal lncRNAs was performed using Iso-Seq in *F. graminearum* (Lu et al. 2021), identifying lncRNAs generally larger than 1 kb. Therefore, we excluded from our analysis Iso-Seq transcripts overlapping with TEs and smaller than 1 kb in length. We also excluded Iso-Seq transcripts with an ORF longer than 300 bp (100 amino acids). We chose these stringent criteria to select reliable lncRNAs, and to avoid false lncRNAs encoding putative "coding genes" not retained by InGenAnnot. This process led to 55 lncRNAs, among which 3 were labelled as "coding" based on their coding potential and 1 contained an ORF with a pfam domain. Finally, 51 transcripts were classified as lncRNAs according to our criteria among which 35 (68%) were differentially expressed (p-value 0.05).

Half of these lncRNAs were differentially expressed between infection and *in vitro* culture, including 5 that were up-regulated and 12 that were down-regulated during infection (log2FC > 2). Most lncRNAs that were down-regulated during infection were antisense transcripts (83%). The lncRNA PB1188.1 that was down-regulated during infection compared to all culture conditions (Table S9), was an antisense transcript of ZtIPO323_016330, encoding a secreted Subtilisin-like protein. ZtIPO323_016330 was up regulated during infection and down regulated during *in vitro* culture.

Another RGM (ZtIPO323_037670) encoding a TTL protein (Tubulin tyrosine ligase involved in tubulin posttranslational modifications) and its antisense lncRNA PB.2709.1 displayed opposite expression patterns during infection (Table S9), as lncRNA PB.2709.1 was up regulated during infection, while ZtIPO323_037670 was down regulated.

Identification of polycistronic mRNAs

Alignment of Iso-Seq transcripts with RGMs identified 2,625 putative polycistronic transcripts. Multiple stop codons were present in these polycistronic transcripts, excluding the possibility of errors in annotated genes for a larger single ORF, as observed for polycistronic transcripts described in Agaricomycetes (Gordon et al. 2015), and *F. graminearum* (Lu et al. 2021) or *Cordyceps militaris* (Chen et al. 2019). Overall, 224 putative polycistronic transcripts contained two-three RGMs on the same strand. For example, adjacent RGMs ZtIPO323_010430 and ZtIPO323_010440 were transcribed on the same strand with overlapping 3' UTRs and 5' UTRs (Figure 5). Iso-Seq polycistronic single-transcript molecules covering these two RGMs were detected, as well as single-RGM Iso-Seq transcripts (Figure 5). Assembled RNA-Seq reads supported a transcript covering the two RGMs (Figure 5). However, RNA-Seq coverage strongly decreased in the overlap between the two RGMs, suggesting two independent transcripts (Figure 5). RNA-Seq coverage showed that the abundance of the polycistronic transcript was low compared to single-gene transcripts. This analysis suggested that these polycistronic transcripts were likely rare read-through transcripts.

Iso-Seq transcripts encoding fungal mycoviruses

A total of 2,203 Iso-Seq transcripts did not map to the *Z. tritici* genome. These transcripts were combined into two clusters of highly related sequences. The larger cluster (1919 sequences) was identical to Fusarivirus 1 (ZtFV1) (Gilbert et al. 2019). The second cluster gathered 17 independent Iso-Seq transcripts closely related to narnavirus 4 of *Sclerotinia sclerotiorum* (SsNV4) (Jia et al. 2021), and named ZtNV1 (*Zymoseptoria tritici* NarnaVirus 1). As these viral Iso-Seq transcripts were probably obtained by internal polyA priming, they did not cover the full sequence of the viruses. RNA-Seq reads corresponding to these two fungal viruses were detected in all our cDNA libraries. ZtFV1 Iso-Seq transcript was confirmed to be a full-length viral sequence. Assembled ZtNV1 RNA-Seq reads led to the reconstruction of a full-length viral sequence of 3091 nucleotides encoding a protein of 986 amino acids corresponding to an RNA-dependent RNA polymerase. ZtNV1 was as long as SsNV4 (3105 bp), and its encoded protein displayed 71% identity at the nucleotide level and 67% identity (79% similarity) at the protein level with SsNV4. The phylogenetic tree of viral RNA-dependent RNA polymerases confirmed that ZtNV1 was highly related to narnaviruses identified in *S. sclerotiorum, Plasmopara viticola,* and *Fusarium asiaticum* (Figure S10). IPO323 ZtNV1 sequence was detected in many publicly available *Z. tritici* RNA-Seq data (few reads per library), validating the ubiquitous presence of this virus in *Z. tritici*. ZtFV1 was also detected in these RNA-Seq data in higher amounts compared to ZtNV1 (70,000 fold).

Comparison of InGenAnnot to other gene prediction tools (funannotate, BRAKER3 and Helixer) using Annotation Edit Distance

Two integrated gene prediction tools (funannotate, BRAKER3) and a deep-learning-based software (Helixer) were used to annotate the *Z. tritici* genome. Funannotate integrates four *ab initio* tools (Augustus, SNAP, GeneMark, CodingQuarry) and uses EvidenceModeler to select the best gene model (funannotate n.d.). BRAKER3 integrates two *ab initio* tools (Augustus, GeneMark) and utilizes TSEBRA to select the best gene model (Gabriel et al. 2024). Helixer is a deep-learning tool that was trained on fungal gene models (Stiehler et al. 2021). These tools were run with the same transcript and protein evidence as InGenAnnot. The novel gene models were scored with AED using the same transcript and protein evidence as InGenAnnot (Figure S11-S13). Comparison of these gene models to RGMs highlighted 6,389 identical CDS predicted by the four tools (47% of RGMs, Figure S14). The number of identical CDS predicted by funannotate, BRAKER3 and InGenAnnot (RGMs) was higher (8,220 CDS, 61% of RGMs, Figure S14). Individually, Helixer displayed the lowest number of gene models similar to RGMs (8,086, 60% of RGMs, Figure S14). A higher number of funannotate and BRAKER3 gene models

were similar to RGMs (67 and 73% of RGMs for funannotate and BRAKER3 respectively, Figure S14). Helixer was the only tool to predict a large number of unique CDS (6,190), including 4,103 CDS originating from shared loci. The other 2,087 gene models that were unique to Helixer originated from loci at which no gene model was predicted by other tools, among which 1,358 had no evidence. Overall, this comparison highlighted a high number of discrepancies in gene model prediction between different tools, as already observed during RGM selection (Table 2).

In terms of cumulative AED scores, InGenAnnot (RGMs) gave the best results, closely followed by BRAKER3 (Figure S15). This could be explained by the strong weight assigned to transcriptomic data to obtain the InGenAnnot and BRAKER3 gene models. Indeed, since BRAKER3 predicted isoforms at some loci (14,833 transcripts for 12,293 genes), it likely increased the number of gene models with transcript evidence. AED plots were used to compute metrics on the dispersion of annotations for the four gene sets (Table S10). InGenAnnot and BRAKER3 showed the best agreement with transcriptomic evidence (median transcript AED scores: 0.12 and 0.14, respectively, for InGenAnnot and BRAKER3, Table S10) compared to funannotate (median transcript AED score: 0.15) and Helixer (median transcript AED score: 0.26), but BRAKER3 surpassed all tools in protein evidence. Finally, BRAKER3 showed the best AED scores for its gene annotation set (best score relative to the ideal point, median: 0.338, Table S10), closely followed by InGenAnnot (median: 0.398), while funannotate and Helixer displayed higher values (median: 0.469 and 1.000 respectively), suggesting that their gene models were less fit to the evidence. BRAKER3 was more specific, since it predicted only 12,293 genes compared to InGenAnnot (13,414 genes) and funannotate (13,423 genes). This suggested that BRAKER3 selected mostly gene models with evidence, while funannotate and InGenAnnot allowed the selection of gene models with less or no transcript or protein evidence, but strong gene signals from *ab initio* prediction, thereby increasing their sensitivity.

Discussion

Improvement of the *Z. tritici* IPO323 genome annotation

The production of an Iso-Seq library of full-length transcript sequences corresponding to a wide array of growth conditions was essential to improve *Z. tritici* IPO323 genome annotation. Indeed, the assembly of RNA-Seq short reads frequently leads to artefacts such as chimeras corresponding to adjacent genes with overlapping transcripts (Raghavan et al. 2022), which are frequent in compact genomes (Testa et al. 2015). Iso-Seq long-read data bypasses these artefacts, as it produces sequences from single cDNA molecules without assembly. Iso-Seq also provides transcript isoforms corresponding to alternative start, stop and intron splicing events. Still, Iso-Seq has pitfalls since it is not quantitative. Indeed, we identified rare, long Iso-Seq transcripts likely corresponding to intron retention events and polycistronic transcripts. Filtering out low-abundance Iso-Seq transcripts using short-read RNA-Seq quantification reduced such drawbacks. Overall, filtered Iso-Seq transcripts were highly reliable in selecting the best gene model among those predicted by different *ab initio* gene prediction software using AED transcript scores (transcript evidence). Protein evidence was also helpful for genes not expressed under the conditions used for producing mRNAs. We observed that the combination of six *ab initio* gene prediction software was needed to improve annotation. First, a diversity of software was needed to produce a sufficient number of gene models at each locus to be selected by InGenAnnot. Indeed, none of the *ab initio* gene prediction software was able to independently predict all the RGMs (Table 2). For example, Eugene, the most efficient *ab initio* software with our dataset, only predicted 76% of the selected RGMs. Second, the use of different *ab initio* software allowed the rescue of gene models without evidence (1,846 rescued RGMs with AED scores over the thresholds). Most of these rescued RGMs were not conserved across fungi and were not expressed under the available conditions (Figure 2). They typically included candidate fungal effectors that could be important for plant-fungal interactions (File 1). Yet some rescued RGMs could be artefacts of *ab initio* prediction, and they should be validated manually.

Overall, our strategy significantly improved the annotation of the *Z. tritici* IPO323 genome, and missing genes encoding effectors such as Avr-Stb6 were predicted correctly. In addition, it revealed different biases from previous annotations. Among the 13,414 RGMs, 2,047 were either different from all previous gene models (1,376 modified RGMs, Table S6) or not predicted in previous annotations (671 RGM-specific, Table S6). Transcripts and protein evidence supported these RGMs. The most frequent discrepancy was the occurrence of fused genes in previous annotations that were split into distinct RGMs. These fused genes corresponded to RGMs with overlapping transcripts (Figures S6, S7). Indeed, for such genes, RNA-Seq read assembly likely generated chimeric transcripts, providing erroneous evidence to the *ab initio* software. Changes in parameters used for RNA-Seq read assembly could reduce the number of chimeric transcripts. However, Iso-Seq long-read sequencing clearly avoided this artefact and its use as transcript evidence likely explains the improvement observed in RGMs. To our knowledge, only two previous studies demonstrated improved fungal gene prediction using Iso-Seq transcript long-read sequences: *C. militaris* (Chen et al. 2019); and *F. graminearum* (Lu et al. 2021). We further improved the method used in these papers by filtering Iso-Seq transcripts according to their abundance, and by creating a method to select the best gene model according to different *ab initio* annotations and evidence.

Iso-Seq long reads reveal the complexity of transcripts in *Z. tritici*

Iso-Seq long-read sequencing allowed the identification of alternative transcripts in *Z. tritici*. However, Iso-Seq is not quantitative and minor transcripts with long UTRs or IR without strong support from RNA-Seq data were identified (Figure 4, Figure 5, Figure S7). These low-abundance transcript isoforms could be produced by the transcriptional machinery either as by-products, or to regulate gene expression. The best strategy to detect such transcripts was to quantify Iso-Seq transcript isoforms using RNA-Seq data. As observed in other fungal genomes (Lu et al. 2021); (Jeon et al. 2022), most alternative splicing events were intron retention (IR, Table 1). IR events could generate premature termination codons (PTCs) that were likely degraded by the NMD pathway. However, NMD signals are

difficult to predict with current bioinformatics tools in filamentous fungi. DIU analysis revealed a few RGMs with differentially expressed transcript isoforms during infection compared to *in vitro* culture conditions. As discussed before, the small amounts of RNA-Seq reads available for infection makes such statistical comparisons difficult. Manual inspection of several loci did not reveal clear patterns of DIU for alternative transcripts.

Compact genomes, such as that of *Z. tritici*, are suitable for polycistronic transcription. Iso-Seq was successful in identifying polycistronic mRNAs in *Z. tritici* as reported in Agaromycotina (Gordon et al. 2015), *F. graminearum* (Lu et al. 2021) and *C. militaris* (Chen et al. 2019). However, polycistronic-specific RNA-Seq reads were always detected in low abundance compared to single-gene transcripts. These RNA-Seq data also showed that polycistronic transcripts mostly corresponded to genes with transcripts overlapping those from adjacent genes. As Iso-Seq is sensitive enough to detect low-abundance transcripts, it is possible that these polycistronic transcripts are rare read-through transcripts. This hypothesis is supported by the fact that *in vitro* culture conditions of yeast known to be associated with increased transcriptional read-through led to more polycistronic transcripts (Hadar et al. 2022). Alternatively, these polycistronic transcripts could be an additional level of transcriptional control.

lncRNAs are differentially expressed during wheat infection

LncRNAs are important components of transcriptional and translational regulation (Till et al. 2018). They can act in *cis* or *trans* of target genes, and either up-regulate or down-regulate target gene expression (Till et al. 2018). Most studies on fungal lncRNAs have used assembled RNA-Seq reads (Liu et al. 2022), likely leading to artefacts from assembly. Iso-Seq bypassed this problem and facilitated the identification of full length, non-chimeric lncRNAs. Using stringent criteria (size > 1000 bp, no ORF > 100 aa, no overlap with TEs), we identified 51 lncRNAs in *Z. tritici*. This number is far lower than those identified in other fungi (939 in *N. crassa* (Arthanari et al. 2014), 352 in *Verticillium dahliae* (Li et al. 2022), and 427-819 in *F. graminearum* (Lu et al. 2021)). This difference could be due to the stringent criteria used for this study. In fact, when using similar criteria to previous studies, such as keeping all ORFs with no coding potential independently of their size, we identified 398 lncRNAs. In addition, many lncRNAs identified in these fungi were detected under specific conditions corresponding to stresses (Arthanari et al. 2014; Cemel et al. 2017) and sexual development (Lu et al. 2021) which we did not survey in our RNA samples. We identified 17 lncRNAs as differentially expressed during plant infection, mostly as antisense transcripts (Table S9). Two displayed expression patterns opposed to their coding genes. lncRNA PB1188.1 was down-regulated during infection compared to *in vitro* culture. This lncRNA is an antisense transcript of ZtIPO323_016330 encoding a secreted Subtilisin-like protein, that is up-regulated during infection. Subtilisin-like proteins are secreted proteases that play a role in plant infection (Li et al. 2010; Muszewska et al. 2011). This negative correlation suggested that the down regulation of lncRNA PB1188.1 during infection allowed the full expression of ZtIPO323_016330 in infected leaves. The second lncRNA (lncRNA PB.2709.1) was up-regulated during infection compared to *in vitro* culture (Table S8), while its corresponding transcript (ZtIPO323_037670) was down-regulated during infection. This transcript encodes a tubulin tyrosine ligase (TTL), a protein involved in the post-translational modification of tubulin. Its reduced expression could alter tubulin turnover. These negative correlations suggested that antisense lncRNAs could control fungal gene expression during infection. Our observations hint at the existence of co-regulation networks between coding and non-coding transcripts in *Z. tritici* and suggest that they could be important for infection, as observed during the infection of rice leaves by *M. oryzae* (Li et al. 2021). These examples stress the importance of including lncRNAs in future studies to have a comprehensive picture of the expression regulation landscape of *Z. tritici*.

RNA mycoviruses are widespread in *Z. tritici*

We detected two RNA mycoviruses in Iso-seq transcripts unmapped to the *Z. tritici* genome. Fusarivirus 1 (Zt-FV1) was previously identified in *Z. tritici* by the systematic screening of unmapped fungal RNA-Seq reads (Gilbert et al. 2019). We also identified a novel mycovirus, Zt-NV1 (Figure S10),

related to the narnavirus 4 of *Sclerotinia sclerotiorum* (SsNV4) (Jia et al. 2021). RNA-Seq reads corresponding to these two mycoviruses were detected in all our IPO323 RNA-Seq libraries, as well as in all publicly available *Z. tritici* RNA-Seq data, showing that these mycoviruses are widespread in *Z. tritici*. Zt-FV1 was the most abundant mycovirus, while Zt-NV1 was only detected in low abundance compared to Zt-FV1 (1/70,000). As mycovirus are known to induce strong phenotypic defects in other fungi, additional studies are needed to evaluate the role of these mycoviruses in the life cycle of *Z. tritici*, in particular its growth, sporulation and pathogenicity (Myers and James 2022).

InGenAnnot is a novel tool for improving gene structure prediction

Many tools (Stanke et al. 2006; Dubarry et al. 2016; Testa et al. 2015; Holt and Yandell 2011; Min et al. 2017; Lukashin 1998; Reid et al. 2014; Sallet et al. 2019) and protocols (Campbell et al. 2014) were established to predict gene models in eukaryotic genomes. Some were dedicated to fungal genome annotation (Haas et al. 2011; Testa et al. 2015; Reid et al. 2014) and were incorporated in bioinformatics workflows (Min et al. 2017). Evaluation of the reliability of an annotation is not an easy task. One of the most frequently used tools is BUSCO, based on the detection of genes encoding conserved proteins to evaluate the completeness of the annotation (Manni et al. 2021). More recently, new datasets and methods were proposed to test the reliability of gene annotations, taking into account intron and exon structures (Scalzitti et al. 2020). However, this evaluation was still based on selected datasets, representing a conserved and partial view of gene content of a genome.

InGenAnnot used the AED metrics (Eilbeck et al. 2009) to select the best gene model with transcript or protein evidence. We improved AED metrics by computing scores for each evidence (transcript, protein) and used a distinct score for Iso-Seq transcripts when available. We also introduced penalty scores for specific discrepancies between the gene model and evidence, in particular for unsupported intron splicing sites. This annotation strategy required an in-depth analysis of data provided as evidence to eliminate artefacts such as wrongly assembled RNA-Seq transcripts or rare Iso-Seq transcripts (see before). As each *ab initio* gene prediction software implements specific ML models with different specificity/sensibility for each data source, their implementation and training parameters are more or less tolerant to particularities such as short CDS or non-canonical splicing sites. The combination of different *ab initio* gene prediction software with distinct intrinsic characteristics has proved essential to avoid drawbacks from each software. Indeed, none of the *ab initio* gene prediction software used individually was able to predict more than 70-76% of the final gene models (Table 2).

Other tools than InGenAnnot have integrated the selection of the best gene models. EvidenceModeler (Haas et al. 2008) and TSEBRA (Gabriel et al. 2021) select the best gene models according to transcript evidence using other metrics than AED. EvidenceModeler is integrated in funannotate, which was already used to annotate the *Z. tritici* genome (MPI annotation). It did not perform better than the single *ab initio* gene prediction software used for InGenAnnot (Table 2), but it was not run with the same evidence as our study. BRAKER3 (Gabriel et al. 2024) was released after the completion of our work. Additionally, we used Helixer, a novel gene prediction software based on deep neural networks and hidden Markov models (Stiehler et al. 2021). Funannotate, BRAKER3 and Helixer were run to annotate of *Z. tritici* genome using the same transcript and protein evidence as InGenAnnot. We then compared the gene models predicted by each tool with RGMs using the AED metric. BRAKER3 predicted/selected gene models with the best overlap with RGMs (73%, Figure S14), followed by funannotate (67%) and Helixer (60%). Helixer appeared less specific and sensitive than the other tools, as it predicted a large number of unique genes (6,190 CDS, Figure S14) mostly without evidence. No single *ab initio* gene predictor (see Table 2, Helixer), nor pipelines selecting the best gene model predicted by two to four *ab initio* gene prediction software (funannotate, BRAKER3), were able to accurately predict all the gene models we selected (RGMs). Overall, this comparison showed that the combination of different *ab initio* gene prediction software is essential to generate a large diversity of gene models to select the best one according to evidence. The AED metric is efficient for this selection process, since it identified more gene models with evidence than funannotate or BRAKER3

(Figure S14). However, the accurate comparison of the InGenAnnot AED-based selection to Evidence modeler (funannotate) and TSEBRA (BRAKER3) requires the use of a fully curated annotated genome as a reference.

Conclusion

In the era of massive sequencing of eukaryotic genomes, inferring gene models by transcript and protein evidence is essential. In this article, we used the Iso-Seq technology to obtain a large dataset of full-length transcripts of the fungal pathogen of wheat *Z. tritici*. We also developed a novel software, InGenAnnot, to improve drastically gene annotation by selecting the best gene model according to transcript and protein evidence across gene models predicted by different software. We expect that our strategy will be useful for improving eukaryotic gene prediction, particularly in fungi with compact genomes. For species with only few previous annotations, we suggest the use of at least three independent *ab initio* gene prediction software to provide a sufficient number of gene models at each locus obtained by different pipelines. Transcriptomic datasets are also important. Without Iso-Seq data, the assembly of RNA-seq reads into transcripts should be performed carefully to avoid fusing transcripts due to their frequent overlap.

Data and materials availability

All raw sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession GSE218898 with data accessions: GSM6758342 to GSM6758379. Processed data files of assembled RNA-Seq transcripts and filtered Iso-Seq reads were associated to the submission. Sequence of the new mycovirus ZtNV1 was deposited to NCBI under accession OP903463. Previous *Z. tritici* IP0323 gene annotations, new annotations (RGMs, Isoforms, LncRNAs) and the annotation file, denoted file S1 (z.tritici.IP0323.annotations.txt), are available at: https://doi.org/10.57745/CVIRIB.

A genome browser with all annotations and evidence was set up at: https://bioinfo.bioger.inrae.fr/portal/genome-portal/12/.

A new IPO323 genome web site at (https://mycocosm.jgi.doe.gov/Zymtr1/Zymtr1.home.html) was released with new genome annotations.

The InGenAnnot code and project is available at: https://forgemia.inra.fr/bioger/ingenannot
Licensed under GNU GPL v3. InGenAnnot documentation is available at https://bioger.pages.mia.inra.fr/ingenannot

Acknowledgments

We are grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing help and/or computing and/or storage resource. BIOGER benefits from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007). We also thank the BARIC workgroup (https://www.cesgo.org/catibaric/) for providing storage and computational resources. Syngenta Crop Protection is acknowledged for funding the sequencing of the cDNA libraries. Rothamsted Research receives strategic funding from the Biotechnology and Biological Sciences Research Council of the United Kingdom (BBSRC). We acknowledge support from the Growing Health [BB/X010953/1], Delivering Sustainable Wheat [BB/X011003/1] and the Resilient Farming Futures [BB/X010961/1] Institute Strategic Programs. The work conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, under proposal (10.46936/10.25585/60008023), is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

Authorship

N.L. and M.H.L. conceived the strategy used for annotation and supervised the project. C.D., M.A, J.S, M.H.L. and G.S. performed experiments needed for constructing cDNA libraries. G.S. funded the sequencing of the cDNA libraries. L.M. developed tools for annotation and performed initial analyses with an early version of InGeAnnot. N.L. finalized InGeAnnot, and genome annotation. A.F., L.M., N.L., Y.P., A.G., G.S and M.H.L compared RGMs to previous annotations. A.F., L.N.A., D.S, R.K., A.R., A.S.S., S.B.G., G.H.J.K., I.V.G., J.H., J.R., E.S, D.C and G.S provided data for genome annotation. N.L., A.S.S., I.V.G., E.S. and M.H.L. set up the genome browsers. N.L., G.S. and M.H.L wrote the draft of the manuscript. All authors discussed of the results and contributed to the improvement of the manuscript.

Literature Cited

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21

An, D., Cao, H. X., Li, C., Humbeck, K., and Wang, W. 2018. Isoform Sequencing and State-of-Art Applications for Unravelling Complexity of Plant Transcriptomes. Genes (Basel). 9

Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., Von Heijne, G., Elofsson, A., and Nielsen, H. 2019. Detecting sequence signals in targeting peptides using deep learning. Life Sci. Alliance. 2

Arthanari, Y., Heintzen, C., Griffiths-Jones, S., and Crosthwaite, S. K. 2014. Natural Antisense Transcripts and Long Non-Coding RNA in Neurospora crassa K. McCluskey, ed. PLoS One. 9:e91353

Badet, T., Oggenfuss, U., Abraham, L., McDonald, B. A., and Croll, D. 2020. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. BMC Biol. 18

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19:455–477

Beiki, H., Liu, H., Huang, J., Manchanda, N., Nonneman, D., Smith, T. P. L., Reecy, J. M., and Tuggle, C. K. 2019. Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. BMC Genomics. 20:344

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. Genome Res. 14:988–995

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., and Borodovsky, M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinforma. 3:1–11

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. 2014. Genome Annotation and Curation Using MAKER and MAKER-P. Curr. Protoc. Bioinforma. 48:4.11.1-39

Casco, A., Gupta, A., Hayes, M., Djavadian, R., Ohashi, M., and Johannsen, E. 2022. Accurate Quantification of Overlapping Herpesvirus Transcripts from RNA Sequencing Data. J. Virol. 96

Cemel, I. A., Ha, N., Schermann, G., Yonekawa, S., and Brunner, M. 2017. The coding and noncoding transcriptome of Neurospora crassa. BMC Genomics. 18

Chen, H., King, R., Smith, D., Bayon, C., Ashfield, T., Torriani, S., Kanyuka, K., Hammond-Kosack, K., Bieri, S., and Rudd, J. 2023. Combined pangenomics and transcriptomics reveals core and redundant virulence processes in a rapidly evolving fungal plant pathogen. BMC Biol. 21:1–22

Chen, Y., Wu, Y., Liu, L., Feng, J., Zhang, T., Qin, S., Zhao, X., Wang, C., Li, D., Han, W., Shao, M., Zhao, P., Xue, J., Liu, X., Li, H., Zhao, E., Zhao, W., Guo, X., Jin, Y., Cao, Y., Cui, L., Zhou, Z., Xia, Q., Rao, Z., and Zhang, Y. 2019. Study of the whole genome, methylome and transcriptome of Cordyceps militaris. Sci. Rep. 9:1–15

Chiba, Y., Yoshizaki, K., Tian, T., Miyazaki, K., Martin, D., Saito, K., Yamada, A., and Fukumoto, S. 2022. Integration of Single-Cell RNA- and CAGE-seq Reveals Tooth-Enriched Genes. J. Dent. Res. 101

Cook, D. E., Valle-Inclan, J. E., Pajoro, A., Rovenich, H., Thomma, B. P. H. J., and Faino, L. 2019. Long-

Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing. Plant Physiol. 179:38–54

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., Claverie, J., and Gascuel, 0livier. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36

Dhillon, B., Gill, N., Hamelin, R. C., and Goodwin, S. B. 2014. The landscape of transposable elements in the finished genome of the fungal wheat pathogen Mycosphaerella graminicola. BMC Genomics. 15

Donaldson, M. E., Ostrowski, L. A., Goulet, K. M., and Saville, B. J. 2017. Transcriptome analysis of smut fungi reveals widespread intergenic transcription and conserved antisense transcript expression. BMC Genomics. 18

Dubarry, M., Noel, B., Rukwavu, T., Farhat, S., Silva, C. Da, Seeleuthner, Y., Lebeurrier, M., Aury, J.-M., Dubarry, M., Noel, B., Rukwavu, T., Farhat, S., Da Silva, C., Seeleuthner, Y., Lebeurrier, M., and Aury, J.-M. 2016. Gmove a tool for eukaryotic gene predictions using various evidences. F1000Research. 5

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

Ejigu, G. F., and Jung, J. 2020. Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology (Basel). 9

Feurtey, A., Lorrain, C., Croll, D., Eschenbrenner, C., Freitag, M., Habig, M., Haueisen, J., Möller, M., Schotanus, K., and Stukenbrock, E. H. 2020. Genome compartmentalization predates species divergence in the plant pathogen genus Zymoseptoria. BMC Genomics. 21

funannotate.

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., and Stanke, M. 2024. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 34:769–777

Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., and Stanke, M. 2021. TSEBRA: transcript selector for BRAKER. BMC Bioinformatics. 22:566

Gay, E. J., Soyer, J. L., Lapalu, N., Linglin, J., Fudal, I., Da Silva, C., Wincker, P., Aury, J. M., Cruaud, C., Levrel, A., Lemoine, J., Delourme, R., Rouxel, T., and Balesdent, M. H. 2021. Large-scale transcriptomics to dissect 2 years of the life of a fungal phytopathogen interacting with its host plant. BMC Biol. 19

Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32:258D – 261

Gerads, M., and Ernst, J. F. 1998. Overlapping coding regions and trancriptional units of two essential chromosomal genes (CCT8, TRP1) in the fungal pathogen Candida albicans. Nucleic Acids Res. 26

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

Goodwin, S. B., M'Barek, S. Ben, Dhillon, B., Wittenberg, A. H. J., Crane, C. F., Hane, J. K., Foster, A. J., van der Lee, T. A. J., Grimwood, J., Aerts, A., Antoniw, J., Bailey, A., Bluhm, B., Bowler, J., Bristow, J., van der Burgt, A., Canto-Canché, B., Churchill, A. C. L., Conde-Ferràez, L., Cools, H. J., Coutinho, P. M., Csukai, M., Dehal, P., de Wit, P., Donzelli, B., van de Geest, H. C., van Ham, R. C. H. J., Hammond-Kosack, K. E., Henrissat, B., Kilian, A., Kobayashi, A. K., Koopmann, E., Kourmpetis, Y., Kuzniar, A., Lindquist, E., Lombard, V., Maliepaard, C., Martins, N., Mehrabi, R., Nap, J. P. H., Ponomarenko, A., Rudd, J. J., Salamov, A., Schmutz, J., Schouten, H. J., Shapiro, H., Stergiopoulos, I., Torriani, S. F. F., Tu, H., de Vries, R. P., Waalwijk, C., Ware, S. B., Wiebenga, A., Zwiers, L. H., Oliver, R. P., Grigoriev, I. V., and Kema, G. H. J. 2011. Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 7

Gordon, S. P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., Kang, D., Underwood, J., Grigoriev, I. V., Figueroa, M., Schilling, J. S., Chen, F., and Wang, Z. 2015. Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing D. Zheng, ed. PLoS One.

10:e0132628
Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talón, M., Dopazo, J., and Conesa, A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36:3420
Grandaubert, J., Bhattacharyya, A., and Stukenbrock, E. H. 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 Genes, Genomes, Genet. 5:1323–1333
Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., and Wortman, J. R. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9:R7
Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A., and Wortman, J. R. 2011. Approaches to fungal genome annotation. Mycology. 2:118–141
Hadar, S., Meller, A., and Shalgi, R. 2022. Stress-induced transcriptional readthrough into neighboring genes is linked to intron retention. bioRxiv. :2022.03.24.485601
Hansen, K., Birse, C. E., and Proudfoot, N. J. 1998. Nascent transcription from the nmt1 and nmt2 genes of Schizosaccharomyces pombe overlaps neighbouring genes. EMBO J. 17
Holst, F., Bolger, A., Günther, C., Maß, J., Triesch, S., Kindel, F., Kiel, N., Saadat, N., Ebenhöh, O., Usadel, B., Schwacke, R., Bolger, M., Weber, A. P. M., and Denton, A. K. 2023. Helixer–de novo Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model. bioRxiv. :2023.02.06.527280
Holt, C., and Yandell, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 12:491
Hunt, S. P., Jarvis, D. E., Larsen, D. J., Mosyakin, S. L., Kolano, B. A., Jackson, E. W., Martin, S. L., Jellen, E. N., and Maughan, P. J. 2020. A Chromosome-Scale Assembly of the Garden Orach (Atriplex hortensis L.) Genome Using Oxford Nanopore Sequencing. Front. Plant Sci. 11:539054
Ingolia, N. T. 2014. Ribosome profiling: New views of translation, from single codons to genome scale. Nat. Rev. Genet. 15:205–213
Jeon, J., Kim, K. T., Choi, J., Cheong, K., Ko, J., Choi, G., Lee, H., Lee, G. W., Park, S. Y., Kim, S., Kim, S. T., Min, C. W., Kang, S., and Lee, Y. H. 2022. Alternative splicing diversifies the transcriptome and proteome of the rice blast fungus during host infection. RNA Biol. 19
Jia, J., Fu, Y., Jiang, D., Mu, F., Cheng, J., Lin, Y., Li, B., Marzano, S. Y. L., and Xie, J. 2021. Interannual dynamics, diversity and evolution of the virome in Sclerotinia sclerotiorum from a single crop field. Virus Evol. 7
Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., and Hunter, S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30:1236
Kupfer, D. M., Drabenstot, S. D., Buchanan, K. L., Lai, H., Zhu, H., Dyer, D. W., Roe, B. A., and Murphy, J. W. 2004. Introns and splicing elements of five diverse fungi. Eukaryot. Cell. 3
Lelandais, G., Remy, D., Malagnac, F., and Grognet, P. 2022. New insights into genome annotation in Podospora anserina through re-exploiting multiple RNA-seq data. BMC Genomics. 23:859
Li, J., Yu, L., Yang, J., Dong, L., Tian, B., Yu, Z., Liang, L., Zhang, Y., Wang, X., and Zhang, K. 2010. New insights into the evolution of subtilisin-like serine protease genes in Pezizomycotina. BMC Evol. Biol. 10
Li, R., Xue, H. S., Zhang, D. D., Wang, D., Song, J., Subbarao, K. V., Klosterman, S. J., Chen, J. Y., and Dai, X. F. 2022. Identification of long non-coding RNAs in Verticillium dahliae following inoculation of cotton. Microbiol. Res. 257
Li, Z., Yang, J., Peng, J., Cheng, Z., Liu, X., Zhang, Z., Bhadauria, V., Zhao, W., and Peng, Y. L. 2021. Transcriptional Landscapes of Long Non-coding RNAs and Alternative Splicing in Pyricularia oryzae Revealed by RNA-Seq. Front. Plant Sci. 12
Liu, N., Wang, P., Li, X., Pei, Y., Sun, Y., Ma, X., Ge, X., Zhu, Y., Li, F., and Hou, Y. 2022. Long Non-Coding RNAs profiling in pathogenesis of Verticillium dahliae: New insights in the host-pathogen

interaction. Plant Sci. 314:111098
Lorrain, C., Feurtey, A., Ller, M. M., Haueisen, J., and Stukenbrock, E. 2021. Dynamics of transposable elements in recently diverged fungal pathogens: Lineage-specific transposable element content and efficiency of genome defenses. G3 Genes, Genomes, Genet. 11
Lu, P., Chen, D., Qi, Z., Wang, H., Chen, Y., Wang, Q., Jiang, C., Xu, J.-R., and Liu, H. 2021. Landscape, complexity and regulation of a filamentous fungal transcriptome 1 Corresponding author: 8 Running Title: Full-length transcriptome of F. graminearum 12. bioRxiv. :2021.11.08.467853
Lukashin, A. 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26:1107–1115
Manni, M., Berkeley, M., Seppey, M., Simão, F., and Zdobnov, E. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol. Biol. Evol. 38
Min, B., Grigoriev, I. V., and Choi, I.-G. 2017. FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation. Bioinformatics. 33:2936–2937
Mirzadi Gohari, A., Ware, S. B., Wittenberg, A. H. J., Mehrabi, R., Ben M'Barek, S., Verstappen, E. C. P., van der Lee, T. A. J., Robert, O., Schouten, H. J., de Wit, P. P. J. G. M., and Kema, G. H. J. 2015. Effector discovery in the fungal wheat pathogen Zymoseptoria tritici. Mol. Plant Pathol. 16:931–945
Moller, M., Habig, M., Lorrain, C., Feurtey, A., Haueisen, J., Fagundes, W. C., Alizadeh, A., Freitag, M., and Stukenbrock, E. H. 2021. Recent loss of the Dim2 DNA methyltransferase decreases mutation rate in repeats and changes evolutionary trajectory in a fungal pathogen. PLoS Genet. 17:e1009448
Möller, S., Croning, M. D. R., and Apweiler, R. 2001. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics. 17:646–653
Morais do Amaral, A., Antoniw, J., Rudd, J. J., and Hammond-Kosack, K. E. 2012. Defining the Predicted Protein Secretome of the Fungal Wheat Leaf Pathogen Mycosphaerella graminicola G.H. Goldman, ed. PLoS One. 7:e49904
Muszewska, A., Taylor, J. W., Szczesny, P., and Grynberg, M. 2011. Independent subtilases expansions in fungi associated with animals. Mol. Biol. Evol. 28:3395–3404
Myers, J. M., and James, T. Y. 2022. Mycoviruses. Curr. Biol. 32:R150–R155
Nielsen, H. 2017. Predicting secretory proteins with signalP. Pages 59–73 in: Methods in Molecular Biology, Humana Press Inc.
Oggenfuss, U., Badet, T., Wicker, T., Hartmann, F. E., Singh, N. K., Abraham, L., Karisto, P., Vonlanthen, T., Mundt, C., McDonald, B. A., and Croll, D. 2021. A population-level invasion by transposable elements triggers genome expansion in a fungal pathogen. Elife. 10
Petit-Houdenot, Y., Lebrun, M.-H., and Scalliet, G. 2021. Understanding plant-pathogen interactions in Septoria tritici blotch infection of cereals. Pages 263–302 in: Achieving durable disease resistance in cereals, Burleigh Dodds Science Publishing, London.
Quaedvlieg, W., Kema, G. H. J., Groenewald, J. Z., Verkley, G. J. M., Seifbarghi, S., Razavi, M., Mirzadi Gohari, A., Mehrabi, R., and Crous, P. W. 2011. Zymoseptoria gen. nov.: A new genus to accommodate Septoria-like species occurring on graminicolous hosts. Persoonia Mol. Phylogeny Evol. Fungi. 26:57–69
Quinlan, A. R., and Hall, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842
Raghavan, V., Kraft, L., Mesny, F., and Rigerte, L. 2022. A simple guide to *de novo* transcriptome assembly and annotation. Brief. Bioinform. 23:1–30
Reid, I., O'Toole, N., Zabaneh, O., Nourzadeh, R., Dahdouli, M., Abdellateef, M., Gordon, P. M., Soh, J., Butler, G., Sensen, C. W., and Tsang, A. 2014. SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models. BMC Bioinformatics. 15:229
Sallet, E., Gouzy, J., and Schiex, T. 2019. EuGene: An automated integrative gene finder for eukaryotes and prokaryotes. Pages 97–120 in: Methods in Molecular Biology, Humana Press Inc.
Salzberg, S. L. 2019. Next-generation genome annotation: We still struggle to get it right. Genome

Biol. 20
Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O., and Thompson, J. D. 2020. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. BMC Genomics. 21:293
Schotanus, K., Soyer, J. L., Connolly, L. R., Grandaubert, J., Happel, P., Smith, K. M., Freitag, M., and Stukenbrock, E. H. 2015. Histone modifications rather than the novel regional centromeres of Zymoseptoria tritici distinguish core and accessory chromosomes. Epigenetics and Chromatin. 8
Standage, D. S., and Brendel, V. P. 2012. ParsEval: Parallel comparison and analysis of gene structure annotations. BMC Bioinformatics. 13:187
Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34:W435–W439
Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A. P. M., and Denton, A. K. 2021. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. Bioinformatics. 36:5291–5298
Tardaguila, M., De La Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., Martens, L., Rodriguez-Navarro, S., Moreno-Manzano, V., and Conesa, A. 2018. SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 28:396–411
Testa, A. C., Hane, J. K., Ellwood, S. R., and Oliver, R. P. 2015. CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics. 16:170
Till, P., Mach, R. L., and Mach-Aigner, A. R. 2018. A current view on long noncoding RNAs in yeast and filamentous fungi. Appl. Microbiol. Biotechnol. 102:7319–7331
Zhang, G., Sun, M., Wang, J., Lei, M., Li, C., Zhao, D., Huang, J., Li, W., Li, S., Li, J., Yang, J., Luo, Y., Hu, S., and Zhang, B. 2019. PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. Plant J. 97:296–305
Zhang, Y., and Sachs, M. S. 2015. Control of mRNA stability in fungi by NMD, EJC and CBC factors through 3'UTR introns. Genetics. 200:1133–1148
Zhong, Z., Marcel, T. C., Hartmann, F. E., Ma, X., Plissonneau, C., Zala, M., Ducasse, A., Confais, J., Compain, J., Lapalu, N., Amselem, J., McDonald, B. A., Croll, D., and Palma-Guerrero, J. 2017. A small secreted protein in *Zymoseptoria tritici* is responsible for avirulence on wheat cultivars carrying the *Stb6* resistance gene. New Phytol. 214:619–631

Figure captions

Figure 1. Comparison of *Zymoseptoria tritici* reference isolate IPO323 genome annotations. **a**) Upset plot of the gene models from the four annotations of IPO323 (JGI, MPI, RRES and CURTIN). Numbers of gene models with identical coding sequences (CDS). **b**) Comparison of IPO323 gene annotations. Number of CDS in each annotation. Identical CDS: identical CDS at a given locus. Unique Dissimilar CDS: at a given locus, a CDS is predicted by at least one other annotation, but they differ in their structure. Unique Specific CDS: at a given locus, a single CDS is predicted by a single annotation. The highest numbers of identical gene models between two annotations were observed for MPI-RRES (8,442), RRES-CURTIN (8,289), and MPI-Curtin (7,981), while the lowest numbers of identical gene models were observed between JGI and the three other annotations (4,495, 4,621 and 5,276 for JGI-Curtin, JGI-MPI and JGI-RRES respectively).

Figure 2. Selection of the best Re-annotated Gene Models (RGMs) according to their Annotation Edit Distance (AED) scores.

Plot of RGM AED scores. AED scores (0-1) describe how a given gene model fits to transcript and protein evidence (best fit = 0). Transcript evidence was computed from RNA-Seq or Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines represent the AED thresholds to filter out genes (0.3 for transcripts, 0.1 for proteins), except if they are supported by at least four different annotations (1846 RGMs, upper right area of the graph). The numbers of genes in the four areas are displayed in white text boxes. Numbers of transcripts with transcript evidence were plotted on cumulative histograms above the scatter plot (green). Numbers of transcripts with protein evidence were plotted on cumulative histograms on the right of the scatter plot (red).

Figure 3. Comparison of the novel IPO323 genome annotation (Re-annotated Gene Models, RGM) with the four available annotations.

a) Upset plot of RGMs with gene models from the four available annotations (JGI, MPI, RRES and CURTIN). Numbers of shared (identical) gene models for coding sequences (CDS).

b) Numbers of identical CDS between RGMs and each available annotation.

Figure 4. Transcript isoforms of Re-annotated Gene Models (RGMs) (a) ZtIPO323_030030 and (b) ZtIPO323_013330 supported by Iso-Seq and RNA-Seq evidence.

a) Gene ZtIPO323_030030 (chr2: 777930…1778675, 747 b). This RGM has two transcript isoforms (alternative 3' acceptor site). Both encoded Small, Secreted Proteins (SSP 10, File S1). Previous annotations selected the second acceptor site leading to the longest CDS. A single Iso-Seq transcript corresponding to the longest CDS was detected (Iso-Seq track), while both isoforms were detected using RNA-Seq data (RNA-Seq assembled transcript). RNA-Seq coverage identified both isoforms in equal amounts (RNA-Seq coverage Xyl track). Based on read coverage from different RNA-Seq libraries, the isoform corresponding to the shortest CDS was the most frequent. This isoform was likely the canonical form and encoded a protein with a C-terminus that was reduced in length by 34% compared to the other isoform. RGMs with isoforms track: different isoforms. Iso-Seq track: filtered Iso-Seq transcripts. RNA-Seq coverage Xyl track: coverage of strand-specific RNA-Seq reads from the Xylose as

sole carbon source medium library. RNA-Seq assembled transcript track: assembly of strand-specific RNA-Seq reads.

b) ZtIPO323_013330 (chr_1:3420115..3424093, 3.98 Kb). This RGM had four transcript isoforms. The selected RGM had four splicing sites, one of which in the 5' UTR was supported by Iso-Seq transcript (Iso-Seq n°2) and RNA-Seq (RNA-Seq coverage Xyl) data. Two Iso-Seq transcripts with one or two intron retention events were detected as Iso-Seq transcripts (Iso-Seq n°1 and 3) and confirmed by RNA-Seq (RNA-Seq coverage Xyl). One Iso-Seq transcript had an alternative 5' donor splicing site in the 5' UTR (Iso-Seq n°4). This isoform was likely weakly expressed, as it was not supported by RNA-Seq (RNA-Seq coverage Xyl). RGMs with isoforms track: different RGM isoforms. Iso-Seq track: filtered Iso-Seq transcripts. RNA-Seq coverage Xyl track: coverage of strand-specific RNA-Seq reads. RNA-Seq assembled transcript track: assembly of strand-specific RNA-Seq reads.

Figure 5. Examples of polycistronic transcripts shown for Re-annotated Gene Models (RGMs) ZtIPO323_010430 and ZtIPO323_010440.

RGMs ZtIPO323_010430 and ZtIPO323_010440, located at chr_1:2692858...2697168 and chr_1:2692858...2697168, respectively, were transcribed on the same strand with overlapping 3' UTR and 5' UTR (red rectangle). Iso-Seq polycistronic track: evidence of transcripts covering the two RGMs. A strong decrease in RNA-Seq coverage was observed in the region of the overlap (red, dashed rectangle), suggesting two singles, overlapping transcripts. The assembly of RNA-Seq reads led to a polycistronic transcript involving the two RGMs, likely resulting from the wrong assembly of reads from these overlapping transcripts. Iso-Seq track: filtered Iso-Seq transcripts mapping at this locus. Iso-Seq polycistronic track: polycistronic transcripts identified in the Iso-Seq database. RNA-Seq transcript track: assembly of strand-specific RNA-Seq reads mapping at this locus. RNA-Seq coverage Xyl track: coverage of strand-specific RNA-Seq reads from the Xylose as sole carbon source medium library.

Tables

Categories	Counts
Full-splice_match (FSM) [1]	7872
Incomplete-splice_match (ISM)[2]	305
Fusion	45
Genic[3]	664
Intron retention (IR)	1571
novel_in_catalog (NIC)[4]	7
novel_not_in_catalog (NNC)[5]	474
Antisense	395
Intergenic	357

[1] Whole transcripts with possible alternative 3' and 5' ends
[2] Partial overlaps of transcripts fitting with intron coordinates
[3] Partial overlaps of introns and exons not compliant with intron/exon coordinates
[4] Use combination_of_known_splice sites
[5] At_least_one_novel_splice site detected

Table 1. Classification of Iso-Seq transcript isoforms from *Zymoseptoria tritici* isolate IPO323 where filtered Iso-Seq transcripts from different growth conditions were analysed and classified with Sqanti3.

Type	Annotation	Identical CDS[1]			Unique identical CDS[2]	
Available annotations	JGI (FGENESH/Genewise[3])	4865	48 %	11367	157	929
	MPI (Evidence modeler[3])	8431	62 %		91	
	RRES (MAKER-HMM[3])	8317	62 %		175	
	CURTIN (CodingQuarry[3])	9584	71 %		506	
New annotations	Eugene[3]	10224	76 %	11677	1603	1802
	LoReAn[3]	7769	58 %		199	

Table 2. Contribution of each annotation of the *Zymoseptoria tritici* IPO323 genome to Re-annotated Gene Models (RGMs).

[1] Identical coding sequence (CDS): number of CDS identical to RGMs

[2] Unique identical coding sequence (CDS): number of CDS predicted in a single annotation and retained as RGMs.

[3] *ab initio* gene prediction software used for the given annotation

The annotations that contributed the most to RGMs were respectively Eugene (76% Identical CDS*, 1603 Unique identical CDS**) and Curtin (71% Identical CDS, 506 Unique identical CDS). Combining gene models from the four available annotations (JGI, MPI, RRES, CURTIN) showed that 11,367 of their CDS were identical to RGMs (contribution: 84.7%). Combining gene models from the two new annotations (Eugene, LoRean) showed that 11,677 of their CDS were identical to RGMs (contribution: 87%). The combination of the six annotations was needed to predict all the 13,414 RGMs.

Supplementary figure captions

Figure S1. Annotation Edit Distance (AED) computation using InGenAnnot
Annotation Edit Distance (AED) is an annotation quality-control measure, proposed by Maker (Holt and Yandell 2011) to compare annotations based on their overlap. We have used AED scores to quantify the concordance between a gene model and its associated evidence (transcript, protein), as previously described (Eilbeck et al. 2009). Several options for computing this customized AED were implemented, such as restriction to coding sequence (CDS) or penalty on unsupported intron splicing sites. No intron splicing site penalty was applied for introns located in UTRs, nor for introns defined using protein alignments which could be biased by phylogenetically distant proteins. The table displays AED scores for two gene models (G1, G2) located at the same locus (metagene). The orange rectangle (L1) corresponds to a single Iso-Seq transcript detected at this locus. Blue rectangles (T1-T3,) correspond to the different assembled RNA-seq transcripts detected at this locus. Green rectangles (P1-P4) correspond to gene models deduced from protein alignments. As G2 had no UTRs, calculating AED on CDS only or on the complete gene has no impact on the score. However, since the Iso-Seq transcript has UTRs, G1 AED score was better (lower) when the complete gene was used instead of the CDS. G2 contains an intron splicing site not supported by transcript evidence, resulting in a penalty (0.25) added to the raw AED score. In this case, G1 was selected as the best gene model.

Figure S2. Bioinformatics workflow with the different steps of InGenAnnot tool suite
Sources of evidence (Proteins, Iso-Seq and RNA-Seq transcripts) were used to predict new gene models with Eugene and LoReAn. Gene models predicted by novel and previous *ab initio* gene prediction software were submitted to the following workflow. Coding Sequences (CDS) were filtered out if overlapping with known transposable elements (TE). Then, filtered gene models were used to compute AED scores for each source of evidence. The best gene model at a given locus (metagene) was selected based on its AED score (lower than 0.3 for transcripts, lower than 0.1 for proteins). Gene models failing the AED score threshold, but predicted by at least 4 independent gene predictors, were retained. All the gene models without ATG or stop codon were removed. Transcripts without gene models were analysed to infer potential missing effectors (small, secreted proteins). Finally, Iso-Seq transcripts were filtered according to their RNA-Seq support (low abundance Iso-Seq were removed) before being used to define UTRs. All of this workflow was described with associated command lines in the documentation of InGenAnnot (https://bioger.pages.mia.inra.fr/ingenannot/usecases/select_best_gene_models.html). The gene models selected by InGenAnnot (RGMs) were compared to all gene models to compute AED scores and quantify the contribution of each *ab initio* gene prediction software to the final RGM dataset (https://bioger.pages.mia.inra.fr/ingenannot/usecases/annotation_comparison.html).

Figure S3. Comparison of the Annotation Edit Distance (AED) scores of the JGI, MPI, RRES and CURTIN gene models.
AED scores (0-1) described how a given gene model fit the transcript and protein evidence (best fit = 0). Transcript evidence was computed from RNA-Seq and Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines represent the AED thresholds to filter out genes (0.3 for transcripts, 0.1 for proteins), except if they are supported by at least 4 different annotations (upper right area of the graph). The numbers of genes from each area were displayed in white boxes.

Figure S4. Cumulative distributions of the best Annotation Edit Distance (AED) scores for Re-annotated Gene Models (RGMs) and those from previous annotations.
AED scores (0-1) described how a given gene model fit the transcript and protein evidence (best fit = 0). The best AED score (X axis) was computed from either transcript or protein evidence. a) cumulative plot of the number of transcripts; b) cumulative plot of the density of transcripts (normalized). The red line indicated the cutoff used to select the best gene model (0.3 for transcript evidence).

Figure S5. Comparison of the Annotation Edit Distance (AED) scores of split Re-annotated Gene Models (RGMs) and their corresponding fused gene models from the MPI annotation.
AED scores (0-1) described how a given gene model fit the transcript and protein evidence (best fit = 0). Split RGMs were displayed in blue and the corresponding MPI fused genes were displayed in orange. Transcript evidence was computed from RNA-Seq or Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines corresponded to the AED thresholds used to filter out gene models (0.3 for transcripts, 0.1 for proteins). The number of transcripts for each value of AED score was plotted on cumulative histograms above the scatter plot (RGM in blue, MPI in orange). The number of transcripts with protein evidence were plotted on cumulative histograms on the right of the scatter plot (RGM in blue, MPI in orange).

Figure S6. Fused genes from the MPI annotation located at chr_8:940236...948036 split into two Re-annotated Gene Models (RGM-1 and RGM-2).
In the region of chr_8:940236...948036 (7.8 Kb), three RGMs were predicted (red, dashed squares: RGM-1, RGM-2 and RGM-3). The MPI annotation predicted a gene model corresponding to the fusion of RGM-2 and RGM-3 (green rectangle). This fusion had no transcript evidence (Iso-seq, RNA-Seq). Specific transcripts of RGM-2 (Iso-Seq, RNAseq) and RGM-3 (RNAseq) were detected. On the other strand, RGM-1 had correctly predicted intron splicing sites, while the corresponding gene models from the MPI and JGI annotations had incorrect intron splicing sites. RGM-3 was not predicted by JGI, despite encoding a conserved Histone-3 variant protein. RGM track: RGM gene models. MPI track: MPI gene models. JGI track: JGI gene models. Iso-Seq track: filtered Iso-Seq transcripts. RNA-Seq forward track: coverage of forward-strand RNA-Seq reads. RNA-Seq reverse track: coverage of reverse-strand RNA-Seq reads mapping at this locus. RNA-Seq transcript track: assembled RNA-Seq transcripts.

Figure S7. Fused genes from the MPI annotation located at chr_2:2938368...2940768 split into two Re-annotated Gene Models (RGM-1 and RGM-2).
In the region of chr_2:2938368...2940768 (2.4 Kb), two RGMs were predicted (red, dashed squares). The MPI annotation predicted a gene model corresponding to the fusion of RGM-1 and RGM-2 (green rectangle) that was not supported by Iso-Seq. Iso-Seq transcripts supporting RGM-1 were detected. RGM-2 with no Iso-seq support was not predicted by JGI and RRES, while it was predicted as an independent gene in the Curtin annotation). RGM2 (ZtIPO323_034630) encoded a Small, Secreted Protein (SSP, see File S1). Assembled RNA-seq transcripts corresponding to the fused MPI gene model were detected. They were likely artefacts from assembly of overlapping RNA-Seq reads. Indeed, in a specific condition (infection at 13 days post inoculation), only RGM-2 transcript was detected (RNA-seq coverage 13 dpi track), supporting the prediction of RGM-2. RGM track: RGM gene models. MPI track: MPI gene model. JGI track: JGI gene model. RRES track: RRES gene model. Curtin track: Curtin gene models. Iso-Seq track: filtered Iso-Seq transcripts. RNA-Seq transcript track: assembled RNA-Seq transcripts. RNA-seq coverage Xylose: coverage of strand-specific RNA-Seq reads from a Xylose medium library. RNA-seq coverage infection 13 dpi: coverage of strand-specific RNA-Seq reads from a 13-dpi wheat infection library.

Figure S8a. *Avr-Stb6* paralogs located on chromosome 10 predicted by the new annotation.
The two new paralogs of *Avr-Stb6* (ZtIPO323_106210, ZtIPO323_106220) are located head to tail on chromosome 10 between position 534287 and 536486. ZtIPO323_106210 was predicted in the JGI and MPI annotations, but the gene models did not match Iso-Seq and RNA-seq evidence. ZtIPO323_106220 was not predicted by any previous annotations. RGM annotation was only supported by RNA-seq evidence. ZtIPO323_106210 is expressed during *in vitro* growth (RNA-Seq coverage glucose track), and infection stages (RNA-Seq coverage infection 11 dpi track). ZtIPO323_106220 was differentially upregulated during infection (RNA-Seq coverage infection 11 dpi track) to the same level as ZtIPO323_106210. RGM track: RGM gene models. JGI track: JGI gene model. MPI track: MPI gene model. Iso-Seq track: filtered Iso-Seq transcripts. RNA-seq coverage Glucose: coverage of strand-

specific RNA-Seq reads from a Glucose medium library. RNA-seq coverage infection 11 dpi: coverage of strand-specific RNA-Seq reads from a 11-dpi wheat infection library.

Figure S8b. Original *Avr-Stb6* located on chromosome 5.
The original *Avr-Stb6* is located at the end of chromosome 5. It was not predicted by any previous annotations, while it was correctly predicted as RGM (see track RGM track reannotation). Iso-Seq transcripts were detected (see track Iso-Seq long reads), since this gene is highly expressed during *in vitro* growth of *Zymoseptoria tritici*, as well as during infection (see tracks RNA-seq short reads distribution). RGM track: RGM gene models. JGI track: JGI gene model. MPI track: MPI gene model. RRES track: RRES gene model. Curtin track: Curtin gene models. Iso-Seq track: filtered Iso-Seq transcripts. RNA-seq distribution Glucose: coverage of strand-specific RNA-Seq reads from a Glucose medium library. RNA-seq distribution infection 11 dpi: coverage of strand-specific RNA-Seq reads from a 11-dpi wheat infection library.

Figure S9. Isoforms of Re-annotated Gene Model (RGM) ZtIPO323_013090 supported by Iso-Seq.
RGM ZtIPO323_013090 located at chr_1:3358097...3361350 (3.25 Kb) has four different isoforms. One alternative splicing site (red, dashed rectangle) supported by RNA-Seq, was not supported by Iso-Seq. Another alternative splicing site detected both by RNA-Seq and Iso-Seq (black-dashed rectangle), was not used to predict an isoform by *ab initio* software due to a stop codon before the splicing site. Finally, the canonical form retained is the transcript without any introns. This selection could be an artefact of transcripts from AE medium inducing many intron retention events. The canonical isoform is likely the RGM corresponding to the Iso-Seq transcript PB.940.5 with two introns (isoform 3). RGM track: RGM gene model. Iso-Seq track: filtered Iso-Seq transcripts. RNA-Seq coverage Xyl track: coverage of strand-specific RNA-Seq reads from the Xylose as sole carbon source medium library. RNA-Seq coverage AE track: coverage of strand-specific RNA-Seq reads from AE medium library.

Figure S10. Phylogenetic tree of RNA-dependent RNA polymerases of fungal narnaviruses related Zt-NV1 from *Zymoseptoria tritici.*
Iso-Seq transcripts not mapping to the *Z. tritici* IPO323 reference genome were clustered with blastclust. Similarities with known sequences were analysed by *blastn* search against the NCBI nr database. Reconstruction of the full-length sequences of viruses was performed by de-novo assembly with SPAdes (v3.15.4) (Bankevich et al. 2012). RNA-dependent RNA polymerase sequences from narnaviruses related to Zt-NV1 were retrieved from NCBI and analyzed using Phylogeny.fr (Dereeper et al. 2008). Alignment of protein sequences was performed with Muscle 3.8.31 and curated by G-blocks. The phylogenetic analysis was performed using PhyML 3.1 and the phylogenetic tree was drawn with TreeDyn 198.3. Bootstrap values over 50% are indicated on supported branches (1000 replicates).

Figure S11. Plot of Annotation Edit Distance (AED) scores for BRAKER3 gene predictions. Plot of BRAKER v3.0.3 (Gabriel et al. 2024)AED scores. AED scores (0-1) describing how a given gene model fits to transcript and protein evidence (best fit = 0, no fit = 1). Transcript evidence was computed from RNA-Seq or Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines represent the AED thresholds used to filter out genes during RGM selection (0.3 for transcripts, 0.1 for proteins). The numbers of genes in the four areas are displayed in white text boxes, and in blue the number of genes if no AED score penalty on splicing junction is applied. Numbers of gene models with transcript evidence were plotted on cumulative histograms above the scatter plot (green). Numbers of gene models with protein evidence were plotted on cumulative histograms on the right of the scatter plot (red).

Figure S12. Plot of Annotation Edit Distance (AED) scores for funannotate gene predictions. Plot of funannotate (funannotate n.d.) v1.8.17 AED scores. AED scores (0-1) describing how a given gene model fits to transcript and protein evidence (best fit = 0, no fit = 1). Transcript evidence was computed

from RNA-Seq or Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines represent the AED thresholds previously used to filter out genes during RGM selection (0.3 for transcripts, 0.1 for proteins). The numbers of genes in the four areas are displayed in white text boxes, and in blue the number of genes if no AED score penalty on splicing junction is applied. Numbers of gene models with transcript evidence were plotted on cumulative histograms above the scatter plot (green). Numbers of gene models with protein evidence were plotted on cumulative histograms on the right of the scatter plot (red).

Figure S13. Plot of Annotation Edit Distance (AED) scores for Helixer gene predictions. Plot of Helixer v0.3.1 (Holst et al. 2023) AED scores. AED scores (0-1) describing how a given gene model fits to transcript and protein evidence (best fit = 0, no fit = 1). Transcript evidence was computed from RNA-Seq or Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines represent the AED thresholds previously used to filter out genes during RGM selection (0.3 for transcripts, 0.1 for proteins). The numbers of genes in the four areas are displayed in white text boxes, and in blue the numer of genes if no AED score penalty on splicing junction is applied. Numbers of gene models with transcript evidence were plotted on cumulative histograms above the scatter plot (green). Numbers of gene models with protein evidence were plotted on cumulative histograms on the right of the scatter plot (red).

Figure S14. Comparison of *Z. tritici* genome annotations obtained with different tools (InGenAnnot/RGMs, BRAKER3, funannotate and Helixer). a) Upset plot of the gene models obtained with InGenAnnot (RGM), funannotate(funannotate n.d.), BRAKER3 (Gabriel et al. 2024) and Helixer (Stiehler et al. 2021). Intersecting sets of coding sequences (CDS) : number of shared gene models with identical CDS. Unique CDS: number of CDS predicted only by a single tool. **b**) Comparison of gene models. Number of CDS in each annotation. Identical CDS at a given locus. Unique Dissimilar CDS at a given locus: CDS differing in its structure from RGM. Unique Specific CDS at a given locus: CDS predicted only by InGenAnnot (RGM). The highest number of gene models identical to RGMs was observed for BRAKER3 (9,766, 72% of the RGMs). Among the 425 CDS identified by BRAKER3, funannotate and Helixer, but not InGenAnnot (RGM), 331 have either no evidence (AED = 1) or an AED value below the threshold (transcript =0.3 and protein =0.1). Among the 922 CDS identified by BRAKER3 and funannotate, but not InGenAnnot (RGM), 829 have either no evidence (AED = 1) or an AED value below the threshold (transcript =0.3 and protein =0.1).

Figure S15. Cumulative distributions of the best Annotation Edit Distance (AED) scores for RGM (InGenAnnot) and gene models obtained with three other tools (BRAKER3, funannotate and Helixer). AED scores (0-1) described how a given gene model fit to transcript or protein evidence (best fit = 0, no fit = 1). The best AED score (X axis) was computed from either transcript or protein evidence. a) cumulative plot of the number of transcripts; b) cumulative plot of the density of transcripts (normalized). The red line indicated the cutoff used to select the best gene model (0.3 for transcript evidence).

Supplementary table titles

Table S1. CDS features of the four available gene annotations of the *Z. tritici* IPO323 genome (JGI, MPI, RRES and CURTIN).

Table S2. Chromosome localization of gene models of the four available annotations of the *Z. tritici* IPO323 genome (JGI, MPI, RRES and CURTIN).

Table S3 (A and B). RNA-Seq and Iso-Seq cDNA libraries from *Z. tritici* IPO323.

Table S4. Features of Re-annotated Gene Models (RGMs) of the *Z. tritici* IPO323 genome.

Table S5. Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis of Re-annotated Gene Models (RGM) and gene models from previous annotations of the *Z. tritici* IPO323 genome.

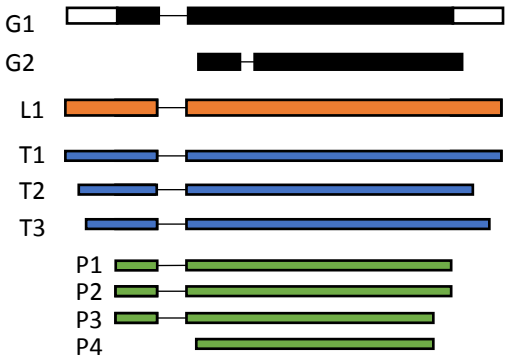
Table S6. Distribution of Re-annotated Gene Models (RGMs) on *Z. tritici* IPO323 chromosomes

Table S7. Identification of fused/split genes in the *Z. tritici* IPO323 genome annotations.

Table S8. Numbers of transcript isoforms detected for Re-annotated Gene Models (RGMs) in *Z. tritici* IPO323.

Table S9. Long, non-coding RNAs (lncRNA) from *Z. tritici* IPO323 differentially expressed during infection.

Table S10. Metrics calculated from AED scores for the four gene annotations obtained with InGenAnnot (RGM), funannotate, Helixer and BRAKER3.



Gene model	Coverage	Evidence	Best evidence	AED score (raw)	Penalty	Final AED score (penalty)
G1	CDS-only	Transcript (short-reads)	T2	0.06	no	0.06
		Transcripts (long-reads)	L1	0.11	no	0.11
		Proteins	P1	0.001	-	0.001
	Full gene	Transcript (short-reads)	T1	0.001	no	0.001
		Transcripts (long-reads)	L1	0.0	no	0.0
		Proteins	P1	0.001	-	0.001
G2	CDS-only	Transcript (short-reads)	T2	0.15	yes	0.40
		Transcripts (long-reads)	L1	0.19	yes	0.44
		Proteins	P4	0.068	-	0.068
	Full gene	Transcript (short-reads)	T2	0.15	yes	0.40
		Transcripts (long-reads)	L1	0.19	yes	0.44
		Proteins	P4	0.068	-	0.068

Figure S1. Annotation Edit Distance (AED) computation using InGenAnnot
Annotation Edit Distance (AED) is an annotation quality-control measure, proposed by Maker (Holt and Yandell 2011) to compare annotations based on their overlap. We have used AED scores to quantify the concordance between a gene model and its associated evidence (transcript, protein), as previously described (Eilbeck et al. 2009). Several options for computing this customized AED were implemented, such as restriction to coding sequence (CDS) or penalty on unsupported intron splicing sites. No intron splicing site penalty was applied for introns located in UTRs, nor for introns defined using protein alignments which could be biased by phylogenetically distant proteins. The table displays AED scores for two gene models (G1, G2) located at the same locus (metagene). The orange rectangle (L1) corresponds to a single Iso-Seq transcript detected at this locus. Blue rectangles (T1-T3,) correspond to the different assembled RNA-seq transcripts detected at this locus. Green rectangles (P1-P4) correspond to gene models deduced from protein alignments. As G2 had no UTRs, calculating AED on CDS only or on the complete gene has no impact on the score. However, since the Iso-Seq transcript has UTRs, G1 AED score was better (lower) when the complete gene was used instead of the CDS. G2 contains an intron splicing site not supported by transcript evidence, resulting in a penalty (0.25) added to the raw AED score. In this case, G1 was selected as the best gene model.

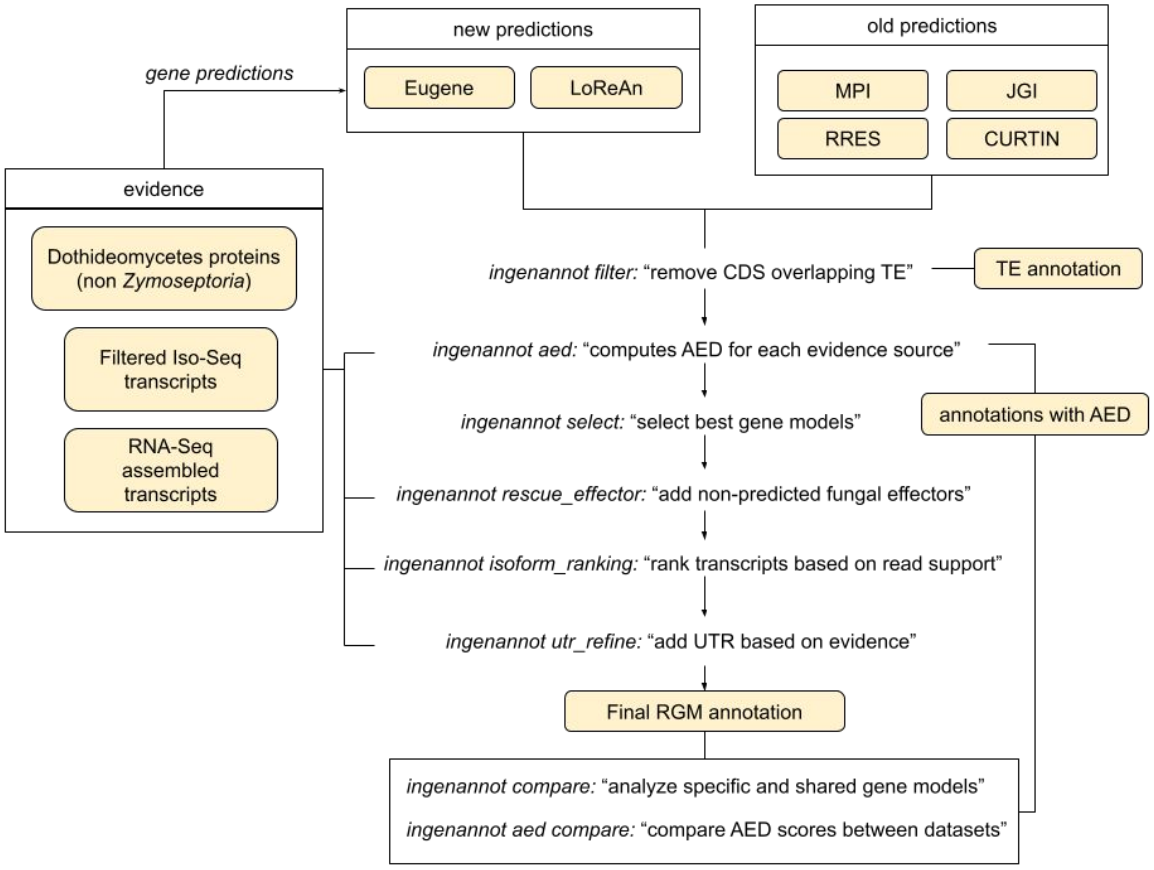


Figure S2. Bioinformatics workflow with the different steps of InGenAnnot tool suite
Sources of evidence (Proteins, Iso-Seq and RNA-Seq transcripts) were used to predict new gene models with Eugene and LoReAn. Gene models predicted by novel and previous *ab initio* gene prediction software were submitted to the following workflow. Coding Sequence (CDS) were filtered out if overlapping with known transposable elements (TE). Then, filtered gene models were used to compute AED scores for each source of evidence. The best gene model at a given locus (metagene) was selected based on its AED score (lower than 0.3 for transcripts, lower than 0.1 for proteins). Gene models failing the AED score threshold, but predicted by at least 4 independent gene predictors, were retained. All the gene models without ATG or stop codon were removed. Transcripts without gene models were analysed to infer potential missing effectors (small secreted proteins). Finally, Iso-Seq transcripts were filtered according to their RNA-Seq support (low abundance Iso-Seq were removed) before being used to define UTRs. All of this workflow was described with associated command lines in the documentation of InGenAnnot (https://bioger.pages.mia.inra.fr/ingenannot/usecases/select_best_gene_models.html). The gene models selected by InGenAnnot (RGMs) were compared to all gene models to compute AED scores and quantify the contribution of each *ab initio* gene prediction software to the final RGM dataset (https://bioger.pages.mia.inra.fr/ingenannot/usecases/annotation_comparison.html).

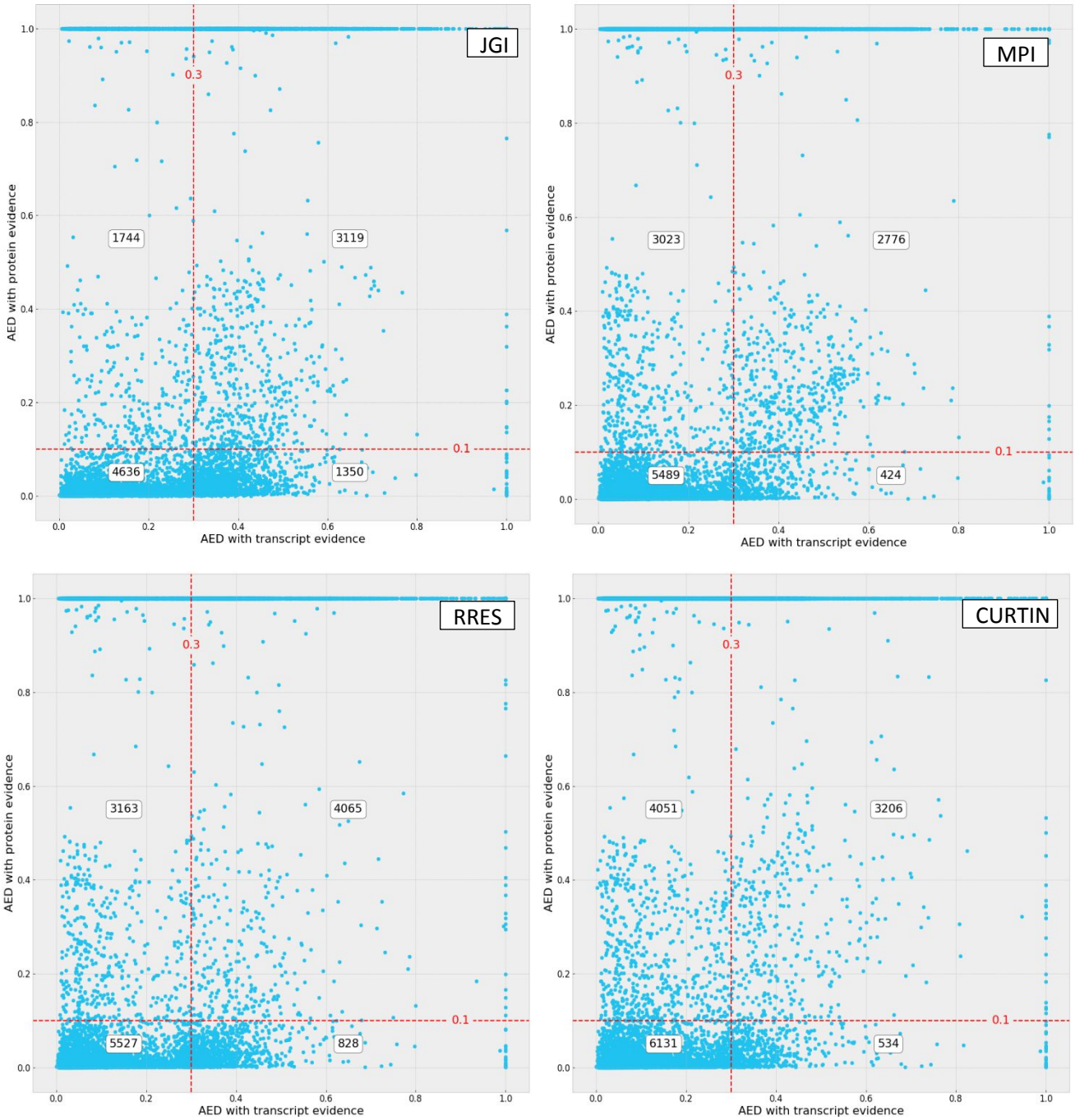


Figure S3. **Comparison of the Annotation Edit Distance (AED) scores of the JGI, MPI, RRES and CURTIN gene models** AED scores (0-1) described how a given gene model fit the transcript and protein evidence (best fit = 0). Transcript evidence was computed from RNA-Seq and Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines represent the AED thresholds to filter out genes (0.3 for transcripts, 0.1 for proteins), except if they are supported by at least 4 different annotations (upper right area of the graph). The numbers of genes from each area were displayed in white boxes.

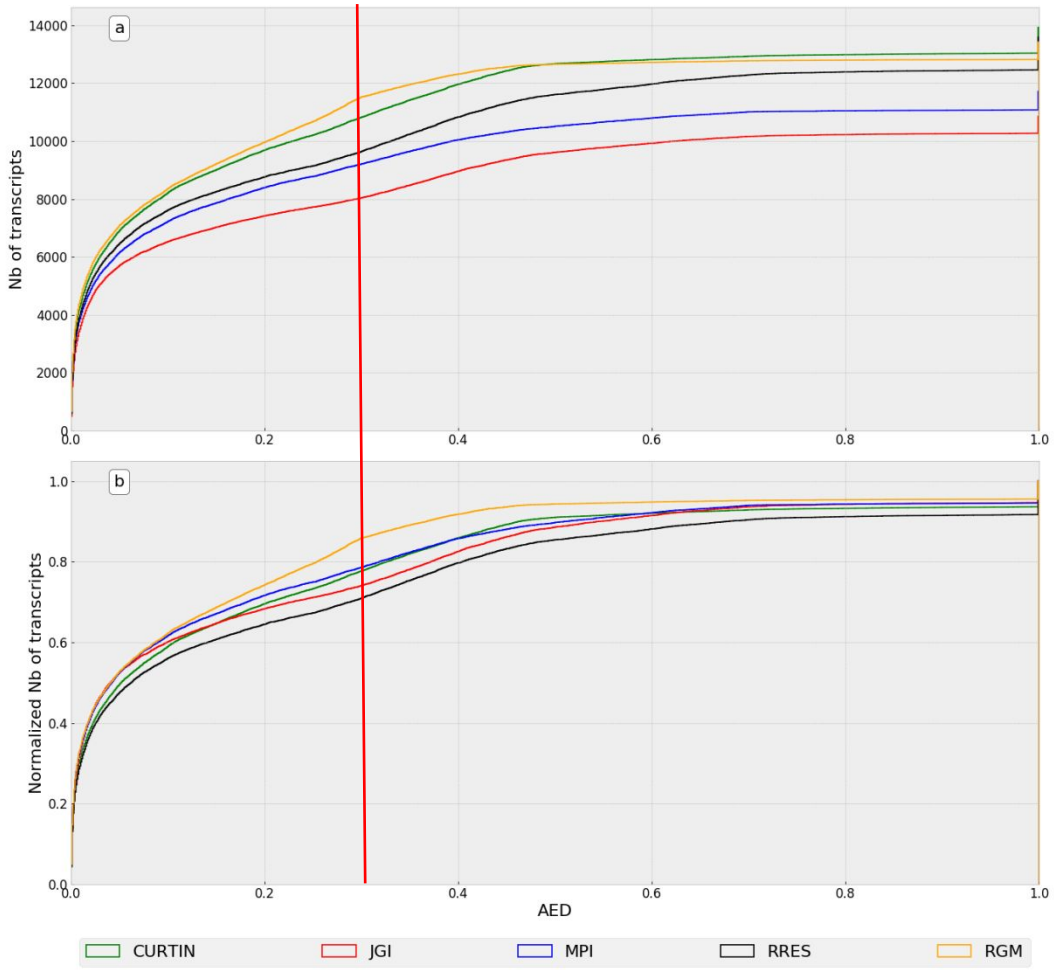


Figure S4. Cumulative distributions of the best Annotation Edit Distance (AED) scores for Re-annotated Gene Models (RGMs) and those from previous annotations.
AED scores (0-1) described how a given gene model fit the transcript and protein evidence (best fit = 0). The best AED score (X-axis) was computed from either transcript or protein evidence. a) cumulative plot of the number of transcripts, b) cumulative plot of the density of transcripts (normalized). The red line indicated the cutoff used to select the best gene model (0.3 for transcript evidence).

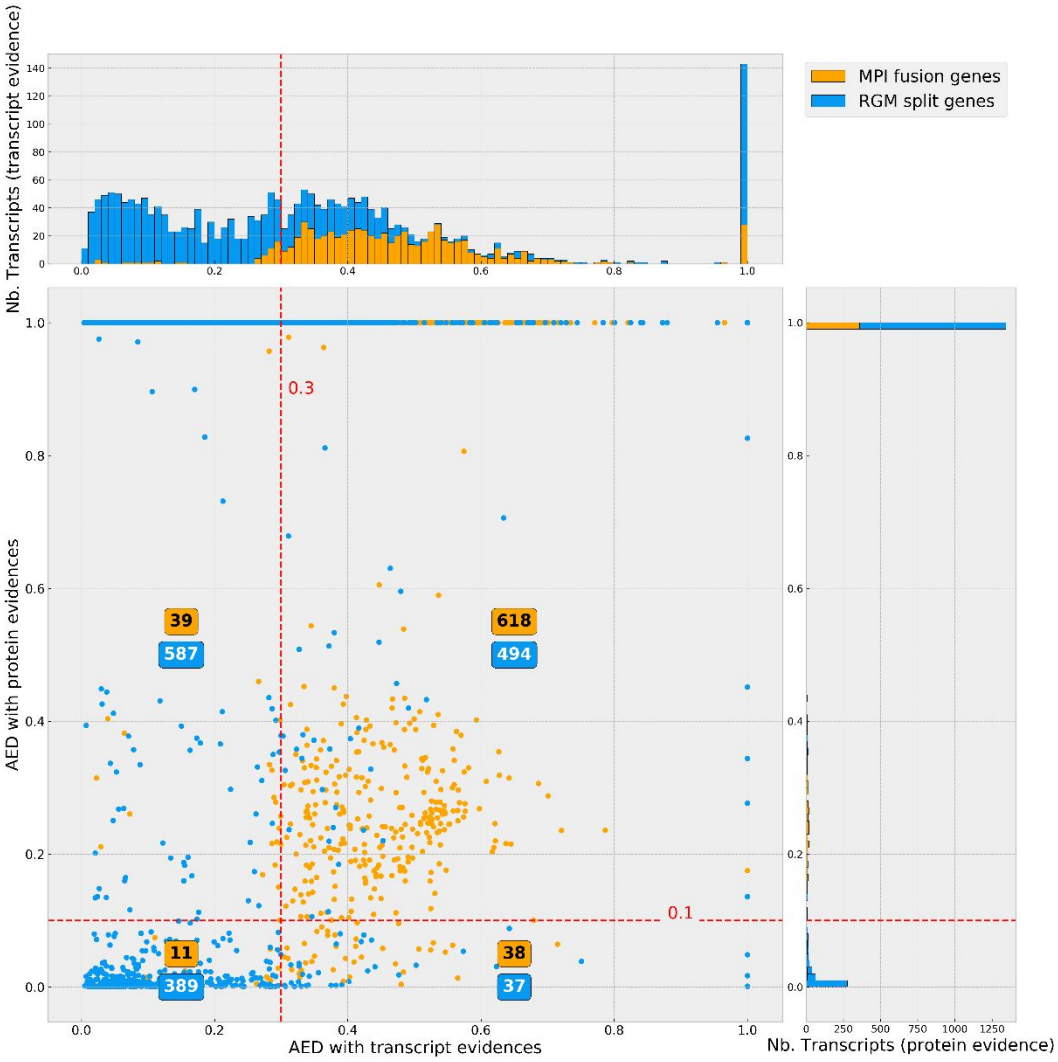


Figure S5. Comparison of the Annotation Edit Distance (AED) scores of split Re-annotated Gene Models (RGMs) and their corresponding fused gene models from the MPI annotation

AED scores (0-1) described how a given gene model fit the transcript and protein evidence (best fit = 0). Split RGMs were displayed in blue and the corresponding MPI fused genes were displayed in orange. Transcript evidence were computed from RNA-Seq or Iso-Seq data (X axis). Protein evidence were computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines corresponded to the AED thresholds used to filter out gene models (0.3 for transcripts, 0.1 for proteins). The number of transcripts for each value of AED score was plotted on cumulative histograms above the scatter plot (RGM in blue, MPI in orange). The number of transcripts with protein evidence were plotted on cumulative histograms on the right of the scatter plot (RGM in blue, MPI in orange).

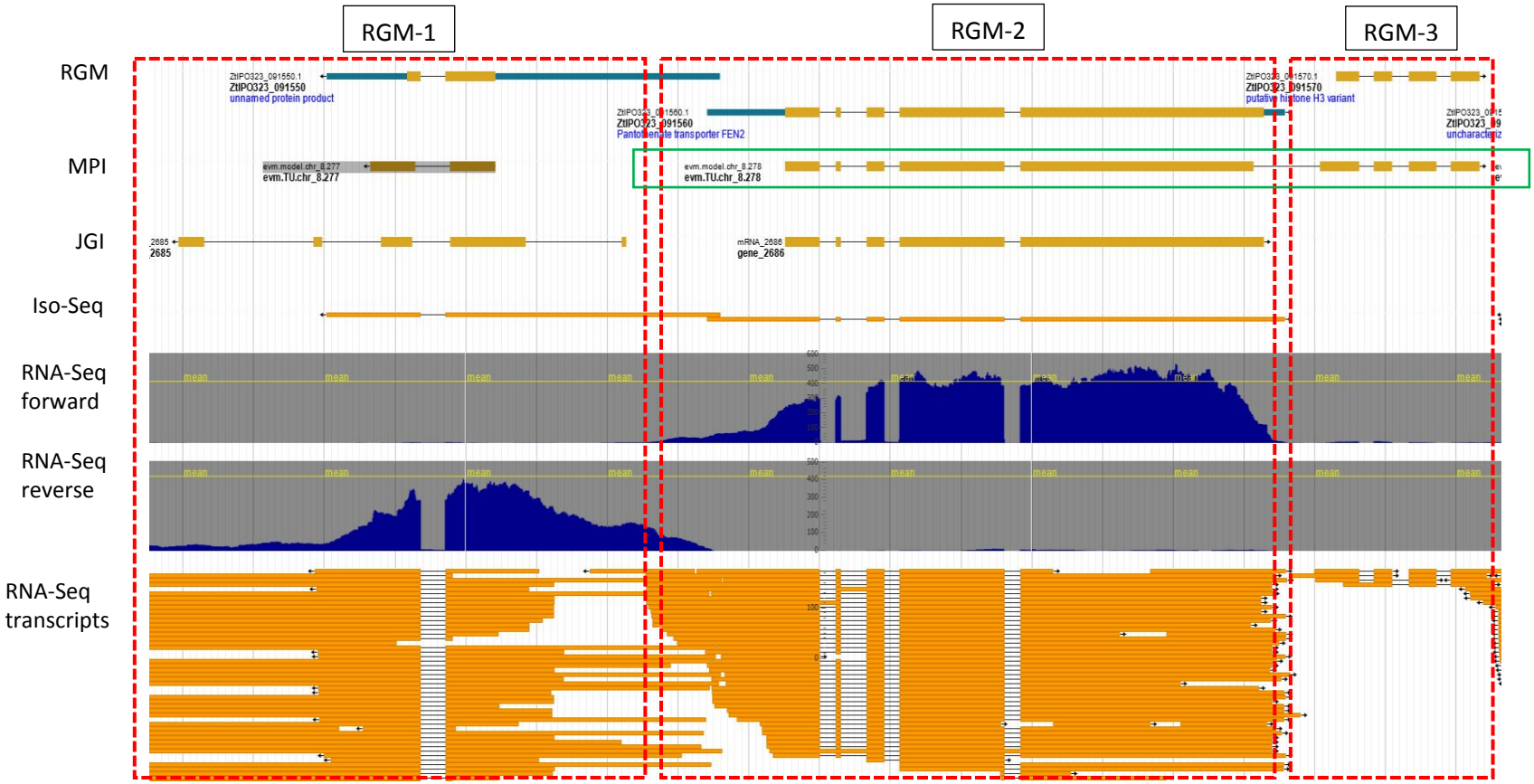


Figure S6. Fused genes from the MPI annotation located at chr_8:940236...948036 split in two Re-annotated Gene Models (RGM-1 and RGM-2).
In the region of chr_8:940236...948036 (7.8 Kb), three RGMs were predicted (red-dashed squares: RGM-1, RGM-2 and RGM-3). The MPI annotation predicted a gene model corresponding to the fusion of RGM2 and RGM-3 (green rectangle). This fusion had no transcript evidence (Iso-seq, RNA-Seq). Specific transcripts of RGM-2 (Iso-Seq, RNAseq) and RGM-3 (RNAseq) were detected. On the other strand, RGM-1 had correctly predicted intron splicing sites, while the corresponding gene models from the MPI and JGI annotations had incorrect intron splicing sites. RGM-3 was not predicted by JGI, despite encoding a conserved Histone-3 variant protein. RGM track: RGM gene models. MPI track: MPI gene models. JGI track: JGI gene models. Iso-Seq track: filtered Iso-Seq transcripts. RNA-Seq forward track: coverage of forward-strand RNA-Seq reads. RNA-Seq reverse track: coverage of reverse-strand RNA-Seq reads mapping at this locus. RNA-Seq transcript track: assembled RNA-Seq transcripts.

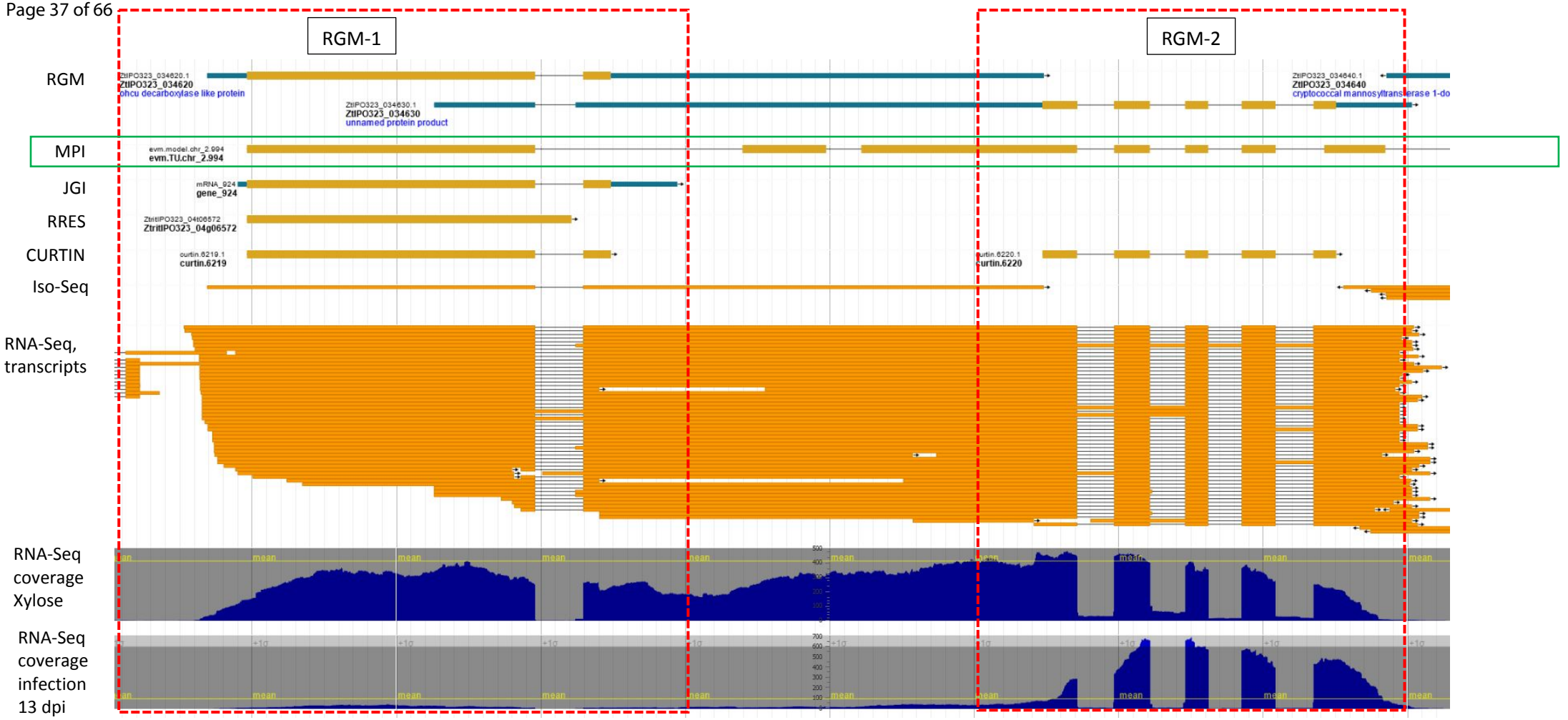


Figure S7. Fused genes from the MPI annotation located at chr_2:2938368...2940768 split in two Re-annotated Gene Models (RGM-1 and RGM-2)

In the region of chr_2:2938368...2940768 (2.4 Kb), two RGMs were predicted (red-dashed squares). The MPI annotation predicted a gene model corresponding to the fusion of RGM-1 and RGM-2 (green rectangle) not supported by Iso-Seq. Iso-Seq transcripts supporting RGM-1 were detected. RGM-2 with no Iso-Seq support was not predicted by JGI and RRES, while it was predicted as an independent gene in Curtin annotation.). RGM2 (ZtIPO323_034630) encoded a Small Secreted Protein (SSP, see File S1). Assembled RNA-seq transcripts corresponding to the fused MPI gene model were detected. They were likely artefacts from assembly of overlapping RNA-Seq reads. Indeed, in a specific condition (infection at 13 days post inoculation), only RGM-2 transcript was detected (RNA-seq coverage 13 dpi track), supporting the prediction of RGM-2. RGM track: RGM gene models. MPI track: MPI gene model. JGI track: JGI gene model. RRES track: RRES gene model. Curtin track: Curtin gene models. Iso-Seq track: filtered Iso-Seq transcripts. RNA-Seq transcript track: assembled RNA-Seq transcripts. RNA-seq coverage Xylose: coverage of strand-specific RNA-Seq reads from a Xylose medium library. RNA-seq coverage infection 13 dpi: coverage of strand-specific RNA-Seq reads from a 13 dpi wheat infection library.

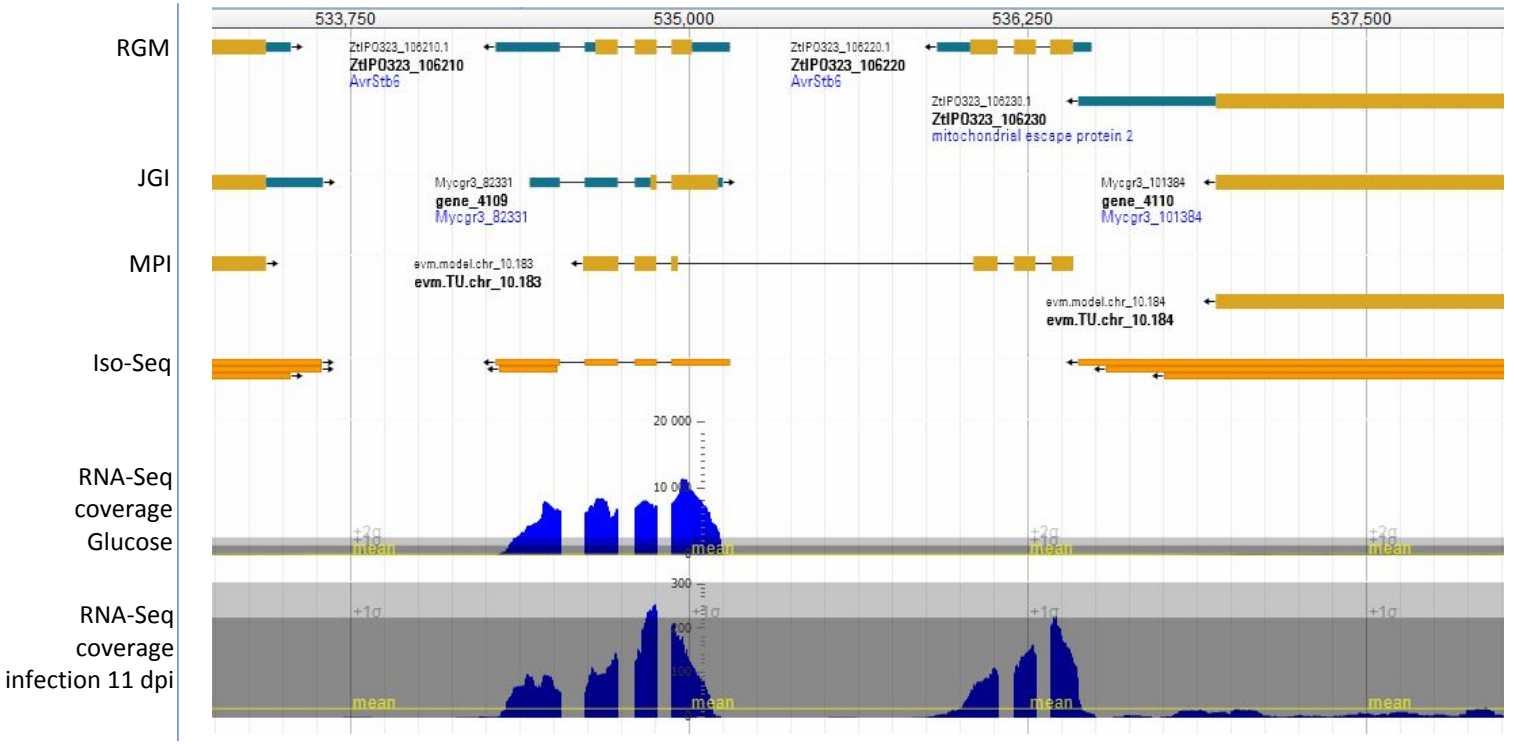


Figure S8a. *Avr-Stb6* paralogs located on chromosome 10 predicted by the new annotation.
The two new paralogs of *Avr-Stb6* (ZtIPO323_106210, ZtIPO323_106220) are located head to tail on chromosome 10 between position 534287 and 536486. ZtIPO323_106210 was predicted in JGI and MPI annotation, but the gene models did not match Iso-Seq and RNA-seq evidence. ZtIPO323_106220 was not predicted by any previous annotations. RGM annotation was only supported by RNA-seq evidence. ZtIPO323_106210 is expressed during *in vitro* growth (RNA-Seq coverage glucose track), and infection stages (RNA-Seq coverage infection 11 dpi track). ZtIPO323_106220 was differentially upregulated during infection (RNA-Seq coverage infection 11 dpi track) to the same level as ZtIPO323_106210. RGM track: RGM gene models. JGI track: JGI gene model. MPI track: MPI gene model. Iso-Seq track: filtered Iso-Seq transcripts. RNA-seq coverage Glucose: coverage of strand-specific RNA-Seq reads from a Glucose medium library. RNA-seq coverage infection 11 dpi: coverage of strand-specific RNA-Seq reads from a 11 dpi wheat infection library.

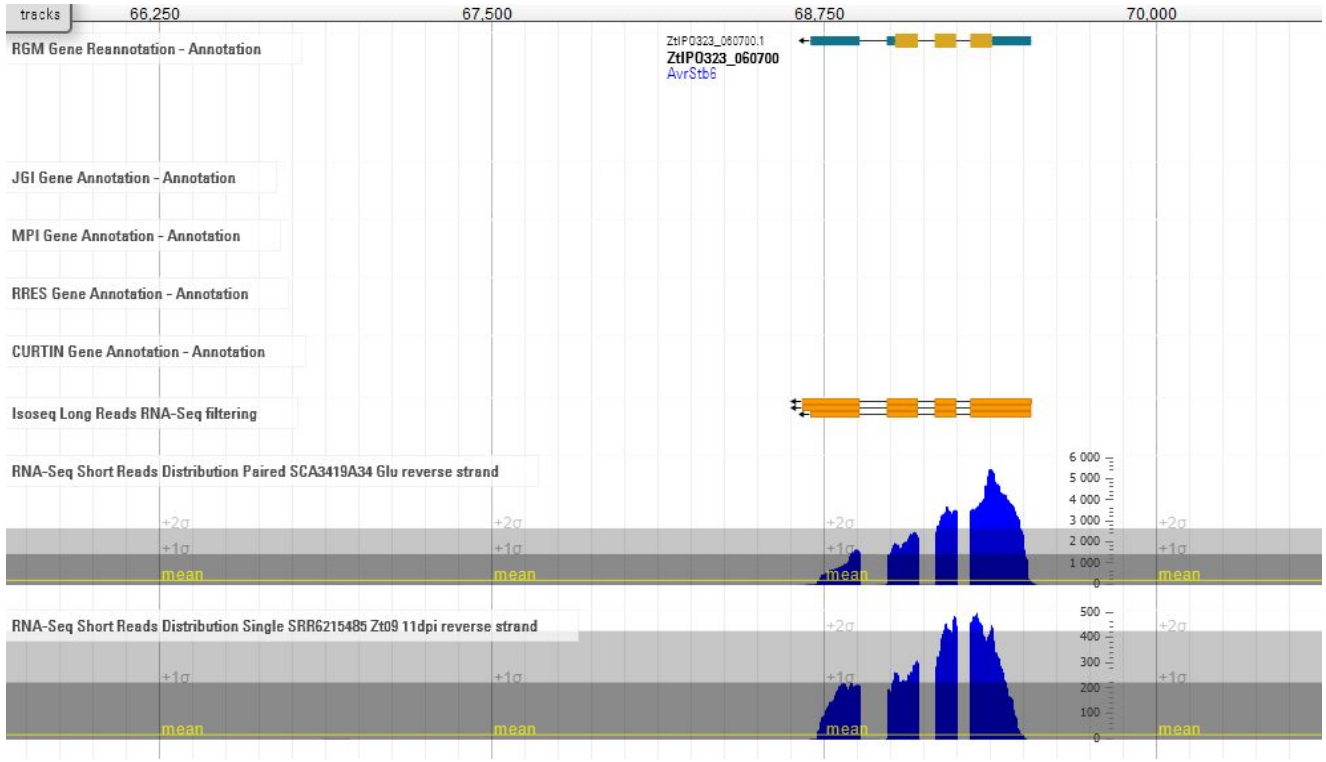


Figure S8b. Original *Avr-Stb6* located on chromosome 5

The original *Avr-Stb6* is located at the end of chromosome 5. It was not predicted by any previous annotations, while it was correctly predicted as RGM (see track RGM track reannotation). Iso-Seq transcripts were detected (see track Isoseq long reads), since this gene is highly expressed during *in vitro* growth of *Zymoseptoria tritici*, as well as during infection (see tracks RNA-seq short reads distribution). RGM track: RGM gene models. JGI track: JGI gene model. MPI track: MPI gene model. RRES track: RRES gene model. Curtin track: Curtin gene models. Iso-Seq track: filtered Iso-Seq transcripts. RNA-seq distribution Glucose: coverage of strand-specific RNA-Seq reads from a Glucose medium library. RNA-seq distribution infection 11 dpi: coverage of strand-specific RNA-Seq reads from a 11 dpi wheat infection library.

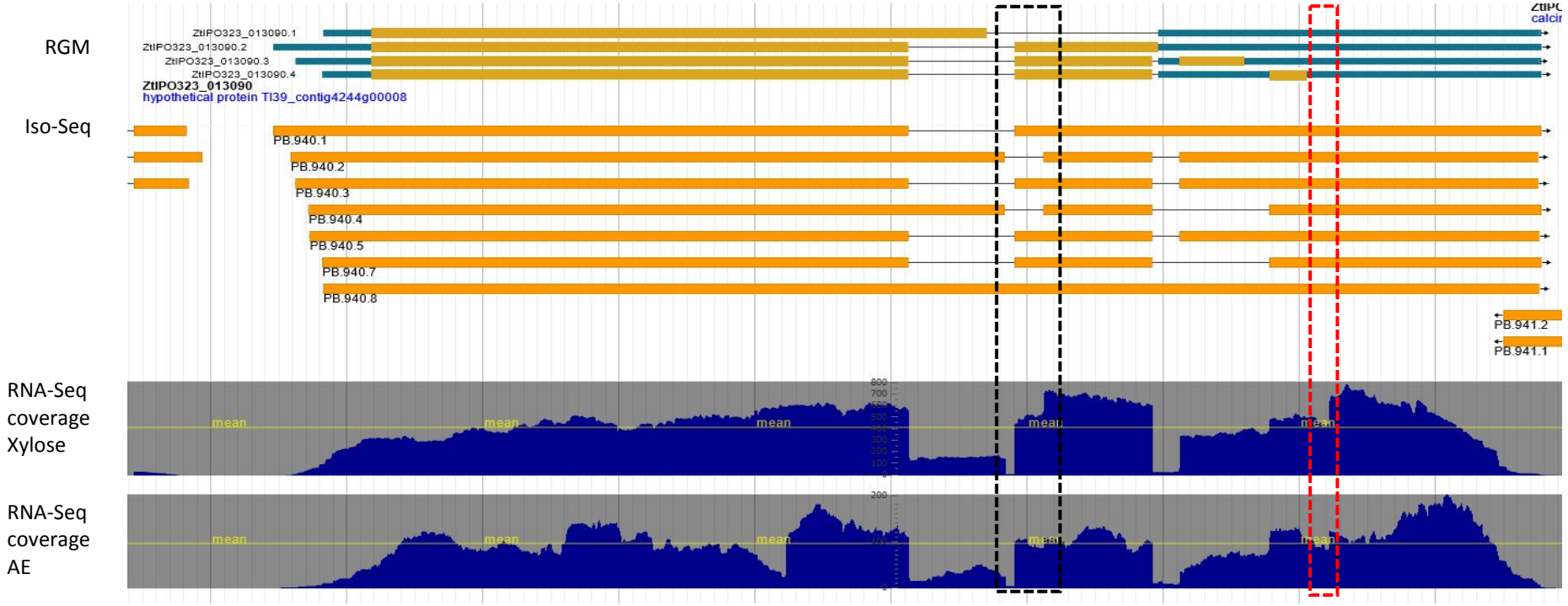


Figure S9. Isoforms of Re-annotated Gene Model (RGM) ZtIPO323_013090 supported by Iso-Seq

RGM ZtIPO323_013090 located at chr_1:3358097...3361350 (3.25 Kb) has four different isoforms. One alternative splicing site (red-dashed rectangle) supported by RNA-Seq, was not supported by Iso-Seq. Another alternative splicing site detected both by RNA-Seq and Iso-Seq (black-dashed rectangle), was not used to predict an isoform by ab initio software due to a stop codon before the splicing site. Finally, the canonical form retained is the transcript without any introns. This selection could be an artefact of transcripts from AE medium inducing many intron retention events. The canonical isoform is likely the RGM corresponding to the Iso-Seq transcript PB.940.5 with two introns (isoform 3). RGM track: RGM gene model. Iso-Seq track: filtered Iso-Seq transcripts. RNA-Seq coverage Xyl track: coverage of strand-specific RNA-Seq reads from the Xylose medium library. RNA-Seq coverage AE track: coverage of strand-specific RNA-Seq reads from AE medium library.

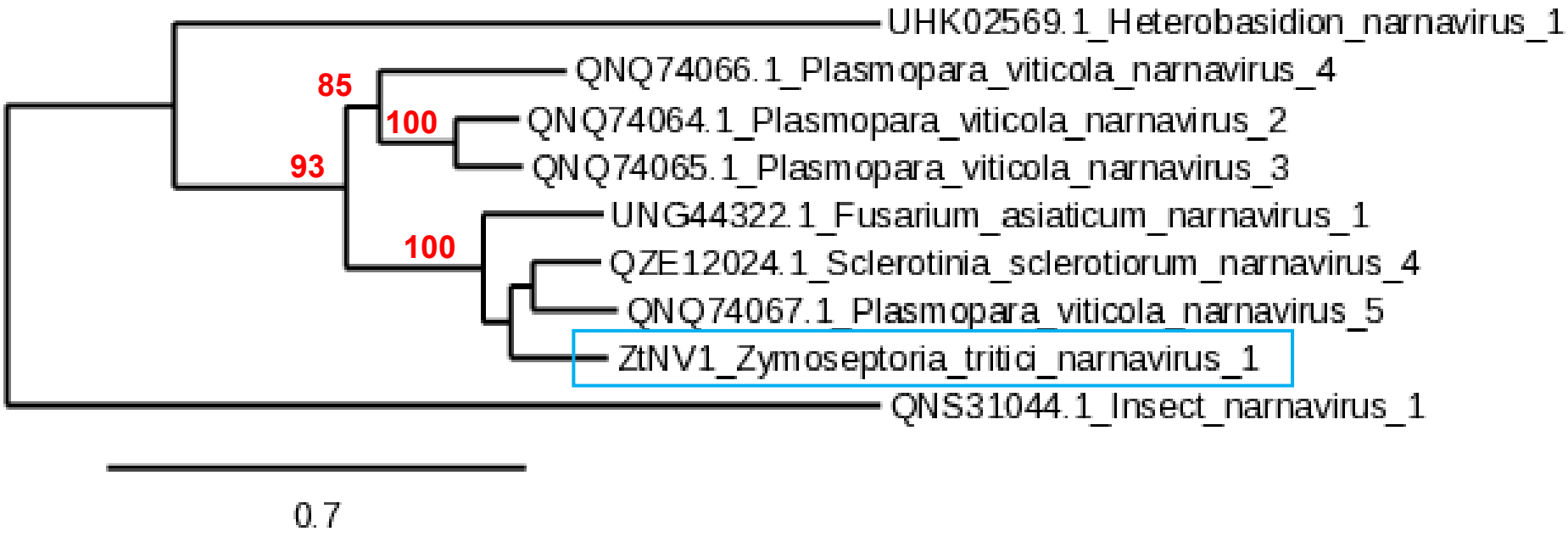


Figure S10. Phylogenetic tree of RNA-dependent RNA polymerases of fungal narnaviruses related Zt-NV1 from *Zymoseptoria tritici*

Iso-Seq transcripts not mapping to the *Z. tritici* IPO323 reference genome were clustered with blastclust. Similarities with known sequences were analysed by *blastn* search against the NCBI nr database. Reconstruction of the full-length sequences of viruses was performed by de-novo assembly with SPAdes (v3.15.4) (Bankevich et al. 2012). RNA-dependent RNA polymerase sequences from narnaviruses related to Zt-NV1 were retrieved from NCBI and analyzed using Phylogeny.fr (Dereeper et al. 2008). Alignment of protein sequences was performed with Muscle 3.8.31 and curated by G-blocks. The phylogenetic analysis was performed using PhyML 3.1 and the phylogenetic tree was drawn with TreeDyn 198.3. Bootstrap values over 50% are indicated on supported branches (1000 replicates).

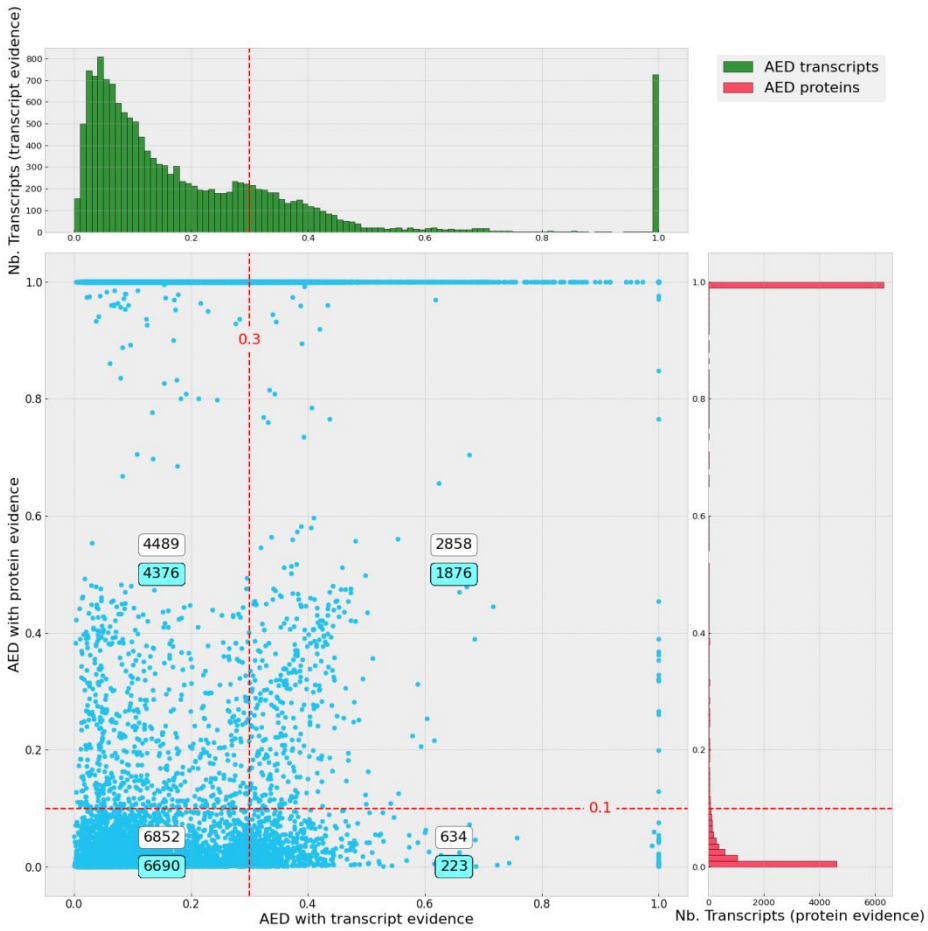


Figure S11. Plot of Annotation Edit Distance (AED) scores for BRAKER3 gene predictions. Plot of BRAKER v3.0.3 (Gabriel et al. 2024) AED scores. AED scores (0-1) describing how a given gene model fits to transcript and protein evidence (best fit = 0, no fit = 1). Transcript evidence was computed from RNA-Seq or Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines represent the AED thresholds used to filter out genes during RGM selection (0.3 for transcripts, 0.1 for proteins). The numbers of genes in the four areas are displayed in white text boxes, and in blue the number of genes if no AED score penalty on splicing junction is applied. Numbers of gene models with transcript evidence were plotted on cumulative histograms above the scatter plot (green). Numbers of gene models with protein evidence were plotted on cumulative histograms on the right of the scatter plot (red).

Braker3 command launched to obtain annotations:

```
braker.pl --genome Mygr_323_reformat_with_mito.clean.fsa --prot_seq
UNIPROTKB.Dothideomycetes.15072020.NoZymo.clean.fasta  --bam
SCA3419A44.bam,SCA3419A47.bam,SCA3419A31.bam,SCA3419A40.bam,SCA3419A32.bam,SCA3419A48.bamSCA3419A3
3.bam,SCA3419A49.bam,SCA3419A34.bam,SCA3419A50.bam,SCA3419A35.bam,SCA3419A57.bam,SCA3419A36.bam,SC
A3419A89.bam,SCA3419A37.bam,SCA3419A58.bam,SCA3419A38.bam,SCA3419A81.bam,SCA3419A39.bam,SCA3419A82
.bam,SCA3419A41.bam,SCA3419A65.bam,SCA3419A42.bam,SCA3419A66.bam,SCA3419A43.bam,SCA3419A73.bam,SCA
3419A45.bam,SCA3419A74.bam,SCA3419A46.bam,SCA3419A90.bam --threads 16 --fungus
```

InGenAnnot command launched to add AED annotations:

```
ingenannot -v 2 -p 10 aed braker3.gff IPO323.braker3.aed.cdsonly.gff braker3
all_transcripts_counts_filter.sort.gff.gz exonerate_no_zymo.sort.gff3.gz --longreads
isoforms.top.sort.gff.gz --penalty_overflow 0.25 --evtrstranded --evprstranded --aed_tr_cds_only -
-aedtr 0.3 --aedpr 0.1
```

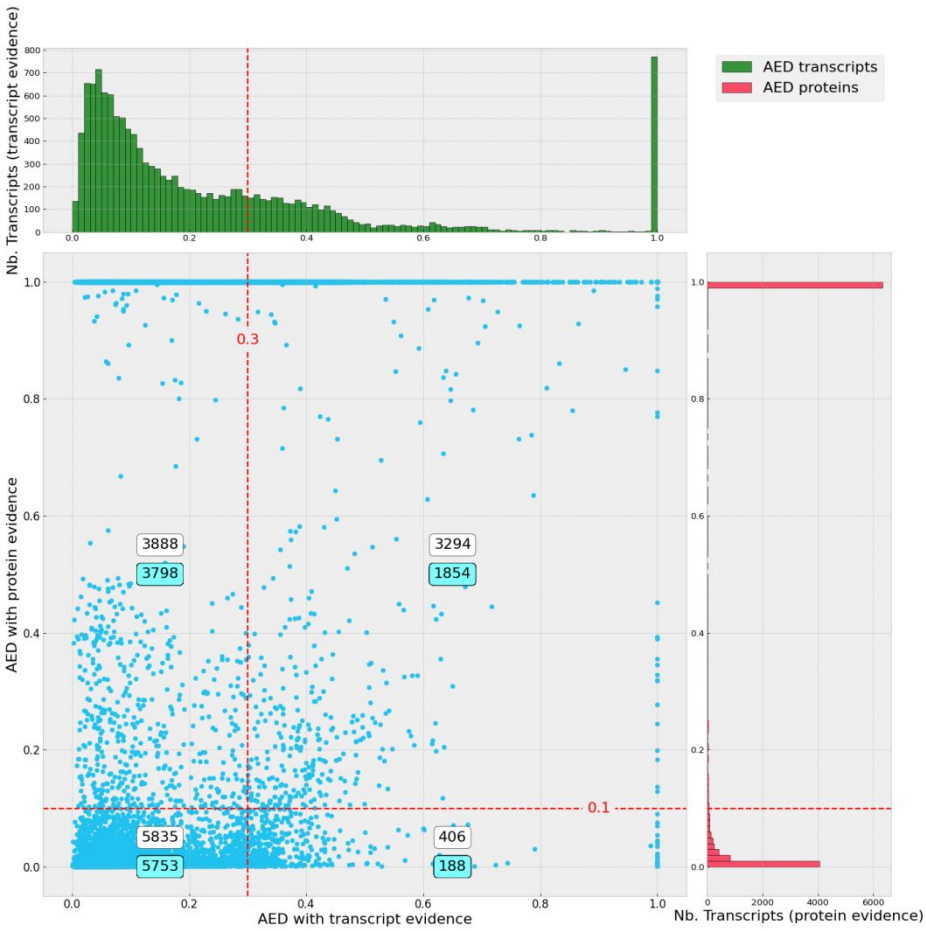


Figure S12. Plot of Annotation Edit Distance (AED) scores for funannotate gene predictions. Plot of funannotate v1.8.17 (Palmer and Stajich 2020) AED scores. AED scores (0-1) describing how a given gene model fits to transcript and protein evidence (best fit = 0, no fit = 1). Transcript evidence was computed from RNA-Seq or Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines represent the AED thresholds previously used to filter out genes during RGM selection (0.3 for transcripts, 0.1 for proteins). The numbers of genes in the four areas are displayed in white text boxes, and in blue the number of genes if no AED score penalty on splicing junction is applied. Numbers of gene models with transcript evidence were plotted on cumulative histograms above the scatter plot (green). Numbers of gene models with protein evidence were plotted on cumulative histograms on the right of the scatter plot (red).

<u>funannotate v1.8.17 command launched to obtain annotations:</u>

```
funannotate predict -o results -i Mygr_323_reformat_with_mito.clean.fsa --species "Zymoseptoria
tritici" --isolate IPO323 --transcript_evidence all_transcripts.fasta
all_samples.chained.refine.reclusterize.fasta --rna_bam all.merge.bam  --protein_evidence
UNIPROTKB.Dothideomycetes.15072020.NoZymo.clean.fasta --cpus 8
```

<u>InGenAnnot command launched to add AED annotations:</u>

```
ingenannot -v 2 -p 10 aed Zymoseptoria_tritici_IPO323.gff3 IPO323.funannotate.aed.cdsonly.gff
funannotate all_transcripts_counts_filter.sort.gff.gz exonerate_no_zymo.sort.gff3.gz -longreads
isoforms.top.sort.gff.gz --penalty_overflow 0.25 --evtrstranded --evprstranded --aed_tr_cds_only -
-aedtr 0.3 --aedpr 0.1
```

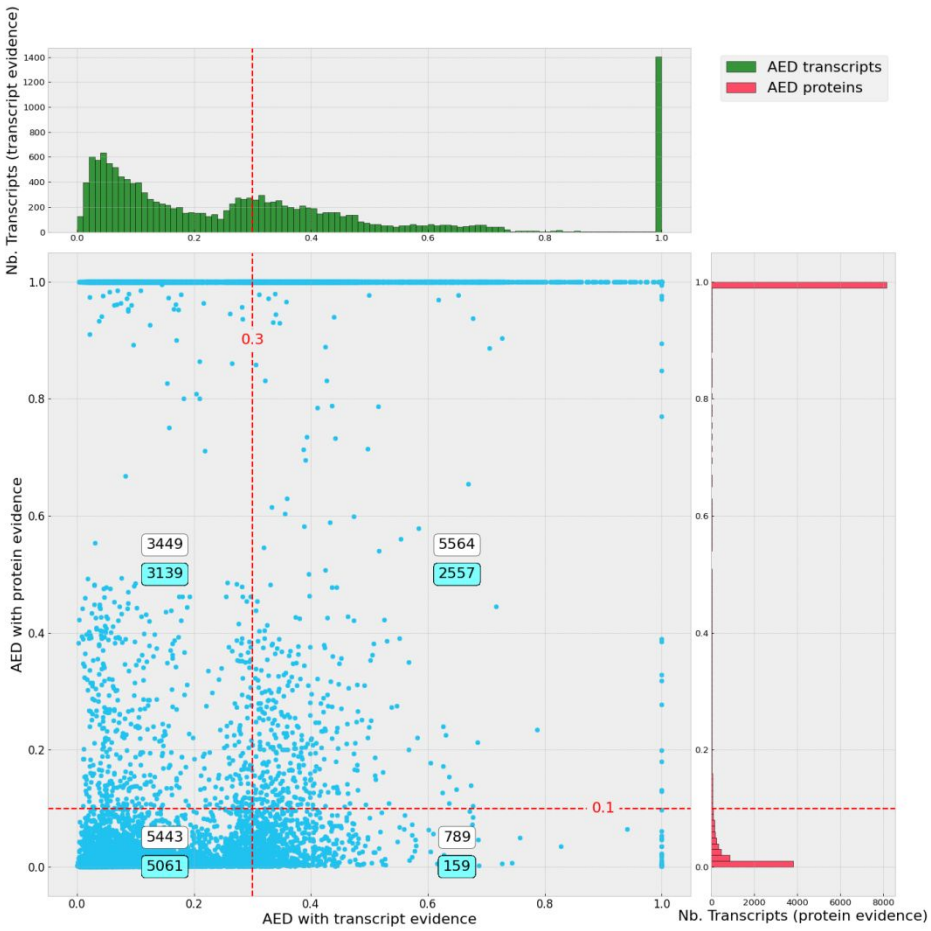


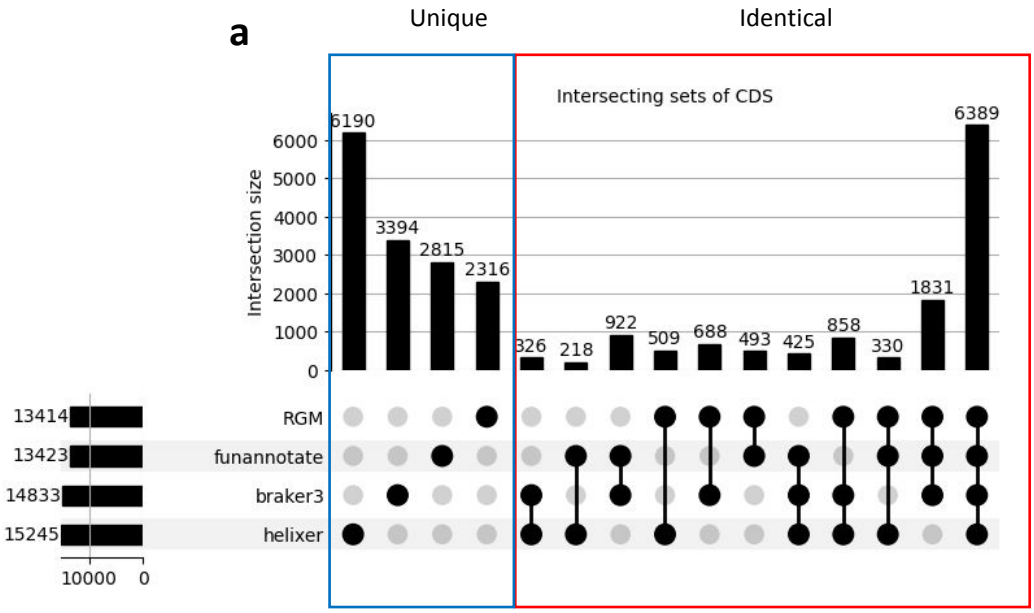
Figure S13. Plot of Annotation Edit Distance (AED) scores for Helixer gene predictions. Plot of Helixer v0.3.1 (Holst et al. 2023) AED scores. AED scores (0-1) describing how a given gene model fits to transcript and protein evidence (best fit = 0, no fit = 1). Transcript evidence was computed from RNA-Seq or Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines represent the AED thresholds previously used to filter out genes during RGM selection (0.3 for transcripts, 0.1 for proteins). The numbers of genes in the four areas are displayed in white text boxes, and in blue the numer of genes if no AED score penalty on splicing junction is applied. Numbers of gene models with transcript evidence were plotted on cumulative histograms above the scatter plot (green). Numbers of gene models with protein evidence were plotted on cumulative histograms on the right of the scatter plot (red).

Helixer v0.3.1 command launched to obtain annotations:

```
Helixer.py --lineage fungi --fasta-path Mygr_323_reformat_with_mito.clean.fsa --species
Zymoseptoria_tritici --gff-output-path Zt_IPO323_helixer.gff3
```

InGenAnnot command launched to add AED annotations:

```
ingenannot -v 2 -p 10 aed Zt_IPO323_helixer.gff3 IPO323.helixer.aed.cdsonly.gff helixer
all_transcripts_counts_filter.sort.gff.gz exonerate_no_zymo.sort.gff3.gz --longreads
isoforms.top.sort.gff.gz --penalty_overflow 0.25 --evtrstranded --evprstranded --aed_tr_cds_only -
-aedtr 0.3 --aedpr 0.1
```



b

		RGM
Identical CDS	RGM	*13414*
	funannotate	9043
	Braker3	9766
	Helixer	8086
Unique:		2316
	Dissimilar CDS	1797
	Specific CDS	519

Figure S14. Comparison of *Z. tritici* genome annotations obtained with different tools (InGenAnnot/RGMs, BRAKER3, funannotate and Helixer). a) Upset plot of the gene models obtained with InGenAnnot (RGM), funannotate (Palmer and Stajich 2020), BRAKER3 [Gabriel et al. 2024] and Helixer (Holst et al. 2023]. Intersecting sets of coding sequences (CDS) : number of shared gene models with identical CDS. Unique CDS: number of CDS predicted only by a single tool. **b**) Comparison of gene models. Number of CDS in each annotation. Identical CDS at a given locus. Unique Dissimilar CDS at a given locus: CDS differing in its structure from RGM. Unique Specific CDS at a given locus: CDS predicted only by InGenAnnot (RGM). The highest number of gene models identical to RGMs was observed for BRAKER3 (9,766, 72% of the RGMs). Among the 425 CDS identified by BRAKER3, funannotate and Helixer, but not InGenAnnot (RGM), 331 have either no evidence (AED = 1) or an AED value below the threshold (transcript =0.3 and protein =0.1). Among the 922 CDS identified by BRAKER3 and funannotate, but not InGenAnnot (RGM), 829 have either no evidence (AED = 1) or an AED value below the threshold (transcript =0.3 and protein =0.1).

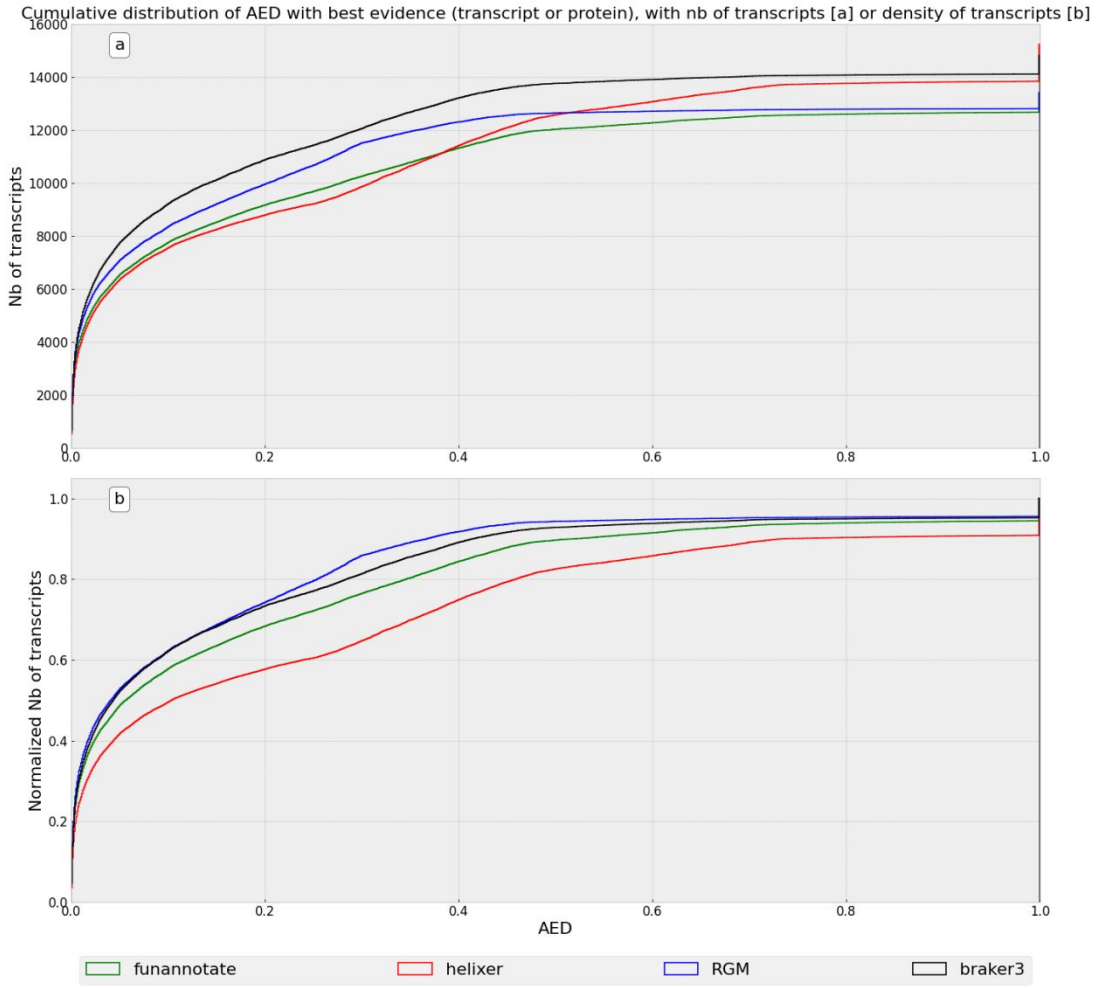


Figure S15. Cumulative distributions of the best Annotation Edit Distance (AED) scores for RGM (InGenAnnot) and gene models obtained with three other tools (BRAKER3, funannotate and Helixer). Funannotate (Palmer and Stajich 2020), BRAKER3 (Gabriel et al. 2024) and Helixer (Stiehler et al. 2021). AED scores (0-1) described how a given gene model fit to transcript or protein evidence (best fit = 0, no fit = 1). The best AED score (X-axis) was computed from either transcript or protein evidence. a) cumulative plot of the number of transcripts, b) cumulative plot of the density of transcripts (normalized). The red line indicated the cutoff used to select the best gene model (0.3 for transcript evidence).

Category	JGI	MPI	RRES	CURTIN
nb_CDS	10849	11712	13583	13922
average_CDS_length, bp	1307	1465	1293	1287
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	5233	5594
nb_introns	17464	18016	17189	16642
average_introns_per_transcript	1.6	1.5	1.2	1.2
average_intron_length	133	93	109	92
min_intron_length	11	23	4	10
max_intron_length	42135	7292	59574	5000

Table S1. CDS features of the four available gene annotations of the *Z. tritici* IPO323 genome (JGI, MPI, RRES and CURTIN).

The first annotation of *Z. tritici* genome , with 10,933 gene models, was developed in 2011 by the Joint Genome Institute with *ab initio* tools FGENESH and Genewise (Birney et al. 2004) using EST (expressed sequence tag) and proteome evidence (JGI, (Goodwin et al. 2011)). The second annotation was performed in 2015 by the Max Planck Institute, resulting in 11,839 gene models (MPI, Germany, (Grandaubert et al. 2015)) identified with the Fungal Genome Annotation pipeline (Haas et al. 2011). This pipeline uses *ab initio* tools GeneMark-ES, GeneMark-HMM (Lukashin 1998) and Augustus (Stanke et al. 2006) combined by EVidenceModeler (Haas et al. 2008) with RNA-Seq evidence and keeping as much as possible of the first annotation provided by JGI. The third annotation was generated in 2015 by the Rothamsted Research Experimental Station (UK) with 13,862 gene models (RRES, (Chen et al. 2023)) obtained with the *ab initio* tool MAKER-HMM (Holt and Yandell 2011) and RNA-Seq evidence. The fourth annotation published in 2015 by the Centre for Crop & Disease Management, Curtin University. (CURTIN, Australia) with 13,260 gene models, was obtained with *ab initio* tool CodingQuarry (Testa et al. 2015) and RNA-Seq evidence. Gene models were filtered out for transposable elements. nb : number, min : minimum, max : maximum.

	JGI		MPI		RRES		CURTIN	
Chr	#genes	%	#genes	%	#genes	%	#genes	%
1	1975	18,2%	2107	18,0%	2321	17,1%	2326	17,8%
2	1127	10,4%	1236	10,6%	1377	10,1%	1380	10,5%
3	1067	9,8%	1138	9,7%	1297	9,5%	1261	9,6%
4	818	7,5%	889	7,6%	998	7,3%	993	7,6%
5	776	7,2%	848	7,2%	986	7,3%	971	7,4%
6	685	6,3%	716	6,1%	820	6,0%	810	6,2%
7	758	7,0%	842	7,2%	975	7,2%	957	7,3%
8	685	6,3%	749	6,4%	843	6,2%	817	6,2%
9	601	5,5%	620	5,3%	703	5,2%	689	5,3%
10	513	4,7%	551	4,7%	624	4,6%	616	4,7%
11	482	4,4%	523	4,5%	604	4,4%	607	4,6%
12	405	3,7%	446	3,8%	514	3,8%	523	4,0%
13	325	3,0%	362	3,1%	411	3,0%	401	3,1%
14	111	1,0%	108	0,9%	163	1,2%	115	0,9%
15	83	0,8%	84	0,7%	143	1,1%	99	0,8%
16	86	0,8%	102	0,9%	169	1,2%	98	0,7%
17	76	0,7%	84	0,7%	131	1,0%	74	0,6%
18	60	0,6%	75	0,6%	121	0,9%	91	0,7%
19	84	0,8%	78	0,7%	141	1,0%	90	0,7%
20	78	0,7%	89	0,8%	135	1,0%	92	0,7%
21	54	0,5%	65	0,6%	107	0,8%	84	0,6%
Total	10849		11712		13583		13094	

Table S2. Chromosome localization of gene models of the four available annotations of the *Z. tritici* IPO323 genome (JGI, MPI, RRES and CURTIN).

Chr: Chromosome, of which the first 13 are the core and the remaining 8 accessories.
#genes: ratio compared to the whole dataset.

A:

	Synthetic media		Complex media
	Nitrogen	Carbon	
18°C	Nitrate	Glucose	YPD
	Nitrate	Xylose	PDB
	Nitrate	Mannitol	AE
	Nitrate	Galactose	
	Nitrate	Sucrose	
25°C			YPD
			PDB

Conditions used for preparing IsoSeq and RNA-seq libraries

B:

Condition	Replicate	Type	# transcripts - post stringtie	# transcripts - post jaccardclip	# transcripts - post cov filter	RUN accession	Iso-Seq accession	# Iso-Seq transcripts post filtering	# Iso-Seq genes
AE (Yeast extract Glycerol)	1	PE-stranded	11686	12528	12369	GSM6758342	GSM6758369	6867	3896
AE (Yeast extract Glycerol)	2	PE-stranded	10937	11803	11644	GSM6758343			
AE (Yeast extract Glycerol)	3	PE-stranded	10967	11666	11544	GSM6758344			
MMZt NO3 + Glucose	1	PE-stranded	10780	11568	11406	GSM6758345	GSM6758370	10083	5669
MMZt NO3 + Glucose	2	PE-stranded	9713	10432	10303	GSM6758346			
MMZt NO3 + Glucose	3	PE-stranded	10807	11796	11595	GSM6758347			
MMZt NO3 + Sucrose	1	PE-stranded	10796	11683	11522	GSM6758348	NA		
MMZt NO3 + Sucrose	2	PE-stranded	10876	11648	11511	GSM6758349			
MMZt NO3 + Sucrose	3	PE-stranded	10496	11428	11301	GSM6758350			
MMZt NO3 + Xylose	1	PE-stranded	10881	11913	11722	GSM6758351	GSM6758371	9888	5685
MMZt NO3 + Xylose	2	PE-stranded	10575	11296	11135	GSM6758352			
MMZt NO3 + Xylose	3	PE-stranded	10636	11310	11165	GSM6758353			
MMZt NO3 + Mannitol	1	PE-stranded	10534	11275	11142	GSM6758354	GSM6758372	9049	5502
MMZt NO3 + Mannitol	2	PE-stranded	11008	12038	11832	GSM6758355			
MMZt NO3 + Mannitol	3	PE-stranded	10848	11693	11539	GSM6758356			
MMZt NO3 + Galactose	1	PE-stranded	10831	12045	11831	GSM6758357	NA		
MMZt NO3 + Galactose	2	PE-stranded	10939	11950	11794	GSM6758358			
MMZt NO3 + Galactose	3	PE-stranded	11000	12064	11853	GSM6758359			

MMZt NO3_Glucose +SAHA	1	PE-stranded	11141	12344	12132	GSM6758360	GSM6758373	10922	6317
MMZt NO3_Glucose +SAHA	2	PE-stranded	11199	12422	12226	GSM6758361			
MMZt NO3_Glucose +SAHA	3	PE-stranded	11283	13039	12767	GSM6758362			
MMZt NO3_Glucose +TSA	1	PE-stranded	11473	13061	12785	GSM6758363	GSM6758374	10063	5828
MMZt NO3_Glucose +TSA	2	PE-stranded	11762	13325	13048	GSM6758364			
MMZt NO3_Glucose +TSA	3	PE-stranded	11566	13287	13021	GSM6758365			
YPD - 18 °C	1	PE-stranded	10964	12468	12240	GSM6758366	GSM6758378	2187	1699
YPD - 18 °C	2	PE-stranded	11018	12442	12201	GSM6758367			
YPD - 18 °C	3	PE-stranded	10923	12462	12228	GSM6758368			
Wheat infection - 4 dpi	1	SE-stranded	10349	10431	10395	SRR6215483			
Wheat infection - 4 dpi	2	SE-stranded	8661	8671	8665	SRR6215484			
Wheat infection - 11 dpi	1	SE-stranded	9135	9141	9133	SRR6215485			
Wheat infection - 11 dpi	2	SE-stranded	9803	9816	9804	SRR6215486			
Wheat infection - 13 dpi	1	SE-stranded	11675	11688	11672	SRR6215487			
Wheat infection - 13 dpi	2	SE-stranded	11357	11362	11348	SRR6215488			
Wheat infection - 20 dpi	1	SE-stranded	12215	12226	12205	SRR6215489			
Wheat infection - 20 dpi	2	SE-stranded	12196	12202	12180	SRR6215490			
YMS - 18 °C	1	SE-stranded	9183	11100	10952	SRR8788920			
YMS - 18 °C	2	SE-stranded	8885	10930	10797	SRR8788921			
YMS - 18 °C kmt1 mutant	1	SE-stranded	9696	12054	11875	SRR8788922			
YMS - 18 °C kmt1 mutant	2	SE-stranded	9554	11725	11559	SRR8788923			
YMS - 18°C kmt6 mutant	1	SE-stranded	8688	10876	10720	SRR8788924			
YMS - 18°C kmt6 mutant	2	SE-stranded	9001	11152	11006	SRR8788925			
YMS - 18 °C kmt1/6 mutant	1	SE-stranded	10408	13168	12929	SRR8788926			
YMS - 18 °C kmt1/6 mutant	2	SE-stranded	10415	13127	12914	SRR8788927			
YPD-25°C							GSM6758379	5759	4158
PDB-18°C							GSM6758375	1654	1392
PDB-25°C							GSM6758376	3107	2490

Table S3. RNA-Seq and Iso-Seq cDNA libraries from *Z. tritici* IPO323

A: Summary of the conditions used to produce RNAs

B: Ten growth conditions were used to produce RNAs for Iso-seq and RNA-seq libraries. The reference isolate of *Z. tritici* IPO323 (Goodwin et al. 2011) was stored at -80°C as a yeast-like cell suspension (10^7 cells/mL in 30% glycerol). *Z. tritici* was grown at 18°C in the dark on solid (Yeast extract Peptone Dextrose (YPD) agar) or liquid (Potato Dextrose Broth (PDB)) media. For RNA production, *Z. tritici* isolate IPO323 (4-day-old yeast-like cells diluted to 10^5 cells/mL final) was cultivated in 75-mL agitated liquid cultures (500 mL Erlen flasks, 150 rpm) at 18°C in the dark for 4 days. Different media were used (Table S3) including Glucose-NO_3 synthetic medium defined as MM-Zt (Marchegiani et al. 2015). MM-Zt was modified by replacing glucose (10 g/L) by different carbon sources (Xylose, Mannitol, Galactose, Sucrose at 10 g/L)). Histone Deacetylase inhibitors such as

trichostatin ((TSA, Sigma T8552, 1 µM final) and SAHA (SAHA, Sigma SML0061, 1 mM final) were added to MM-Zt to express genes located in genomic regions with repressive chromatin marks (Meile et al. 2020). The composition of complex media (Yeast-Peptone-Dextrose: YPD, Potato-Dextrose-Both: PDB, Glycerol-Nitrate: AE) was already described (Scalliet et al. 2012). Cultures of IPO323 in YPD and PDB were performed at 18°C and 25°C, while AE cultures were performed only at 18°C. A total of 14 culture conditions was used for RNA production (Table S3). All cultures for RNA-Seq were performed in triplicate. Cultures were centrifuged at 3000 rpm for 10 minutes and mycelium pellets were washed with water and frozen with liquid nitrogen. Frozen mycelium was lyophilized and kept at -80°C until extraction. RNAs were extracted using the Qiagen Plant RNeasy Kit according to the manufacturer's protocol (Ref. 74904, Qiagen France SAS, Courtaboeuf, France). Preparation and sequencing of PacBio Iso-Seq libraries were performed by the INRAE platform Gentyane (http://gentyane.clermont.inrae.fr). The SMARTer PCR cDNA Synthesis Kit (ref 634926, Clontech, Mountain View, CA, USA) was used for polyA-primed first-strand cDNA synthesis followed by optimized PCR amplification and library preparation using the SMRTbell Template Prep Kit (ref 101-357-000, Pacific Bioscience, Menlo Park, CA, USA) according to manufacturer protocols. The cDNA libraries were prepared without size selection and bar coded for multiplexing. Sequencing was performed on a PacBio SEQUEL (version 1). Illumina RNA-seq single-stranded libraries were prepared using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB #E7490, New England BioLabs, Ipswich, Massachusetts, USA) and the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB #E7765, New England BioLabs, Ipswich, Massachusetts, USA). Custom 8-bp barcodes were added to each library during the preparation process. Pooled samples were cleaned with magnetic beads included in the library preparation kit. Each pool was run on a lane of Illumina HiSeqX (Illumina, San Diego, California, USA) using a 150-cycle paired-end run. Additional single-stranded RNA-seq data were obtained from public databases (see below). RNA-Seq sequences were used for transcript assemblies and differential expression analyses. Iso-Seq sequences were used for gene annotation and isoform analyses. Wheat infection data for SRR6215483- SRR6215490 and YMS mutants SRR8788920- SRR8788927 were downloaded from the Sequence Read Archive (SRA) repository. Other RNA-Seq data generated in this study were submitted to the SRA database (see columns accession).

RNA-Seq data were cleaned and trimmed with Trimmomatic (v 0.36) (Bolger et al. 2014). The cleaned sequences were then mapped to the *Z.tritici* IPO323 genome using STAR (v 2.5.1b, --alignIntronMin 4 --alignIntronMax 5000 -- alignMatesGapMax 5000) (Dobin et al. 2012). Wig files of uniquely mapped reads were converted to BigWig files with wigToBigWig (v4). StringTie (v2.1.1) (Pertea et al. 2015) was then used to assemble the mapped RNA-Seq reads into transcripts with different parameters depending on the depth of sequencing of libraries and their type (-m 150 --rf --g 0 -f 0.1 -a 10 -j 2 or -j 4). The Trinity script inchworm_transcript_splitter.pl (version 2.8.5) (Haas et al. 2013) was used to split the transcripts with non-uniform coverage based on the Jaccard clip method. Clipped transcripts were extracted with home-made scripts and clustered with Stringtie and associated bam files to obtain transcripts per million (TPM) counts. All libraries were concatenated into one gff file without merge to avoid loss of information by fusion of small transcripts into larger ones due to the large number of genes in the *Z. tritici* genome with overlapping untranslated regions (UTRs).

Iso-Seq raw data were processed with the Iso-Seq V3.2 pipeline from PacBio generating polished Circular Consensus Sequences (CCS). CCS were then mapped to the *Z. tritici* IPO323 genome with Gmap (2019-01-31) (Wu and Watanabe 2005) and unmapped, low-mapping-quality (≤0) or multi-mapped CCS were filtered out. The CupCake package (v10.0.0, https://github.com/Magdoll/cDNA_Cupcake) filtered the isoforms, removing the less-expressed and degraded transcripts using the following tools: *collapse_isoforms_by_sam.py*, *get_abundance_post_collapse.py*, *filter_by_count.py*, *filter_away_subset.py*. Readthrough transcripts were removed using the previous annotations (MPI, JGI, CURTIN, RRES) with BEDTools intersect (Quinlan and Hall 2010) with an an overlap of 100% for full coding sequences (CDS) (-F 1.0) and the same strand (-s)) of at least 2 CDS. Transcripts mapped on the mitochondrial genome were filtered out as well. Subsequently, all libraries were processed with *chain_samples.py* from CupCake and clustered for stringent selection. Splicing junctions obtained by STAR (SJ.out.tab files) from Illumina RNA-Seq libraries were used to filter out isoform transcripts with unsupported junctions. Finally, long-read transcripts fully spanning transposable elements were removed with BEDTools, giving the final set of transcript evidence.

## statistics ##	
nb_genes	13414
average_gene_length	1900
median_gene_length	1635
min_gene_length	102
max_gene_length	17065
nb_transcripts	13414
average_transcripts_per_gene	1
average_transcript_length	1900
median_transcript_length	1635
min_transcript_length	102
max_transcript_length	17065
nb_exons	30946
average_exons_per_transcript	2.3
average_exon_length	782
median_exon_length	431
min_exon_length	1
max_exon_length	16680
nb_transcript_mono_exon	4850
nb_introns	17532
average_introns_per_transcript	1.3
average_intron_length	73
median_intron_length	57
min_intron_length	5
max_intron_length	3166
nb_CDS	13414
average_CDS_length	1287
median_CDS_length	1041
min_CDS_length	102
max_CDS_length	16506
nb_transcripts_with_utr	9856
average_five_prime_utr_length	315
median_five_prime_utr_length	156
min_five_prime_utr_length	1
max_five_prime_utr_length	7053
average_three_prime_utr_length	389
median_three_prime_utr_length	220
min_three_prime_utr_length	1
max_three_prime_utr_length	8647

Table S4. Features of Re-annotated Gene Models (RGMs) of *Z. tritici* IPO323 genome

nb : number, min : minimum, max : maximum, utr : untranslated region from transcripts

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

Table S5. Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis of Re-annotated Gene Models (RGM) and gene models from previous annotations of the *Z. tritici* IPO323 genome.

This comparison was performed using BUSCO and the ascomycota_odb as reference genes. Higher BUSCO scores (99.4 % identical) were obtained with RGMs compared to the JGI, MPI and CURTIN annotations (95.7-98.5% identical), while scores obtained with RRES gene models were similar to RGMs (99.1 % identical). The JGI annotation had the highest number of missing BUSCO genes. The eight missing BUSCOs in RGMs were reduced to six after manual inspection. These six RGMs that were missing in BUSCO encoded a Leucyl-tRNA synthetase, a WD40-repeat-containing domain protein, a Zinc finger protein, a Heavy metal-associated domain protein, a protein with an HMA domain, a PHD-type protein and a GTP binding domain protein. Their conservation across fungi is questionable, since a blastp search showed that they are missing from numerous genomes.

Chromosome	# RGMs	# dissimilar RGMs	# specific RGMs	# RGMs with no evidence
1	2365	184	100	47
2	1392	132	54	28
3	1291	130	61	32
4	1021	70	54	29
5	997	98	49	18
6	835	78	41	12
7	952	88	39	224
8	843	79	33	27
9	708	68	34	25
10	624	56	21	29
11	604	53	28	9
12	542	51	39	11
13	423	42	23	12
14	133	41	18	8
15	115	43	18	7
16	103	28	10	15
17	89	24	9	8
18	94	26	5	8
19	103	30	19	9
20	93	23	7	8
21	87	32	9	8
Total	13414	1376	671	574

Table S6. Distribution of Re-annotated Gene Models (RGMs) on *Z. tritici* IPO323 chromosomes

RGMs: number of Re-annotated Gene Models (RGMs).
dissimilar RGMs: number of RGMs with the following properties, at a given locus, a RGM was predicted by at least one other annotation, but they differed in their structure, here RGMs differing from all previous annotations.
specific RGMs: number of RGMs with the following properties, at a given locus, a single RGM is predicted by a single annotation, here RGM specific.
RGMs with no evidence: number of RGMs with the following properties, RGMs without transcript or protein evidence, but rescued since they were predicted by at least four different annotations.

fusion / split	JGI	MPI	CURTIN	RRES	RGM
JGI	—	312/674	626/1385	425/946	**558/1258**
MPI	286/584	—	801/1702	533/1127	**706/1507**
CURTIN	113/230	31/63	—	133/278	**83/176**
RRES	102/206	51/103	445/929	—	**333/701**
RGM	**92/186**	**19/38**	**177/363**	**98/200**	—

Table S7. Identification of fused/split genes in the *Z. tritici* IPO323 genome annotations.

These comparisons were performed in a pairwise manner. The first value is the number of fused genes for the horizontal entry and the second is the number of split genes for the vertical entry.

Isoforms per RGM	1	2	3	4	5	6+
Number of RGMs	11672	1342	274	77	29	20

Table S8. Number of transcript isoforms detected for Re-annotated Gene Models (RGMs) in *Z. tritici* IPO323

The annotation of transcript isoforms was performed with sqanti3 (Tardaguila et al. 2018) using Iso-Seq transcripts, previously established to infer UTRs, filtered for UTR length isoforms and low expression levels (less than 10% of total RNA-Seq reads), using the *ingenannot isoform_ranking* tool. RNA-Seq reads were mapped to Iso-Seq transcripts with RSEM v1.3.3 (Li and Dewey 2011) and Differential Isoform Usage (DIU) performed with tappAS (De La Fuente et al. 2020) with annotations obtained from sqanti3. The gene with the highest number of isoforms detected by Iso-Seq is ZtIPO323_108820 (chr_10:1162483...1166667 - 4.19 Kb) with 15 isoforms, among which only a few were supported quantitatively by RNA-Seq.

LncRNA	Type	Up in planta log2FC > 2	Down in planta log2FC < 2	Gene (antisense)	Gene annotation
PB.854.1	intergenic		Y		
PB.927.1	intergenic		Y		
PB.1188.1	antisense		Y	ZtIPO323_016330	subtilisin-like protein
PB.1594.1	antisense		Y	ZtIPO323_022040	related to allantoase permease
PB.2214.1	antisense		Y	ZtIPO323_030630/ ZtIPO323_030640	NA
PB.2569.1	antisense		Y	ZtIPO323_035460	alpha/beta like hydrolase
PB.2709.1	antisense	Y		ZtIPO323_037670	TTL-domain containing protein
PB.4776.1	antisense		Y	ZtIPO323_066740	NA
PB.5130.1	intergenic	Y			
PB.5328.2	antisense		Y	ZtIPO323_074720	NA
PB.5366.1	antisense		Y	ZtIPO323_075280	HSP20-like chaperone
PB.6002.1	antisense		Y	ZtIPO323_084180	P450 monooxygenbase
PB.6788.1	antisense	Y		UTR overlap	
PB.7120.1	antisense		Y	ZtIPO323_102870	Phosphomevalonate kinase
PB.7618.1	antisense		Y	ZtIPO323_110030	glycoside hydrolase family 45 protein
PB.8769.1	intergenic	Y			
PB.8778.1	intergenic	Y			

Table S9. Long non-coding RNAs (lncRNA) from *Z. tritici* IPO323 differentially expressed during infection

Iso-Seq transcripts annotated as antisense and intergenic with sqanti3 were selected as long non-coding (lnc) RNAs. Then transcripts shorter than 1 Kb in length (Novikova et al. 2012), overlapping with TEs and containing an open reading frame (ORF) longer than 100 amino acids predicted with getorf by EMBOSS (Rice et al. 2000) were discarded. The resulting "non-coding" transcripts were annotated with CPC2 (Kang et al. 2017), and only transcripts without an ORF with a PFAM domain were kept as lncRNAs. featureCounts (v1.5.1) (Liao et al. 2014) was used to count reads per transcript, followed by differential expression analysis by edgeR (Robinson et al. 2010) with the SARTools package (v1.6) (Varet et al. 2016). List of lncRNAs annotated from Iso-Seq data differentially expressed during infection with a strong fold change (4X). Antisense lncRNAs are preferentially involved in cis regulation, whereas intergenic lncRNAs are preferentially involved in trans regulation

	transcripts			proteins			genometric median				distance to (0,0)	
	mean	median	stdev	mean	median	stdev	x	y	mean distance	median distance	mean	median
funannotate	0.245	0.151	0.255	0.497	0.258	0.483	0.185	0.301	0.526	0.348	0.604	0.469
Helixer	0.316	0.261	0.288	0.557	1.0	0.481	0.309	0.657	0.546	0.59	0.69	1.0
RGM	0.2	0.125	0.226	0.488	0.176	0.486	0.144	0.249	0.514	0.301	0.578	0.398
BRAKER3	0.223	0.141	0.237	0.453	0.09	0.478	0.155	0.176	0.493	0.227	0.559	0.338

Table S10. Metrics calculated from AED scores for the four gene annotations obtained with InGenAnnot (RGM), funannotate, Helixer and BRAKER3. The geometric median corresponds to the centroid of all points of each AED plot. The mean and median distances were computed for all AED values relative to the geometric median and are used as proxies for the dispersion of the AED scores of all annotations. The same values are computed relative to the "ideal" point at coordinates (x=0, y=0), where all annotations perfectly match transcript and protein evidence. BRAKER3 (Gabriel et al. 2024) proposes isoforms for well-annotated genes, which introduces a bias in terms of the number of genes considered compared to funannotate (Palmer and Stajich 2020), Helixer (Stiehler et al. 2021), and RGM, for which only one transcript per gene is considered.

References

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19:455–477

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. Genome Res. 14:988–995

Bolger, A. M., Lohse, M., and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–20

Chen, H., King, R., Smith, D., Bayon, C., Ashfield, T., Torriani, S., Kanyuka, K., Hammond-Kosack, K., Bieri, S., and Rudd, J. 2023. Combined pangenomics and transcriptomics reveals core and redundant virulence processes in a rapidly evolving fungal plant pathogen. BMC Biol. 21:1–22

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., Claverie, J., and Gascuel, 0livier. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. 2012. STAR: ultrafast universal RNA-seq aligner. Bioinformatics.

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

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., and Stanke, M. 2024. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 34:769–777

Goodwin, S. B., M'Barek, S. Ben, Dhillon, B., Wittenberg, A. H. J., Crane, C. F., Hane, J. K., Foster, A. J., van der Lee, T. A. J., Grimwood, J., Aerts, A., Antoniw, J., Bailey, A., Bluhm, B., Bowler, J., Bristow, J., van der Burgt, A., Canto-Canché, B., Churchill, A. C. L., Conde-Ferràez, L., Cools, H. J., Coutinho, P. M., Csukai, M., Dehal, P., de Wit, P., Donzelli, B., van de Geest, H. C., van Ham, R. C. H. J., Hammond-Kosack, K. E., Henrissat, B., Kilian, A., Kobayashi, A. K., Koopmann, E., Kourmpetis, Y., Kuzniar, A., Lindquist, E., Lombard, V., Maliepaard, C., Martins, N., Mehrabi, R., Nap, J. P. H., Ponomarenko, A., Rudd, J. J., Salamov, A., Schmutz, J., Schouten, H. J., Shapiro, H., Stergiopoulos, I., Torriani, S. F. F., Tu, H., de Vries, R. P., Waalwijk, C., Ware, S. B., Wiebenga, A., Zwiers, L. H., Oliver, R. P., Grigoriev, I. V., and Kema, G. H. J. 2011. Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 7

Grandaubert, J., Bhattacharyya, A., and Stukenbrock, E. H. 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 Genes, Genomes, Genet. 5:1323–1333

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., Macmanes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., Leduc, R. D., Friedman, N., and Regev, A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., and Wortman, J. R. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9:R7

Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A., and Wortman, J. R. 2011. Approaches to fungal genome annotation. Mycology. 2:118–141

Holst, F., Bolger, A., Günther, C., Maß, J., Triesch, S., Kindel, F., Kiel, N., Saadat, N., Ebenhöh, O., Usadel, B., Schwacke, R., Bolger, M., Weber, A. P. M., and Denton, A. K. 2023. Helixer–de novo Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model. bioRxiv. :2023.02.06.527280

Holt, C., and Yandell, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 12:491

Kang, Y. J., Yang, D. C., Kong, L., Hou, M., Meng, Y. Q., Wei, L., and Gao, G. 2017. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 45:W12–W16

De La Fuente, L., Arzalluz-Luque, Á., Tardáguila, M., Del Risco, H., Martí, C., Tarazona, S., Salguero, P., Scott, R., Lerma, A., Alastrue-Agudo, A., Bonilla, P., Newman, J. R. B., Kosugi, S., McIntyre, L. M., Moreno-Manzano, V., and Conesa, A. 2020. TappAS: A comprehensive computational framework for the analysis of the functional impact of differential splicing. Genome Biol. 21:119

Li, B., and Dewey, C. N. 2011. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 12:323

Liao, Y., Smyth, G. K., and Shi, W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 30:923–930

Lukashin, A. 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26:1107–1115

Marchegiani, E., Sidhu, Y., Haynes, K., and Lebrun, M. H. 2015. Conditional gene expression and promoter replacement in Zymoseptoria tritici using fungal nitrate reductase promoters. Fungal Genet. Biol. 79:174–179

Meile, L., Peter, J., Puccetti, G., Alassimone, J., McDonald, B. A., and Sánchez-Vallet, A. 2020. Chromatin dynamics contribute to the spatiotemporal expression pattern of virulence genes in a fungal plant pathogen. MBio. 11:1–18

Novikova, I. V., Hennelly, S. P., and Sanbonmatsu, K. Y. 2012. Sizing up long non-coding RNAs: Do lncRNAs have secondary and tertiary structure? Bioarchitecture. 2:189–199

Palmer, J. M., and Stajich, J. 2020. Funannotate v1.8.1: Eukaryotic genome annotation (v1.8).

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33:290–295

Quinlan, A. R., and Hall, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842

Rice, P., Longden, L., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16:276–277

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26:139–140

Scalliet, G., Bowler, J., Luksch, T., Kirchhofer-Allan, L., Steinhauer, D., Ward, K., Niklaus, M., Verras, A., Csukai, M., Daina, A., and Fonné-Pfister, R. 2012. Mutagenesis and Functional Studies with Succinate Dehydrogenase Inhibitors in the Wheat Pathogen Mycosphaerella graminicola O. Lespinet, ed. PLoS One. 7:e35429

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34:W435–W439

Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A. P. M., and Denton, A. K. 2021. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. Bioinformatics. 36:5291–5298

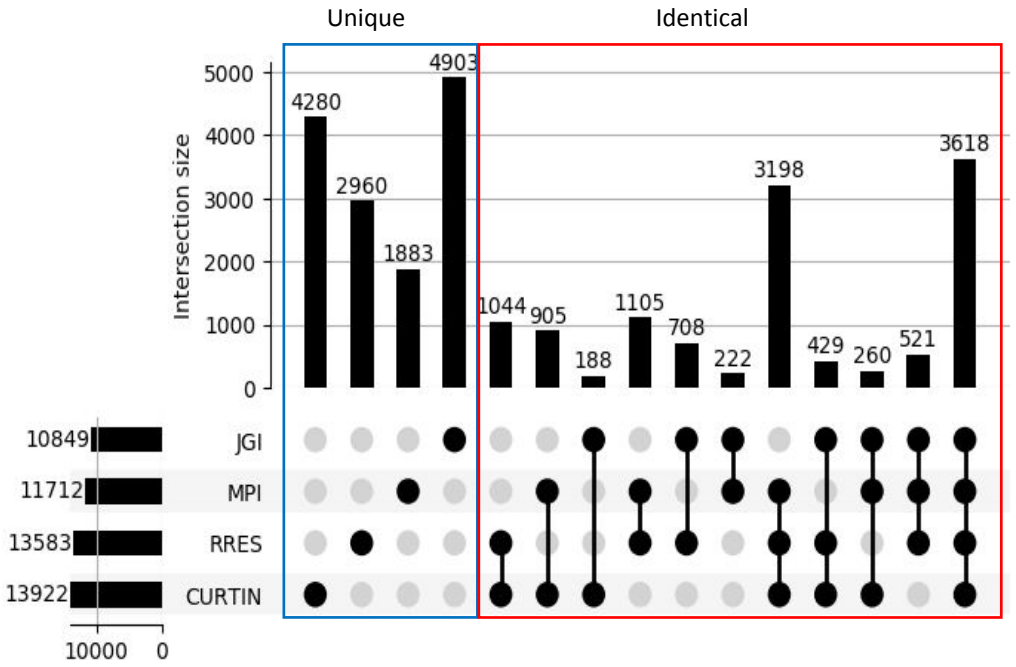
Tardaguila, M., De La Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., Martens, L., Rodriguez-Navarro, S., Moreno-Manzano, V., and Conesa, A. 2018. SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 28:396–411

Testa, A. C., Hane, J. K., Ellwood, S. R., and Oliver, R. P. 2015. CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics. 16:170

Varet, H., Brillet-Guéguen, L., Coppée, J.-Y., and Dillies, M.-A. 2016. SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data K. Mills, ed. PLoS One. 11:e0157022

Wu, T. D., and Watanabe, C. K. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 21:1859–1875

a



b

		JGI	MPI	RRES	CURTIN
	JGI	*10849*			
Identical	MPI	4621	*11712*		
CDS	RRES	5276	8442	*13583*	
	CURTIN	4495	7981	8289	*13922*
Unique:		4903	1883	2960	4280
Dissimilar CDS		4752	1871	2367	3844
Specific CDS		151	12	593	436

Figure 1. Comparison of *Zymoseptoria tritici* reference isolate IPO323 genome annotations. **a**) Upset plot of the gene models from the four annotations of IPO323 (JGI, MPI, RRES and CURTIN). Number of gene models with identical coding sequences (CDS). **b**) Comparison of IPO323 gene annotations. Number of CDS in each annotation. Identical CDS: identical CDS at a given locus. Unique Dissimilar CDS: at a given locus, a CDS is predicted by at least one other annotation, but they differ in their structure. Unique Specific CDS: at a given locus, a single CDS is predicted by a single annotation. The highest numbers of identical gene models between two annotations were observed for MPI-RRES (8,442), RRES-CURTIN (8,289), and MPI-Curtin (7,981), while the lowest numbers of identical gene models were observed between JGI and the three other annotations (4,495, 4,621 and 5,276 for JGI-Curtin, JGI-MPI and JGI-RRES respectively).

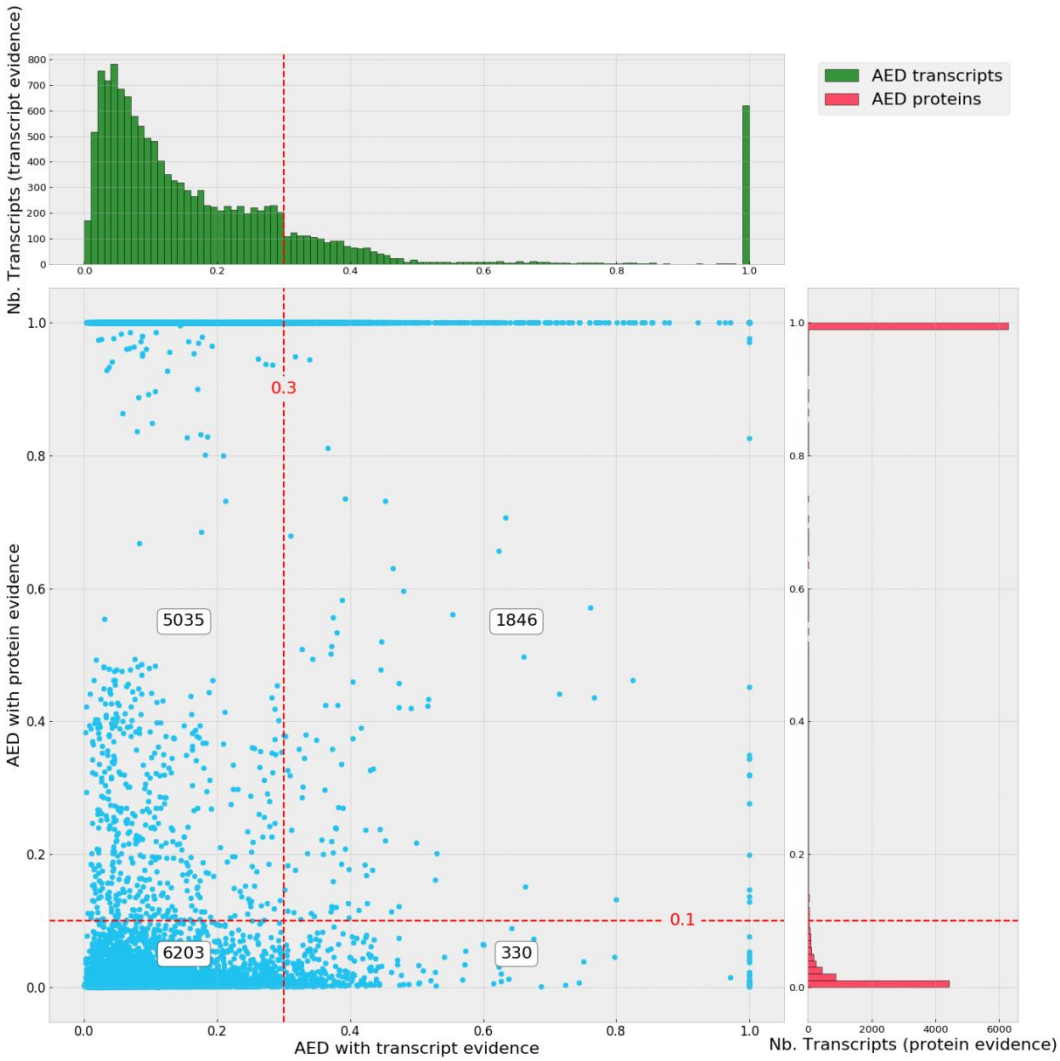
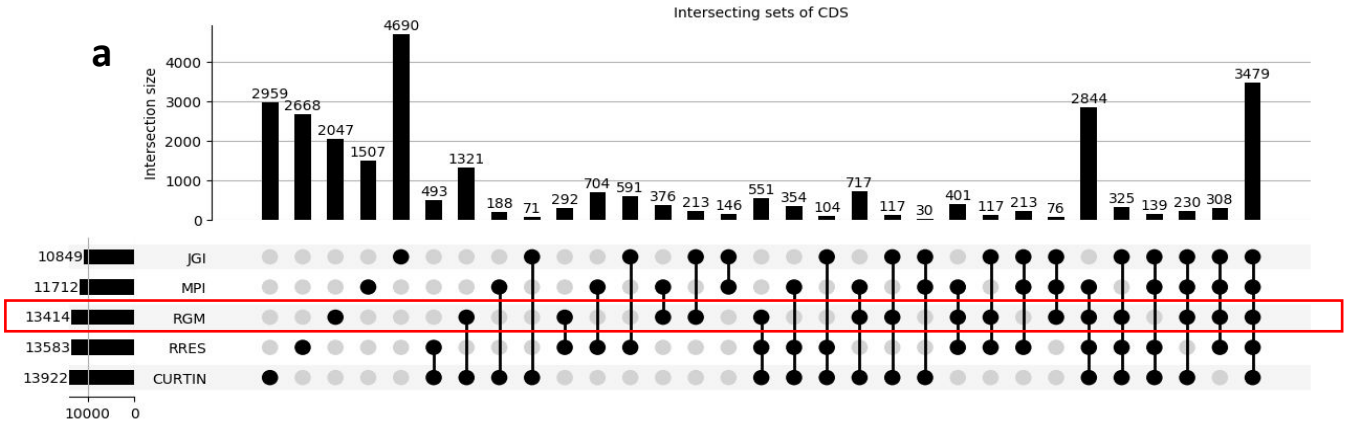


Figure 2. Selection of the best Re-annotated Gene Models (RGMs) according to their Annotation Edit Distance (AED) scores.

Plot of RGM AED scores. AED scores (0-1) describe how a given gene model fits to transcript and protein evidence (best fit = 0). Transcript evidence was computed from RNA-Seq or Iso-Seq data (X axis). Protein evidence was computed from fungal protein sequences excluding *Zymoseptoria* species (Y axis). The red, dashed lines represent the AED thresholds to filter out genes (0.3 for transcripts, 0.1 for proteins), except if they are supported by at least four different annotations (1846 RGMs, upper right area of the graph). The numbers of genes in the four areas are displayed in white text boxes. Numbers of transcripts with transcript evidence were plotted on cumulative histograms above the scatter plot (green). Numbers of transcripts with protein evidence were plotted on cumulative histograms on the right of the scatter plot (red).



b

	JGI	MPI	RRES	CURTIN	Any
RGM	4865	8431	8317	9584	11367

Figure 3. Comparison of the novel IPO323 genome annotation (Re-annotated Gene Models, RGM) with the four available annotations

a) Upsetplot of RGMs with gene models from the four available annotations (JGI, MPI, RRES and CURTIN). Number of shared (identical) gene models for coding sequences (CDS).

b) Number of identical CDS between RGMs and each available annotation.

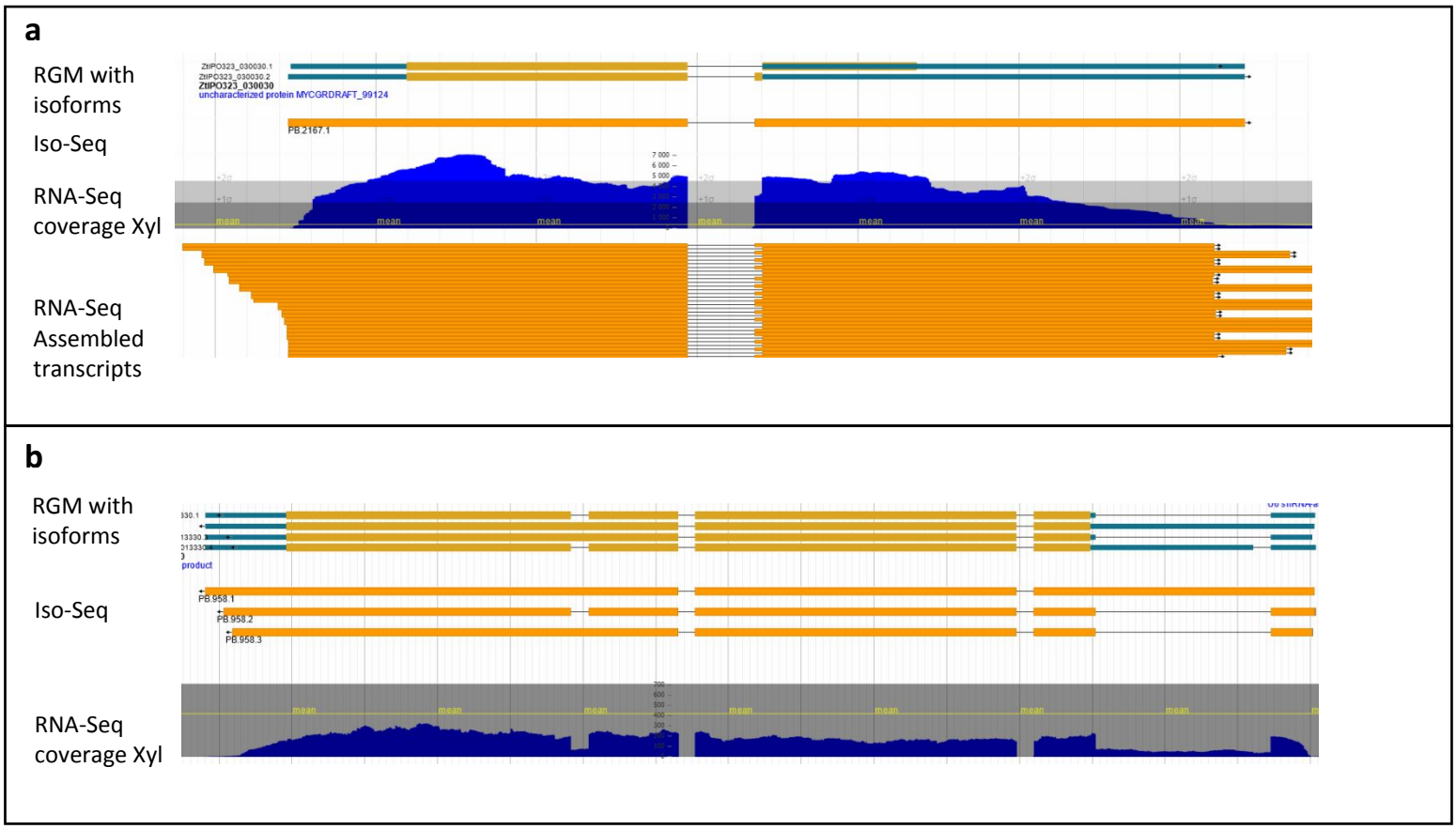


Figure 4. Transcript isoforms of Re-annotated Gene Models (RGMs) ZtIPO323_030030 (a) and ZtIPO323_013330 (b) supported by Iso-Seq and RNA-Seq evidence.

a) Gene ZtIPO323_030030 (chr2: 777930…1778675, 747 b). This RGM has two transcript isoforms (alternative 3' acceptor site). Both encoded Small Secreted Proteins (SSP 10, File S1). Previous annotations selected the second acceptor site leading to the longest CDS. A single Iso-Seq transcript corresponding to the longest CDS was detected (Iso-Seq track), while both isoforms were detected using RNA-Seq data (RNA-Seq assembled transcript). RNA-Seq coverage identified both isoforms in equal amounts (RNA-Seq coverage Xyl). Based on read coverage from different RNA-Seq libraries, the isoform corresponding to the shortest CDS was the most frequent. This isoform was likely the canonical form and encoded a protein with a C-terminus that was reduced in length by 34% compared to the other isoform. RGMs with isoforms track: different isoforms. Iso-Seq track: filtered Iso-Seq transcripts. RNA-Seq coverage Xyl track: coverage of strand-specific RNA-Seq reads. RNA-Seq assembled transcript track: assembly of strand-specific RNA-Seq reads.

b) ZtIPO323_013330 (chr_1:3420115..3424093, 3.98 Kb). This RGM had four transcript isoforms. The selected RGM had four splicing sites, one of which in the 5' UTR was supported by Iso-Seq transcript (Iso-Seq n°2) and RNA-Seq (RNA-Seq coverage Xyl). Two Iso-Seq transcripts with one or two intron retention events were detected as Iso-Seq transcripts (Iso-Seq n°1 and 3) and confirmed by RNA-Seq (RNA-Seq coverage Xyl). One Iso-Seq transcript had an alternative 5' donor splicing site in the 5' UTR (Iso-Seq n°4). This isoform was likely weakly expressed, as it was not supported by RNA-Seq (RNA-Seq coverage Xyl). RGMs with isoforms track: different RGM isoforms. Iso-Seq track: filtered Iso-Seq transcripts. RNA-Seq coverage Xyl track: coverage of strand-specific RNA-Seq reads. RNA-Seq assembled transcript track: assembly of strand-specific RNA-Seq reads.

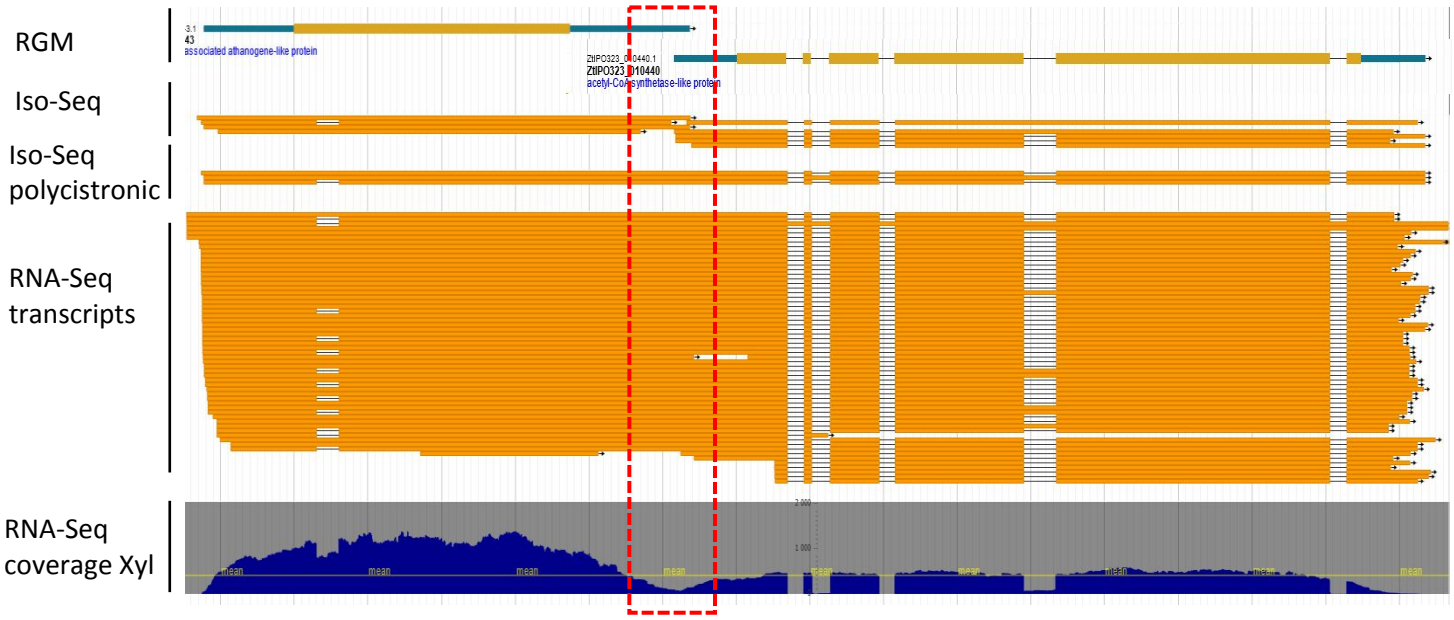


Figure 5. Examples of polycistronic transcripts shown for Re-annotated Gene Models (RGMs) ZtIPO323_010430 and ZtIPO323_010440

RGMs ZtIPO323_010430 and ZtIPO323_010440, located at chr_1:2692858...2697168 and chr_1:2692858...2697168, respectively, were transcribed on the same strand with overlapping 3'UTR and 5'UTR (red rectangle). Iso-Seq polycistronic track: evidence of transcripts covering the two RGMs. A strong decrease in RNA-Seq coverage was observed in the region of the overlap (red dashed rectangle), suggesting two singles, overlapping transcripts. The assembly of RNA-Seq reads led to a polycistronic transcript involving the two RGMs, likely resulting from the wrong assembly of reads from these overlapping transcripts. Iso-Seq track: filtered Iso-Seq transcripts mapping at this locus. Iso-Seq polycistronic track: polycistronic transcripts identified in the Iso-Seq database. RNA-Seq transcript track: assembly of strand-specific RNA-Seq reads mapping at this locus. RNA-Seq coverage Xyl track: coverage of strand-specific RNA-Seq reads mapping at this locus.