MPMI Vol. X, No. X, XXXX, pp. X-X, https://doi.org/10.1094/MPMI-07-25-0077-TA

TECHNICAL ADVANCE

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,[8] Gert H. J. Kema,[9] Igor V. Grigoriev,[7,10] James Hane,[6] Jason Rudd,[5] Eva Stukenbrock,[11,12] Daniel Croll,[3] Gabriel Scalliet,[2] and 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] Department of Protecting Crops and the Environment, Rothamsted Research, Harpenden, Herts AL52JQ, U.K.
[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, U.S.A.
[8] USDA-Agricultural Research Service, West Lafayette, IN 47907-2054, U.S.A.
[9] Wageningen University and Research, Laboratory of Phytopathology, Wageningen 6700 AA, The Netherlands
[10] Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94720, U.S.A.
[11] Environmental Genomics, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany
[12] Christian-Albrechts University of Kiel, 24118 Kiel, Germany

Accepted for publication 12 September 2025.

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 reannotated gene models (RGMs) were predicted, improving previous annotations. Iso-Seq transcripts outlined 5′ and 3′ untranslated regions 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: gene prediction, genome annotation, isoforms, Septoria tritici blotch, transcripts

[†]Corresponding author: N. Lapalu; nicolas.lapalu@inrae.fr

Author contributions: 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 the experiments needed for constructing cDNA libraries. G.S. funded the sequencing of the cDNA libraries. L.L. developed tools for annotation and performed initial analyses with an early version of InGenAnnot. N.L. finalized InGenAnnot and genome annotation. A.F., L.L., N.L., Y.P., A.G., G.S., and M.-H.L. compared RGMs with 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 the results and contributed to the improvement of the manuscript.

Funding: BIOGER benefits from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007). Syngenta Crop Protection funded 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 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 DE-AC02-05CH11231.

***e*-Xtra:** Supplementary material is available online.

The author(s) declare no conflict of interest.

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 the eukaryotic gene structure, including difficulties in intron or start codon prediction, and the quality of the 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; Gerads and Ernst 1998; Hansen et al. 1998), leading to incorrect gene predictions, such as gene fusions (Testa et al. 2015). In addition, fungi have short introns (70 to100 bp; Kupfer et al. 2004) and frequently fragmented genome assemblies. These particularities have led to the development of fungal-specific annotation pipelines (Birney et al. 2004; Brůna et al. 2021;

Holt and Yandell 2011; Lukashin and Borodovsky 1998; Min et al. 2017; Reid et al. 2014; Sallet et al. 2019; Scalzitti et al. 2020; Stanke et al. 2006). Long-read sequencing now provides fungal genome assemblies with almost no fragmentation. Experimental transcript evidence has also been improved by using large datasets of assembled Illumina single-stranded RNA-Seq short reads. Recently, Iso-Seq long-read sequencing has provided 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, because 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 or cap-analysis gene expression sequencing are available to define transcript start sites (Casco et al. 2022; Chiba et al. 2022) but are still rarely used in fungi.

We chose 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 and 8 accessory chromosomes. Twenty-two additional fully assembled (long-read) genome sequences of *Z. tritici* isolates are available (Badet et al. 2020; Feurtey et al. 2020; Möller et al. 2021), as well as genome sequences from four related *Zymoseptoria* species (*Z. ardabiliae*, *Z. brevis*, *Z. passerinii*, and *Z. pseudotritici*) (Feurtey et al. 2020). Around 14 to 22% of *Z. tritici* genomes are composed of transposable elements (TEs) (Badet et al. 2020; Dhillon et al. 2014; Grandaubert et al. 2015; Lorrain et al. 2021; Oggenfuss et al. 2021). The interest in *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 programs. This selection relies on a customized annotation edit distance (AED) metric (Eilbeck et al. 2009). InGenAnnot computes an AED for each piece of evidence, with penalties for unsupported intron splicing sites (Supplementary Fig. 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 (Supplementary 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 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 and Borodovsky 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 (RRES, U.K.) (Chen et al. 2023) with 13,862 gene models 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 in a dedicated INRAE genome browser (https://bioinfo.bioger.inrae.fr/portal/genome-portal/12) or the *Z. 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 agar) or liquid (potato dextrose broth) media. For RNA production, different media were used (Supplementary Table S3). Additional single-stranded RNA-Seq data were obtained from public databases (Supplementary Table S3). Novel and public RNA-Seq data were cleaned and mapped to the *Z. tritici* IPO323 genome (see Supplementary Table S3 for methods). Processed Iso-Seq data were also mapped to the *Z. tritici* IPO323 genome (see Supplementary Table S3 for methods).

Gene prediction and selection of the best gene models

The two ab initio gene prediction software programs Eugene v.1.6.1 (Sallet et al. 2019) and LoReAn v.2.0 (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 (Supplementary 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 Supplementary Table S3 for methods). The new (Eugene

and LoReAn) and previous (JGI, MPI, RRES, and CURTIN) gene models were analyzed with *ingenannot filter* to filter out TE-encoding genes.

Filtered gene models were analyzed with *ingenannot aed* to provide AED (Eilbeck et al. 2009) scores for each gene model compared with 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 (Supplementary Fig. S1). AED scores were calculated with "--aed_tr_cds_only" to avoid bias between datasets with or without untranslated region (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 indicative of good annotations, whereas values of 0.3 or below are classified as high-quality annotations (Holt and Yandell 2011; Hunt et al. 2020). 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 gene structure accurately 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 four independent ab initio gene prediction software programs, were also retained. However, all the gene models without an ATG or stop codon were removed. The relatively high number of annotation sources (six) and the selection of loci detected by four independent gene predictors allowed us to use stringent AED thresholds, leading to well-supported gene structures (see Supplementary Fig. 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 three frames to analyze the predicted peptides with the same criteria as those used to detect small secreted proteins (SSPs) 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 *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 *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 Supplementary Table S5 for details and comments).

Functional annotation of reannotated gene models (RGMs)

RGM protein sequences were analyzed 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 (v.4.1) (Nielsen 2017), and TargetP (v.1.1b) (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 (v.2.0).

Analysis of Iso-Seq transcript isoforms, antisense transcripts, and lncRNAs

The annotation of transcript isoforms was performed with sqanti3 (Tardaguila et al. 2018) using Iso-Seq transcripts (see Supplementary Table S8 for methods). Iso-Seq transcripts annotated as antisense and intergenic with sqanti3 were selected as lncRNAs and further filtered (see Supplementary Table S9 for methods; Supplementary 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 reliable. Detection of overlaps between transcripts and annotations was performed with intersect from BEDTools (Quinlan and Hall 2010).

AED as a metric for comparing gene models predicted by different tools

InGenAnnot RGMs were compared with gene models obtained with either funannotate (Palmer and Stajich 2025), Helixer (Holst et al. 2023; Stiehler et al. 2020), or BRAKER3 (Gabriel et al. 2024). Gene models obtained with these three tools were scored with AED using the same evidence as for RGMs, and their AED scores were plotted for both transcript and protein evidence (Supplementary Figs. 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* (Supplementary Fig. 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 CDSs. The comparison of the different gene models occurring at each "locus" highlighted three categories of metagenes: (i) identical gene models (same CDS), (ii) dissimilar gene models (same metagene but different CDS), and (iii) specific gene models (predicted by a single gene predictor). Only 3,618 identical gene models were shared across the four annotations (27%; Fig. 1). The MPI, RRES, and CURTIN annotations share more identical gene models, reaching 6,816 (51%; Fig. 1). The JGI and CURTIN annotations displayed the highest numbers of dissimilar gene models (4,752 and 3,844, respectively) compared with RRES and MPI (2,367 and 1,871, respectively; Fig. 1). On the other side, the RRES, CURTIN,

and JGI annotations displayed higher numbers of specific gene models (593, 436, and 151, respectively; Fig. 1) compared with 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 (Supplementary 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 (Supplementary Table S2). In addition, the average sizes of gene models differed between MPI (1,465 bp) and the other annotations (approximately 1,300 bp). This difference could result from the occurrence of incorrect 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 with 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 or new transcript sequences obtained from long-read PacBio sequencing (Iso-Seq) and short-read, single-stranded RNA-Seq (Supplementary 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 (Supplementary 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 transcriptional starting site/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%), whereas other loci had either two to five isoforms (42%) or at least six isoforms (8%). Transcripts from each single-stranded RNA-Seq library were assembled separately, and those with weak expression levels (transcripts per million reads < 1) were removed (Supplementary 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 for 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, and CURTIN). All these gene models were clustered into 17,147 metagenes.

InGenAnnot computed an 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

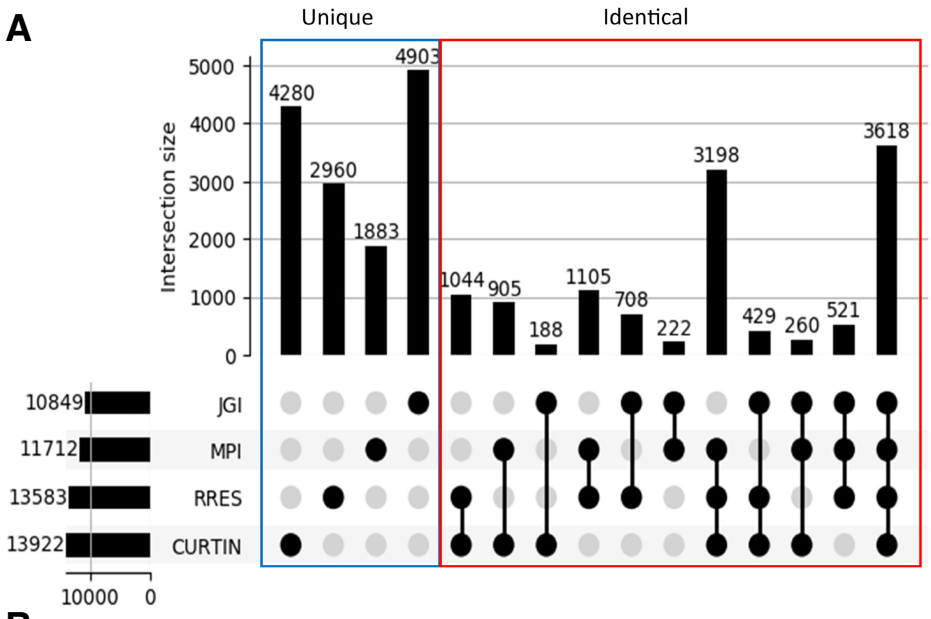


Fig. 1. Comparison of *Zymoseptoria tritici* reference isolate IPO323 genome annotations. **A,** UpSet plot of the gene models from the four annotations of IPO323 (JGI, MPI, RRES, and CURTIN). Numbers of gene models with identical coding sequences (CDSs). **B,** Comparison of IPO323 gene annotations. Number of CDSs 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), whereas 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).

B

		JGI	MPI	RRES	CURTIN
Identical CDS	JGI	*10849*			
	MPI	4621	*11712*		
	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

(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 piece of 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 (Fig. 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 Fig. 2 corresponding to 1,846 gene models). These rescued genes models were mostly not conserved across fungi and had low transcriptional support (Fig. 2). For CDSs 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 RGMs (Supplementary File S1; Supplementary 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 length of 5′ UTRs was 315 bp, whereas it was 389 bp for 3′ UTRs (Supplementary Table S4), similar to what was reported for the fungus *Podospora anserina* (5′ UTR 275 bp, 3′ UTR 303 bp) (Lelandais et al. 2022). A small proportion of genes displayed

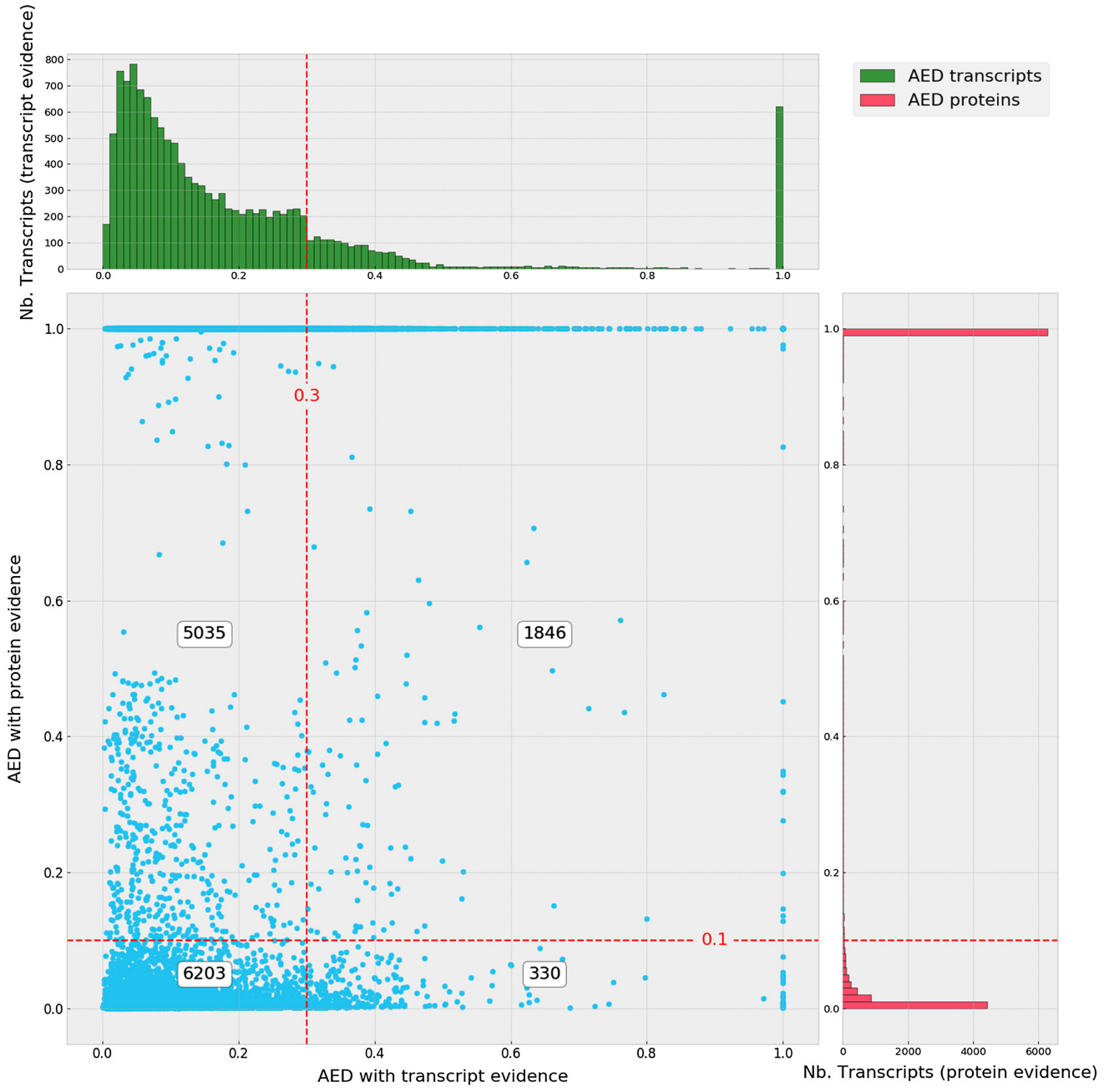


Fig. 2. Selection of the best reannotated gene models (RGMs) according to their annotation edit distance (AED) scores. Plot of RGM AED scores. AED scores (0 to 1) describe how a given gene model fits to transcript and protein evidence (best fit = 0). Transcript evidence was computed from RNA-Seq or Iso-Seq 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 (1,846 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).

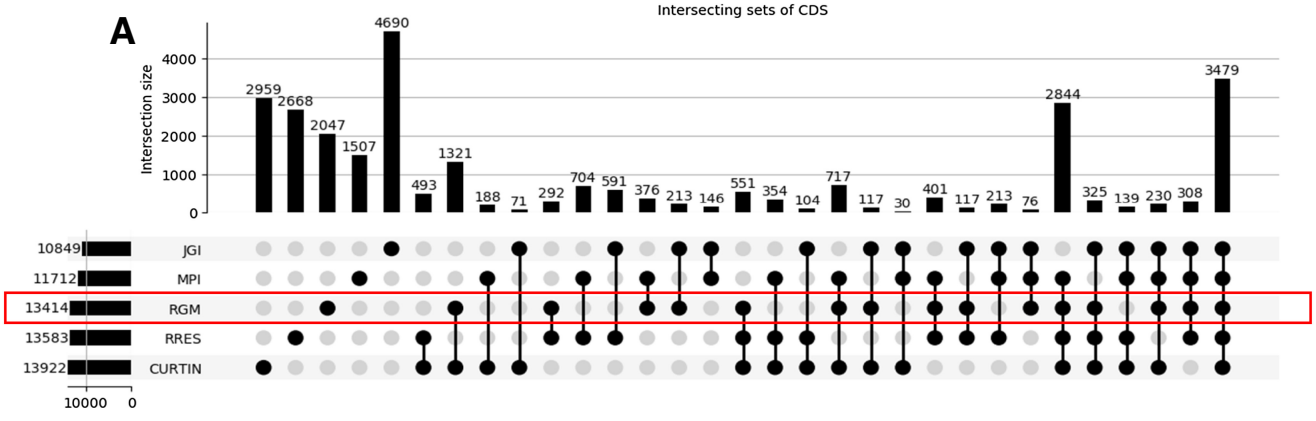
long 5′ UTRs (1,000 to 7,000 bp, 6%) and/or long 3′ UTRs (1,000 to 8,600 bp, 8.6%).

Comparison of the reannotated IPO323 gene models with available genome annotations

The 13,414 IPO323 RGMs were compared with previous gene models (JGI, MPI, RRES, and CURTIN) using BUSCO and the ascomycota_odb as references. High BUSCO scores were obtained with the RGM, RRES, MPI, and CURTIN annotations (98.4 to 99.4%; Supplementary Table S5), whereas the score obtained with the JGI annotation was lower (95.7%), likely due to a high number of fragmented and missing BUSCO genes (Supplementary Table S5). The comparison between annotations was then performed using AED scores (Fig. 2; Supplementary Figs. 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 (Fig. 2). CURTIN had a high level of support (10,716 gene models; Supplementary Fig. S3), followed by RRES (9518 gene models) and MPI (8936 gene models), whereas 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 Supplementary 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 (Supplementary Table S6). Almost none of these RGMs were 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 artifact from annotation pipelines. The other fully ab initio RGMs were enriched in genes localized on accessory chromosomes (Supplementary Table S6).

Among the 13,414 RGMs, 7,888 were identical to at least one gene model from another annotation (Fig. 3), and 3,479 RGMs were identical to all the gene models from the four previous annotations (Fig. 3). Because 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 (ORF) 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). Two thousand and forty-seven RGMs were either different from (1,376 modified RGMs; Supplementary Table S6) or not predicted by (671, specific RGMs; Supplementary Table S6) previous annotations. 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 (Supplementary Table S6). Of these specific RGMs, 117 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 (Supplementary 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; Supplementary Table S7) and in a lower number in the RRES annotation (701), whereas 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, because their AED scores were higher than the cutoff value (>0.3; Supplementary Fig. S5). Even though most transcript AED scores of split RGMs (65%) were lower than the cutoff value (<0.3; Supplementary Fig. S5), a significant number of split RGMs (494, 35%) had low support from both transcript and protein evidence (Supplementary Fig. S5). These split RGMs were rescued because they were also identified in annotations other than MPI. The transcript evidence of two randomly chosen MPI fused genes and their corresponding split RGMs are shown in Supplementary Figures S6 and S7. Both MPI fused genes had no Iso-Seq transcript support, whereas Iso-Seq transcripts supported the split RGMs. Assembled RNA-Seq transcripts supporting split RGMs were also observed for RGM-1 and RGM-2 from Supplementary Figure S6. However, large



B

	JGI	MPI	RRES	CURTIN	Any
RGM	4865	8431	8317	9584	11367

Fig. 3. Comparison of the novel IPO323 genome annotation (reannotated gene models, RGMs) 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 (CDSs). **B,** Numbers of identical CDSs between RGMs and each available annotation.

assembled RNA-Seq transcripts supporting fused MPI genes were observed (Fig. 4). We hypothesized that these long transcripts were artifacts 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 postinoculation, wheat infection; Supplementary Fig. S5) in which RGM-2, but not RGM-1, was strongly expressed.

Functional annotation of reannotated IPO323 gene models

Functional annotation of RGM proteins was performed using Blast2Go and InterProScan. Of the RGMs, 5,593 exhibited a Gene Ontology term or an InterPro annotation, and 2,838 were annotated with at least one Enzyme Code. Several tools (Grandaubert et al. 2015; Morais do Amaral et al. 2012) were used to identify 1,895 RGMs encoding putative secreted proteins, including effectors (Supplementary File S1). 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 SSPs 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 CDSs inferred from transcripts not associated with a gene model. This strategy allowed for the rescue of genes encoding SSPs that were difficult to predict by ab initio gene prediction software. Four of these 43 novel SSPs (ZtIPO323_001210, ZtIPO323_072700, ZtIPO323_105940, and ZtIPO323_123970) displayed a significant upregulation during infection compared with 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 (Supplementary Fig. S8b). Two additional *Avr-Stb6* paralogs located on chromosome 10 were also predicted as RGM-specific SSPs (Supplementary Fig. 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). Of these RGMs, 7,872 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). Two thousand seven hundred and sixteen Iso-Seq transcripts identified as alternative splice variants (25% of coding transcripts) were classified

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

Category	Count
Full-splice_match (FSM)[a]	7,872
Incomplete-splice_match (ISM)[b]	305
Fusion	45
Genic[c]	664
Intron retention (IR)	1,571
novel_in_catalog (NIC)[d]	7
novel_not_in_catalog (NNC)[e]	474
Antisense	395
Intergenic	357

[a] Whole transcripts with possible alternative 3′ and 5′ ends.
[b] Partial overlaps of transcripts fitting with intron coordinates.
[c] Partial overlaps of introns and exons not compliant with intron/exon coordinates.
[d] Use combination_of_known_splice sites.
[e] At_least_one_novel_splice site detected.

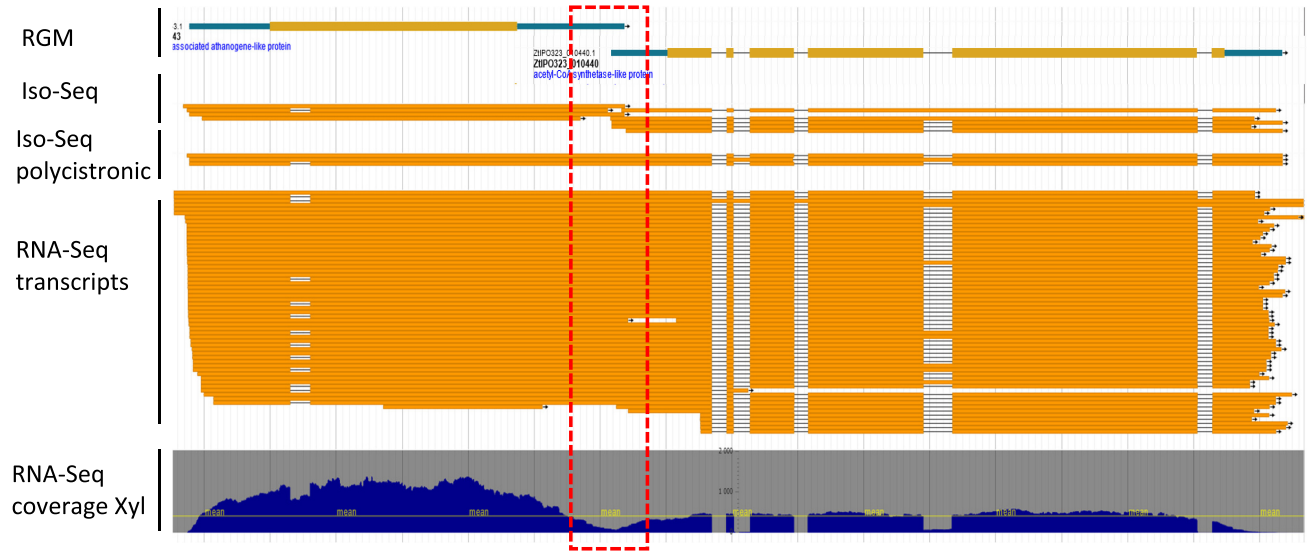


Fig. 4. Single gene and polycistronic transcripts shown for reannotated gene models (RGMs) ZtIPO323_010430 and ZtIPO323_010440. RGMs ZtIPO323_010430 and ZtIPO323_010440, located at chr_1: 2692944...2694593 and chr_1: 2694544...2697087, respectively, were transcribed on the same strand with overlapping 3′ untranslated region (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.

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 IR events (IR, 75%). Transcripts with a premature termination codon 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 (Supplementary 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) (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 (Fig. 5A). The 1,753 remaining transcript isoforms with IR 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 (Fig. 5B). Its main transcript (Iso-Seq 2), corresponding to the selected RGM, had four splicing sites, one being in the 5′ UTR. Two alternative Iso-Seq transcripts (Iso-Seq 1 and 2) with one or two IR 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 (Fig. 5A) or ZtIPO323_013090 (Supplementary Fig. S9). These examples illustrate the difficulty in choosing between gene models with complex alternative splicing events, leading to transcript isoforms with similar expression levels (Fig. 5A). 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 ($P < 0.01$) between galactose/sucrose and mannose/xylose culture conditions (Supplementary File S1). A total of 163 RGMs displayed a significant DIU between infection and culture conditions (Supplementary File S1), including 23 secreted proteins. Some of these RGMs were highly up- or downregulated during infection, such as ZtIPO323_042160 (unknown), ZtIPO323_042360 (unknown), ZtIPO323_043800 (PHD/RING finger protein), and two secreted proteins (ZtIPO323_016670 and ZtIPO323_043500) that were significantly upregulated during late infection (13 and 21 days postinoculation). ZtIPO323_016670 encoded a carbohydrate esterase from family 8 involved in cell wall modifications, and ZtIPO323_043500 encoded an 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 with in vitro culture conditions, leading to a bias in DIU analyses.

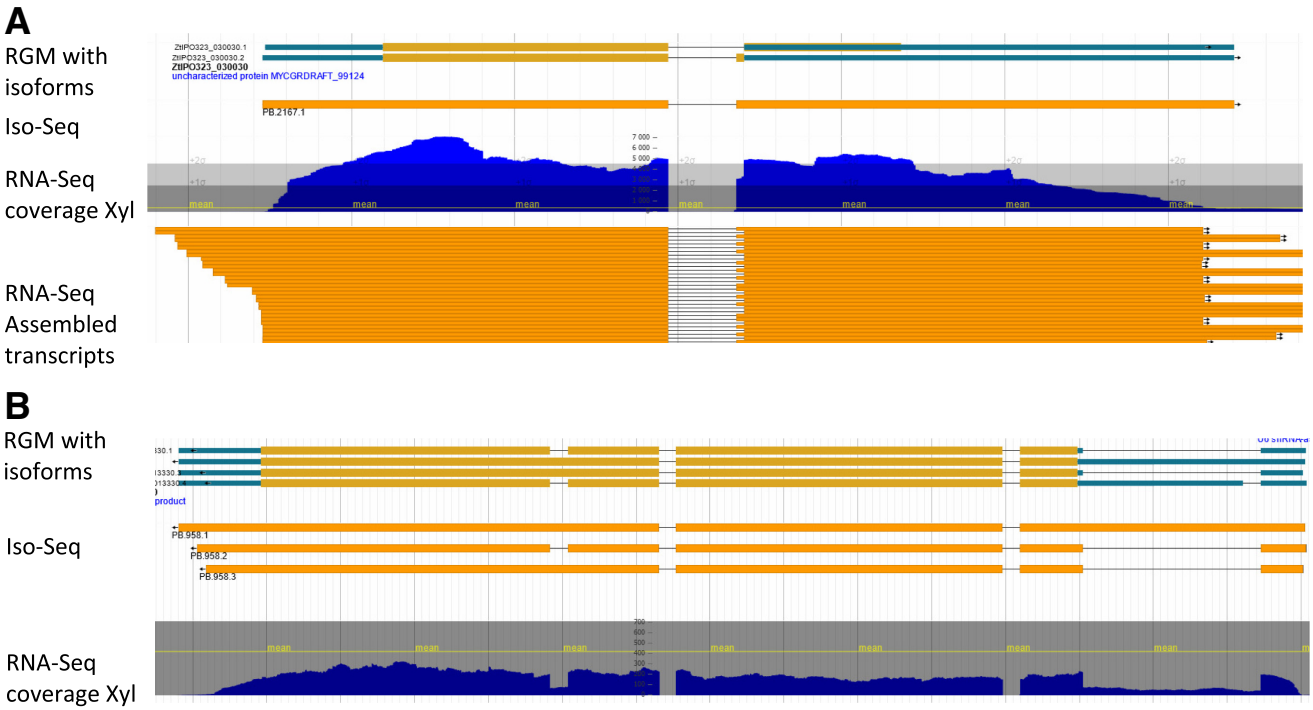


Fig. 5. Transcript isoforms of reannotated gene models (RGMs) **A,** ZtIPO323_030030 and **B,** ZtIPO323_013330 supported by Iso-Seq and RNA-Seq evidence. A, Gene ZtIPO323_030030 (chr2: 1777934...1778652, 0.718 Kb). This RGM has two transcript isoforms (alternative 3′ acceptor site). Both encoded small secreted proteins (SSPs, 10; Supplementary File S1). Previous annotations selected the second acceptor site leading to the longest coding sequence (CDS). A single Iso-Seq transcript corresponding to the longest CDS was detected (Iso-Seq track), whereas 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 with 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: 3420265...3424017, 3.752 Kb). This RGM had four transcript isoforms. The selected RGM had four splicing sites, one of which in the 5′ untranslated region (UTR) was supported by Iso-Seq transcript (Iso-Seq no. 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 nos. 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 no. 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.

Identification of lncRNAs

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 *Fusarium graminearum* (Lu et al. 2022), 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 aa). 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 three were labeled as "coding" based on their coding potential and one 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 ($P = 0.05$).

Half of these lncRNAs were differentially expressed between infection and in vitro culture, including 5 that were upregulated and 12 that were downregulated during infection ($\log_2$FC > 2). Most lncRNAs that were downregulated during infection were antisense transcripts (83%). The lncRNA PB1188.1 that was downregulated during infection compared with all culture conditions (Supplementary Table S9) was an antisense transcript of ZtIPO323_016330, encoding a secreted subtilisin-like protein. ZtIPO323_016330 was upregulated during infection and downregulated 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 (Supplementary Table S9), as lncRNA PB.2709.1 was upregulated during infection, whereas ZtIPO323_037670 was downregulated.

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. 2022) or *Cordyceps militaris* (Chen et al. 2019). Overall, 224 putative polycistronic transcripts contained two to 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 (Fig. 4). Iso-Seq polycistronic single-transcript molecules covering these two RGMs were detected, as well as single-RGM Iso-Seq transcripts (Fig. 4). Assembled RNA-Seq reads supported a transcript covering the two RGMs (Fig. 4). However, RNA-Seq coverage strongly decreased in the overlap between the two RGMs, suggesting two independent transcripts (Fig. 4). RNA-Seq coverage showed that the abundance of the polycistronic transcript was low compared with 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 (1,919 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 (*Z. 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. The 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 3,091 nucleotides encoding a protein of 986 amino acids corresponding to an RNA-dependent RNA polymerase. ZtNV1 was as long as SsNV4 (3,105 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* (Supplementary Fig. S10). The 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 with ZtNV1 (70,000-fold).

Comparison of InGenAnnot with other gene prediction tools (funannotate, BRAKER3, and Helixer) using AED

Two integrated gene prediction tools (funannotate and 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, and CodingQuarry) and uses EVidenceModeler to select the best gene model (Palmer and Stajich 2025). BRAKER3 integrates two ab initio tools (Augustus and 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. 2020). 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 (Supplementary Figs. S11, S12, and S13). Comparison of these gene models to RGMs highlighted 6,389 identical CDSs predicted by the four tools (47% of RGMs; Supplementary Fig. S14). The number of identical CDSs predicted by funannotate, BRAKER3, and InGenAnnot (RGMs) was higher (8,220 CDSs, 61% of RGMs; Supplementary Fig. S14). Individually, Helixer displayed the lowest number of gene models similar to RGMs (8,086, 60% of RGMs; Supplementary Fig. S14). A higher number of funannotate and BRAKER3 gene models were similar to RGMs (67 and 73% of RGMs for funannotate and BRAKER3, respectively; Supplementary Fig. S14). Helixer was the only tool to predict a large number of unique CDSs (6,190), including 4,103 CDSs 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 (Supplementary Fig. S15). This could be explained by the strong weight assigned to transcriptomic data to obtain the InGenAnnot and BRAKER3 gene models. Indeed, because 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 (Supplementary 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; Supplementary Table S10) compared with 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; Supplementary Table S10), closely followed by InGenAnnot (median 0.398), whereas 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, because it predicted only 12,293 genes, compared with InGenAnnot (13,414 genes) and funannotate (13,423 genes). This suggested that BRAKER3 selected mostly gene models with evidence, whereas funannotate and InGenAnnot allowed for 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 artifacts 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 bypass these artifacts, as they produce 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 because it is not quantitative. Indeed, we identified rare, long Iso-Seq transcripts likely corresponding to IR 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 programs 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 programs 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 programs 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 for 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 (Fig. 2). They typically included candidate fungal effectors that could be important for plant–fungal interactions (Supplementary File S1). Yet, some rescued RGMs could be artifacts 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; Supplementary Table S6) or not predicted in previous annotations (671 RGM-specific; Supplementary 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 (Supplementary Figs. S6 and 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 artifact, 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. 2022). 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 for 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 (Figs. 4 and 5; Supplementary Fig. 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 (Jeon et al. 2022; Lu et al. 2022), most alternative splicing events were IR events (Table 1). IR events could generate premature termination codons 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

Table 2. Contribution of each annotation of the *Zymoseptoria tritici* IPO323 genome to reannotated gene models (RGMs)

		Identical CDSs[b]			Unique identical CDSs[c]	
Type	**Annotation[a]**	**Per annotation**	**Percent**	**Combined**	**Per annotation**	**Combined**
Available annotations	JGI (FGENESH/Genewise[d])	4,865	48%	11,367	157	929
	MPI (EVidenceModeler[d])	8,431	62%		91	
	RRES (MAKER-HMM[d])	8,317	62%		175	
	CURTIN (CodingQuarry[d])	9,584	71%		506	
New annotations	Eugene[d]	10,224	76%	11,677	1,603	1,802
	LoReAn[d]	7,769	58%		199	

[a] The annotations that contributed the most to RGMs were, respectively, Eugene (76% identical CDSs*, 1,603 unique identical CDSs**) and Curtin (71% identical CDSs, 506 unique identical CDSs). Combining gene models from the four available annotations (JGI, MPI, RRES, and CURTIN) showed that 11,367 of their CDSs were identical to RGMs (contribution: 84.7%). Combining gene models from the two new annotations (Eugene and LoRean) showed that 11,677 of their CDSs were identical to RGMs (contribution: 87%). The combination of the six annotations was needed to predict all the 13,414 RGMs.
[b] Identical coding sequences (CDSs): number of CDSs identical to RGMs.
[c] Unique identical coding sequence (CDSs): number of CDSs predicted in a single annotation and retained as RGMs.
[d] Ab initio gene prediction software used for the given annotation.

transcript isoforms during infection compared with 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 Agaricomycotina (Gordon et al. 2015), *F. graminearum* (Lu et al. 2022), and *C. militaris* (Chen et al. 2019). However, polycistronic-specific RNA-Seq reads were always detected in low abundance compared with 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 upregulate or downregulate 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 artifacts from assembly. Iso-Seq bypassed this problem and facilitated the identification of full-length, non-chimeric lncRNAs. Using stringent criteria (size > 1,000 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 *Neurospora crassa* [Arthanari et al. 2014], 352 in *Verticillium dahliae* [Li et al. 2022], and 427 to 819 in *F. graminearum* [Lu et al. 2022]). 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. 2022), which we did not survey in our RNA samples. We identified 17 lncRNAs as differentially expressed during plant infection, mostly as antisense transcripts (Supplementary Table S9). Two displayed expression patterns opposed to their coding genes. lncRNA PB1188.1 was downregulated during infection compared with in vitro culture. This lncRNA is an antisense transcript of ZtIPO323_016330 encoding a secreted subtilisin-like protein that is upregulated 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 downregulation of lncRNA PB1188.1 during infection allowed for the full expression of ZtIPO323_016330 in infected leaves. The second lncRNA (lncRNA PB.2709.1) was upregulated during infection compared with in vitro culture (Supplementary Table S8), whereas its corresponding transcript (ZtIPO323_037670) was downregulated during infection. This transcript encodes a 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 (Supplementary Fig. S10), related to the narnavirus 4 of *S. 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, whereas Zt-NV1 was only detected in low abundance compared with 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 (Dubarry et al. 2016; Holt and Yandell 2011; Lukashin and Borodovsky 1998; Min et al. 2017; Reid et al. 2014; Sallet et al. 2019; Stanke et al. 2006; Testa et al. 2015) and protocols (Campbell et al. 2014) have been established to predict gene models in eukaryotic genomes. Some were dedicated to fungal genome annotation (Haas et al. 2011; Reid et al. 2014; Testa et al. 2015) 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 the 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 artifacts such as wrongly assembled RNA-Seq transcripts or rare Iso-Seq transcripts (see before). As each ab initio gene prediction software program implements specific machine learning models with different specificity/sensibility for each data source, their implementation and training parameters are more or less tolerant to particularities such as short CDSs or non-canonical splicing sites. The combination of different ab initio gene prediction software programs with distinct intrinsic characteristics has proved essential to avoid drawbacks from each software. Indeed, none of the ab initio gene prediction software programs used individually was able to predict more than 70 to 76% of the final gene models (Table 2).

Tools other 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 metrics other 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. 2020). Funannotate, BRAKER3, and Helixer were run to annotate the *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%; Supplementary Fig. 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 CDSs; Supplementary Fig. 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 programs (funannotate, BRAKER3), was 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 programs 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, because it identified more gene models with evidence than funannotate or BRAKER3 (Supplementary Fig. S14). However, the accurate comparison of the InGenAnnot AED-based selection with EVidenceModeler (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 study, 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 gene annotation drastically by selecting the best gene model according to transcript and protein evidence across gene models predicted by different software programs. 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 programs 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 with the submission. The sequence of the new mycovirus ZtNV1 was deposited to NCBI under accession OP903463. Previous *Z. tritici* IP0323 gene annotations, new annotations (RGMs, isoforms, and lncRNAs), and the annotation file, denoted Supplementary 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 website (https://mycocosm.jgi.doe.gov/Zymtr1/Zymtr1.home.html) was released with new genome annotations.

The InGenAnnot code and project are available at https://forgemia.inra.fr/bioger/ingenannot licensed under GNU GPL v.3. InGenAnnot documentation is available at https://bioger.pages.mia.inra.fr/ingenannot.

Acknowledgments

We thank the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing help, computing, and/or storage resources and the BARIC workgroup (https://www.cesgo.org/catibaric/) for providing storage and computational resources.

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. 21:30.

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 9:43.

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:e201900429.

Arthanari, Y., Heintzen, C., Griffiths-Jones, S., and Crosthwaite, S. K. 2014. Natural antisense transcripts and long non-coding RNA in *Neurospora crassa*. 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. 18:12.

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 Bioinform. 3::lqaa108.

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

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. 2014. Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinform. 48:4.11.1-4.11.39.

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:e01635-21.

Cemel, I. A., Ha, N., Schermann, G., Yonekawa, S., and Brunner, M. 2017. The coding and noncoding transcriptome of *Neurospora crassa*. BMC Genomics 18:978.

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. 21:24.

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. 9:898.

Chiba, Y., Yoshizaki, K., Tian, T., Miyazaki, K., Martin, D., Genomics and Computational Biology Core, Saito, K., Yamada, A., and Fukumoto, S. 2022. Integration of single-cell RNA- and CAGE-seq reveals tooth-enriched genes. J. Dent. Res. 101:542-550.

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.

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 15:1132.

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 18:340.

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. https://doi.org/10.7490/f1000research.1111735.1

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

Ejigu, G. F., and Jung, J. 2020. Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology 9:295.

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 21:588.

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 Bioinform. 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. 19:55.

Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32:D258-D261.

Gerads, M., and Ernst, J. F. 1998. Overlapping coding regions and transcriptional units of two essential chromosomal genes (*CCT8*, *TRP1*) in the fungal pathogen *Candida albicans*. Nucleic Acids Res. 26:5061-5066.

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:e0219207.

Goodwin, S. B., M'Barek, S. B., 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:e1002070.

Gohari, A. M., 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.

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. 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. 36:3420-3435.

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 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., Saida, N., and Shalgi, R. 2022. Stress-induced transcriptional readthrough into neighboring genes is linked to intron retention. iScience 25:105543.

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. 17:3066-3077.

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 527280.

Holt, C., and Yandell, M. 2011. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinform. 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: 624.

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. 19:373-386.

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:veab032.

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 30:1236-1240.

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 3:1088-1100.

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. 10:68.

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. 257:126962.

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. 12:723636.

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., Möller, 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 11:jkab068.

Lu, P., Chen, D., Qi, Z., Wang, H., Chen, Y., Wang, Q., Jiang, C., Xu, J.-R., and Liu, H. 2022. Landscape and regulation of alternative splicing and alternative polyadenylation in a plant pathogenic fungus. New Phytol. 235:674-689.

Lukashin, A. V., and Borodovsky, M. 1998. GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res. 26:1107-1115.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. 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. 38:4647-4654.

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.

Möller, 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*. 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: Protein Function Prediction: Methods and Protocols. D. Kihara, ed. Humana, New York, U.S.A.

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: e69249.

Palmer, J. M., and Stajich, J. 2025. Funannotate v1.8.17. https://github.com/nextgenusfs/funannotate

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. R. Oliver, ed. Burleigh Dodds Science Publishing, London, U.K.

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 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. K., 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 Bioinform. 15:229.

Sallet, E., Gouzy, J., and Schiex, T. 2019. EuGene: An automated integrative gene finder for eukaryotes and prokaryotes. Pages 97-120 in: Gene Prediction: Methods and Protocols: Methods in Molecular Biology, vol. 1962. M. Kollmar, ed. Humana, New York, NY, U.S.A.

Salzberg, S. L. 2019. Next-generation genome annotation: We still struggle to get it right. Genome Biol. 20:92.

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 Chromatin 8:41.

Standage, D. S., and Brendel, V. P. 2012. ParsEval: Parallel comparison and analysis of gene structure annotations. BMC Bioinform. 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. 2020. 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.