BMC Genomics

RESEARCH

Open Access

New insights into genome annotation in *Podospora anserina* through re-exploiting multiple RNA-seq data

Gaëlle Lelandais[†], Damien Remy[†], Fabienne Malagnac and Pierre Grognet[*]

Abstract

Background: Publicly available RNA-seq datasets are often underused although being helpful to improve functional annotation of eukaryotic genomes. This is especially true for filamentous fungi genomes which structure differs from most well annotated yeast genomes. *Podospora anserina* is a filamentous fungal model, which genome has been sequenced and annotated in 2008. Still, the current annotation lacks information about cis-regulatory elements, including promoters, transcription starting sites and terminators, which are instrumental to integrate epigenomic features into global gene regulation strategies.

Results: Here we took advantage of 37 RNA-seq experiments that were obtained in contrasted developmental and physiological conditions, to complete the functional annotation of *P. anserina* genome. Out of the 10,800 previously annotated genes, 5'UTR and 3'UTR were defined for 7554, among which, 3328 showed differential transcriptional signal starts and/or transcriptional end sites. In addition, alternative splicing events were detected for 2350 genes, mostly due alternative 3'splice sites and 1732 novel transcriptionally active regions (nTARs) in unannotated regions were identified.

Conclusions: Our study provides a comprehensive genome-wide functional annotation of *P. anserina* genome, including chromatin features, cis-acting elements such as UTRs, alternative splicing events and transcription of non-coding regions. These new findings will likely improve our understanding of gene regulation strategies in compact genomes, such as those of filamentous fungi. Characterization of alternative transcripts and nTARs paves the way to the discovery of putative new genes, alternative peptides or regulatory non-coding RNAs.

Keywords: Transcriptome, RNA-seq, Fungal genome, Functional annotation, Alternative splicing, nTARs

Introduction

If coding sequences define protein primary structures, messenger RNAs (mRNAs) direct their cytoplasmic expression. From pre-mRNA processing to translation initiation, their untranslated regions (UTRs) control most of the post-transcriptional gene regulation aspects, including nucleo-cytoplasmic transport, subcellular localization, mRNA stability and translation efficiency [1, 2]. To initiate gene expression at transcriptional start sites (TSS), transcriptional factors, histone chaperones [3] and chromatin remodelers [4] bind to cis-acting DNA sequences known as core-promoter, to recruit the RNA polymerase II complex. Conversely, transcription termination at specific transcription end sites (TES) prevent read-through transcription into adjacent genes, an acute concern in fungal compact genomes [5]. Both 5'UTR and 3'UTR present a variety of canonical cis-acting elements that are bound by

[†]Gaëlle Lelandais and Damien Remy contributed equally to this work.

*Correspondence: pierre.grognet@universite-paris-saclay.fr

Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198 Gif-sur-Yvette, France

trans-acting elements [6, 7]. In addition, upstream ORF present in the 5'UTR are key regulators of translation [8]. This combinatory repertoire tunes the composition of proteome (entire set of proteins) in accordance with developmental and/or metabolic needs of the cell. Several evidence suggests that the UTRs may harbor mutations that drives human traits and diseases [9], including cancer pathogenesis [10].

At a given core-promoter, the transcription may start from one of several TSS. Extensive studies performed on various human tissues established high-resolution transcription start sites maps [11]. In animals, compilation of TSS localizations relative to gene expression identified two categories of core-promoters (reviewed in [12]). Core-promoters that show sharp initiation patterns, i.e. one main TSS, are found active in adult tissue-specific genes or terminally differentiated cell-specific genes, whereas core-promoters that show dispersed initiation patterns, i.e. multiple equally used TSS, are found active either for broadly expressed housekeeping genes or for developmental genes. In unicellular eukaryotes multiple or alternative TSS are often used to cope with changing environmental conditions. In the budding yeast, in vivo translation activities of alternative 5'UTR isoforms can vary by more than 100-fold [13].

In eukaryotes, chromatin accessibility is also a way to regulate gene expression. Heterochromatin is less prone to transcription than euchromatin. To combine genome-scale functional information coming from both cis-acting elements (i.e. enhancers, promoters, TSS and TES) and histone modification patterns, schematic representations of model genes emerged for animals [14], plants [15] and some yeast species [16]. Although a fairly large number of complete annotated fungal genome sequences is available [17], no such gene model has been built to date for filamentous fungi. Still, a recent assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) performed in *Neurospora crassa* highlights the diversity of promoter structures and evidenced that histone acetylation and small RNA production are correlated with accessible chromatin, whereas some histone methylations are correlated with inaccessible chromatin [18].

Alternative splicing (AS) also regulates gene expression of eukaryotes. In animals, AS allows the generation of tissue- and time-specific isoforms, especially in brains. In *Drosophila*, the *Dscam* gene can generate over 38,000 distinct mRNA isoforms [19], which is more transcripts than the total number of genes in this organism (~14,500). Notably, AS frequency is far less frequent in fungi than in animals, ranging from less than 1% in the budding yeast to 18% in the human pathogen *Cryptococcus neoformans* [20]. Due to genomic features (few and short introns), intron retention (IR) is the most prevalent splicing type found in fungi (reviewed in [21]). However, studies performed in non-yeast fungi are limited.

P. anserina is a coprophilous ascomycete fungus that has been used as a model organism for almost a century [22]. Its genome has been sequenced multiple times and watchfully annotated [23–25]. However, no integrative genome-wide transcriptional landscape of *P. anserina* has been published yet. To do so, we took advantage of large and diversified sets of transcriptomic data and developed a customized annotation pipeline to map the 5′ and 3'UTRs genome-wide. We also evidenced the existence of alternative 5'and 3′ UTRs and described distinct types of alternative splicing events. Finally, novel transcriptionally active regions (nTARs) were searched and annotated, on which functional domain predictions were conducted to discover several putative new genes. We finally build a gene model that integrate the canonical *P. anserina* transcriptional features and the epigenomic landscape [26] in relation with gene expression status.

Results

Collection of multiple RNA-seq data from various experimental conditions

A search for *P. anserina* in the SRA and BioProject repositories [27, 28] returned 44 RNA-seq data from different studies on *P. anserina*'s life cycle [25], adaptation to carbon sources [29], response to bacteria [30] and senescence [31]. Because it was generated by the SuperSAGE technology, this last dataset was excluded from the analyses. This left us with 37 RNA-seq from three studies, referred to as datasets A, B and C (Table 1). Altogether, these data cover a large variety of developmental states and growth conditions, which is important to increase the rate of transcriptionally active genes one might observe. Out of 1,054,787,963 reads in the 37 fastq files, 82.19% were mapped to the reference genome for which 10,800 CDS were annotated [23, 24]. Only 13 genes had no read mapped and 126 genes had only between 1 and 10 aligned reads (Fig. 1). Reads from dataset A alone, covering the entire life cycle, covered more than 99.7% of annotated CDS (respectively 31, 101 and 96 genes were not mapped in dataset A, B and C). This pool of dataset is then an interesting starting point to infer the transcript characteristics in *P. anserina*.

Detection of TSS and TES for transcript related to already annotated CDS

Our first goal was to get a more accurate annotation of the *P. anserina*'s transcripts, related to the trustworthy annotated CDS in the genome. In this context, our rationale was to consider that more accurate prediction for TSS and TES positions can be obtained with a high coverage of reads along transcripts. Therefore, the

Table 1 Composition of the RNA-seq databank. 37 RNA-seq in total parted in 3 datasets were collected from public databases (SRA and BioProject, accession number provided). The *Cs* strain genome from dataset C only differs from the strain *S* at the *het-s* VI locus. Overall, the 37 RNA-seq used in the analysis represent 19 unique experimental conditions. The heterogeneity of this pool of dataset provides better chance to observe genes in their active expression state

Name	Number of dataset	SRA/BioProject identifiers	Strain	Sequencing technology	Reads library	Reads size	Growth conditions	Reference
A	6	ERR2224046 to ERR224051	S	NextSeq 500	Paired-end	42	Sexual developement	Silar P, et al. 2019 [25]
B	19	PRJNA442509 to PRJNA442527	S	Illumina Hiseq 2500	Single-end	101	Multiple carbon sources	Benocci T, et al. 2018 [29]
C	12	SRR3197700 to SRR3197711	Cs	Illumina Hiseq 1000	Paired-end	100	Response to bacteria	Lamacchia M. et al. 2016 [32]

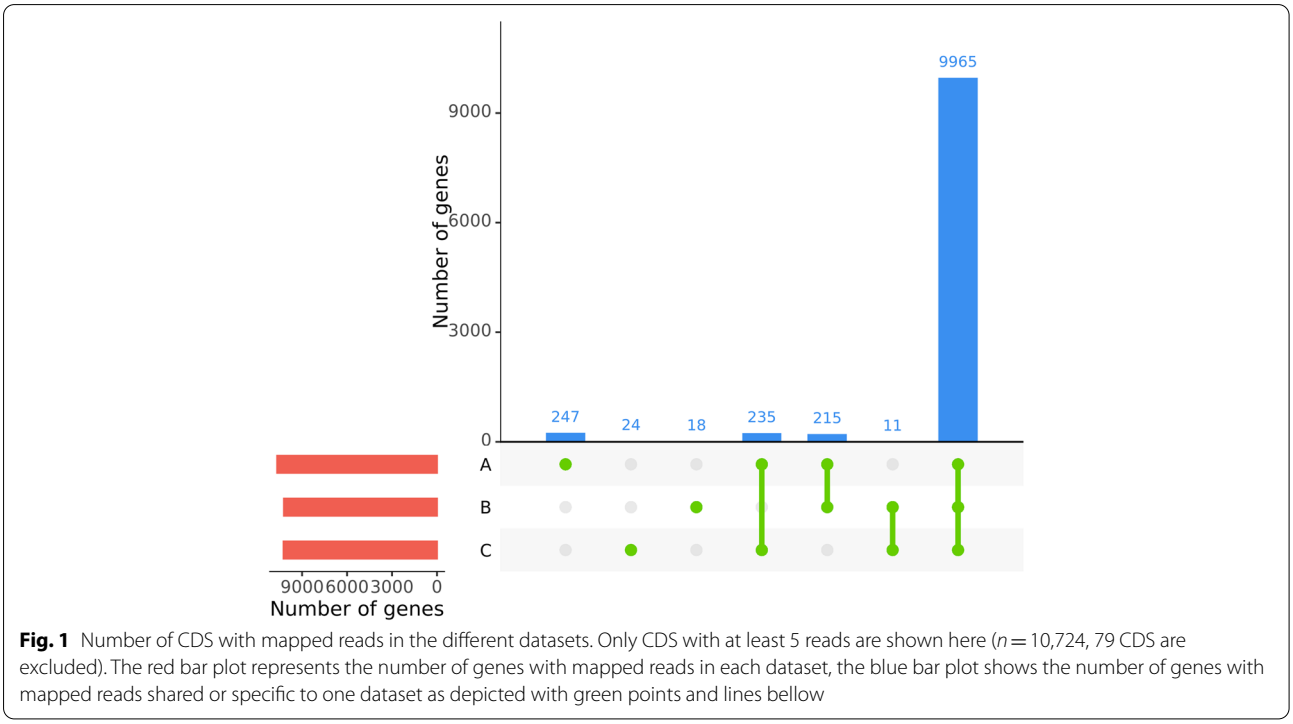

Fig. 1 Number of CDS with mapped reads in the different datasets. Only CDS with at least 5 reads are shown here ($n = 10{,}724$, 79 CDS are excluded). The red bar plot represents the number of genes with mapped reads in each dataset, the blue bar plot shows the number of genes with mapped reads shared or specific to one dataset as depicted with green points and lines bellow

doubtfulness of the prediction decreases as the coverage increases (Fig. 2A). With that in mind, we developed a strategy in which we selected the most reliable transcript annotation, according to the read coverage (Fig. 2B, C). For a given gene, the multiple transcript annotations obtained from the 37 samples were sorted according to the average coverage value. Only those above a given threshold were next selected. The process was repeated for each gene to select the most accurate annotations. Hence, by using this Successive Coverage Values (SCV) method, only the most reliable annotations from all datasets were conserved.

Applying this strategy on 10,803 predicted CDS of *P. anserina*, we could predict the transcript annotations for 7554 genes (69.9% of the all set of CDS) (Table S1). The other CDS, for which no transcript prediction could be assigned, had very low coverage of reads. Thanks to the already available CDS annotations, we could get insight into the 5′ and 3′ UTRs characteristics. Of note, while 4219 transcripts were predicted to have both a single TSS and TES, 3335 genes got multiple transcript annotations (Fig. 3A-B). Most of the variations originated from both TSS and TES positions (Fig. 3C). Note that each dataset contributes

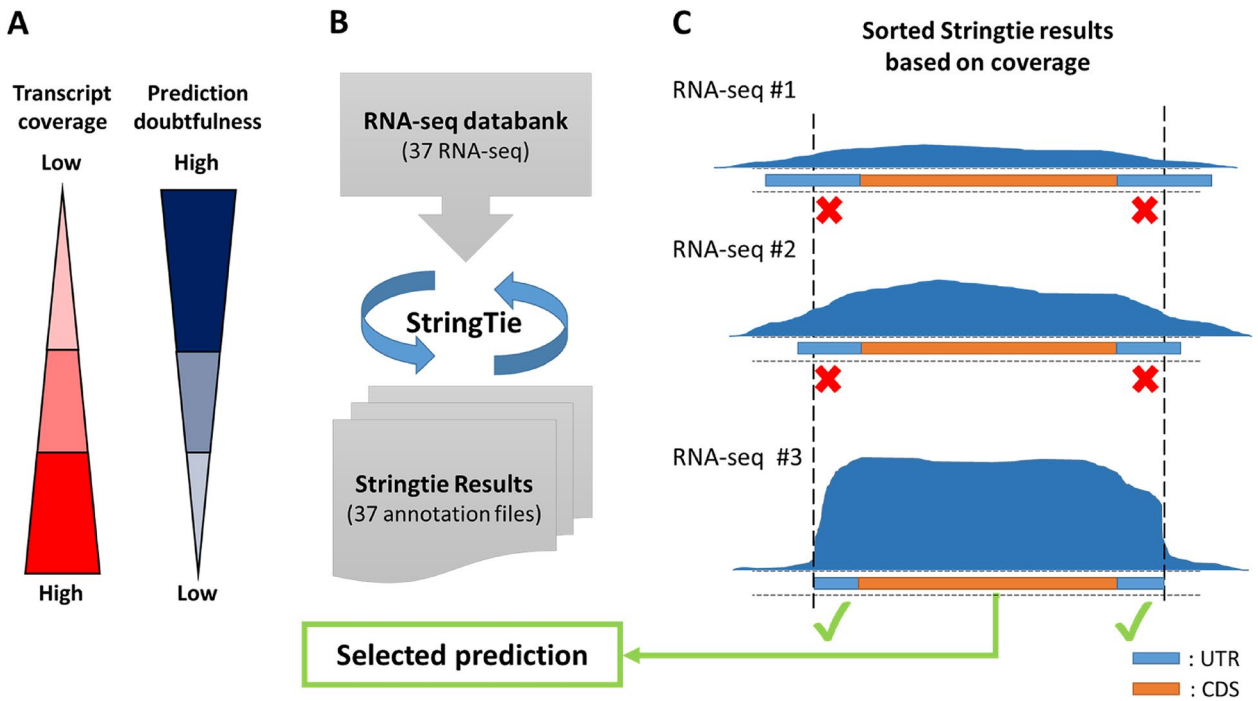

Fig. 2 Schematic description of the Successive Coverage Values (SCV) methodology. **a**) This strategy is based on the assumption that for a given gene, the highest the coverage value, the less doubtful is the transcript prediction. **b**) All RNA-seq alignments are processed by the StringTie tool providing transcriptome annotation files in output. **c**) All transcriptome annotation files are compared gene by gene. First, transcript predictions are validated if they fully cover their associated CDS. Then, the average coverage from each validated prediction are compared and those above the most restrictive value are finally selected

significantly to the global annotation (Fig. 3D), highlighting the importance to work with a diversity of conditions, to get broadest transcriptional landscapes. We compare our annotation with the output of StringTie Merge and found our results more accurate as, in our case, StringTie Merge tends to fuse transcripts of closely located genes (Fig. S1).

The average sizes for the 5′ and 3′ UTRs were 275 bp and 303 bp respectively. When genes with single or multiple UTR are considered separately, the average size of UTRs does not extensively vary (Fig. 4A, Table 2). Indeed, we observed the most distant multiple TSS and TES are spaced with 156 bp and 114 bp in average respectively (Fig. 4B, Table 2), suggesting that if there are multiple transcription initiation or end sites, the transcripts do not display very different sizes. We also search for enriched sequence patterns located upstream of the defined TSS. Consistent with other fungal species, no clear TATA box was found.

This allows us to describe the first average gene model in *P. anserina* shown Fig. 5. The 5′ and 3′ UTR are 275 bp and 303 bp long, CDS is 1483 bp long with an 80 bp long intron and the genes are spaced with 1581 bp on average (Fig. 5).

Genome-wide schematic representation of average patterns of histone modifications in relation with transcription initiation

In order to validate our 5'UTR predictions, we took advantage of the ChIP-seq data that have been generated on histone marks in *P. anserina* [26]. In mammals and plants, it has been established that H3K4me3 is enriched in the promoter region of active genes, whereas transcriptionally inactive gene promoters are rather marked with H3K27me3. We thus combined the enrichment of these two marks with our annotation for both transcriptionally active and inactive genes (Fig. 6). As a result, we could clearly observe that our predicted TSS positions fit with the peaks of H3K4me3 for expressed genes. Furthermore, the signal drop observed before the TSS, corresponds to the well described nucleosome free region [33]. These observations support our predictions and show that data integration (RNA-seq and ChIP-seq) brings important information on gene organisation and epigenetic regulations of gene expression.

Detection of splicing sites and alternative splicing

In addition to the new annotations of UTRs, we used our RNA-seq dataset to validate the positions of introns in

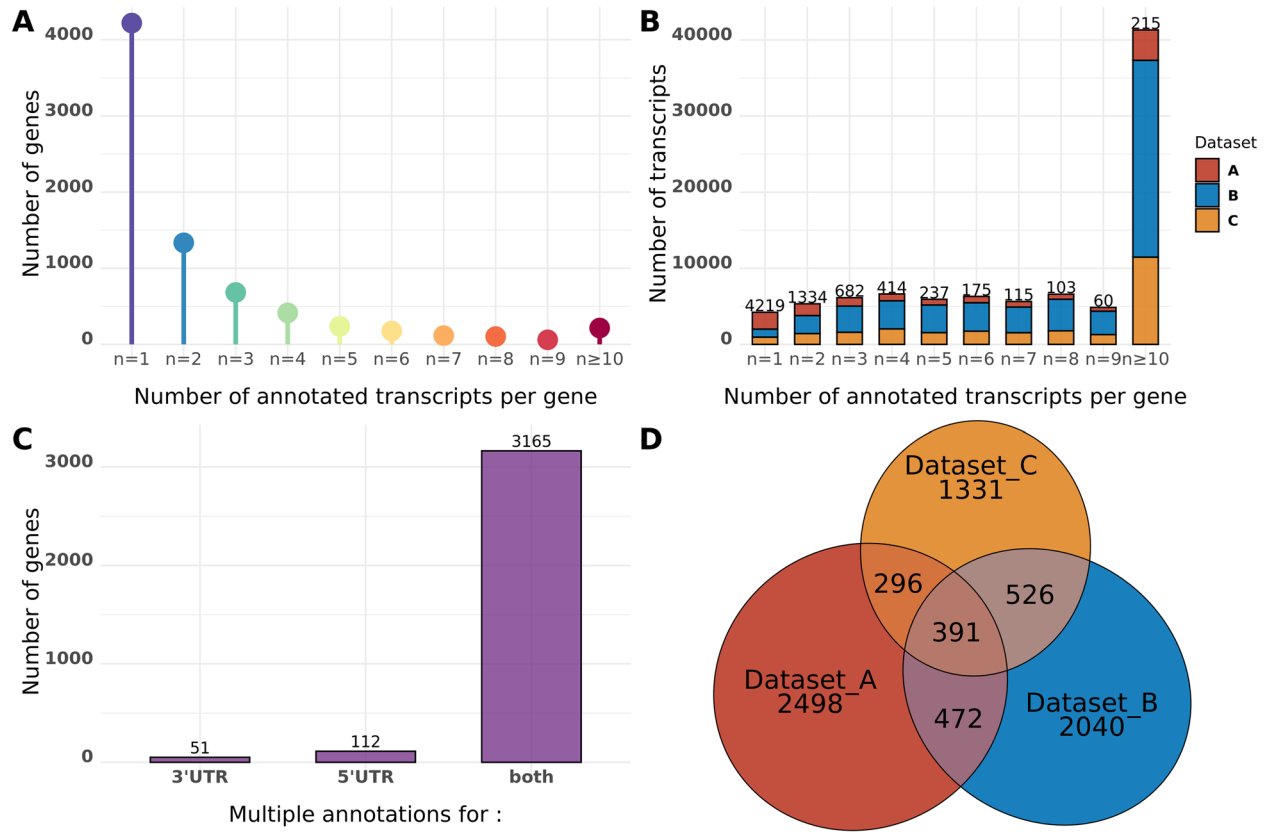

Fig. 3 Transcript predictions of *P. anserina*'s genes. **a**) Distribution of genes with multiple transcript predictions. Genes are divided in categories according to the number of predictions given by the SCV methodology. **b**) Distribution of the number of transcripts for each category of the genes. The number of genes in each category is written on top of the bar. **c**) Distribution of transcripts with different TSS and/or TES annotated for genes with multiple detected transcripts. The number of genes in each category is written on top of the bar. **d**) Venn diagram of genes according to the dataset from which their transcripts are predicted. More than 75% of the genes with predicted transcripts have only one associated dataset ($n = 5869$ genes)

the current annotation of *P. anserina* genome. We could detect introns in the annotated UTRs for an important number of genes: 923 genes have at least one intron in their 5'UTR and 344 genes in the 3'UTR. Among them, 43 have introns in both UTRs. Furthermore, no information regarding the possible ASEs were available. We thus used the collected data to predict these ASEs. All four kinds of ASEs were detected: intron retention (IR), alternative 5'splice site (A5SS), alternative 3'splice site (A3SS), and exon skipping (ES) (Fig. 7) (Table. S2). A total of 2350 genes were found subjected to at least one ASE. IR is the most frequent event; however, if the gene number is considered, A3SS represents the most frequent ASE detected in *P. anserina* with 1016 associated genes, followed by A5SS, IR and ES with respectively 758, 438 and 138 genes. A total of 278 genes could have isoforms with high combinatorial complexity (more than one ASE detected).

Identification of new transcripts, outside already annotated CDS

About 50% of reads mapped on the *P. anserina* genome were located in intergenic regions. They most likely correspond to novel transcriptionally active regions (named "nTARs" as in [35]). Therefore, we were able to detect 3203 nTARs i.e. transcripts that do not fully cover already annotated gene. A significant part of them were very short (32% of nTARs shorter than 500 bp with a mean size of 1 kbp, while CDS length is app. 1.5 kbp long in average). Among all nTARs, 1732 did not overlap any already annotated feature (Fig. 8) (with an average size of 1043 bp and 32% of them smaller than 500 bp). The 1471 others were partially overlapping genes. Interestingly, we could detect introns in 55.8% of these 1732 nTARs ($N = 968$) (Fig. 9) demonstrating production of processed transcripts by these potentially novel genes. Analysis of 332 nTARs longer than 1.5 kbp with the FGENESH

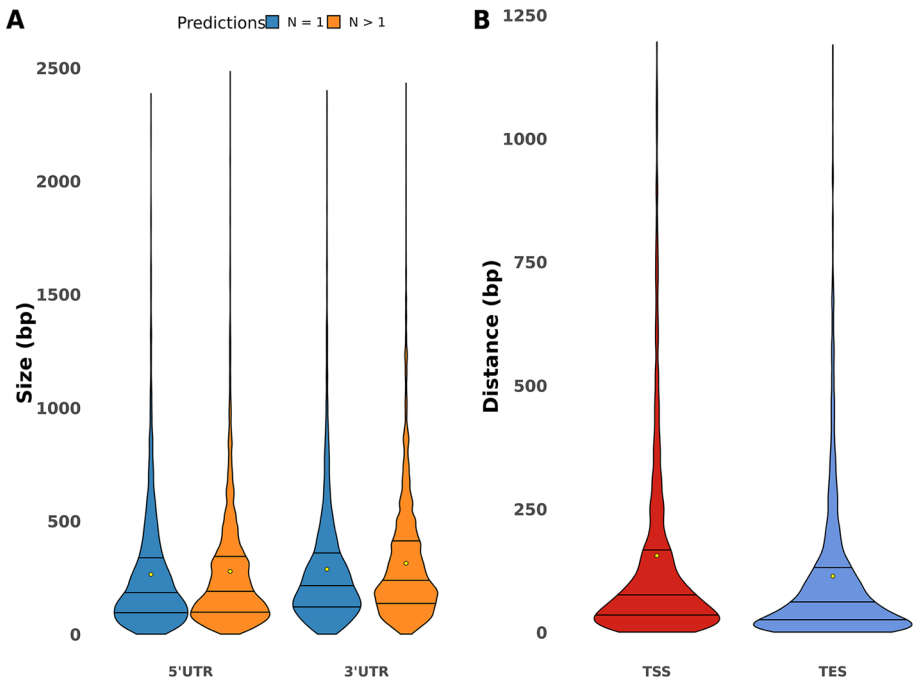

Fig. 4 Size variation of UTRs. **a**) Summary of UTR sizes and violin plots representing the UTR size distribution. 5′ and 3′UTR size predictions are divided between genes with only one prediction. The yellow dots show the mean. For visualization purposes, the UTRs bigger than 2500 bp have been removed ($n=4$ for unique 5′UTR, $n=19$ for multiple 5′UTR; $n=2$ for unique 3′UTR and n = 19 for multiple 3′ UTR). **b**) Distance between the most distant TSS and TES. The mean distance values is marked with a yellow dot

Table 2 Summary of UTRs characteristics. A) Summary of UTRs sizes for both unique and multiple TSS/TES prediction When multiple UTRs, data are calculated from all UTRs from all genes. B) Summary of distance between most distant TSS and TES for each gene with multiple 5′UTR and/or 3′UTR. Mean, median and maximum sizes are expressed in base pairs

A	UTRs size variations			
		Size (bp)		
	UTR	Median	Mean	Max
Single transcript genes $n=4219$	5′UTR	178	265	10,995
	3′UTR	212	288	5620
Multiple transcripts $n=88,728$	5′UTR	191	279	4340
	3′UTR	239	314	4104
B	Multiple transcripts variations			
		Distance (bp)		
	#genes	Median	Mean	Max
TSS	3277	70	156	3987
TES	3216	55	114	3763

gene prediction program yield 20 putative new protein-coding genes (Table 3). Domain prediction found 1 predicted gene with a putative rhodopsin C-terminal tail, transmembrane domains in 3 predicted genes and signal peptides in 2 predicted genes. One transcript was overlapping two sequences recently annotated as pseudogenes [25]. No other protein domain was detected.

One of the RNA-seq datasets was "stranded" (dataset B). This means that one knows from which strand the RNA molecule, which has been sequenced, originated. We thus used this dataset to seek nTARs overlapping previously annotated CDS but transcribed in the other direction, which we termed NATs (Noncoding Antisense Transcripts). NATs are long ncRNAs transcribed from the strand opposite to a protein-coding transcript, thus exhibiting sequence complementarity to mRNAs. We found 1472 NATs overlapping 452 genes (including 2 rRNA genes), 4 pseudogenes and 18 repeated sequences. Among these NATs on repeats, 7 were overlapping transposable elements which rules out a potential role of these NATs in silencing TEs, the other were found on segmental duplications.

Discussion

With this work, we completed the annotation of *P. anserina*'s genome by estimating transcripts size and variations using multiple RNA-seq data. Among the genes with mapped reads, we could make a trustworthy prediction of transcripts for more than two third of them, using a robust method, hence ensuring the reliability of the results. Although we detected multiple transcripts in 44%

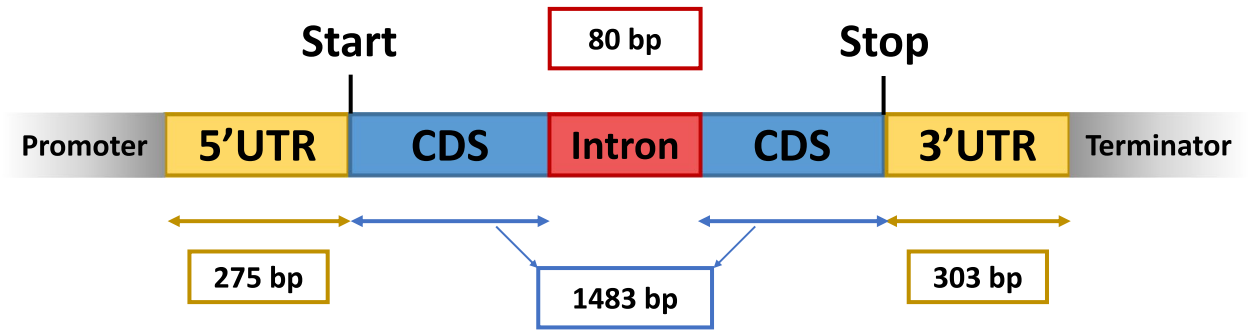

Fig. 5 The average gene model of *P. anserina*. An average gene model was designed that includes all the new information generated here. TSS is expected at 275 bp upstream from the start codon and TES at 303 bp downstream from the stop codon. There are on average 1.49 introns per gene with a size of 80 bp. The mean CDS size is 1483 bp. Yet, little is still known on promoter and terminator

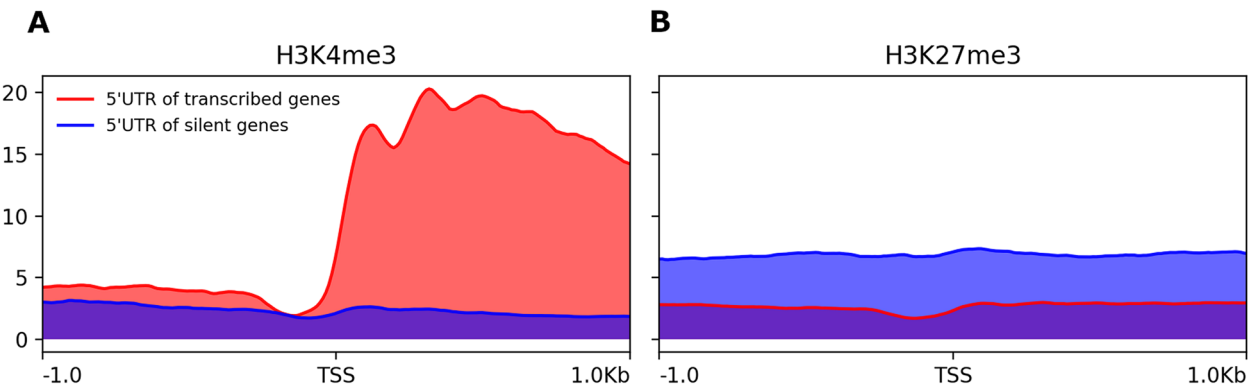

Fig. 6 Integration of ChIP-seq data with the TSS annotations. **a**) H3K4me3 and **b**) H3K27me3 are mapped around the TSS of transcriptionally active and inactive genes. The x-axis represents a window span of 1 kbp upstream and downstream of the average TSS position. The y-axis represents the average value of the normalized number of reads mapped per bins (BPM normalization, bin size = 10)

of the genes, we didn't observe much variation in transcript size even with multiple TSS/TES. This is actually in line with previous observations. For example the *Masc1* gene in *Ascobolus immersus* has two TSS separated with 43 bp [36], whereas the *NiaD* gene in *Aspergillus nidulans* has two TSS separated with 72 bp [37]. Usage of alternative TSSs in filamentous fungi has been described as transcriptional regulator in response of carbon source in *Aspergillus oryzae* [38] or translational regulator in regulating pathogenesis in *Metarhizium robertsii* [39]. However, knowledge about how alternative TSSs affect gene expression is still nascent in filamentous fungi in contrast of what has been uncovered in mammals [40]. In budding yeast (YeasTSS, [41]), a median of 26 transcript isoforms per gene were detected during regular growth conditions [42] and variable UTR sizes in different strains is linked with phenotypic variation [43]. Usage of alternative TSSs and TESs is also involved in budding yeast cell fate transition. High resolution transcriptomic analysis evidenced elevated expression of alternative TSS and TES clusters in a stage-specific manner during yeast gametogenesis program and the mitotic cell cycle [44]. Because, unlike yeasts, filamentous fungi present a syncytial organisation that cannot be synchronized, in-depth description of alternative TSSs and TESs remains challenging. However, our results show that over one third of *P. anserina* genes displays alternative TSS and/or TES usage. Moreover, when present, these alternative transcripts are specific of only one of the environmental conditions tested in this study. This suggests that the use of alternative TSS and/or TES also participates in *P. anserina* stage-specific gene expression and more generally to the resourceful ability of fungi for adaptation.

The genome average length of 5'UTR is quite similar across the diverse eukaryotic taxa, ranging from 100 to 200 bp with the size increasing during eukaryotes evolution [6], while genome average length of 3'UTR seems much more variable, ranging from 200 bp in plants and fungi to 800 bp in humans and 1000 bp in some vertebrates [45, 46]. Thereafter, analyses performed on larger

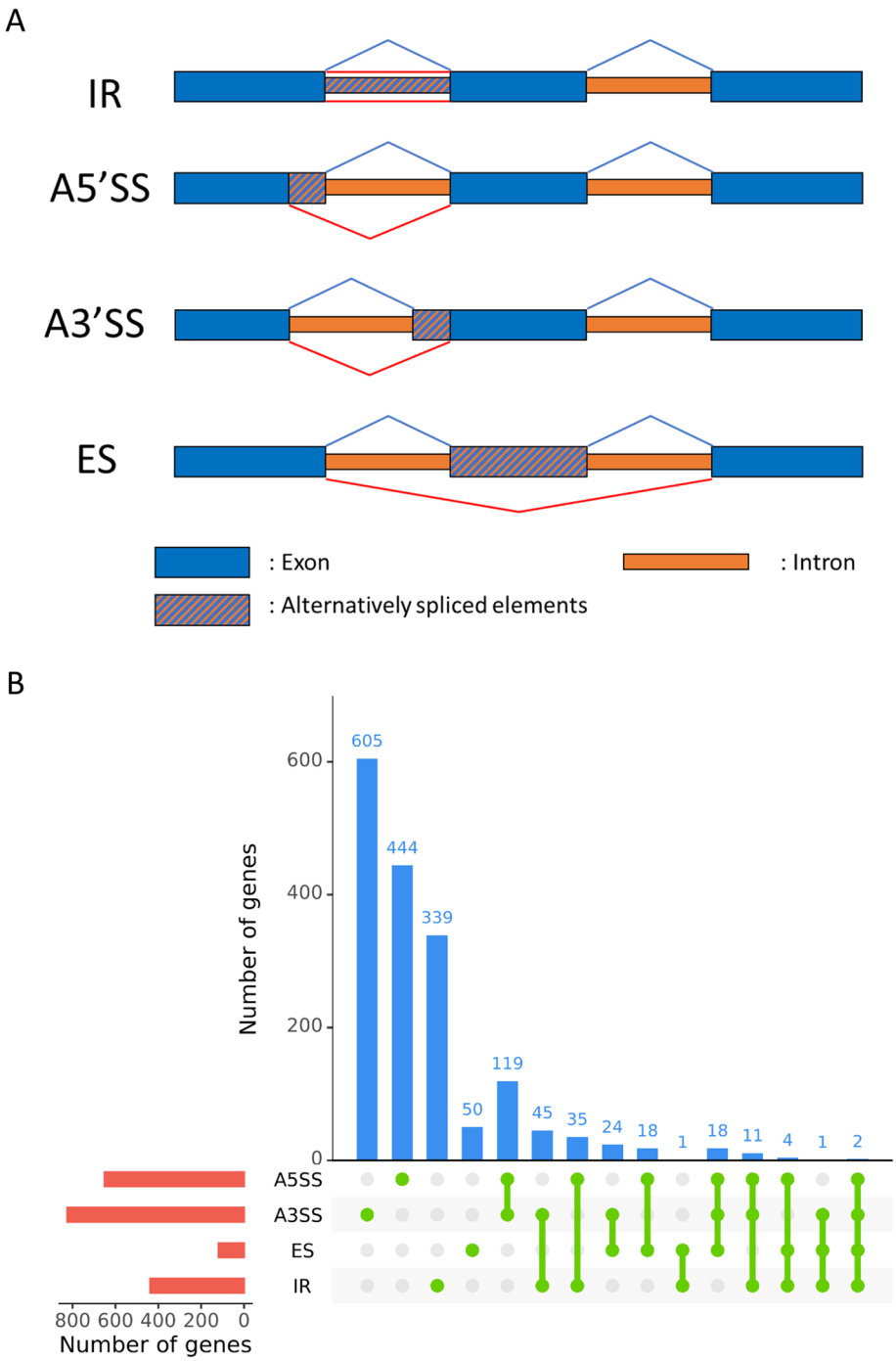

Fig. 7 ASE detected in *P. anserina* coding transcripts. **a**) Representative example of four categories of ASE detected in *P. anserina* transcriptome. IR = intron retention, A5SS = alternative 5′ splicing site, A3SS = alternative 3′ splicing site, and ES = exon skipping (inspired from Kempken, [34]). **b**) Statistics of genes associated with ASE: The red bar plot represents the number of genes undergoing each type of ASE, the blue bar plot shows the number of genes undergoing the combination of ASEs depicted with green points and lines bellow

sets of eukaryotic transcripts showed that although more variable than originally described, 5'UTRs average length is not as diverse as that of 3'UTRs in eukaryotic genomes [47]. The *P. anserina* average size of both 5′ and 3'UTRs that we measured in this study were similar to those established for other fungal and non-fungal eukaryotes.

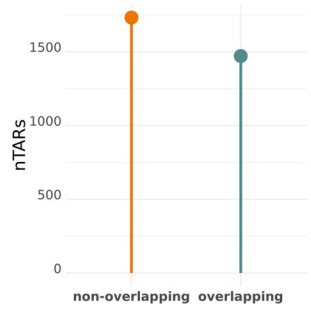

Fig. 8 Quantification of nTARs overlapping or not previously annotated features

In eukaryotes, the low size variability of UTRs contrasts with the very large size increase of intergenic regions during evolution. This intergenic space extension might be correlated with the necessity of a conserved "core promoter" structure, including the TATA box element. In *P. anserina's* 5'UTRs we did not detect clear TATA box signature. This finding is consistent with previous observations showing that most of fungal promoters do not contain a canonical TATA box [48]. As a result, *P. anserina* likely uses the "scanning initiation" mode to start transcription rather than the "classic model" where most TSSs locate ~30 bp downstream from the TATA box. Again, these different ways of initiating transcription might be correlated to the intergenic regions size, where a large sequence does not allow scanning and requires well defined sequences to recruit the polymerase. This new UTR annotation led us to search for UTR introns. As expected several UTRs were found to possess at least one intron although in a lesser extent than in human and plants [49, 50]. As important as the UTRs in gene expression are the introns present in these regions. Their splicing may affect both positively or negatively gene expression through various mechanisms such as mRNA export or nonsense-mediated decay (NMD). In eukaryotes, NMD degrades mRNAs containing premature stop codons as well as those containing an intron downstream of a stop codon, i.e., aberrantly spliced transcripts or 3'UTR intron-containing transcripts. Regulation of expression by mRNAs degradation is functional in *N. crassa* [51] and is expected to be functional as well in *P. anserina* since the NMD core components and the exon junction complex (EJC) are present (Table S3). The set of *P. anserina* genes for which we detected introns in their 3'UTR (~3%) could therefore be prone to regulation by NMD.

We also looked at ASEs genome wide. In our prediction, the proportion of the different patterns of ASEs is in accordance with what has been observed in other filamentous fungi [52]. Regarding the prevalence, we found almost 30% of the genes subjected to AS, which is far more than the 6% found in average for ascomycete fungi [20]. However, this later estimation is based on ESTs and might underestimate the real number of ASEs [21]. Indeed, recent RNA-seq showed ASEs in 24% of expressed genes in an oomycete [53] and 38% in a plant pathogenic ascomycete [54]. One IR event was recently evidenced for the *PaKmt1* gene [26]. This event can be used as a positive control and has been indeed detected in our analysis supporting the robustness of our results. The physiological relevance of alternative splicing is still to be assessed in syncytial organisms but discovery of stage specific splicing events such as that of PaKmt1 suggests a finely regulated process in relation with developmental programs. In search for reliable ASEs we selected those present in at least two independent RNA-seq. Lifting this rules would allow us to detect stage specific ASEs but also expose to false positive.

In this study, we also identified thousands of novel transcripts. Some of them potentially encode functional proteins but the vast majority does not. Other comparable transcriptomic analyses expanded the annotated protein sets of *A. nidulans* and *U. maydis* by 2.9 and 2.5%, respectively [55, 56]. By comparison, the potential 29 new encoding proteins uncovered in this study represent only 0.3% of the previously annotated *P. anserina* CDS. This may be indicative of the good quality of its genome annotation. Among the non-coding nTARs, we detected potential antisense RNA that could also contribute to regulate gene expression. In fungi, non-coding RNAs, including natural antisense transcripts (NATs) are involved in development, metabolism, pathogenesis [57–59], etc. and can be expressed in a cell-specific manner [60]. Some ncRNA/NAT are evolutionary conserved among related smut fungi, which suggests conservation of the corresponding ncRNA/NAT functions [55]. In *N. crassa* and *A. nidulans*, antisense transcripts represent ~5% and ~14% of the annotated protein-coding loci, respectively [56, 61]. Since only one of the 37 RNA-seq is strand-specific and therefore suitable for antisense transcripts, it is too preliminary to quantify the importance of ncRNA/NAT contribution to gene expression. However, this study revealed the first evidence of expression anticorrelation between asRNAs and downstream CDSs.

By collecting information on UTRs and alternative splice sites, as well as identifying novel protein-coding genes and new isoforms, this study, among others, contributes to a better understanding of the molecular basis that governs gene expression in fungi. We propose here the first filamentous fungus average gene model, as to what already exists in animals and plants and show that it fits to already available epigenomic

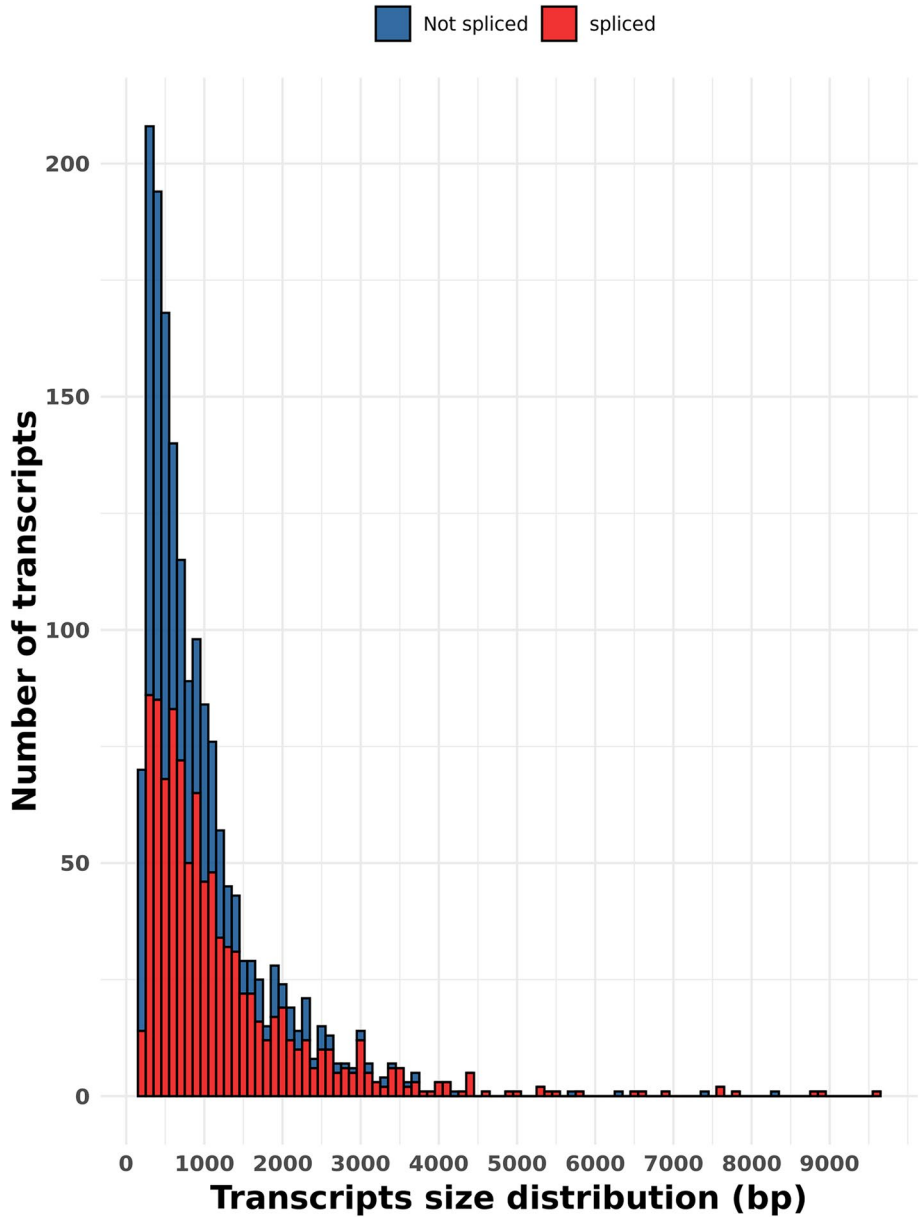

Fig. 9 Transcripts predicted in non-coding regions. Size distribution of the 1732 predicted transcripts in non-coding regions. The color shows the detection of splicing events in these transcripts

data. This model will be useful for the future dataset to be generated.

Methods

Collection, alignment of RNA-seq data and transcriptome annotation

RNA-Seq data were downloaded from the Sequence Read Archive (SRA) [28] at the accession number ERR2224046 to ERR2224051 [25], SRR3197700 to SRR3197711 [32], SRR6960207 to SRR6960225 [29]. Each fastq file was mapped onto the genome of *Podospora anserina* S *mat*+ [23–25] using HISAT2 version 2.1.0 [62] with default parameters. In order to make sure that all the data are of equivalent quality (e.g. no RNA extremity degradation in one dataset) raw read coverage was checked on constitutively expressed genes (Fig. S2) and verified to be consistent across studies at least for the beginning and end of the transcripts. Output alignments files were respectively processed by the StringTie program version 1.3.5 [63]

Table 3 Gene prediction from the putative new transcripts. The table shows the chromosome, position, number of exons, length and sequence of the putative new transcripts. In addition, InterPro prediction and observation from BLAST against NCBI nr database are indicated

Predicted_CDS	Chromosome	Position	exon	length (aa)	sequence	InterPro	Observation
1	1	4,418,621	2	63	MARKGSNKREIYTRSSLSASLYSWLAGGGGSSGGVVAVDHDADAADDSDDAGDDAGDAGD GRG	nothing	
2	1	5,197,730	2	170	MSSTERLLTLAMPLKHEMELMELDVSPKPISVTSSEDNDATAKPHHNEQNRRYSTTNRLD PQEAEPAALEETSYAQQTTDPETTWSPDGADDEGFHVEDDNECASTLPPSPELEAVEDEE MAGWVKEQQTSQPSLRLYRGSPGDILVMARTAAPSFPTYYPRELNMDDEC	nothing	
3	1	5,382,147	3	165	MKTAILLAVLFVGASSLPVAPSLEAKVYLSLSFPTVDATSTRAQWVSSYGNKPPKKAQEN EPTVDGTSTRGQWVSSHGNNPPKKFQEGPTVDATSTRAQCVSSYGNKPPKKGASFSDHWL PVGPGYLDLDALESGYHAHLMIMHDSTSWLQSGMTAIKRFMRRFA	Signal peptide, transmembrane domain	
4	1	5,504,377	1	395	MADLIDLSQQSPSSLSPWEPSVLLQLWNPDVQAWSCLGWTRAERRCRRVLSQAKREATMR ILPDLGLSGSHDVDFETELLGELSHECLCRYHSDDETAKTLVEQWKTALKKARSQYQKTE RTTSKESLSTDSVAHHARTQSPPLSSTESEETTEVMLKDHPEDSTVLKQEAIVEQTIEQS TSASSPRLNSTESASPPAGKTPPKLGFSTPVPGATPSKSPEVSKSTPLFDFTQLFQTPGS HKPTVNPHKGTPAPAAFKYDQSRTPGTSSTADMSHSSVASLSPNENTPAFVFTSSSTPSR PPATTEGSLPKSEDPFRYSGSPIREAGQRIIKSLREIGDMEVPNELDGDISGLGQSIERL RLRLEKGRSLCLSGPDAGSEGDDETSDQDGKRRME	nothing	nothing significant found with BLASTp against NCBI database

Table 3 (continued)

Predicted_CDS	Chromosome	Position	exon	length (aa)	sequence	InterPro	Observation
5	1	7,676,610	1	212	MPASDITIDLDRSSTPF-PRSRRSERAHEQVEDFN-RRSSSPRRFTSTPTRSRRLSP-SPTRY RSAATIARSPPRYRSPTRIT-SAPRPRRISPSPLRHR-SPIRLPSSTSRRAPFPSPY-FKETR ITEARRTTTYHPTSPVS-RYTETRTTYRTPISRPRTPPRG-SIGSPSRRITSPRLVDIRSSS EIYSSSRRYDSYPST-TRSPGRSTDRSLFTRRY	nothing	
6	1	7,835,241	3	94	MSGYPPQQGGYYPQAP-PQGYPPQGYPPPDG-GYPPQGYPPQGYPP-PQQQMQYQQAPPPKEE KSHGCLYTCVAAMCCCWL-CGETCECCLECLDCCF	rhodopsin C-terminal tail	nothing significant found with BLASTp against NCBI database
7	2	1,526,413	2	214	MANIKTETADEGVTAADP-GAIKKAPFSMTESELREILV-LAIDRHPAIHPIVQRHLDRLRD NNLGGFQDDFEKIRCEVY-ACASEPCFSDPKAIGACIKRY-FEKLLNVATRESPYETRYSAV EWFLRVLNLLVFTSDPH-DVRKEIWSHTDGSCLKLVM-LVCRFRTAERGRLLRDHN-SLIMLK MDLITANAKDLPELLREFEP-TIHVIKSWRAESRG	nothing	
8	3	822,817	1	357	MGYEWNGDPTILIVVIACS-VCFGWVPIITVVSIVRHCRAR-LRAKRGSNGTNSDAESQGGR PSTAPDVPKPLQTYHPSST-KGLERSASSRTRSSADGY-DLKRVDTNSSWNPIRHSF-HYDNE SLWGGDGLSRSNSRHRP-PYFPTHVHNTTPSLSRPA-SIRSVASSHRQQSRSRRSS-MASNSD NAPAAFQINDTYYDTTPLP-NVTRTVNPVVASSSTPTSSK-GPGQAPQQRQQKQPKQD-NPHP PQRNRRTRHSL-DARGDSDSLTRD-ISRPNTSMTRREVEEYEDLDN-QKQKATHRSHRPPRPG SASRRGSHSAPGGSEETD-DDLSMAGALPPAKLPPRRA-SLHAQTFERPAWLHEEPHAM	transmembrane domain	nothing significant found with BLASTp against NCBI database

Table 3 (continued)

Predicted_CDS	Chromosome	Position	exon	length (aa)	sequence	InterPro	Observation
9	3	1,121,553	2	317	MHDCEFEENPAGFCCAVET-VELHAAGRSYFYSSFEGAS-CYRQDFAFFRNLQHIS-LRNFFD DPNRSRQQTVQLLRHSPNL-HRLELGLSAKAVVRQLEREGS-FGVFVHFFDRLCDEYAESGG QPLRLTHLGLFDAM-WVWKPESLRKKPAD-LAFLQEVRLNTETIEDCITDN-LVDLFDSEALS GYAVLVETDRGSKYG-PAYLVGARELEMRRPRTP-MQLAEMSLVLGGTWGN-QKLLAATTRHS LQGLVVNMNRPDPRRSLD-FLLAPLQNMHRLARLWIVS-ANMYKDLPLLTKAAQKG-GCRVSC LALHRDRVALLGGNWQN	nothing	nothing significant found with BLASTp against NCBI database
10	3	3,727,340	4	215	MEPRETYREFGSRAA-GRHRRKTLGSTRQTVRDSCK-VNGGNDDTSWLRTPPPHCP-CQKLPV TTGLREEEYAGKQTQAREES-DGMWVVVERKEGFDRQPG-GDGCFQSGERKRGFSGGWR-RLS STTTTTTTTTTATTAT TTTTTTTDEEGEHREE-QQETEETGGCGG GSKARPHSLGTVEEKK PNKKMAGDYPEVLRKF-SLPLSVQGIGSMGLGFPMP	nothing	
11	4	2,328,575	2	69	MVISMTQRNIPGI-WRSGGGRGQDNSAPLPQLQ QQQQQQQQQQQQQQQQ QQQQQQQQQQQQ PQPQPRLQQ	nothing	
12	4	3,371,922	2	373	MCQGTIYDFWCPCIFHAPST-SFYLQFDIHPPDFNYT-FTRRPTTNPLKAHLSKSSH-SIVYS QHCAAYKFCDDYLHSEGFN-PGDVFDMGGLCPAGHQV-TYEREAFISSRLCDACISGK-CEEN MEFAGVKTVRRSRYG-WRSREEEREGKRRSRSR-PGRGVSPAGSVRSFDSTGR-GRSSSVGST RTVKGRDMGVEKGG-VAGEGEGKTLGAMNLKN-LVDKMVQTVSALRVGG-GAERQDQPRVMPA SDLEAMAEESMPTPLP-SRHKPSGKNLEDMFDNS-GRPEYDSDQDTVVGASKT-TEKKSKVNG KTIAADEISGVMQEIPTGR-SKSRKRRMWTDPRTDEEAS-RVLRFLRRGKGAAPVET-GNSRE RSRGQGYERITIE	nothing	nothing significant found with BLASTp against NCBI database

Table 3 (continued)

Predicted_CDS	Chromosome	Position	exon	length (aa)	sequence	InterPro	Observation
13	5	484,874	2	126	MQLTRSLSTALVALLLS-SIATGHRIPAQSEELQL-RDAAPAEVNETGTPPV-VLPVDDTLSA DVIVDETEHGSLVGRAVH-PRQLGKGKGGGKGKGGGK-GKGGAKGKGGGKGKGGKG-KGGKGK GGKGKG	signal peptide	
14	5	1,699,000	5	503	MTPDVDKPNRTIPN-LQKQLSVEREEKELKEAQYQ-FRIQELQDEINSLRDNEHE-SISTGCP QPEPGTTSVNREDIVVRAM-LRGTSPAMLHQEGTIAL-PLSESPRLSVSHSEDHYEWK-DNIT ALALTSQGEDTPKVAYKVEEG-SQNDESDFDEVDYRIPMKG-KEKWKAAVTSERYKYREQKD REYREALNKQHVDGSDRIL-RMDELVAEGNQPWSTFN-MRHTLKATTAHDIPLQSSK-SHPVE SHDVPLSDHDWISGKH-PDDPRAEDRLAPEDVD-VKLAPLKDDTAMGSVP-DLGYGLPRELSI RPQNESKTDDGNIQEDQSD-NQTVYSDDGSIDGDILNVCK-TELADSLANHIRQLEVGPEGY ANITRKLPPLLKAFAL-RVGIQAHRGCRGMLCSLCT-SIATKPASEYNGTSTEALG-SRIINW IQHDEDSDSTGLEQQPSKET-PDELPVEVEADGINFLPD-DHQGLAHQRNHHSSNFG-EHSGV SWLLALWCETPKLWSFG-MVIVAI	nothing	nothing significant found with BLASTp against NCBI database
15	6	2,341,619	2	81	MTEDLHRDITERLRCLELQIR-ITSHMFIGVAQNAGDDPTN-LVKVKDEMLGKLQEMRYEEE RLARERLAALKQRVPSA-GNSD	nothing	

Table 3 (continued)

Predicted_CDS	Chromosome	Position	exon	length (aa)	sequence	InterPro	Observation
16	6	2,662,998	1	382	MSAPLLMHPAEPAT-ADNTKPRLACPFFRYDP-CRHYACASYELKGFEAVK-KHLERKHILKN HCARCFRSFESEDARNNHIV-SECCSIALGRDEITYDEWTR ARRCPRTKSCEVKWKWLWTT FFKLPALPRELVYFQ-DAVVEAKNVLIDPVTIQSV-LKARLHLDQQEISSVADEV-REALLRK NSGARPYRVCDSEGGG-DNGIPANLKASGYGSMGG-GAAEMEAEAVAFALPPARH-ALLPEEP CLPIIGESSPHPAAAVSPVT-PLPTSFSLGPILVPQQPAST-SGGGPETNTFDAWRTVCLVP WATADGILARLMEDPISWFK-PDGPKWSDVYDHIDRDAL-RKFWALGNTPAVQVSIPIRSTH VQSLAAIESKLFDFEVAGIRPS	nothing	
17	6	2,780,217	1	214	MDTKDEDSAQQQSSPLL-PISNHPPSSRPRTPILLKLETN-LPLVTPAQPPETTPQETWDYP TSLRQLTALLLFTLQLLILI-TYHPSFLSLLPIPGPLSNHH-CLLLADTIITCLAIIISSYV HFCIASLDCELLEQGWKPVY-FYIMAADETVILLAAASS-GLENVCSWGLFVVTVGSWY-VGW RLGAVEVLSRRLFRAEGWEF-GQGEGEEGRGLRVV	transmembrane domain	
18	6	4,158,427	4	64	MACDSHGRQPSE-FALVHEALPRDIHLPTCI-HASPKRKTVSSSDTKPRRFLL-HTQGVTSGP RACG	nothing	

Table 3 (continued)

Predicted_CDS	Chromosome	Position	exon	length (aa)	sequence	InterPro	Observation
19	7	440,886	8	626	MVEGVRAFDKLDWKDDVAF-CSLTEDMEEAVGPGDEVFVC-SNQDGMTGSWEMIHNS-SSFGA PPITSELFENANEEPMIDPAV-LGDTWSQMKAWATLCGIKD-DPIAPGIAELLEIEEQESGD GGFCCYGTISHAEVKLVGN-LAESRDRLLNNEHVQSFA-VIKHDDYLMVIFSDNHI-FAQVNE AVSQALTSLFNKFKFFEVKA-FAQIGKIQSLFYQSHTPGQAK-LRVDINIYGSAADADAVGL YLGSTAKLYLQDPE-YGTENIEYLNRQLIHFPG-FEEPKVFAGPGADFANKT-SKALQGVRSQ REHFDQTLSQILLTSRSH-HVLVVGRNQKRPQTTLFKAA-CEIRANFGWCLTATPIQNRLEE LGSPLAFLPIDQLQNRAMFK-KKIMDASSPDAHTMLELP-PIEERYHYITLSQEERNRYDKT AADMSNWINHKTGLHVLT-PNSGDDNNDKVDHFDLSG-VSSKIEVLIRHLQQTPRDT-KRYVG SARLAEVLENQAYINSPSIVF-SCWTKTLDLVALHLTRMKIL-HQRIDGRQKLAERQHNMSR FVSDEGTSVPVLLTTTGVGAF-GLNLTAANHVYILEPQWNPS-VESQALSRVARRGQKKTVL VTRYLVHGTVEILRKMRLAE-AGWATP	transcription facteur SNF2 related, DNA binding domain, ATP binding site	Transcript overlap two features now annotated as pseudogenes
20	7	3,133,570	2	52	MPPKILSEKHEALRQD-VNAKMNKFELRINRKVDDH-MQLRDMFHDRREATSFS	nothing	

with parameters:–g 5 –c 10 (—-rf only for RNA-seq from dataset B) and default values for the remaining parameters.

Reads quantifications

Reads counts were performed for each alignment files using htseqcount version 0.6.0 with the following parameters: --stranded no (RNA-seq A and C) --stranded reverse (RNA-seq B) --mode intersection-non-empty.

Transcript annotation

The annotation files were processed by the Successive Coverage Values (SCV) method using custom-made scripts. The SCV algorithm selects transcript predictions that fully overlap only one annotated CDS by their genomic coordinates. Then, respectively for each gene, transcript predictions from different RNA-seq are compared by their average coverage value to a discrete scale of threshold (from 10 to 20,000 reads). Transcript predictions above the most restrictive threshold are selected to annotate a new single gene model. In order to keep valuable information without redundancy for genes with several transcripts predictions, the longest UTRs are selected and all alternative start and end signal of transcription are annotated considering their RNA-seq dataset.

Integration of ChIP-seq data

ChIP-seq data normalized coverage value were from [26]. Active genes and inactive genes were selected respectively as the 800 most expressed and 800 less expressed in the M4 condition from dataset A as calculated in [25]. This sample was chosen because it is the closet from the conditions used for the ChIP experiments.

Detection of new transcripts

All StringTie transcript predictions with an average coverage equal or higher to 10 that did not intersect any coding or non-coding element of the current annotation were annotated as novel non-coding transcripts. Then, nested transcripts were merged using bedtools version 2.29.2 [64] with default parameters. Spliced non-coding transcripts were retrieved by intersection of TopHat junctions (TopHat2 version 2.1.1, Trapnell et al., 2009) with genomic coordinates of non-coding transcripts.

Gene prediction was made with FGENESH [65] and domain prediction with InterProScan version 5.52–86.0 [66].

Detection of motif in promoters region

The 200 bp sequences upstream of each TSSs have been extracted using bedtools' flank and getfasta tools [64]. Motif search has been performed using MEME with default parameters [67].

Alternative splicing events detection

To obtain information of alternative splicing events that occur in *P. anserina*, we used TopHat2 version 2.1.1 [68] with all default parameters but --min-intron-length 30, −-max-multihits 5 and specifically –segment-length 21 for dataset A and --library-type fr-firstrand for dataset B. Exon-exon junctions annotations were then processed to filter out low-confidence exon-exon junctions (independent RNA-seq ≥ 2 and coverage ≥ 5 in at least one RNA-seq for A5SS, A3SS and ES).

For IR, we quantified aligned reads on CDS and intron annotations following the method above (see Reads quantifications). Intron annotations were segmented by 8 bp bins to assess coverage variability. Then, retained introns were selected applying four thresholds (T1, T2, T3 and T4): 1) average coverage per bin higher than T1, 2) standard deviation of coverage per bin lower than T2, 3) overall expression of the associated CDS higher than T3 and finally 4) ratio between average coverage per bin of the intron and the overall expression of the associated CDS higher than T4. Different association between threshold values were tested, to finally retain: $T1 = 30$, $T2 = 20$, $T3 = 200$ and $T4 = 0.1$. These values allowed to properly select a positive control, i.e. the intron Pa_6_990.G_intron_1 which was expected to be retained in the experiments ERR2224048 and ERR2224049 [26].

Detection of the NATs

Potential NAT (Noncoding Antisense Transcripts) were extracted from the StringTie outputs, obtained with dataset B (see section "Collection, alignment of RNA-seq data and transcriptome annotation"). They are transcripts which 1) have coordinates that overlap (entirely or partially) only one annotated CDS in *Podospora anserina* genome and 2) are found on the opposite strand.

Abbreviations

mRNA: Messenger RNA; UTR: Untranslated region; TSS: Transcription start site; TES: Transcription end site; CDS: Coding sequence; IR: intron retention; A5'SS: Alternative 5′ splicing site; A'3SS: Alternative 3′ splicing site; ES: Exon skipping; ASE: Alternative splicing event; nTARs: Novel transcriptionally active regions; SCV: Successive coverage values.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-09085-4.

Additional file 1.

Additional file 2. (PPTX 85 kb)

Additional file 3.

Additional file 4.

Additional file 5. (PPTX 737 kb)

Acknowledgements

We thank Robert Debuchy for fruitful discussions and precious experimental help. We thank Daniel Gautheret for his suggestions and critical reading of the manuscript. We are also thankful to Cecile Fairhead, Philippe Silar and Stephane Le Crom for their expertise and fruitful discussions.

Authors' contributions

Design analyses: GL, FM, PG. Analyze data: DR, GL, PG. Wrote the paper: FM, PG.

Funding

This work was supported by the department of Genome Biology of the I2BC (https://www.i2bc.paris-saclay.fr/genome-biology-department/). Damien Remy is a recipient of a "contrat doctoral" from Paris-Saclay University.

Availability of data and materials

All scripts and annotation files can be found at: https://github.com/Podospora-anserina/transcript_annotation_2022

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 16 September 2022 Accepted: 16 December 2022
Published online: 29 December 2022

References

1. Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs. Genome Biol. 2002 Feb;3(3):REVIEWS0004.
2. Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. Nat Rev Genet. 2014 Jul;15(7):469–79.
3. Hammond CM, Strømme CB, Huang H, Patel DJ, Groth A. Histone chaperone networks shaping chromatin function. Nat Rev Mol Cell Biol. 2017 Mar;18(3):141–58.

4. Clapier CR, Iwasa J, Cairns BR, Peterson CL. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. Nat Rev Mol Cell Biol. 2017 Jul;18(7):407–22.
5. Hildreth AE, Ellison MA, Francette AM, Seraly JM, Lotka LM, Arndt KM. The nucleosome DNA entry-exit site is important for transcription termination and prevention of pervasive transcription. eLife. 2020 Aug 26;9:e57757.
6. Leppek K, Das R, Barna M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol. 2018 Mar;19(3):158–74.
7. Szostak E, Gebauer F. Translational control by 3′-UTR-binding proteins. Brief Funct Genomics. 2013 Jan;12(1):58–65.
8. Hood HM, Neafsey DE, Galagan J, Sachs MS. Evolutionary roles of upstream open Reading frames in mediating gene regulation in Fungi. Annu Rev Microbiol. 2009 Oct;63(1):385–409.
9. Griesemer D, Xue JR, Reilly SK, Ulirsch JC, Kukreja K, Davis JR, et al. Genome-wide functional screen of 3′UTR variants uncovers causal variants for human disease and evolution. Cell. 2021;184(20):5247–5260.e19.
10. Schuster SL, Hsieh AC. The untranslated regions of mRNAs in Cancer. Trends Cancer. 2019 Apr 1;5(4):245–62.
11. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007 Jun;447(7146):799–816.
12. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol. 2018 Oct;19(10):621–37.
13. Rojas-Duran MF, Gilbert WV. Alternative transcription start site selection leads to large differences in translation activity in yeast. RNA. 2012 Dec;18(12):2299–305.
14. Kooistra SM, Helin K. Molecular mechanisms and potential functions of histone demethylases. Nat Rev Mol Cell Biol. 2012 May;13(5):297–311.
15. Xiao J, Lee US, Wagner D. Tug of war: adding and removing histone lysine methylation in Arabidopsis. Curr Opin Plant Biol. 2016 Dec;34:41–53.
16. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, et al. Single-Nucleosome Mapping of Histone Modifications in *S. cerevisiae*. Becker P, editor. PLoS Biol. 2005; 3(10):e328.
17. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, et al. MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 2014 Jan;42(D1):D699–704.
18. Ferraro AR, Ameri AJ, Lu Z, Kamei M, Schmitz RJ, Lewis ZA. Chromatin accessibility profiling in *Neurospora crassa* reveals molecular features associated with accessible and inaccessible chromatin. BMC Genomics. 2021 Dec;22(1):459.
19. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, et al. Drosophila *Dscam* is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell. 2000 Jun 9;101(6):671–84.
20. Grutzmann K, Szafranski K, Pohl M, Voigt K, Petzold A, Schuster S. Fungal alternative splicing is associated with multicellular complexity and virulence: a genome-wide multi-species study. DNA Res. 2014 Feb 1;21(1):27–39.
21. Fang S, Hou X, Qiu K, He R, Feng X, Liang X. The occurrence and function of alternative splicing in fungi. Fungal Biol Rev. 2020 Dec 1;34(4):178–88.
22. Silar P. *Podospora anserina*. 2020. https://hal.archives-ouvertes.fr/hal-02475488. https://hal.archives-ouvertes.fr/hal-02475488/file/Podospora%20anserina.pdf.
23. Espagne E, Lespinet O, Malagnac F, Da Silva C, Jaillon O, Porcel BM, et al. The genome sequence of the model ascomycete fungus *Podospora anserina*. Genome Biol. 2008;9(5):R77.
24. Grognet P, Bidard F, Kuchly C. Chan ho Tong L, Coppin E, Benkhali JA, et al. maintaining two mating types: structure of the mating type locus and its role in Heterokaryosis in *Podospora anserina*. Genetics. 2014 May 1;197(1):421–32.
25. Silar P, Dauget JM, Gautier V, Grognet P, Chablat M, Hermann-Le Denmat S, et al. A gene graveyard in the genome of the fungus *Podospora comata*. Mol Gen Genomics. 2019;294(1):177–90.
26. Carlier F, Li M, Maroc L, Debuchy R, Souaid C, Noordermeer D, et al. Loss of EZH2-like or SU(VAR)3–9-like proteins causes simultaneous perturbations in H3K27 and H3K9 tri-methylation and associated developmental defects in the fungus *Podospora anserina*. Epigenetics Chromatin. 2021 May 7;14(1):22.
27. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. 2012 Jan 1;40(D1):D57–63.
28. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39(SUPPL. 1):2010–2.
29. Benocci T, Daly P, Aguilar-Pontes MV, Lail K, Wang M, Lipzen A, et al. Enzymatic adaptation of *Podospora anserina* to different plant biomass provides leads to optimized commercial enzyme cocktails. Biotechnol J. 2019;14(4):1800185.
30. Lamacchia M, Dyrka W, Breton A, Saupe SJ, Paoletti M. Overlapping *Podospora anserina* transcriptional responses to bacterial and fungal non self indicate a multilayered innate immune response. Front Microbiol. 2016;7:471.
31. Philipp O, Hamann A, Servos J, Werner A, Koch I, Osiewacz HD. A genome-wide longitudinal transcriptome analysis of the aging model *Podospora anserina*. PLoS One. 2013 Dec 20;8(12):e83109.
32. Lamacchia M, Dyrka W, Breton A, Saupe SJ, Paoletti M. Overlapping *Podospora anserina* transcriptional responses to bacterial and fungal non self indicate a multilayered innate immune response. Front Microbiol. 2016;7(APR):1–18.
33. Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, et al. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet. 2007 Oct;39(10):1235–44.
34. Kempken F. Alternative splicing in ascomycetes. Appl Microbiol Biotechnol. 2013;97(10):4235–41.
35. Zhao C, Waalwijk C, de Wit PJGM, Tang D, van der Lee T. RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen *Fusarium graminearum*. BMC Genomics. 2013 Jan 16;14(1):21.
36. Malagnac F, Wendel B, Goyon C, Faugeron G, Zickler D, Rossignol JL, et al. A gene essential for de novo methylation and development in Ascobolus reveals a novel type of eukaryotic DNA methyltransferase structure. Cell. 1997 Oct 17;91(2):281–90.
37. Berger H, Pachlinger R, Morozov I, Goller S, Narendja F, Caddick M, et al. The GATA factor AreA regulates localization and in vivo binding site occupancy of the nitrate activator NirA. Mol Microbiol. 2006;59(2):433–46.
38. Inoue T, Toji H, Tanaka M, Takama M, Hasegawa-Shiro S, Yamaki Y, et al. Alternative transcription start sites of the enolase-encoding gene enoA are stringently used in glycolytic/gluconeogenic conditions in *aspergillus oryzae*. Curr Genet. 2020;66(4):729–47.
39. Guo N, Qian Y, Zhang Q, Chen X, Zeng G, Zhang X, et al. Alternative transcription start site selection in Mr-OPY2 controls lifestyle transitions in the fungus *Metarhizium robertsii*. Nat Commun. 2017 Dec;8(1):1565.
40. The FANTOM. Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. Nature. 2014 Mar;507(7493):462–70.
41. McMillan J, Lu Z, Rodriguez JS, Ahn TH, Lin Z. YeasTSS: an integrative web database of yeast transcription start sites. Database. 2019;2019:baz048.
42. Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature. 2013 May;497(7447):127–31.
43. Sardu A, Treu L, Campanaro S. Transcriptome structure variability in *Saccharomyces cerevisiae* strains determined with a newly developed assembly software. BMC Genomics. 2014 Dec 1;15(1):1045.
44. Chia M, Li C, Marques S, Pelechano V, Luscombe NM, van Werven FJ. High-resolution analysis of cell-state transitions in yeast suggests widespread transcriptional tuning by alternative starts. Genome Biol. 2021 Dec;22(1):34.
45. Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S. Structural and functional features of eukaryotic mRNA untranslated regions. Gene. 2001 Oct 3;276(1–2):73–81.
46. Mignone F, Pesole G. mRNA Untranslated Regions (UTRs). eLS. John Wiley & Sons, Ltd; 2018. https://doi.org/10.1002/9780470015902.a0005009.pub3.
47. Wang W, Fang D. Hui, Gan J, Shi Y, Tang H, Wang H, et al. evolutionary and functional implications of 3′ untranslated region length of mRNAs by comprehensive investigation among four taxonomically diverse metazoan species. Genes Genomics. 2019 Jul;41(7):747–55.
48. Sakekar AA, Gaikwad SR, Punekar NS. Protein expression and secretion by filamentous fungi. J Biosci. 2021 Dec;46(1):5.
49. Bicknell AA, Cenik C, Chua HN, Roth FP, Moore MJ. Introns in UTRs: why we should stop ignoring them. BioEssays. 2012;34(12):1025–34.

50. Chung BY, Simons C, Firth AE, Brown CM, Hellens RP. Effect of 5'UTR introns on gene expression in *Arabidopsis thaliana*. BMC Genomics. 2006 May 19;7(1):120.
51. Zhang Y, Sachs MS. Control of mRNA Stability in Fungi by NMD, EJC and CBC Factors Through 3′UTR Introns. Genetics. 2015;200(4):1133–48. https://doi.org/10.1534/genetics.115.176743.
52. McGuire AM, Pearson MD, Neafsey DE, Galagan JE. Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biol. 2008 Mar 5;9(3):R50.
53. Burkhardt A, Buchanan A, Cumbie JS, Savory EA, Chang JH, Day B. Alternative Splicing in the Obligate Biotrophic Oomycete Pathogen *Pseudoperonospora cubensis*. Mol Plant-Microbe Interactions®. 2015 Mar;28(3):298–309.
54. Liu XY, Fan L, Gao J, Shen XY, Hou CL. Global identification of alternative splicing in *Shiraia bambusicola* and analysis of its regulation in hypocrellin biosynthesis. Appl Microbiol Biotechnol. 2020 Jan 1;104(1):211–23.
55. Donaldson ME, Ostrowski LA, Goulet KM, Saville BJ. Transcriptome analysis of smut fungi reveals widespread intergenic transcription and conserved antisense transcript expression. BMC Genomics. 2017 Dec;18(1):340.
56. Sibthorp C, Wu H, Cowley G, Wong PWH, Palaima P, Morozov IY, et al. Transcriptome analysis of the filamentous fungus *aspergillus nidulans* directed to the global identification of promoters. BMC Genomics. 2013 Dec;14(1):1–18.
57. Chacko N, Lin X. Non-coding RNAs in the development and pathogenesis of eukaryotic microbes. Appl Microbiol Biotechnol. 2013 Sep;97(18):7989–97.
58. Li N, Joska TM, Ruesch CE, Coster SJ, Belden WJ. The frequency natural antisense transcript first promotes, then represses, frequency gene expression via facultative heterochromatin. Proc Natl Acad Sci. 2015 Apr 7;112(14):4357–62.
59. Xue Z, Ye Q, Anson SR, Yang J, Xiao G, Kowbel D, et al. Transcriptional interference by antisense RNA is required for circadian clock function. Nature. 2014 Oct 30;514(7524):650–3.
60. Donaldson ME, Saville BJ. *Ustilago maydis* natural antisense transcript expression alters stability and pathogenesis. Mol Microbiol. 2013 Jul;89(1):29–51.
61. Arthanari Y, Heintzen C, Griffiths-Jones S, Crosthwaite SK. Natural antisense transcripts and long non-coding RNA in *Neurospora crassa*. PLoS One. 2014 Mar 12;9(3):e91353.
62. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015 Apr;12(4):357–60.
63. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5.
64. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
65. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. 2006 Aug 7;7(1):S10.
66. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.
67. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015 Jul 1;43(W1):W39–49.
68. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.

Publisher's Note