Computational Molecular Biology 2025, Vol.15, No.1, 1-12
http://bioscipublisher.com/index.php/cmb

Research Article Open Access

Genome-Wide Identification and Analysis of Alternative Splicing in *Aspergillus niger*

Xiangjia Min [1], Courtney Jones [1], Burrows Logan [1], Maria Campean [1], Bilal Wekhyan [1], Chetan Dasana [1], Aaryan Patel [1], Pasan Mudiyanselage [2], Kolade Adeyemo [1], Feng Yu [2]

1 Department of Chemical and Biological Sciences, Youngstown State University, OH 44555

2 Department of Computer Science, Information, and Engineering Technology, Youngstown State University, OH 44555

Corresponding author: xmin@ysu.edu

Computational Molecular Biology, 2025, Vol.15, No.1 doi: 10.5376/cmb.2025.15.0001

Received: 08 Nov., 2024

Accepted: 23 Dec., 2024

Published: 15 Jan., 2025

Preferred citation for this article:

Min X.J., Jones C., Logan B., Campean M., Wekhyan B., Dasana C., Patel A., Mudiyanselage P., Adeyemo K., and Yu F., Genome-wide identification and analysis of alternative splicing in *Aspergillus niger*, Computational Molecular Biology, 15(1): 1-12 (doi: 10.5376/cmb.2025.15.0001)

Abstract *Aspergillus niger* is a widely used fungal species in fermentation industry. Identifying genes having RNA transcripts undergoing alternative splicing (AS) is important for understanding the gene expression regulations and finding novel enzymes for bioprocessing applications. In this study, we combined genome mapping information of all available RNA sequences and expressed sequence tags in the public database with RNA-seq data collected from 303 publicly available samples for identification of AS events in *A. niger*. We identified a total of 63,715 AS events including 10,097 (15.8%) alternative acceptor sites (AltA), 6,063 (9.5%) alternative donor sites (AltD), 12,469 (19.6%) intron retention sites (IntronR), 1,945 (3.1%) exon skipping sites (ExonS), and 33,141 (52.0%) complex events which contained two or more basic events in pairs of compared isoform transcripts. These AS events were identified from 4,972 genes involving 43,156 unique transcripts. The AS rate among all expressed genes was estimated to be ~50.0% in *A. niger*. Protein coding genes having protein family matches were estimated having 68.0% AS rate, including 52 of 84 genes coding for carbohydrate degrading enzymes (CAZymes) alternatively spliced. The functions of these proteins encoded by alternatively splicing generated isoforms need to be further investigated. We also identified a total of 1,592 new genomic loci with 3,388 transcripts that were not annotated in the reference genome. The AS data and genomic mapping data collected in this study provide a resource for further exploration of novel genes and enzymes in *A. niger*.

Keywords *Aspergillus niger*; Alternative splicing; mRNA; RNA-seq; Carbohydrate active enzymes

1 Introduction

Alternative splicing (AS) is a common process which generates more than one RNA transcript from an intron containing gene in eukaryotic organisms. AS plays important biological roles in regulation of biological development and adaptions to the changing environments through increasing both the diversities of transcriptome and proteome (Chaudhary et al., 2019). It is estimated more than 90% genes in humans and ~65-70% genes in plants, such as Arabidopsis and tomato, are subject to alternative splicing (Pan et al., 2008; Zhang et al., 2017; Clark et al., 2019). A recent survey reveals AS events in fungi ranged from 0.2% in non-pathogenic yeast *Saccharomyces cerevisiae* to 38.44% of expressed transcripts in *Shiraia bamlusicola*, a parasitic fungus on bamboo twigs (Fang et al.; 2020; Liu et al., 2020).

Aspergillus niger, a filamentous fungal species, is widely used in fermentation industry for producing citric acid, glucoamylase and some other enzymes (Cairns et al., 2018). Identifying alternatively spliced genes in this species may help to improve industrial strains for enzyme production. Glucoamylase mRNA transcripts in *A. niger* were among the earliest AS cases reported in fungi (Boel et al., 1984). One 169 bp intron was involved in differential mRNA processing leading to two different glucoamylase enzymes G1 and G2 (Boel et al., 1984). Assembling expressed sequence tags (ESTs) identified 56 alternatively spliced genes including glucoamylase genes in *A. niger* (Semova et al., 2006). Mapping mRNA and EST sequences with spliced transcript-genome alignments further revealed 9.5% AS rate in *A. niger* (Grützmann et al., 2014). In last ten years RNA sequencing (RNA-seq) technology has been widely used to quantify RNA transcripts of a transcriptome as well as to identify AS events. A number of RNA-seq experiments have been reported in *A. niger*, such as, Xu et al. (2024) identified a total of 23 out of the 56 lignocellulose-degrading enzyme genes which had AS events with intron retention as the main

type of AS events. Clearly, there is a lack of systematic genome-wide identification of alternatively spliced genes in this important species. In considering the importance of the organism in industrial applications, we carry out a systematic genome-wide identification and analysis of AS events in *A. niger* by integrating available RNA transcripts with RNA-seq data from multiple published projects. The aim is to generate a catalog of genes subjecting to AS in *A. niger*. Such a collection of these genes with their respective transcript isoform annotation information may serve as a foundation for further characterizing the biological functions and regulations of these genes in this important fungal species for the fermentation industry.

2 Materials and Methods

2.1 Genome, mRNA sequences, and RNA-seq datasets

A. niger reference genome sequences with annotation GFF (Gene feature format) file (CBS 513.88, assembly ASM285v2) and other related files were downloaded from the genome database of the National Center for Biotechnology Infomation (NCBI, https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002855.4/) (Pel et al., 2007). A total of 78,361 *A. niger* mRNA sequences which includes 46,938 ESTs were downloaded from NCBI nucleotide database. The RNA-seq data was down-loaded from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra/docs/sradownload/) using SRA Toolkit. The RNA-seq datasets were selected from seven projects which were recently published in eight RNA-seq publications (Table 1). A total of 303 RNA-seq samples generated from diverse treatments were collected (Table 1). The data from the project PRJNA250529 were generated and analyzed by Daly et al. (2017) and van Munster et al. (2020). Daly et al. (2017) investigated the responses of *A. niger* to ionic liquid (IL) or hydrothermally (HT) pretreated knife-milled wheat straw (KMS) over a time course using RNA-seq and proteomics. van Munster et al. (2020) further analyzed the responses of *A. niger* to the feedstock *Miscanthus* and compared the results on wheat straw. Other data were collected including *A. niger* cultured in peanut or cashew nut flour- based media (Mattison et al., 2021), in steam-exploded sugarcane bagasse (Borin et al., 2017), in sugar beet pulp (Garrigues et al., 2022), in sucrose or inulin (Kun et al., 2023), in glucose or wheat straw (Xu et al., 2024), and wildtype and different deletion strains cultured in glucose (van Leeuwe et al., 2020). We also tested RNA-seq datasets reported in project PRJNA316878 and PRJNA148183 and found the RNA-seq data mapping rate <50%, those data were not included for further analysis.

Table 1 RNA-seq data sources and related reference

RNA	Projects	SRA data	Treatments	References
I	PRJNA250529	137	Responses to wheat straw or *Miscanthus*	Daly et al. (2017); van Munster et al. (2020)
II	PRJNA553205	12	Responses to peanut or cashew nut flour	Mattison et al. (2021)
III	PRJNA636647	4	Comparing of a wildtype with a deletion strain	van Leeuwe et al. (2020)
IV	PRJNA350271	8	Responses to sugarcane bagasse	Borin et al. (2015)
V	Multiple projects*	90	Sugar beet pulp utilization	Garrigues et al. (2022)
VI	Multiple projects*	46	Sucrose and inulin utilization	Kun et al. (2023)
VII	PRJNA1067358	6	Responses to glucose or wheat straw	Xu et al. (2024)
Total samples		303	-	-

Note: * The accession numbers of the data can be found in related references and in supplementary files

2.2 mRNA sequence mapping, RNA-seq reads mapping, and AS identification

The procedure for mRNA sequences cleaning and further assembling into a non-redundant set of unique transcripts were described in our previous work (Clark et al., 2019). The final cleaned transcripts consisting of 78,194 sequences were further assembled into a non-redundant set of 23,853 sequences. The assembled nucleotide sequences were mapped to *A. niger* genome sequences using cutoff values of a minimum 95% identity and >75% length coverage using ASFinder and Sim4 programs (Florea et al., 1998; Min, 2013).

The RNA-seq reads were mapped to the reference genome sequences using TopHat (v2.2.6) with default parameters (Kim et al., 2013). TopHat2 is designed to handle a relatively low error rate, typically considered

around 1-2% for most RNA-Seq data, allowing it to accurately map reads even with minor sequencing errors present in the data. The transcript alignment bam file together with annotation GFF file were used as input for Cufflinks (v2.2.1) (Trapnell et al., 2010). The transcript GTF (Gene Transfer Format) files generated from each RNA-seq dataset after running Cufflinks were merged using cuffmerge script within the Cufflinks package for each project. The GTF file generated from merged RNA-seq GTF files in each RNA-seq project was further merged using Cuffcompare script. Astalavista was used for AS event classification (Foissac and Sammeth, 2007). AS events are generally classified as exon skipping (ExonS), alternative donor site (AltD), alternative accepter site (AltA), intron retention (IntronR) and complex event. The complex AS events were counted when two or more basic events occurred in comparing a pair of isoforms. RNA-seq data processing was carried out in our facility and in the Ohio Supercomputer Center.

2.3 Functional annotation of transcripts and data availability

The transcript sequences were retrieved using gtf_to_fasta tool in the TopHat package based on the GTF file generated by Cuffcompare program after merging all GTF files. These transcripts were functionally annotated, including open reading frame (ORF) prediction, BLASTX against UniProt-Swiss-Prot database, protein family (Pfam), and comparison with reference gene models (Min et al., 2005a; Min et al., 2005b). The transcripts sequences, detailed information of AS events, and supplementary files are available at our bioinformatics site (http://bioinformatics.ysu.edu/publication/data/Aniger/).

3 Results

3.1 Mapping mRNA assembled transcripts and RNA-seq data to the genome

Beginning with a total of 78,361 mRNA sequences, after going through the cleaning procedure including trimming poly(A/T) ends and removing contaminants and repetitive sequences, we obtained 78,194 sequences that were further assembled into a non-redundant set of 23,853 transcripts for genome mapping. A total of 19,571 (82.0%) assembled transcripts were mapped to the reference genome.

We mapped a total of 303 RNA-seq datasets to *A. niger* reference genome (Table 2). The accession numbers and detailed mapping information of these RNA-seq data can be found in a supplementary file (supplementary Table 1). The mapping rates varied from 70.2% to 90.0% with 0.4% to 4.0% reads being mapped to more than one location in the datasets collected from different projects. In total 12.7 billion reads were collected with 10.3 billion reads (~81.0%) being mapped to the genome. Among the mapped reads, 2.7% reads (~0.35 billion) were mapped to two or more genomic loci (Table 2).

Table 2 Mapping summary of RNA-seq datasets obtained from different projects

Data source	Total reads	Mapped reads	MA reads *	Mapped (%)	MA (%)
I	8 773 346 255	7 309 217 470	274 072 924	83.3	3.7
II	278 954 306	207 899 991	8 412 714	74.5	4.0
III	185 792 326	157 657 498	165 537	84.9	0.1
IV	391 730 912	352 553 672	9 121 779	90.0	2.6
V	2 042 840 126	1 501 446 874	38 297 236	73.5	2.6
VI	797 080 340	559 903 468	14 453 144	70.2	2.6
VII	276 043 016	234 248 824	902 089	84.9	0.4
Total	12 745 787 281	10 322 927 797	345 425 423	81.0	2.7

Note: * MA reads: reads mapped to more than one genomic locus

3.2 Identification of AS events

Mapping assembled mRNA transcripts, including ESTs, to the genome we identified a total of 2,098 AS events including 74 ExonS, 213 AltD, 397 AltA, 723 IntronR, and 691 complex events (Table 3). These AS events were generated from 1,804 genes involving 3,835 transcripts.

Table 3 Classification of alternative splicing events in different datasets in *A. niger*

Data sources	AltA	AltD	IntronR	ExonS	Complex	Total
I	5679	3244	5194	1126	12322	27565
II	1188	625	814	159	952	3738
III	577	254	318	91	301	1541
IV	554	273	413	128	420	1788
V	3763	2024	2994	596	5402	14779
VI	2321	1154	1556	353	2385	7769
VII	568	229	519	139	756	2211
RNA-seq Merged	9506	5749	11706	1781	30943	59685
mRNA data	397	213	723	74	691	2098
RNA-seq and mRNA merged	10097	6063	12469	1945	33141	63715
(%)	15.8	9.5	19.6	3.1	52.0	100.0

For identification of AS events in RNA-seq data, we first identified AS events in each project by merging mapping information of all samples within each project, then we merged all mapping information of seven projects. Finally, we merged mRNA mapping information with RNA-seq data mapping information to generate the final list of AS events (Table 3). Since each project had different numbers of samples and associated reads, thus the AS events varied greatly among them. Among the basic AS events, AltA was the predominant AS type followed by IntronR type in all the RNA-seq projects (Table 3). However, we noticed that IntronR became predominant type when all RNA mapping data were merged. Another interesting observation was when RNA-seq data were combined with mRNA data, the total numbers of AS events were more than the addition of the two datasets analyzed individually, since AS events were identified by pair-wise comparisons of isoforms generated from a gene undergoing AS (Table 3). In short, in this work we have identified a total of 63,715 AS events including 10,097 (15.8%) AltA, 6,063 (9.5%) AltD, 12,469 (19.6%) IntronR, 1,945 (3.1%) ExonS, and 33,141 (52.0%) complex events. ExonS is the least basic AS type in *A. niger*, suggesting the splicing mechanism in fungal species is similar with plant species. These AS events were identified from 4,972 genomic loc involving 43,156 unique transcripts.

Combining all the mapping data we obtained a total of 9,939 genomic loci with a total of 66,007 transcripts assembled by Cufflinks tool (Table 4). Among these genomic loci, 7,026 (70.7%) loci produced two or more transcripts and 2,913 loci generated one transcript each locus. However, based on the loci mapping of the isoform transcripts with AS events, 4,972 loci were identified for generation of AS events. Thus, the AS rate based on current data collection at the genome level for all genes was estimated to be ~50.0% in *A. niger* (Supplementary Table 2). However, as there were 1,032 gens consisting of only a single exon, i. e., no intron, AS rate among intron containing genes was 55.8%. To our knowledge this is the highest AS rate ever reported in a fungal species (reviewed by Fang et al., 2020). Comparing with gene models in the reference genome annotation, 8,347 genomic loci in our work were mapped to the reference genomic loci. However, current *A. niger* reference genome was annotated with 10,828 genomic loci (Pel et al., 2007), interestingly, these loci were mapped to 8,347 genomic loci generated in our data. Clearly some of reference genomic loci were merged into longer, and, thus resulting in, fewer loci in our data. In addition, there were 1,592 genomic loci unmapped with annotated genomic loci, representing newly identified genomic loci with the supporting evidence from RNA-seq data. The newly identified genomic loci generated 3,388 transcripts and 2,795 ORFs were predicted from these transcripts.

3.3 Functional annotation of transcripts

A total of 66,007 RNA transcript sequences were retrieved and further annotated, including ORF prediction, functional annotation based on BLASTX against UniProt-Swiss-Prot database, and protein family (Pfam) prediction. These basic features of these transcripts were summarized (Table 4). The transcripts have an average length of 3,735 bp and 22,162 (33.6%) transcripts had a BLASTX match against Swiss-Prot data. A total of 61,153 (92.6%) were predicted to have an ORF region which contained a start codon with a minimum length of

20 amino acids encoded. The average length of predicted proteins was 285 amino acids. The current reference protein dataset in NCBI consists of 14,086 sequences with an average length 440 amino acids. In addition, using BLASTN no-gap search with a cut-off of ≥97% identity and a minimum length of align 60 bp 27,756 (41.8%) transcripts were matched with transcripts of gene models in *A. niger*.

Table 4 Basic features of assembled RNA transcripts and functional annotation in *A. niger*

Total genomic loci	9939
Loci having one transcript	2913 (29.3%)
Loci having more than one transcript	7026 (70.7%)
Mapped to gene model loci	8347
New genomic loci	1592
Total unique transcripts	66007
Average transcript length (bp)	3735
Transcripts match to gene model transcripts	27566 (41.8 %)
BLASTX match against Swiss-Prot dataset	22162 (33.6%)
Total predicted ORFs	61153 (92.6%)
Average ORF length (amino acids)	285
Total ORFs with a Pfam match	19177 (31.4%)

The predicted proteins from retrieved mRNA transcripts were annotated to Pfam, that facilitates examination if the functional domain in proteins encoded by different transcript isoforms is maintained. Since the isoforms in alternatively spliced genes may encode a truncated protein, thus resulting in a domain loss, due to a pre-mature stop codon or may not be able to translate to a protein due to a translation frame shift. A total of 19,177 predicted ORFs had a Pfam match (Table 4). Among 9,939 total genomic loci identified in this work, 2,941 loci encoded proteins had Pfam matched, and among them 2,000 genomic loci were alternatively spliced. We compared the Pfam of protein sequences encoded by the isoforms in these genomic loci subject to AS with at least one isoform having Pfam. Within these loci there were 13,914 transcripts encoded ORFs had Pfam match, however, among them 4,867 transcripts encoded ORFs with different Pfam in the same loci, and 4,485 transcripts encoded ORFs lost the Pfam, i. e. a functional domain (Supplementary Table 3). However, the impacts of AS events on the functionalities of different protein isoforms need to be validated experimentally.

To compare the AS rates of genes encoding different Pfam we extracted only one Pfam annotation for genes having multiple isoforms. Among a total of 9,939 genes (genomic loci), 2941 of them encoded at least one ORF matching to protein families. Among them 2,000 (68.0%) genes were alternatively spliced, though different gene families had variable AS rates (Table 5; Supplementary Table 4). The observed much higher AS rates in these protein coding genes having Pfam matches, particularly carbohydrate-active enzymes (CAZymes), indicate AS playing important roles in regulation of various types of cellular processes. For example, based on the CAZy classification we identified 84 genes encoding different families of CAZymes and found 52 of them were alternatively spliced (Table 6) (http://www.cazy.org/Home.html) (Drula et al., 2022). The functions of isoforms need to be further investigated in regards of carbohydrate metabolism in the fermentation process such as for biofuel production (Borin et al, 2017; Daly et al. 2017).

3.4 Dynamic changes of AS events in response to different growth conditions

Gene expression is dynamically regulated by the compositional changes in the growth media. Daly et al. (2017), Borin et al. (2017) and van Munster et al. (2020) reported the gene expression changes of CAZymes, sugar transporters, and transcription factors and other proteins related to lignocellulose degradation in response to different growth substrates including wheat straw, feedstock *Miscanthus*, and sugarcane bagasse, respectively. Here we use data collected by Daly et al. (2017) and van Munster et al. (2020) to demonstrate the dynamic changes of AS events in gene expression. The treatments of growth substrates included glucose-rich conditions (GLU, control), hydrothermal pretreated *Miscanthus* (HTM), hydrothermal pretreated wheat straw (HTS), ionic liquid pretreated *Miscanthus* (ILM), ionic liquid pretreated straw (ILS), knife-milled *Miscanthus* (KMM), and

knife-milled wheat straw (KMS) (Daly et al. 2017; van Munster et al. 2020). We combined all samples collected over a 5-day time course for each treatment for identification AS events, the results were summarized (Table 7). The differences in total numbers of AS events were mostly caused by the data size differences in these treatments (R^2 = 0.9603) (Figure 1). The similar trend of positive relationship between numbers of AS events and mapped RNA-seq reads was reported in our previous analysis with data collected from potato plants (Lee and Min 2023). To detect the effects of different treatments on AS events, we carried out pairwise comparisons of AS events in these treatments of combined data (Figure 2-4). Clearly, AS events were dynamically changed with different treatment of growth substrates in glucose, differently pretreated *Miscanthus*, and differently pretreated wheat straws. There were conserved AS events among all treatments as well as treatments specifically generated AS events (Figure 2-4). Daly et al. (2017) and van Munster et al. (2020) reported dynamic transcriptomic expressions of CAZymes at gene levels in different pretreated growth substrates over a time course. Our analysis demonstrated AS events were also dynamically regulated in these treatments. More detailed analysis would allow to identify treatment specific AS events and novel isoforms for further exploration of novel enzymatic activities for applications in the biofuel production.

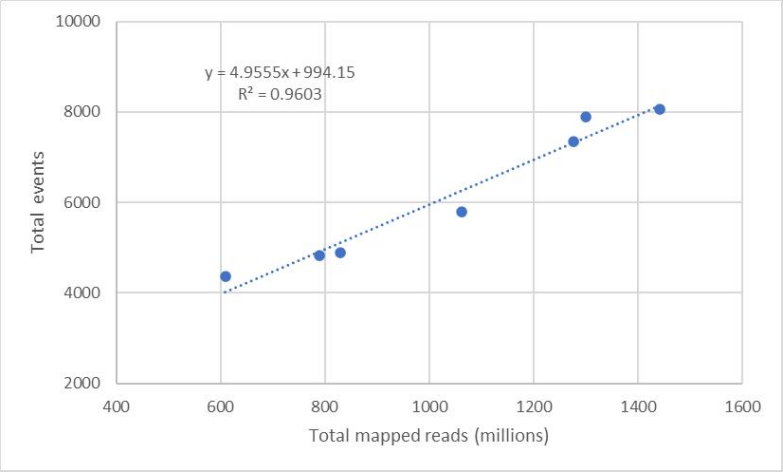

Figure 1 Relationships of total mapped reads and total alternative splicing events identified in seven different treatments of growth substrates in *A. niger*

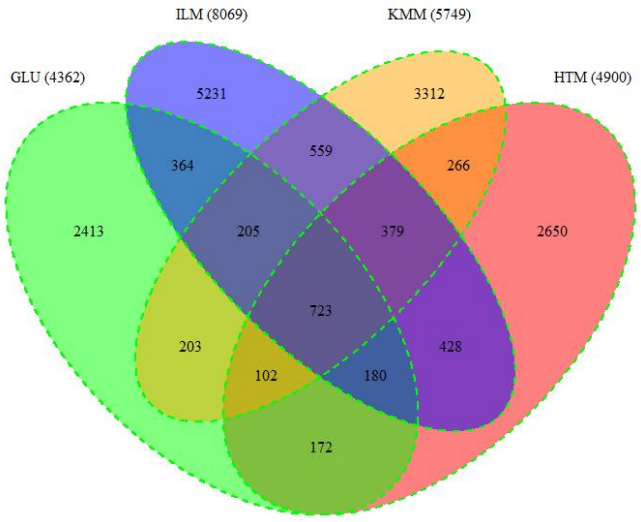

Figure 2 Alternative splicing events identified in glucose (GLU), hydrothermal pretreated *Miscanthus* (HTM), ionic liquid pretreated *Miscanthus* (ILM), and knife-milled *Miscanthus* (KMM) as substrates for *A. niger* culture

Table 5 Genomic loci encoding major protein families have variable alternatives splicing rates *

Pfam	Domain	Domain Description	Total loci	AS loci	AS (%)
pfam07690	MFS_1	Major facilitator superfamily	101	64	63
pfam04082	Fungal_trans	Fungal specific transcription factor	75	60	80
pfam00067	p450	Cytochrome P450	58	38	66
pfam00069	Pkinase	Protein kinase domain	45	32	71
pfam00083	Sugar_tr	Sugar (and other) transporter	44	32	73
pfam00106	adh_short	short chain dehydrogenase	41	22	54
pfam12796	Ank_2	Ankyrin repeats (3 copies)	36	24	67
pfam00172	Zn_clus	Fungal Zn(2)-Cys(6) binuclear cluster domain	29	25	86
pfam11951	Fungal_trans_2	Fungal specific transcription factor	29	26	90
pfam01494	FAD_binding_3	FAD binding domain	23	16	70
pfam00109	ketoacyl-synt	Beta-ketoacyl synthase, N-terminal	22	17	77
pfam00501	AMP-binding	AMP-binding enzyme	20	12	60
pfam08240	ADH_N	Alcohol dehydrogenase GroES-like domain	20	16	80
pfam00135	COesterase	Carboxylesterase family	18	10	56
pfam13561	adh_short_C2	Enoyl-(Acyl carrier protein) reductase	17	11	65
pfam09770	PAT1	Topoisomerase II-associated protein PAT1	16	13	81
pfam00005	ABC_tran	ABC transporter	15	12	80
pfam05199	GMC_oxred_C	GMC oxidoreductase	15	7	47
pfam05368	NmrA	NmrA-like family	15	7	47
pfam00107	ADH_zinc_N	Zinc-binding dehydrogenase	14	9	64
pfam00743	FMO-like	Flavin-binding monooxygenase-like	14	10	71
pfam01565	FAD_binding_4	FAD binding domain	14	10	71
pfam00076	RRM_1	RNA recognition motif	12	11	92
pfam00153	Mito_carr	Mitochondrial carrier protein	12	9	75
pfam00155	Aminotran_1_2	Aminotransferase class I and II	12	6	50
pfam00270	DEAD	DEAD/DEAH box helicase	12	6	50
pfam04479	RTA1	RTA1 like protein	12	5	42
pfam06985	HET	Heterokaryon incompatibility protein (HET)	12	8	67
pfam07519	Tannase	Tannase and feruloyl esterase	12	2	17
pfam13520	AA_permease_2	Amino acid permease	12	10	83
pfam13649	Methyltransf_25	Methyltransferase domain	12	6	50
pfam00171	Aldedh	Aldehyde dehydrogenase family	11	6	55
pfam01490	Aa_trans	Transmembrane amino acid transporter protein	11	7	64
pfam00400	WD40	WD domain, G-beta repeat	10	8	80
pfam00561	Abhydrolase_1	Alpha/beta hydrolase fold	10	6	60
pfam01425	Amidase	Amidase	10	8	80
pfam07992	Pyr_redox_2	Pyridine nucleotide-disulphide	10	5	50
Total	-	-	2941	2000	68

Note: * Partial list. The full list can be found in supplementary table 4

Table 6 Majority of genes of CAZymes related to biomass degradation undergo alternative splicing in *A. niger*

Pfam	Domain	Domain description	Total loci	AS loci
pfam00734	CBM_1	Fungal cellulose binding domain	2	0
pfam01607	CBM_14	Chitin binding Peritrophin-A domain	1	0
pfam14683	CBM-like	Polysaccharide lyase family 4, domain III	1	1
pfam17132	Glyco_hydro_106	alpha-L-rhamnosidase	1	1
pfam00457	Glyco_hydro_11	Glycosyl hydrolases family 11	1	0
pfam03537	Glyco_hydro_114	Glycoside-hydrolase family GH114	1	1
pfam00722	Glyco_hydro_16	Glycosyl hydrolases family 16	3	2
pfam00704	Glyco_hydro_18	Glycosyl hydrolases family 18	4	3
pfam00703	Glyco_hydro_2	Glycosyl hydrolases family 2	1	0
pfam02156	Glyco_hydro_26	Glycosyl hydrolase family 26	1	0
pfam00295	Glyco_hydro_28	Glycosyl hydrolases family 28	7	4
pfam00933	Glyco_hydro_3	Glycosyl hydrolase family 3 N terminal	5	3
pfam01915	Glyco_hydro_3_C	Glycosyl hydrolase family 3 C-terminal	5	3
pfam01055	Glyco_hydro_31	Glycosyl hydrolases family 31	6	5
pfam00251	Glyco_hydro_32N	Glycosyl hydrolases family 32	2	1
pfam01301	Glyco_hydro_35	Glycosyl hydrolases family 35	2	1
pfam04616	Glyco_hydro_43	Glycosyl hydrolases family 43	2	2
pfam01532	Glyco_hydro_47	Glycosyl hydrolase family 47	1	1
pfam07745	Glyco_hydro_53	Glycosyl hydrolase family 53	1	1
pfam01341	Glyco_hydro_6	Glycosyl hydrolases family 6	2	2
pfam03443	Glyco_hydro_61	Glycosyl hydrolase family 61	3	2
pfam03664	Glyco_hydro_62	Glycosyl hydrolase family 62	1	0
pfam03200	Glyco_hydro_63	Glycosyl hydrolase family 63 C-terminal	1	1
pfam03636	Glyco_hydro_65N	Glycosyl hydrolase family 65	1	1
pfam07477	Glyco_hydro_67C	Glycosyl hydrolase family 67	1	1
pfam00840	Glyco_hydro_7	Glycosyl hydrolase family 7	2	2
pfam03659	Glyco_hydro_71	Glycosyl hydrolase family 71	4	3
pfam03198	Glyco_hydro_72	Glucanosyltransferase	3	1
pfam07335	Glyco_hydro_75	Fungal chitosanase of glycosyl hydrolase	1	0
pfam03663	Glyco_hydro_76	Glycosyl hydrolase family 76	2	0
pfam07470	Glyco_hydro_88	Glycosyl Hydrolase Family 88	1	0
pfam07971	Glyco_hydro_92	Glycosyl hydrolase family 92	2	0
pfam11790	Glyco_hydro_cc	Glycosyl hydrolase catalytic core	1	0
pfam01793	Glyco_transf_15	Glycolipid 2-alpha-mannosyltransferase	1	1
pfam00982	Glyco_transf_20	Glycosyltransferase family 20	2	2
pfam01755	Glyco_transf_25	Glycosyltransferase family 25 (LPS)	1	1
pfam03033	Glyco_transf_28	Glycosyltransferase family 28	2	2
pfam05637	Glyco_transf_34	galactosyl transferase GMA12/MNN10	2	1
pfam13844	Glyco_transf_41	Glycosyl transferase family 41	1	1
pfam04666	Glyco_transf_54	N-Acetylglucosaminyltransferase-IV	1	1
pfam05686	Glyco_transf_90	Glycosyl transferase family 90	1	0
pfam14099	Polysacc_lyase	Polysaccharide lyase	1	1
Total	-	-	84	52

Table 7 Identification of alternative splicing events in different growth conditions in *A. niger*

Treatment	AltA	AltD	IntronR	ExonS	Complex	Total
GLU	1340	599	983	215	1225	4362
HTM	1494	696	1056	272	1382	4900
HTS	1445	732	1033	267	1344	4821
ILM	2166	1101	1702	399	2701	8069
ILS	2214	1146	1605	402	2520	7887
KMM	1724	881	1249	295	1600	5749
KMS	2105	1073	1552	364	2253	7347
HTM+ILM+KMM	4057	2250	4808	720	8107	19942
HTS+ILS+KMS	4330	2394	4992	769	8757	21242

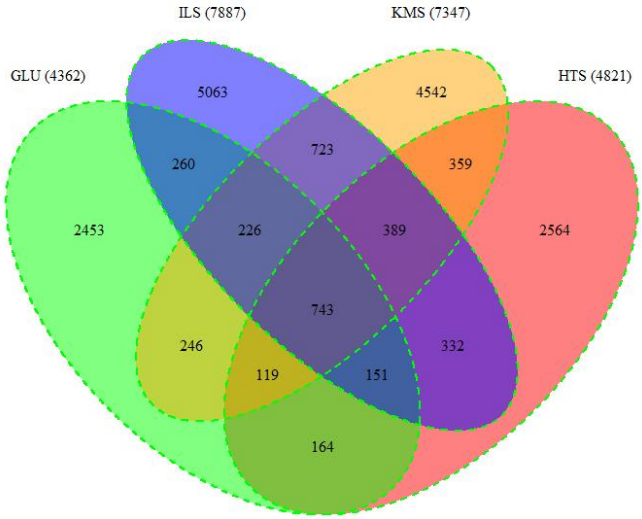

Figure 3 Alternative splicing events identified in glucose (GLU), hydrothermal pretreated wheat straw (HTS), ionic liquid pretreated straw (ILS), and knife-milled wheat straw (KMS) as substrates for *A. niger* culture

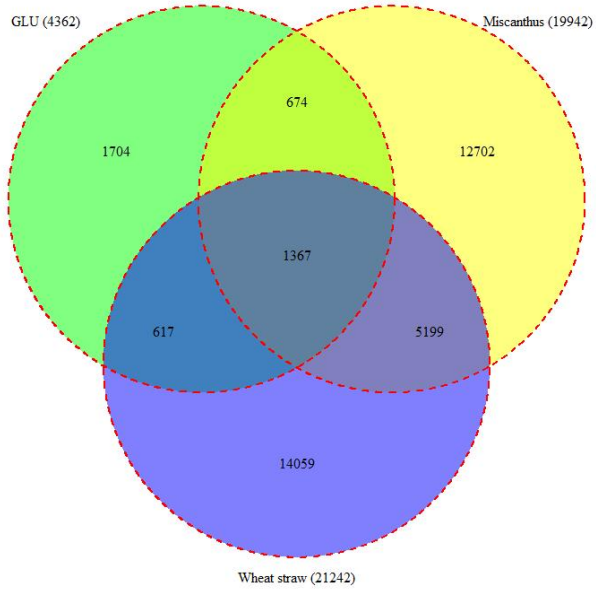

Figure 4 Alternative splicing events identified in glucose (GLU), wheat straw, and *Miscanthus* as substrates for *A. niger* culture

4 Discussion

We integrated mapping information of more than 10 billion of RNA-seq reads generated from 303 samples collected from eight recently published articles with 19,571 assembled transcripts of mRNA mapping information for identification of genome-wide AS events in *A. niger*. To our knowledge, this is the first large scale genome-wide meta-analysis of AS events in *A. niger* using RNA-seq data collected from various growth substrates. Combining all the mapping data generated a total of 9,939 genomic loci with a total of 66,007 transcripts assembled. A total of 63,715 AS events were identified from 4,972 genomic loc involving 43,156 unique transcripts. The AS rate based on current work was estimated to be about 50.0% in *A. niger*. We expect that more AS events and a higher AS rate can be obtained when more RNA-seq or transcripts data are integrated for genome mapping in *A. niger* in future. The impact of AS events on the encoded protein functions including enzymes needs to be evaluated individually. Our data including the assembled transcript sequences and mapping files are publicly available for the community to experimentally verify the identified isoform sequences and explore their functional novelties of enzymes for bioprocessing applications.

The work represents a large scale of genome-wide systematic identification of alternatively spliced genes and isoforms in *A. niger*. As in our analysis showed AS in genes in fungal species may be a common process, we recommend that researchers working in fungal species consider AS analysis when performing transcriptomic studies. However, it should be noted that these isoform transcript sequences were assembled by Cufflinks, validation by RT-PCR or cloning the full-length of mRNA transcripts are needed for further detailed functional analysis. The work carried out by Xu et al. (2024) in lignocellulos-degrading enzyme genes and enzyme variants in *A. niger* can serve as an example for this type of analysis. Computational identification of genes undergoing AS and annotation of their associated transcript isoform sequences are useful for researchers to design more specific experiments to examine the functions of genes of interests. The current work provides an important resource for investigating alternatively spliced genes and their associated functions of protein isoforms in *A. niger*. The data are expected to be useful in identifying homologous alternatively spliced genes in other fungal species in future research.

Author Contributions

XM designed the experiments. XM and FY provided the methodology, software support, and data analysis. XM, CJ, BL, MC, BW, CD, AP, PM, and KA carried out RNA-seq and mRNA transcript data collection and genome mapping. XM and FY prepared the manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

The Ohio Supercomputer Center provided computational resources for part of data processing.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Boel E., Hjort I., Svensson B., Norris F., Norris K.E., Fiil N.P., 1984, Glucoamylases G1 and G2 from *Aspergillus niger* are synthesized from two different but closely related mRNAs, EMBO J., 3(5): 1097-1102.
https://doi.org/10.1002/j.1460-2075.1984.tb01935.x

Borin G.P., Sanchez C.C., de Souza A.P., de Santana E.S., de Souza A.T., Leme A.F., Squina F.M., Buckeridge M., Goldman G.H., Oliveira J.V., 2015, Comparative secretome analysis of *Trichoderma reesei* and *Aspergillus niger* during growth on sugarcane biomass, PLoS One, 10(6): e0129275.
https://doi.org/10.1371/journal.pone.0129275

Cairns T.C., Nai C., Meyer V., 2018, How a fungus shapes biotechnology: 100 years of *Aspergillus niger* research, Fungal Biol. Biotechnol., 5: 1-4.
https://doi.org/10.1186/s40694-018-0054-5

Chaudhary S., Khokhar W., Jabre I., Reddy A.S., Byrne L.J., Wilson C.M., and Syed N.H., 2019, Alternative splicing and protein diversity: plants versus animals, Front. Plant Sci., 10: 708.
https://doi.org/10.3389/fpls.2019.00708

Clark S., Yu F., Gu L., and Min X.J., 2019, Expanding alternative splicing identification by integrating multiple sources of transcription data in tomato, Front. Plant Sci., 10: 689.
https://doi.org/10.3389/fpls.2019.00689

Daly P., Van Munster J.M., Blythe M.J., Ibbett R., Kokolski M., Gaddipati S., Lindquist E., Singan V.R., Barry K.W., Lipzen A., Ngan C.Y., 2017, Expression of *Aspergillus niger* CAZymes is determined by compositional changes in wheat straw generated by hydrothermal or ionic liquid pretreatments, Biotechnol. Biofuels Bioprod., 10(1): 1-9.
https://doi.org/10.1186/s13068-017-0700-9

Drula E., Garron M.L., Dogan S., Lombard V., Henrissat B., Terrapon N., 2022, The carbohydrate-active enzyme database : functions and literature, Nucleic Acids Res., 50: D571-D577.
https://doi.org/10.1093/nar/gkab1045

Fang S., Hou X., Qiu K., He R., Feng X., Liang X., 2020, The occurrence and function of alternative splicing in fungi, Fungal Biol. Rev., 34(4): 178-188.
https://doi.org/10.1016/j.fbr.2020.10.001

Florea L., Hartzell G., Zhang Z., Rubin G.M., Miller W., 1998, A computer program for aligning a cDNA sequence with a genomic DNA sequence, Genome Res., 8: 967-974.
https://doi.org/10.1101/gr.8.9.967

Foissac S., and Sammeth M., 2007, Astalavista: dynamic and flexible analysis of alternative splicing events in custom gene datasets, Nucleic Acids Res., 35: W297-299.
https://doi.org/10.1093/nar/gkm311

Garrigues S., Kun R.S., Peng M., Bauer D., Keymanesh K., Lipzen A., Ng V., Grigoriev I.V., de Vries R.P., 2022, Unraveling the regulation of sugar beet pulp utilization in the industrially relevant fungus *Aspergillus niger*, Iscience, 25(4): 104065.
https://doi.org/10.1016/j.isci.2022.104065

Grützmann K., Szafranski K., Pohl M., Voigt K., Petzold A., Schuster S., 2014, Fungal alternative splicing is associated with multicellular complexity and virulence: a genome-wide multi-species study, DNA Res., 21(1): 27-39.
https://doi.org/10.1093/dnares/dst038

Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R., and Salzberg S.L., 2013, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, Genome Biol., 14: R36.
https://doi.org/10.1186/gb-2013-14-4-r36

Kun R.S., Salazar-Cerezo S., Peng M., Zhang Y., Savage E., Lipzen A., Ng V., Grigoriev I.V., de Vries R.P., Garrigues S., 2023, The amylolytic regulator AmyR of *Aspergillus niger* is involved in sucrose and inulin utilization in a culture-condition-dependent manner, J. Fungi., 9(4): 438.
https://doi.org/10.3390/jof9040438

Lee J.A., Min X., 2023, Comparative analysis of alternative splicing events in foliar transcriptomes of potato plants inoculated with *Phytophthora infestans*, Comput. Mol. Biol., 13(1): 1-8.
https://doi.org/10.5376/cmb.2023.13.0001

Liu X.Y., Fan L., Gao J., Shen X.Y., Hou C.L., 2020, Global identification of alternative splicing in *Shiraia bambusicola* and analysis of its regulation in hypocrellin biosynthesis, Appl. Microbiol. Biotechnol., 104: 211-223.
https://doi.org/10.1007/s00253-019-10189-3

Mattison C.P., Mack B.M., Cary J.W., 2021, Comparative transcriptomic analysis of *Aspergillus niger* cultured in peanut or cashew nut flour based media, J. Appl. Biol. Biotechnol., 9(5): 56-63.

Min X.J., 2013, ASFinder: a tool for genome-wide identification of alternatively spliced transcripts from EST-derived sequences, International J. Bioinformatics Res. Appl., 9: 221-226.
https://doi.org/10.1504/IJBRA.2013.053603

Min X.J., Butler G., Storms R., Tsang A., 2005a, OrfPredictor: predicting protein-coding regions in EST-derived sequences, Nucleic Acids Res., 33: W677-680.
https://doi.org/10.1093/nar/gki394

Min X.J., Butler G., Storms R., Tsang A., 2005b, TargetIdentifier: a web server for identifying full-length cDNAs from EST sequences, Nucleic Acids Res., 33: W669-W672.
https://doi.org/10.1093/nar/gki436

Pan Q., Shai O., Lee L.J., Frey B.J., Blencowe B.J., 2008, Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing, Nat. Genet., 40: 1413-1415.
https://doi.org/10.1038/ng.259

Pel H.J., de Winde J.H., Archer D.B., Dyer P.S., Hofmann G., Schaap P.J., Turner G., de Vries R.P., Albang R., Albermann K., Andersen M.R., 2007, Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88, Nat. Biotechno., 25(2): 221-231.
https://doi.org/10.1038/nbt1282

Semova N., Storms R., John T., Gaudet P., Ulycznyj P., Min X.J., Sun J., Butler G., Tsang A., 2006, Generation, annotation, and analysis of an extensive *Aspergillus niger* EST collection, BMC Microbiol., 6: 1-10.
https://doi.org/10.1186/1471-2180-6-7

Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., Van Baren M.J., Salzberg S.L., Wold B.J., and Pachter L., 2010, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol., 28: 511-515.
https://doi.org/10.1038/nbt.1621

Van Leeuwe T.M., Arentshorst M., Forn-Cuní G., Geoffrion N., Tsang A., Delvigne F., Meijer A.H., Ram A.F., Punt P.J., 2020, Deletion of the *Aspergillus niger* pro-protein processing protease gene kexB results in a pH-dependent morphological transition during submerged cultivations and increases cell wall chitin content, Microorganisms, 8(12): 1918.
https://doi.org/10.3390/microorganisms8121918

Van Munster J.M., Daly P., Blythe M.J., Ibbett R., Kokolski M., Gaddipati S., Lindquist E., Singan V.R., Barry K.W., Lipzen A., Ngan C.Y., 2020, Succession of physiological stages hallmarks the transcriptomic response of the fungus *Aspergillus niger* to lignocellulose, Biotechnol. Biofuels., 13: 1-6.
https://doi.org/10.1186/s13068-020-01702-2

Xu Y., Dong F., Wang R., Ajmal M., Liu X., Lin H., Chen H., 2024, Alternative splicing analysis of lignocellulose-degrading enzyme genes and enzyme variants in *Aspergillus niger*, Appl. Microbiol. Biotechnol., 108(1): 1-11.
https://doi.org/10.1007/s00253-024-13137-y

Zhang R., Calixto C.P., Marquez Y., Venhuizen P., Tzioutziou N.A., Guo W., Spensley M., Entizne J.C., Lewandowska D., Ten Have S., Frei Dit Frey N., Hirt H., James A.B., Nimmo H.G., Barta A., Kalyna M., Brown J.W.S., 2017, A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing, Nucleic Acids Res., 45: 5061-5073.
https://doi.org/10.1093/nar/gkx267