Alternative splicing in the pathogenic fungus *Aspergillus fumigatus*

Keila Velazquez-Arcelay[1] and Antonis Rokas[1]

[1] Department of Biological Sciences, Vanderbilt University, Nashville TN

ABSTRACT

In eukaryotic organisms, the total number of genomic coding regions and the amount of expressed mRNA exist at unequal rates. The alternative splicing machinery can assemble different variants from a single gene, resulting in a more diverse and complex proteome. This higher variation can contribute to evolutionary processes, while it also can facilitate the emergence of pathogenic species. The frequency in which this regulatory mechanism occurs is thought to be associated with higher biological complexity. However, the extent of this process has not been widely explored in fungi and remains elusive. Here we examine the depth of alternative splicing in the fungus *Aspergillus fumigatus*, and how this could be associated with the evolution of pathogenicity. Transcriptomic data from a wild-type *A. fumigatus* strain Af293 grown at 37º was collected from a previous RNA-Seq experiment. The samples were mapped to a reference genome using HISAT2. Using StringTie, the mapped reads were assembled into transcripts. Gene isoforms resulting from alternative splicing were quantified and splicing type events were described. Gene ontology (GO) analyses show enrichment for genes involved in adaptation to pathogenic lifestyles. Alternative splicing events, as well as their functional impact, were analyzed by comparing transcripts to the annotated data. This design will estimate the level of splicing in *A. fumigatus*, the expression level of isoforms and their functional impact. It will also direct the identification of pathogenicity genes with higher splice events.

INTRODUCTION

Fungi have diverse and unique lifestyles. For instance, the digestion of their food takes place outside of the fungal body before consuming it. This mechanism is aided by the secretion of enzymes and acids into the environment. After the potential food has been degraded, the fungi can absorb the nutrients [1]. These organisms also employ secondary metabolites to survive in their niche and compete with other organisms. Some of those mycotoxins include polyketides such as aflatoxin, epipolythiodioxopiperazine (ETP) like gliotoxin and sirodesmin, ochratoxins and citrinin [2].

The mechanism and ability to synthesize a variety of molecules that support the survival in hostile environments has probably been one of the bases for the evolution of some of the most pathogenic organisms we know. Fungi in the *Aspergillus* genus are especially recognized as the most pathogenic. Aspergillosis is a range of diseases associated with these molds and spread airborne, and can show some resistance to azole antifungals. *Aspergillus fumigatus* is the most pathogenic of these species, responsible for 90% of the infections [3].

Living inside a host organism poses many challenges to pathogens. In order to evade the immune system, reactive oxidative species (ROS) and heat shock stress, mechanisms to surpass these stress responses are essential [4]. Glucuronoxylomannan (GMX) is involved in the the ability of these organisms to switch from dimorphic to either filamentous or unicellular growth to increase the chances of invading the host and also their survivability against the threats posed to living inside a host. Many other genes involved in pathogenicity in *A. fumigatus* were reported by Abad et al. [5]. The most relevant groups are involved in thermotolerance, cell wall composition, immune response avoidance, toxin production, nutrient acquisition, and response to stress.

Databases such as FSRD (fungal stress response database) [6] and DFVF (database of fungal virulence factors) [7] were developed to provide extensive collections of genomic elements responsible for this fungal lifestyle. These collections contain genes involved in adaptation to stress that could lead to better survival in their environment and inside other organisms. Other tools like FungalRV[8] and FaaPred (fungal adhesins and adhesin-like proteins) [9] were developed to identify putative

adhesin proteins, which help the organism adhere to its target substrate, and can even aid in invasion of host species.

Alternative splicing generates higher variation in expression from what would be a more limited amount of coding genes in a genome. This process has been associated with higher biological complexity and multicellularity. In fungi, alternative splicing is thought to be limited [10]. However, more studies are needed to better understand the extent impact it has on these organisms. A previous study has suggested that alternative splicing not only is involved in enhancing the evolutionary dynamics in fungi, but also in regulating pathogenicity and virulence [4].

To better understand the role of alternative splicing as a mechanism for adaptation and survival to their environment, we aim to analyze the expressed genes from wild-type *A. fumigatus* cultures grown at a physiological temperature. Alternative splicing provides higher variation in expression through the transcription of various of isoforms from a single gene. If this process has been involved in driving the evolution of fungi to acquire their current lifestyle, we should find enrichment for pathogenicity genes isoforms.

METHODS

Data

Reference genome sequences of *Aspergillus fumigatus* (strain Af293) were extracted from the FungiDB web resource (http://fungidb.org/common/downloads/release-2.0/Afumigatus_AF293B/), as well as the genome annotations in GFF format. For compatibility reasons, the GFF file was converted to GTF using AGAT (https://github.com/NBISweden/AGAT). To analyze transcript data, we used RNA-Seq reads published by a previous *Aspergillus* study (11). The data were retrieved from the NCBI Short Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra/?term=SRP080951) using the SRA Toolkit' prefetch (https://github.com/ncbi/sra-tools/wiki). RNA-seq reads were collected from two wild-type samples of fungi cultures grown at 37º (accession number: SRP080951). The SRA files were converted to FASTQ format using the SRA Toolkit's fastq-dump tool (https://doi.org/10.1093/nar/gkq1019).

Genome mapping

A reference index was generated using the HISAT2 package (https://daehwankimlab.github.io/hisat2/). The index was built from the genome annotations and also from splice-site and exon data extracted using the previously mentioned package. The raw reads were aligned to the reference genome using the default parameters. the alignment was sorted and converted to the binary BAM format for both replicate files using Samtools (http://www.htslib.org).

Transcript assembly and quantification of gene expression

The output BAM files obtained from the mapped reads were processed with StringTie (http://ccb.jhu.edu/software/stringtie/, https://github.com/gpertea/stringtie)to assemble them into transcripts. A warning appeared indicating that some transcripts were mapped but not annotated from the reference annotation. These transcripts are referred to as *de novo* transcripts. The resulting GTF outputs from both *A. fumigatus* replicates were merged into a single file, summarizing the mapped loci. The stringtie_merged.gtf file was then used as input to generate transcript abundance table counts for each replicate. Reads shorter than 100bp were removed from the list.

How the transcripts compare with the reference annotation

The GffCompare tool was used to collapse combined transcripts from the two replicates merged in the previous step, making each recorded transcript unique. It takes as input the StringTie merged GTF file, and it generates several outputs, including: merged.annotated.gtf; merged.loci; merged.stats; merged.stringtie_merged.gtf.refmap; merged.stringtie_merged.gtf.tmap; merged.tracking.

***De novo* annotation**

Cufflinks was used to generate a Fasta file for the identification of de novo transcripts through a BLASTn. The gffread tool from this package was used to convert the *de novo* transcripts file from GTF to Fasta format. A local BLAST database was generated in the Vanderbilt ACCRE HPC cluster using the BLAST tool makeblastdb. The database was built using the reference sequences from AspGD (http://www.aspergillusgenome.org/download/sequence/A_fumigatus_Af293). The

transcripts were identified with a BLASTn. Protein coding transcripts were also identified by running a BLASTx.

Alternative splice type distribution

Splice types across all the transcripts were identified using the AStalavista tool (http://genome.crg.es/astalavista/; this tool is no longer available and the analysis will be updated using Sailfish or Salmon, https://doi.org/10.4172/2469-9853.1000140). The input file required to run this algorithm is an annotation file in GTF format; in this case I used the merged transcripts file obtained as an output from StringTie.

GO analysis

GO annotations for *Aspergilli* genes were extracted from the fetGOat website (http://software.broadinstitute.org/fetgoat/index.html). To retrieve the relevant GO IDs, the list of genes from the data analyzed in this project was matched against the downloaded annotations. The resulting list was processed with the GO Slimmer tool from AmiGO (http://amigo1.geneontology.org/cgi-bin/amigo/slimmer) to cluster each gene into their respective functional group.

RESULTS

Genome mapping and transcript assembly

The first sample had 13,010,484 total reads. Of those, 93.79% mapped one time to the genome, while 1.79% mapped more than once. The other 4.45% reads did not map to the genome, for a total of 95.55% alignment rate. The second sample had 13,199,804 total reads. Of those, 93.62% mapped one time to the genome, while 1.88% mapped more than once. The other 4.50% reads did not map to the genome, for a total of 95.50% alignment rate. Statistics from the merged samples show 52,876 exons, 24,337 of those (46%) being novel. The heavier weight of exon distribution falls on transcripts that include one (25%), two (29%), or three exons (20%). Fewer transcripts contain four (12%), five (7%) and six (3%) exons. A more detailed distribution of exon counts is reported in **Figure 1**. Out of 18,301 predicted loci, 8,404 (45.9%) were novel.

Additionally, 32,601 introns were predicted, with 13,978 (42.9%) being novel. These statistics were obtained using the tool GffCompare.

The total count of predicted genes is 10,024. A list of gene names was sorted and processed in TextWrangler to remove redundancy. This number is still higher than the number of genes in *A. fumigatus* and should be verified. A BLASTx search predicted 9,910 proteins from the annotated transcripts. The *de novo* BLAST results returned an incomplete process and an error; total genes 6,948. The sum of both sets of genes does not account of the total count, since they are not exclusive and could be duplicated.

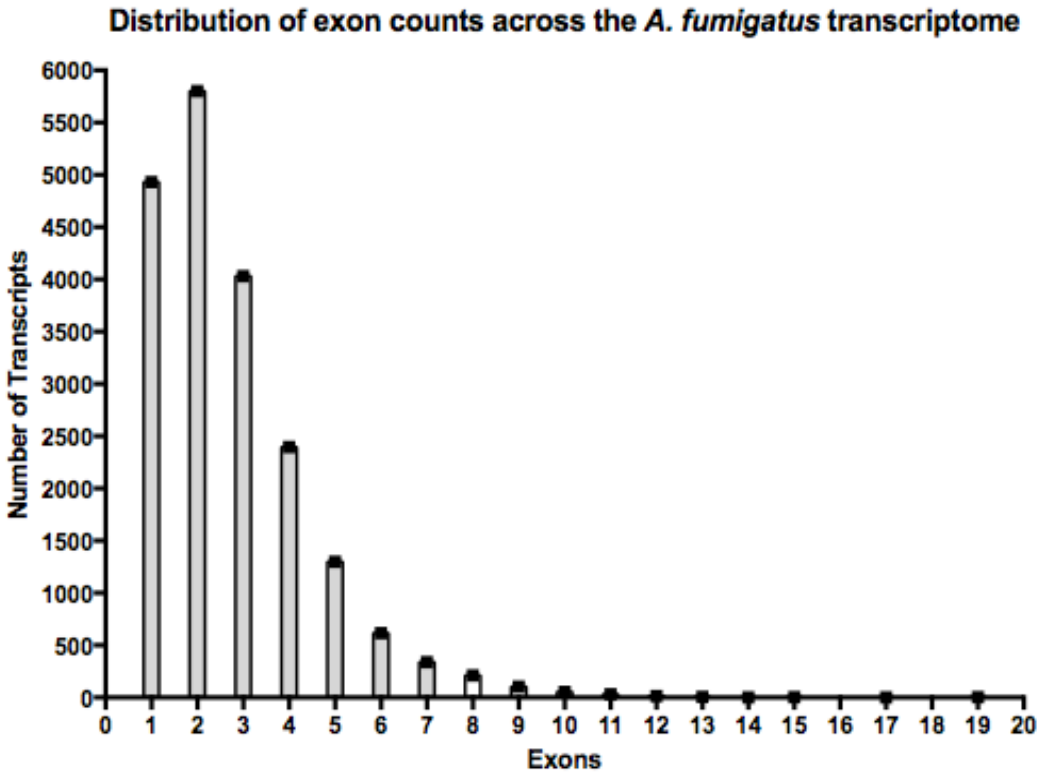

Figure 1. Distribution of exon counts across the *A. fumigatus* transcriptome. Most transcripts are composed of one, two, three or four exons.

De novo annotation

The unannotated transcripts were identified through a BLASTn, for which 6,948 different genes were retrieved from a total of 9,920 transcripts. For 61% of those genes no

isoform was reported, as shown in **Figure 2**. Most of the alternatively spliced genes had an isoform count ranging from two to four. Only 1% had five, six, or seven isoforms. This quantitative analysis was done using Unix command lines on the BLAST output files.

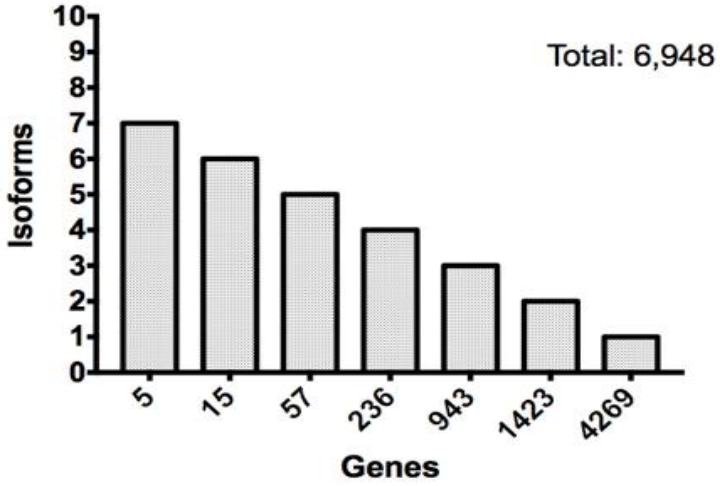

Figure 2. Gene isoform enrichment in de novo transcripts. A total of 6,948 genes represented in the *de novo* transcripts collapsed into

Alternative splice types distribution

The distribution of alternative splice types obtained from this study was verified against the pattern reported by Grützmann *et al*. [4] for different types of Ascomycetes. In both cases (**Figure 3**), around 60% of the splicing events are predicted as intron retention (IR). It also shows consistency in the higher rate of 3'splice junctions (alt acceptor) over 5'splice junctions (alt donor). Exon skipping events are rare in both cases.

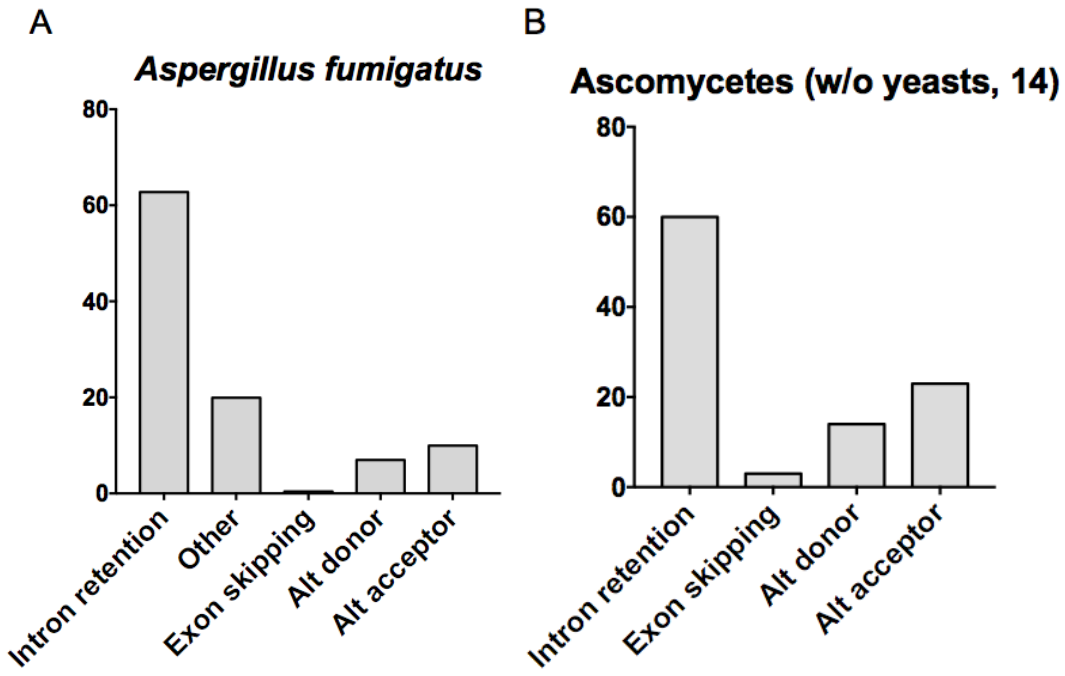

Figure 3. Alternative splice type distribution. (A) Splice type portions for the A. fumigatus samples studied in this project, compared to (B) distribution from a previous study published by Grützmann *et al*. [4].

GO analysis

GO Slimmer predicted enrichment for genes related to stress response, as seen from the biological processes and the molecular functions gene ontology analyses. Oxidoreductase and hydrolase activity related genes aid the pathogen in its adaptation and survival inside the host. A summary is reported in **Figure 4**.

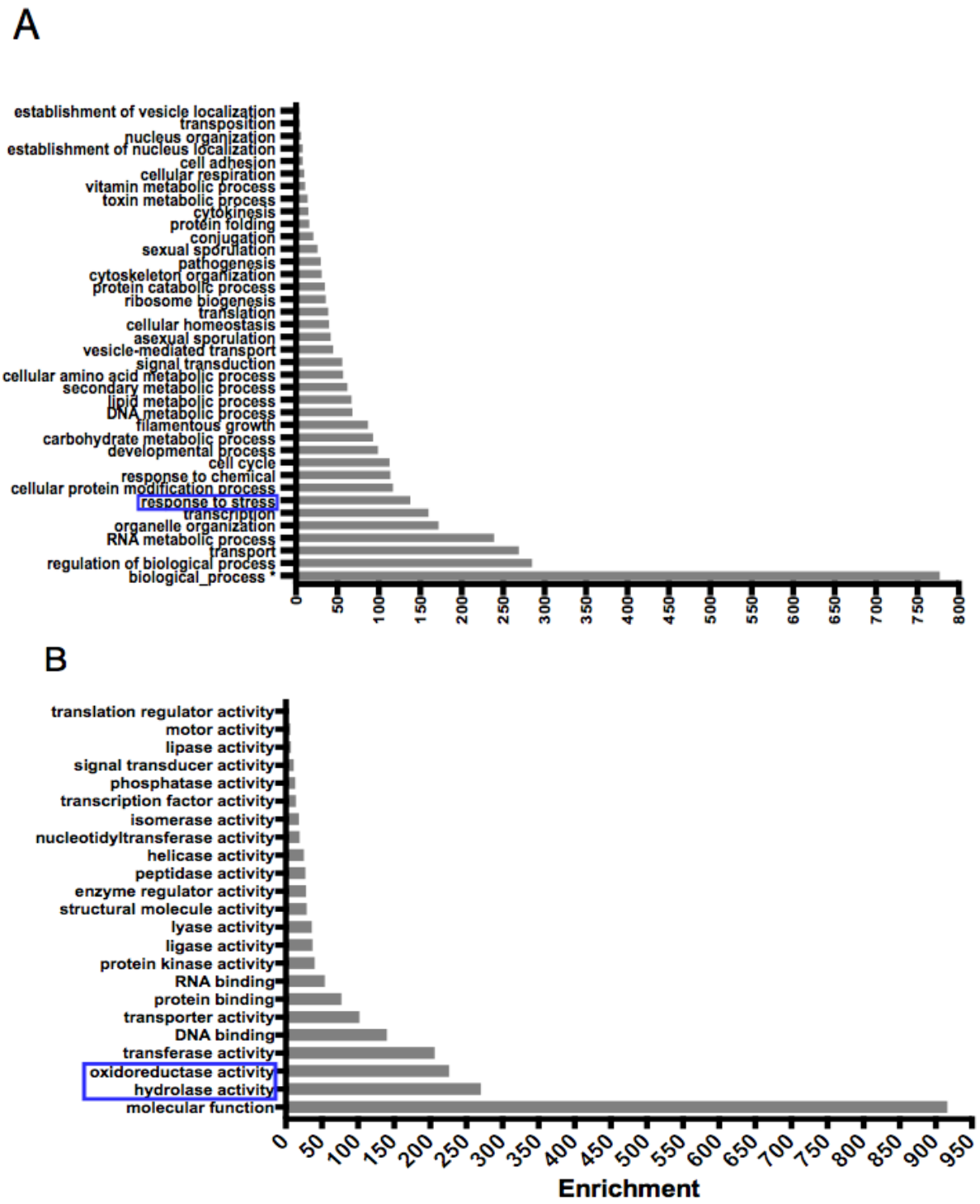

Figure 4. Gene Ontology (GO) enrichment analysis with the GO Slimmer software. (A) Biological process, response to stress enrichment. (B) Molecular function. The stress response categories for oxidoreductase activity and hydrolase activity show enrichment.

DISCUSSION

Aspergillus fumigatus strain Af293 contains a total of 9,916 genes and 57 pseudogenes, according to the NCBI genome database. A total of 10,024 genes were predicted in this study, after curating and annotating the transcripts. The source of error or redundancy that resulted in a higher number of genes should be looked into. Almost half of the transcripts did not map to the reference annotations and were thus classified as *de novo*. This is the first of many results that evidenced for the need of wider annotation coverage for fungi species.

The output files obtained from the initial data processing steps were often modified and reformatted using Unix commands and regular expressions to meet the required needs for each of the subsequent analyses. Future analyses could be accelerated with the development of python scripts for specific steps where there is heavy use of data editing in the way mentioned above.

The number of exons in a gene is representative of the grounds and constraints for alternative splicing. The data for exon count per gene is currently not available, but having it would be useful. It would make possible to see the relationship between that type of constraint and the resulting distribution of transcript types. The observed distribution of exon counts per transcript indicate that there are indeed grounds for alternative splicing in the *Aspergillus* genome. It would be interesting to see the quantification of how many one- or two-exon transcripts originate from three- or more-exon genes.

Ascomycetes, excluding yeasts, have shown a splice type distribution of around 60% intron retention (IR), over 20% alternative 3' splice site (A3'SS), 15% alternative 5' splice sites (A5'SS), and less than 5% exon skipping (SE) events [4]. The results in this analysis results show consistency with the previously reported ratios. This indicates that there is a pattern in the distribution of splice types across fungi species. It also indicates that, although this initial analysis was mainly focused on arranging a pipeline to be used for further analyses rather than stopping more often to check for quality, the data still shows consistent results. Many alternative splicing quantification packages built on R or python are available. It’s necessary to identify the tool that is more efficient for the purpose of this project and run a wider analysis for alternative splicing quantification.

While almost 40% of the genes represented in the *de novo* transcripts have isoforms, not a single transcript matching the genomic annotation showed alternative splicing. Given that there is evidence that alternative splicing does happen in *Aspergilli*, there are reasons to doubt that this outcome is real. For that reason, I have two hypotheses. Perhaps the annotation file acquired from the FungiDB website contains only a limited amount of annotations for gene isoforms. I verified that the annotation database contains no less than 18,600 duplicated gene names (around a half of the total database transcripts), confirming that isoforms are listed in this file. Another possibility lies in the parameters used to run the HISAT2 genome mapping or the StringTie transcript assembly, which could have reduced duplicated gene names. To verify the veracity of these results, the reads should be mapped again. Running HISAT2 without relying on the annotation file will retrieve only *de novo* transcripts, all of which can be annotated together using BLAST.

Future steps in this project should be directed at identifying and annotating pathogenicity genes that are expressed in *A. fumigatus*. Many of those genes have been listed by Abad *et al*. (5). More genes can be identified from the functional clusters generated by GO terms. A relationship between isoform enrichment and identification of functional groups related to response to stress and secondary metabolite production should be introduced to this process, to verify if alternative splicing has a role in the evolution and adaptation of pathogens. Better results will be obtained after the gene, protein and GO annotations for fungi are made more extensive in databases as NCBI and AspGD.

REFERENCES

[1] Bennett, J. W. (2010). An overview of the genus Aspergillus (pp. 1-17). Caiser Academic Press, Portland.

[2] Fox, E. M., & Howlett, B. J. (2008). Secondary metabolism: regulation and role in fungal biology. Current opinion in microbiology, 11(6), 481-487.

[3] Lin, S. J., Schranz, J., & Teutsch, S. M. (2001). Aspergillosis case-fatality rate: systematic review of the literature. Clinical Infectious Diseases, 32(3), 358-366.

[4] Grützmann, K., Szafranski, K., Pohl, M., Voigt, K., Petzold, A., & Schuster, S. (2013). Fungal alternative splicing is associated with multicellular complexity and virulence: a genome-wide multi-species study. DNA research, 21(1), 27-39.

[5] Abad, A., Fernández-Molina, J. V., Bikandi, J., Ramírez, A., Margareto, J., Sendino, J., ... & Rementeria, A. (2010). What makes Aspergillus fumigatus a successful pathogen? Genes and molecules involved in invasive aspergillosis. Revista iberoamericana de micologia, 27(4), 155-182.

[6] Karányi, Z., Holb, I., Hornok, L., Pócsi, I., & Miskei, M. (2013). FSRD: fungal stress response database. Database, 2013, bat037.

[7] Lu, T., Yao, B., & Zhang, C. (2012). DFVF: database of fungal virulence factors. Database, 2012, bas032.

[8] Chaudhuri, R., Ansari, F. A., Raghunandanan, M. V., & Ramachandran, S. (2011). FungalRV: adhesin prediction and immunoinformatics portal for human fungal pathogens. BMC genomics, 12(1), 192.

[9] Ramana, J., & Gupta, D. (2010). FaaPred: a SVM-based prediction method for fungal adhesins and adhesin-like proteins. PLoS One, 5(3), e9695.

[10] Irimia, M., Rukov, J. L., Penny, D., & Roy, S. W. (2007). Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. BMC Evolutionary Biology, 7(1), 188.

[11] Lind, A. L., Smith, T. D., Saterlee, T., Calvo, A. M., & Rokas, A. (2016). Regulation of secondary metabolism by the Velvet complex is temperature-responsive in Aspergillus. G3: Genes, Genomes, Genetics, 6(12), 4023-4033.