ABSTRACT:

The molecular basis of spider silk production is of broad interest because of its possible mechanical applications. For instance, dragline silk, which is produced in the major ampullate gland of certain spiders, has been found to be tougher than nylon and Kevlar®. However, even though there is research on the mechanical and structural properties of spider silk, the gene expression and regulation responsible for spider silk production remains largely unexplored. In this project, we tried to identify the genes that regulate spider silk production by analyzing 8 RNAseq libraries from silk glands of male and female *Dysdera* spiders. Using a reference transcriptome, a differential expression analysis is done to identify statistically relevant expressed genes. The programs BowTie2 and TopHat were used to perform alignment to the reference transcriptome. To perform differential expression analysis, tuxedo tool suite of programs (CuffLinks, CuffMerge, CuffCompare), HTSeq, and DESeq, were used to determine statistically significant differentially expressed genes. This analysis will provide insights regarding the genes that are either upregulated or downregulated during silk production in *Dysdera* spiders.
Introduction

Spider silk has become an important topic in numerous fields due to the potential applications that can be derived from its mechanical properties (reviewed in Ayoub et al. 2007). However, spiders possess more than one type of silk producing glands. It has been shown that Spiders can produce up to six to seven different types of silk from different types of glands, which can vary in tensile strength and elasticity (Lewis, 2006; Fu et al., 2009). One of the spider silk producing glands, the major ampullate gland, has been the focus of spider silk research due to its abundance and ease of access (Xu & Lewis, 1990).

Dragline silk, which is produced from the major ampullate gland, is known for its extremely high tensile strength and toughness (Ayoub et al., 2007). Its mechanical properties are often compared to synthetic fibers, which are known to have high modulus and strength. On the other hand, dragline silk, which possesses average strength, are significantly tougher than both nylon and Kevlar (Fu et al., 2009; Xu & Lewis, 1990). These properties are often associated with the proteins structures that constitutes spider silk.

The majority of research regarding spider silk proteins revolve around major ampullate spidroins 1 and 2 (MaSp1 and MaSp2, respectively) and these proteins are mostly found in silks produced in major ampullate glands (Xu and Lewis, 1990; Hinman and Lewis, 1992). Research has shown that proteins like MaSp2 and MaSp1 have high molecular weight. MaSp1 and MaSp2 is also shown to contain repetitive domains that is glycine-rich and poly-alanine at its end (Gatesy et al., 2001). These proteins contain different amino acid sequence variation in different spidroins are involved on the different strength and toughness of the fibers (Hayashi & Lewis, 1998; Hayashi et al., 1999; Rising et al., 2005; Gaines & Marcotte, 2008).

Although its mechanical and structural properties are extensively being studied, we have yet to discover the complete list of genes that are involved in spider silk gland development and its possible and crucial involvement in spider silk production. Studying the genes that are may be responsible for attaining the uniqueness of the gland may provide insight on how spider silk production is regulated. These genes may be involved in the intricate detailing of how spider silk are manufacture within the gland such as the details of the formation of the silk protein’s secondary structure, and folding and assembly of these silk proteins (Fu et al., 2009).

In this research, transcript activity from various silk glands of male and female Dysdera spiders will be analyzed in in silico. These collections of transcript activities were sequenced using
a next-generation sequencing technique called RNAseq. Differential expression analysis will be done using the pipeline for the tuxedo tools and R programming’s Bioconductor packages. The RNAseq libraries from male and female *Dysdera* spiders are first subjected to whole genome alignment using the programs BowTie2 and TopHat. The resulting data from TopHat was then subjected to the tuxedo pipeline for differential expression analysis: CuffLinks, CuffMerge, and CuffCompare. The results are then counted and visualized through the programs HTSeq and DESeq (Anders, Pyl and Huber, 2014; Anders and Huber, 2010).

Differential expression analysis will be done using the pipeline for tuxedo tools by first performing whole genome alignment using the programs BowTie2 and TopHat, and finally performing differential expression analysis using CuffLinks, CuffMerge, CuffDiff, and CummeRbund. A separate program, an R Bioconductor package named DESeq, will also be used for differential expression analysis. These methodologies will help visualize which genes are utilized during spider silk formation.

**Methods**

RNAseq libraries are formed from the obtained mRNA transcript activities of total glands of male and female *Dysdera* spiders. Due to the limitations of our current technology, sequencing full-length mRNA sequence have yet to be achieved and therefore transcripts are fragmented first before running through RNAseq (Fig 1). RNAseq then generates a library of these fragmented mRNA sequences (via their complementary DNA). Male and female *Dysdera* spiders were separated into different libraries (Fig 2).

Fig. 1. Sequencing of RNA transcript via RNAseq
Figure 2. Male and Female *Dysdera* libraries and their distinct library names

**I. Quality assessment analysis of the FastQ libraries using ShortRead**

The extracted mRNA sequences are then subjected to quality assessment analysis in order to determine if the FastQ libraries have a good quality before running them through the RNAseq analysis pipeline. The QA is performed using a package from R Bioconductor called ShortRead (Morgan et al., 2009).

**II. Analysis of gene expression via the Tuxedo Tools**

After the determination of the sequence quality using the QA analysis of ShortRead, the fragments contained inside the FastQ libraries are then subjected through the RNAseq analysis pipeline for differential gene expression analysis. First, in order to analyze gene expression, these fragments are then aligned with a reference transcriptome. The rationale is that, the more expressed a gene is, the more mRNA it produces, and therefore, the more fragments it will have prior to sequencing via RNAseq. However, since there are no currently existing good quality, fully sequenced genome for spiders, a reference transcriptome was constructed by PI Dr. Arensburger using the data from the RNAseq libraries and *de novo* constructing a reference transcriptome through a program called Trinity (Grabherr et al., 2011).

After the *de novo* construction of a reference transcriptome, the libraries are first subjected to one of the first line of the tuxedo tools. By first using the alignment tool BowTie2 (Langmead
et al., 2012), allows the confirmation of the following: the fragments contained in the FastQ libraries are able to align and the reference transcriptome can be used for the rest of the RNAseq analysis.

After subjecting them into the first alignment tool and confirming the following conditions stated above, the FastQ libraries are then subjected to a more sophisticated alignment tool of the tuxedo tools, TopHat (Trapnell et al., 2012). Using TopHat, the program considers possible exon-exon splicing junctions (Trapnell et al., 2012).

The TopHat results are then used for the following remaining tuxedo tools: CuffLinks, CuffMerge, and CuffCompare (Trapnell et al., 2012), to generate and identify sections of the transcriptome assembly that may be expressed. CuffLinks will use the TopHat aligned matches to quantify expression as well as make a transcript assembly for each library. CuffMerge will merge all these transcript assembly from the eight libraries to construct a master transcriptome. CuffCompare provides additionally comparison of the generated transcriptome to the alignment. After the utilization of the tuxedo tools, differential analysis is conducted.

III. Differential expression analysis using DESeq

After running the RNAseq data through the tuxedo pipeline, it is subjected to differential expression analysis using HTSeq and DESeq (Anders, Pyl and Huber, 2014; Anders and Huber, 2010). DESeq is an R Bioconductor package that is a visualization tool that allows biologists to see differential expression from obtained samples (Anders and Huber, 2010). However, a third-party program is required to count the frequency of the alignment from the alignment program used to the generated transcriptome assembly as DESeq does not possess such program capabilities. One of the most commonly used program in conjunction with DESeq is HTSeq. HTSeq quantifies the alignment of of the result from an alignment program (such as those of TopHat) to a transcriptome or generated transcriptome (which can be done through the Tuxedo Pipeline) (Anders, Pyl and Huber, 2014).
Figure 3. RNAseq Analysis Pipeline Used.

Results

The following libraries were the extracted mRNA sequences from the spiders that were analyzed using RNAseq (Figure 1). Using RNAseq, both single end and paired end were used for RNAseq: single end means the fragment DNA was only read in one direction while paired end sequencing means the fragment DNA was read in both directions. The libraries were first subjected for quality assessment and ran through the first few steps of the tuxedo tools.

<table>
<thead>
<tr>
<th>Unpaired Library</th>
<th># of Reads</th>
<th>Paired Library</th>
<th># of Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAYASHI11-unpaired.fa</td>
<td>1,775,663</td>
<td>HAYASHI11-paired.fq</td>
<td>585,728</td>
</tr>
<tr>
<td>HAYASHI12-unpaired.fa</td>
<td>152,101</td>
<td>HAYASHI12-paired.fq</td>
<td>91,156</td>
</tr>
<tr>
<td>HAYASHI13-unpaired.fa</td>
<td>640,198</td>
<td>HAYASHI13-paired.fq</td>
<td>197,324</td>
</tr>
<tr>
<td>HAYASHI14-unpaired.fa</td>
<td>519,784</td>
<td>HAYASHI14-paired.fq</td>
<td>214,672</td>
</tr>
</tbody>
</table>

Table 1. Table of the RNAseq libraries totaling 8. Four libraries were sequenced single-end and the other four were sequenced paired-end. Number of reads are the amount of sequenced fragmented DNA in that designated library.
I. Quality assessment of RNAseq libraries using ShortRead

![Graphs showing quality assessment of RNAseq libraries](image)

Figure 4. Quality assessment of the paired libraries using Bioconductor package ShortRead. Peaks that lean towards the right indicate good quality library.

Note that the libraries HAYASHI11-unpaired.fa, HAYASHI12—unpaired.fa, HAYASHI13-unpaired.fa, and HAYASHI14-unpaired.fa, were not included on the QA report generated by ShortRead because ShortRead can only recognize FastQ files and not FASTA files (extensions of .fq and .fa respectively). According to ShortRead, Overall Read Quality graphs with peaks at the right panel are good quality reads. The RNAseq libraries show overall good quality reads. Next, these libraries were aligned to a reference genome (via Trinity) using the BowTie2.

II. TopHat Alignment Results

TopHat alignment program is then used to align the unpaired and paired libraries to the reference transcriptome. This is done because TopHat offers a more sophisticated, yet more memory intensive calculations, alignment since it considers possible splicing junctions (Trapnell et al., 2012). Below are the results of the unpaired libraries. Besides the alignment summary results, TopHat also reports accepted hits (mapped reads) and unmapped reads in a .BAM file.
Figure 5. Overall alignment rate of paired and unpaired libraries using TopHat. Overall alignment rate was reported in percentages.

<table>
<thead>
<tr>
<th>Unpaired Library</th>
<th># of Loci</th>
<th>Paired-End Library</th>
<th># of Loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAYASHI11-unpaired.fa</td>
<td>65,804</td>
<td>HAYASHI11-paired.fq</td>
<td>42,377</td>
</tr>
<tr>
<td>HAYASHI12-unpaired.fa</td>
<td>20,306</td>
<td>HAYASHI12-paired.fq</td>
<td>8,236</td>
</tr>
<tr>
<td>HAYASHI13-unpaired.fa</td>
<td>43,661</td>
<td>HAYASHI13-paired.fq</td>
<td>22,605</td>
</tr>
<tr>
<td>HAYASHI14-unpaired.fa</td>
<td>49,8649</td>
<td>HAYASHI14-paired.fq</td>
<td>22,047</td>
</tr>
</tbody>
</table>

Table 2. Table of the number of loci after running the TopHat alignment results into CuffLinks.

III. DEseq and HTSeq results

After aligning the RNAseq libraries to TopHat and running the results of TopHat to the tuxedo pipeline, both HTSeq and DESeq were performed – HTSeq was used for the counting while DESeq was used for visualization. First, DESeq was used to obtain a dispersion curve for all the libraries to determine the amount of variation within the libraries (Figure 5). A heatmap was conducted to see the top 30 most commonly expressed genes for all the eight libraries (Figure 6). Another heatmap was generated to see the relation and graphically represent the variation between the eight libraries (Figure 7). Additionally, a PCA plot was generated in order to see the relationship between males and females (for both paired and unpaired) Dysdera libraries (Figure 8).
Figure 6. Estimated pool dispersion for all 8 samples with a regression curve.

Figure 7. Heatmaps showing the expression data of the 30 most highly expressed genes (transformed data). Clustering of samples does not align with the experimental factor (male vs. female).
Figure 8. Heatmap showing the Euclidean distances between the samples as calculated from the variance stabilizing transformation of the count data.

Figure 9. PCA plot. The 8 samples shown in the 2D plane spanned by their first two principal components.
Discussions

Using RNAseq, the extracted mRNA sequences of male and female sequences of male and female *Dysdera* spiders were sequenced using both single-end and paired-end sequencing. Single-end or unpaired sequencing is when the fragmented DNA is only read one way while paired-end sequencing is when the fragmented DNA is read both ways. The advantage of using paired-end DNA sequencing is that it gives more coverage and a more sensitive sequencing since it detects repetitive sequences, gives an idea of how long the sequence is, and it detects insertions and deletions within the DNA. Also, the difference between single-end and paired-end sequencing is the resulting final FastQ file: single-end has one file while paired-end has two accompanying files for one read, which is both read left and right ways. This mate files, two separate files containing both left and right reads, are DNA sequences of the same fragmented DNA but from sequenced from different directions.

The libraries were then subjected to quality assessment analysis using Bioconductor package ShortRead (Morgan et al., 2009). ShortRead produces a graph as seen in Figure 4., which shows the quality of the paired libraries. The x-axis is the average (calibrated) base quality and the y-axis is the proportion of the reads. The peaks of the graphs leaned towards the right side of the graph, therefore, the peaks were heavily concentrated on a higher base quality (x-axis). This indicates that the libraries possess good quality and therefore can be used for the differential expression analysis.

After determining the quality of the reads, it is first subjected to the alignment program BowTie2 and using a reference transcriptome, which was constructed *de novo* by PI Dr. Arensburger using the program Trinity. BowTie2 was used in order to check if both paired and unpaired libraries are able to align to the reference transcriptome (results not shown). After determining that the libraries are indeed able to align, TopHat was primarily used to perform the alignment.

TopHat offers a more sophisticated and sensitive alignment than BowTie2 – it considers splicing junctions. The resulting overall alignment was indicated in Figure 5. Note that there was a drop in the overall alignment rate compared to BowTie2. This is expected since TopHat considers sections of the reads that might belong to exon-exon splicing junctions as well as introns. TopHat also reports both mapped and unmapped reads in a .bam file (a file that contains sequence alignment data). These files are subjected to differential expression analysis.
Using CuffLinks, the results from TopHat were quantified and used for another transcript assembly. CuffLinks found gene loci from the TopHat alignment results as seen in Table 2. Each library has different number of loci, primarily due to the differences on the size of the libraries which were affected due to other factors such as performing RNAseq and the actual extraction of RNA from the sample spiders. However, to determine if there are overlapping loci between each library, CuffMerge was utilized. CuffMerge merges all the transcript assembly into one master assembly. After the assembly of one whole master transcript, the results from TopHat (the fragments and where they are located), had to be aligned back to determine if they are a part of a gene.

This counting was done through HTSeq and the results from HTSeq were visualized through a R-bioconductor package, DESeq. First DESeq tests the dispersion of the results from HTSeq through a double log graph (Figure 6). In Fig. 6, as mean read count increases, the dispersion decreases. This is important in analyzing the data as more biological variation that the data has, the differences between the data points become significant. However, DESeq tends to underestimate the dispersion of the data and does this base on sample size. Since there’s only 8 libraries for male and female spiders (4 libraries for each sex of Dysdera), it underestimates some of the data points and moves them below the red fitted line as seen in Figure 5.

After the analysis of the dispersion of the data, two heatmaps were produced: one heatmap was for the top 30 genes that were commonly expressed among the 8 libraries of Dysdera (Figure 7) while the second heatmap was for determining the relationship between the different sexes of Dysdera (Figure 8). In Figure 7, most the genes listed were almost as equally expressed in all the libraries of Dysdera regardless of sex. However, there were transcripts to this exception: XLOC_000782, XLOC_002249, XLOC_001415, XLOC_001373, XLOC_1336, XLOC_001368, which were expressed less in HAYASHI12 paired and unpaired libraries (male Dysdera), and HAYASHI14 paired and unpaired libraries (female Dysdera) compared to HAYASHI13 (male) and HAYASHI11 (female) paired and unpaired libraries. XLOC_002634, in addition to having less expression with HAYASHI12 and HAYASH14 libraries, also has less expression in HAYASHI13 libraries. The cause of this difference between the expression, also taking note that it is not a difference in sex, may have been due to the biological sample. Additional data analyzation, such as aligning this set of genes to a non-random database, can be done to determine the nature of the genes that were differentially expressed.
The second heatmap (Fig 8), which pertains to the sexes of the *Dysdera* spiders, illustrates that there are no differences between the expression of genes in the sample. This is because of the Euclidian distances between the samples – four of the male and female libraries are more related to one another than the other four male and female libraries. This was further confirmed from a PCA graph that was generated (Fig 9). As expected, paired and unpaired libraries of each male and female libraries are related to one another. However, both the PCA plot and the Fig 8 heatmap supports that there’s no difference between the genetic expression of both male and female *Dysdera* spiders. This suggests that the silk production between male and female *Dysdera* spiders may use the same gene expression.

Though this study demonstrates the list of genes that are expressed in *Dysdera* spiders, regardless of sex, further research is required in order to evaluate the nature of the genes. This means that these individual genes have to be run through a database using local alignment tools to find any gene similarity to existing genes present in other species. Additionally, the list of the genes may contain novel genes that have yet to be discovered. Further annotation, such as determination of conserved domains (or conserved sequences) from the resulting proteins of these genes in order to give light of their function should be performed.
References:


