Assessing the Utility of the Oxford Nanopore MinION for Snake Venom Gland cDNA Sequencing

Adam Hargraves and John Mulley

Nicholas Perry

3 May 2019

Abstract

Previous short-read technologies have not been very accurate in their sequencing of transcriptomes, as transcriptomes are highly variable. Sequencing them, however, is important, as they can provide insight into the proteins that they create. In their study, the transcriptome that produces venom from the painted, saw-scaled viper was sequenced and assembled by the original researchers with the hopes of eventually developing antivenom. In this research, the venom transcriptome was assembled, polished, and annotated so that venom proteins could be predicted. Most of the original methods could not be used, however, as they have been discontinued. Even though many of the same results could not be achieved, the protein prediction showed that some form of the transcriptome that produces venom was created.

Introduction

Sequencing a transcriptome can be quite difficult due to its high variability. A transcriptome, by definition, is the sum of the mRNA molecules expressed by a single or group of cells (McGettigan, 2013). A transcriptome, therefore, can include all of the genes that encode for proteins (Rudd, 2003). A transcriptome that is of particular interest is the one that makes venom. Venom can be quite dangerous for humans, making it a high priority for research. By studying venom, researchers can learn more about its components and potentially develop antivenom. Venom, however, has proven to be quite difficult to sequence in the past. Since venom is a protein, it is created through the transcription of its transcriptome, which tends to be quite large and highly variable. There have been previous attempts to sequence the transcriptome of venom from certain species of snake using Illumina sequencing, a short-read technology, that have proven to be ineffective. For example, the species Okinawa habu, or *Protobothrops flavoviridis*, and Hime habu, or *Ovophis okinavensis*, have had the cDNA of their transcriptomes sequenced with low accuracy (Aird *et al*., 2013). Short read technologies have proven to not be effective in sequencing transcriptomes as they can not accurately assemble the transcriptome.

A long-read technology may hold the key to sequencing a transcriptome, such as the one that makes venom. Researchers have used the MinION by Oxford Nanopore to sequence the transcriptome of venom from the painted, saw-scaled viper, *Echis coloratus*. Previously, these researchers had tried to sequence this transcriptome using Illumina and Sanger sequencing to little success (Hargreaves *et al*., 2014b; Hargreaves *et al*., 2014a). These researchers were able to successfully sequence the transcriptome that produces venom from the painted, saw-scaled viper using the MinION by Oxford Nanopore with much higher levels of accuracy and precision (Hargreaves and Mulley, 2015).

Hargreaves and Mulley performed their analysis of the transcriptome that produces venom by using four data files. These data files were from 2 samples of the painted, saw-scaled viper. Each sample had a file that was from a 48-hour sequencing run, and each sample had a file where the MinION was remuxed 4 times,

at 8-hour intervals (Hargreaves and Mulley, 2015). Once the sequences were obtained, Hargreaves and Mulley performed sequencing statistics using poretools and poRe (Loman and Quinlan, 2014; Watson *et al*., 2015). Following this, proovread-flex, which is designed for transcriptomes, was used to correct for hybrid errors (Hackl *et al*., 2014). Also, nanocorrect was used to correct for *de novo* errors and nanopolish was used to correct based on the original electrical signal found in the .fast5 files (Loman *et al*., 2015). Hargreaves and Mulley then used BWA-MEM and TransRate to assess the quality of the assembly (Li, 2013; Smith-Unna *et al*., 2015). Following this assessment, the researchers used TransDecoder to predict protein-coding open-reading frames (Haas *et al*., 2013). Finally, the researchers used BLAST+ to identify reads of interest and CLUSTAL to align and manually annotate sequences (Camacho *et al*., 2013; Larkin *et al*., 2007). It is important to note, however, that most of this software has been discontinued and could not be used for this analysis. The only software that could be used was TransRate. It is also important to note that there were no .fastq files available during this analysis, so many sequencing statistics that were previously reported could not be obtained.

Methods

The analysis I used to look at the venom transcriptome data was very different than what the researchers completed originally. First, I used the basecaller DeepNano, which was run using the .fast5 files (Cao *et al*., 2017). Following this, I made 2 assemblies using Canu (Koren *et al*., 2017). One of the assemblies was for the Eco6 4x8-hour run. The other assembly was created using both the Eco8 files. The Eco8 48-hour run and Eco8 4x8-hour run files were merged into the same file so that higher read counts could be used. Following the assemblies, I performed polishing using Pilon by using the original Nanopore reads and Illumina data (Walker *et al*., 2014). Once the polishing was complete, I predicted genes using Prodigal, and then used Blast-2-Go to see what proteins may result from those genes (Hyatt *et al*., 2010; Conesa *et al*., 2005). Finally, I used TransRate and NanoStat to confirm the quality of the assemblies (Smith-Unna *et al*., 2016; De Coster *et al*., 2018).

Results

Read length distributions were created for each of the four sequence files. Specifically, read length distributions were created in R for the Eco6 4x8-hour run, Eco6 48-hour run, Eco8 4x8-hour run, and Eco8 48-hour run. These distributions would show how many reads of a certain length were obtained during the initial sequencing run.

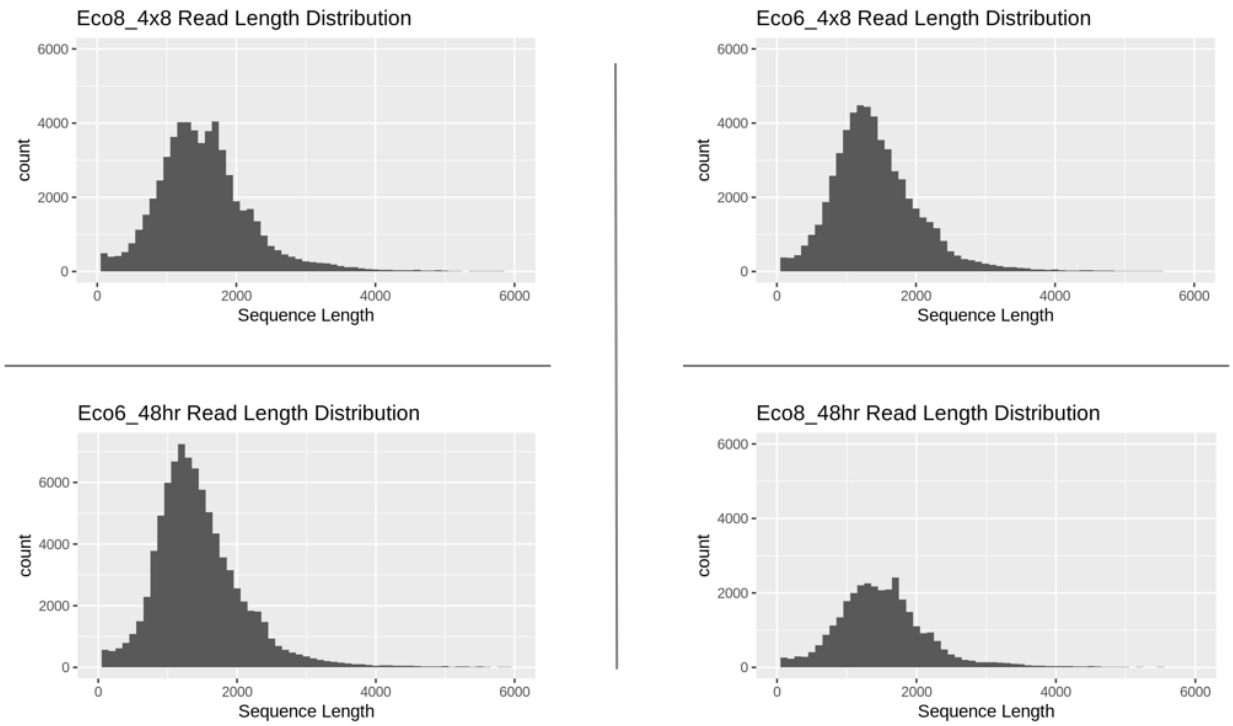

Figure 1. Read length distributions for the four sequencing runs that were originally completed.

Statistics on the reads were collected on the Eco6 4x8-hour and Eco8 files. Some of these statistics included the mean length in base pairs, the total number of reads and base pairs, the maximum and minimum number of base pairs in the reads, and the N50 in base pairs.

Table 1. Sequencing statistics for the Eco6 4x8-hour and Eco8 reads.

Statistic	Eco6 4x8-hour	Eco8
Total Reads	57,884	92,133
Total Bases (Mb)	89,779,650	154,361,646
Max Length (bp)	940,362	5,225
Min Length (bp)	1	1
Mean Length (bp)	1,554	1,675.40
N50 (bp)	1,644	1,785

Sequencing statistics were obtained following the assembly of contigs and polishing using Illumina and Nanopore reads. The same statistics were obtained so that comparisons could be made before and after these processes were completed.

Table 2. Sequencing statistics for Eco6 4x8-hour and Eco8 assemblies.

	Illumina Corrected	Illumina Corrected	Nanopore Corrected	Nanopore Corrected
Statistic	Eco6 4x8-hour	Eco8	Eco6 4x8-hour	Eco8
Total Reads	249	526	249	526
Total Bases (bp)	449,307	971,910	437,856	948,478
Max Length (bp)	4,434	4,956	4,326	4,699
Min Length (bp)	1,040	1,046	1,031	1,018
Mean Read Length	1,804.40	1,848	1758.40	1,803
Contig N50 (bp)	1,882	1,899	1,515	1,862

The proteins were predicted and counted so as to see which proteins would be more prevalent in venom. Proteins with a higher count have been predicted to be in venom more than proteins with a lower count.

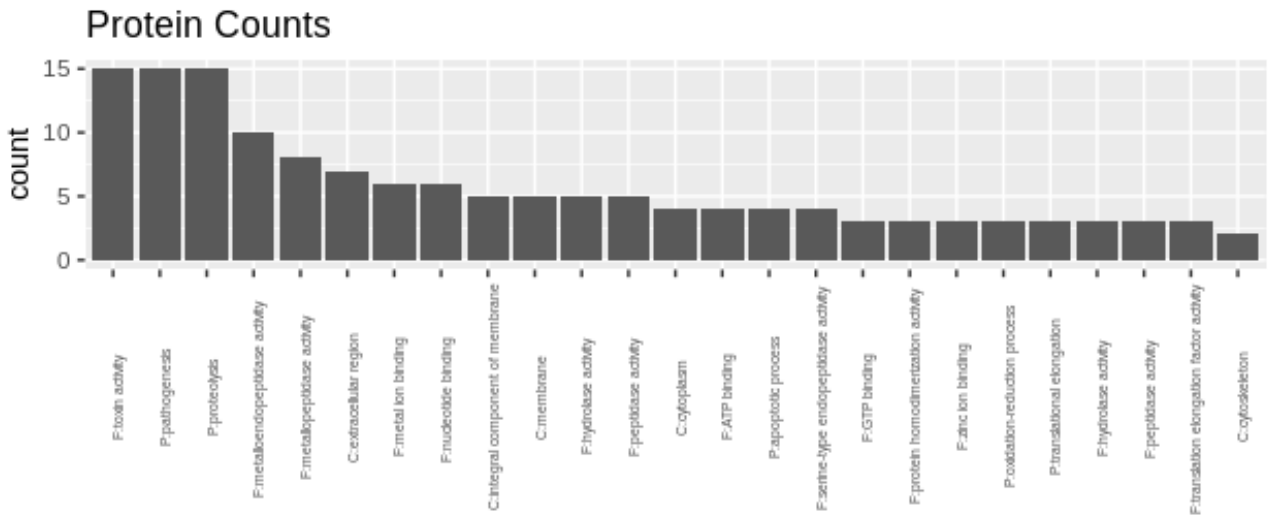

Figure 2. Protein counts for proteins that were predicted to be in venom.

Discussion

The results that were obtained were very different than those of the original study. This could be due to different software being used or not having any .fastq files. First, a different basecaller was used. There is no way to tell how well DeepNano worked as a basecaller. It is not known if DeepNano was making accurate basecalls based on the .fast5 files. It is also not known if the original .fast5 files were only designed to be used with an older basecaller. Without any of the original sequence files, there is no way to tell how accurate the reads being used were.

Looking at the read length distributions (Figure 1), it can be determined that the Eco6 reads were of higher quality than the Eco8 reads. This is due to the higher read counts present for the Eco6 4x8-hour and Eco6 48-hour runs. For this reason, the Eco8 reads were combined into a single file so as to achieve the highest possible read counts. However, even though it appeared to have an even read length distribution, the Eco6 48-hour reads did not assemble. This may be due to poor read quality or DeepNano not accurately basecalling the files. Without any .fastq files, though, there is no way to be sure of this.

When looking at the read counts, it can be shown that none of statistics regarding the read counts are the same as the original study (Table 1). This would be expected for Eco8, as it was a combined reads file in this research but separate in the original study as Eco8 4x8-hour and Eco8 48-hour runs. So, since this file was separate in the original study and combined in this research, it makes sense that these statistics would be different. However, for the Eco6 4x8-hour reads, the statistics were drastically different. For example, the original study had 66,916 reads, while this research had 57,884 reads for Eco6 4x8-hour reads (Table 1; Hargreaves *et al*., 2015). This could be due to the basecaller not getting accurate reads. If DeepNano was not able to get accurate reads for all of the .fast5 files, there would be less reads present, which could alter the total number of reads.

The contigs counts suffered from the same issue as the read counts (Table 2). The contigs counts differed from the original study in that the Eco8 assembly could not be compared since it was from a combined reads file and the Eco6 4x8-hour assembly had a different count than the original study. Since the Eco8 assembly was built off of reads that were from combined Eco8 4x8-hour and Eco8 48-hour reads, the contig counts were going to be different. So, there would be no way to compare any of the statistics for Eco8. The Eco6 4x8-hour assembly had different count since DeepNano may have called the original reads incorrectly. However, there is no way of telling this without any .fastq files.

When looking at the protein counts, it can be shown that there was no protein that was predicted many times (Figure 2). The highest number for a protein count was 15. Looking at which proteins had the highest counts, however, would allow one to see if the transcriptome could possibly be from venom. For example, one of the predicted proteins was toxin activity, with a count of 15. This indicates that there would

be some toxicity in the proteins that were potentially present in the venom. Other proteins that were predicted include peptidases and metallopeptidases. These enzymes indicate that there would be proteins present that would break down proteins in the host that the venom would infect. This provides some information about the accuracy of the transcriptome. One would expect venom to have toxin activity and be full of enzymes to break down proteins. So, this means that the reads and assembly must have had some level of accuracy in them, since the proteins predicted could definitely be present in venom.

Overall, even though many of the original software programs and data were not available for this research, the results do show that some form of a transcriptome that makes venom was assembled. While the reads and contigs counts do not match up with the original study and the read length distributions were not great for some of the samples, the protein predictions do show that there is some level of toxicity and enzymatic activity present in the venom proteins from this transcriptome. Therefore, some form of a transcriptome must have been successfully sequenced. However, the sequence may not be the most accurate since there is a lot that is not known about it due to discontinued software and missing information.

References

Aird, S.D., Watanabe, Y., Villar-Briones, A., Roy, M.C., Terada, K. and Mikheyev, A.S., 2013. Quantitative high-throughput profiling of snake venom gland transcriptomes and proteomes (Ovophis okinavensis and Protobothrops flavoviridis). *BMC genomics*, *14*(1), p.790.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L., 2009. BLAST+: architecture and applications. *BMC bioinformatics*, *10*(1), p.421.

Cao, M.D., Nguyen, S.H., Ganesamoorthy, D., Elliott, A.G., Cooper, M.A. and Coin, L.J., 2017. Scaffolding and completing genome assemblies in real-time with nanopore sequencing. Nature Communications, 8, p.14515.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21(18), pp.3674-3676.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M. and Van Broeckhoven, C., 2018. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics, 34(15), pp.2666-2669.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M. and MacManes, M.D., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, *8*(8), p.1494.

Hackl, T., Hedrich, R., Schultz, J. and Förster, F., 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, *30*(21), pp.3004-3011.

Hargreaves, A.D. and Mulley, J.F., 2015. Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. *PeerJ*, *3*, p.e1441.

Hargreaves, A.D., Swain, M.T., Hegarty, M.J., Logan, D.W. and Mulley, J.F., 2014b. Restriction and recruitment—gene duplication and the origin and evolution of snake venom toxins. *Genome biology and evolution*, *6*(8), pp.2088-2095.

Hargreaves, A.D., Swain, M.T., Logan, D.W. and Mulley, J.F., 2014a. Testing the Toxicofera: comparative transcriptomics casts doubt on the single, early evolution of the reptile venom system. *Toxicon*, *92*, pp.140-156.

Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11(1), p.119.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research, 27(5), pp.722-736.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R. and Thompson, J.D., 2007. Clustal W and Clustal X version 2.0. *bioinformatics*, *23*(21), pp.2947-2948.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Loman, N.J., Quick, J. and Simpson, J.T., 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods*, *12*(8), p.733.

Loman, N.J. and Quinlan, A.R., 2014. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, *30*(23), pp.3399-3401.

McGettigan, P.A., 2013. Transcriptomics in the RNA-seq era. *Current opinion in chemical biology*, *17*(1), pp.4-11.

Rudd, S., 2003. Expressed sequence tags: alternative or complement to whole genome sequences? *Trends in plant science*, *8*(7), pp.321-329.

Smith-Unna, R.D., Boursnell, C., Patro, R., Hibberd, J.M. and Kelly, S., 2015. TransRate: reference free quality assessment of de-novo transcriptome assemblies. bioRxiv. 2015.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K. and Earl, A.M., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one, 9(11), p.e112963.

Watson, M., Thomson, M., Risse, J., Talbot, R., Santoyo-Lopez, J., Gharbi, K. and Blaxter, M., 2014. poRe: an R package for the visualization and analysis of nanopore sequencing data. *Bioinformatics*, *31*(1), pp.114-115.