Chacon, J., Cuajungco, M.P. / Californian Journal of Health Promotion 2018, Volume 16, Issue 1, Pages 46-53.

Comparative De Novo Transcriptome Assembly of *Notophthalmus viridescens* RNA-seq Data using Two Commercial Software Programs

Jonathan Chacon[1] and Math P. Cuajungco[1,2]

[1] *Department of Biological Science, California State University Fullerton*
[2] *Center for Applied Biotechnology Studies, California State University Fullerton*

Abstract

Background and Purpose: The reduction of cost and ease of using core laboratories or commercial sequencing companies have allowed biomedical and health researchers alike to employ reference-based genomic or transcriptomic sequencing (RNA-seq) projects to expand their work. Non-reference based data analysis, in cases of inexperienced researchers, become more challenging despite the availability of many open source and commercial software programs. **Methods:** We performed de novo assembly of RNA-seq data obtained from a non-model organism (Eastern Newt skin) to compare data output of two commercially available software workflows. **Results:** Our results show that the software packages performed satisfactorily albeit with differences in how the annotated and novel transcripts were identified and listed. **Conclusion:** Overall, we conclude that the use of commercial software platforms has a clear advantage to that of open source programs because of convenience with data analysis workflows. One caveat is that users need to know the software's basic algorithm and technical approach, in order to determine the precision and validity of the data output. Thus, it is imperative that researchers fully evaluate the software according to their needs to determine their suitability.

Keywords: RNA-seq, gene annotation, non-model organism, next generation sequencing

Introduction

The dramatic decline in costs over the past decade has ushered the widespread application of next-generation sequencing (NGS) projects like genomic and transcriptomic sequencing (RNA-seq) to traverse across many niches of research that previously lacked the incentive to perform such high-throughput studies. This increase in accessibility has resulted in a massive accumulation of open access NGS data for many model organisms. Reference-based transcriptomic projects have thus become feasible additions to many research programs. On the other hand, the availability of sequence data for non-model organisms, although increasing, is often incomplete and places significant constraints on studies that aim to look at unique characteristics that are devoid in model organisms. During these instances, de novo assemblies have great value, granting researchers some insight into gene expression and characterization for organisms with insufficient genomic data available. Transcriptomic de novo studies also have the added benefit of identifying novel transcript isoforms and alternative splicing events even when reference genomes are available since they, by definition, are not confined to the ever-expanding source of information provided at the time of given study. Despite the potential to unveil new information for both model and non-model organisms, de novo assemblers face many hurdles, especially when dealing with alternatively spliced isoforms and highly variable contig alignment (Wall et al. 2009; Zhao et al. 2011), which often becomes a barrier when end-users have limited bioinformatics experience. Fortunately, NGS bioinformaticians are aware of the issue, and several commercial software packages have

become available in recent years. All NGS data analysis software platforms utilize specialized algorithms that are geared towards streamlining entire RNA-seq experiments; however, although the intent is very helpful, it diminishes some control and oversight on the user's end – the black box effect.

The Present Study

Here, we report a limited evaluation on the streamlined de novo assembly workflows using two commercial software packages, CLC Genomics Workbench (GW) version 10 and Lasergene SeqMan NGen (SMN) version 14, using Illumina® HiSeq RNA-seq data from the skin of Eastern Newt (*Notophthalmus viridescens*). Although a comparison of previous versions of these two assemblers has been performed by Kumar and Blaxter (2010), their study used 454 pyrosequencing data – an older sequencing method that is costly and relatively lower throughput compared to newer sequencing strategies (Liu, et al., 2012). In this study, we generated RNA-seq data using the Illumina® sequencing technology to provide some insight on how commercial de novo assembly workflows perform with shorter sequencing reads. By elucidating relative advantages and disadvantages of de novo assemblers, we aim to help novice researchers decide which program to use for their work.

Methods

Materials

Newt RNA Extraction and cDNA Library Construction. Epidermal tissues from the dorsal torso of *N. viridescens* were kindly provided by Dr. Christopher Tracy at California State University Fullerton. Total RNA was extracted and purified using TRI-zol® reagent (Invitrogen, Carlsbad CA). We used the Ovation® Human FFPE RNA-seq Multiplex System kit (NuGEN Technologies, San Carlos CA) to construct the cDNA library according to the manufacturer's recommendations for low-yield RNA.

Procedures

The Insert Dependent Adaptor Cleavage (InDA-C) method was used to deplete non-transcript RNA contaminants. Selective ribosomal RNA (rRNA)-specific primers were designed from Newt and frog sequence data available at NCBI (www.ncbi.nlm.nih.gov) through the generous help of Denise Stephens (NuGen). The primers were commercially synthesized (Integrated DNA Technologies, Coralville IA). Paired-end sequencing using Illumina® HiSeq 2500 was performed by the Genomics High Throughput Facility at the University of California, Irvine.

De novo Transcriptome Assembly. CLC GW and Lasergene SMN differ most notably in their assembly approach, using de Bruijn graphs and overlap-layout consensus strategies, respectively. The following default parameters for CLC GW software v. 10 (www.qiagenbioinformatics.com) were used: mismatch cost = 2; min contig length = 200; min-max distance = 1-1000. Meanwhile, the following default parameters for the Lasergene SMN software v. 14 (www.dnastar.com) were used: mismatch penalty = 20; min contig seqs = 101; min-max distance = 0-750. Note that Lasergene's SMN de novo transcriptome assembly project wizard allowed users to specify rRNA or other input contaminant sequences prior to assembly. This option is not currently available in the CLC GW de novo transcriptome workflow. For Lasergene SMN, we loaded 5S rRNA sequences from *Xenopus tropicalis* genome obtained from the EnsEMBL database (www.ensembl.org). We also loaded sequences obtained from the NCBI database that included 28S rRNA, 16S mitochondrial

RNA, and 18s rRNA sequences from *Xenopus laevis, Lithobates catesbeianus,* and *Lithobates pipiens,* respectively, as well as 5S and 16S rRNA sequences from partially annotated *N. viridescens* genome.

Analyses. Following sequence assembly, Lasergene SMN produced both annotated and novel transcripts lists. The NCBI RefSeq database was used to obtain a number of known or homologous genes from the assembled transcript sequences. The total count of transcript fragments that aligned and matched RefSeq sequences provides the sequencing coverage and gives us confidence with the resulting data. The CLC GW assembly output contained a list of assembled transcripts and unassembled sequence reads. We then used the "Transcript Detection" plugin to identify open-reading frames and then perform a BLAST-based transcript annotation process using the InterPro protein family database as a reference. Meanwhile, gene ontology (GO) analysis provides functional description of the genes and existing relationship or functional nodes among genes. Lasergene SMN has an integrated tool to perform GO analysis, but not CLC GW.

Results

N. Viridescens Transcript Assembly

As a baseline for comparison, we used the partially annotated Newt transcriptome published recently by Abdullayev et al. (2013), and omics data by Bruckskotten et al. (2012). Our initial Lasergene SMN assembly returned 18,357 transcripts, of which 15,890 were recorded as novel, with a total average contig length of 559 nucleotides (Table 1). Bruckskotten et al. (2012) reported 26,594 novel transcripts while Abdullayev et al. (2013) reported 118,893 transcripts. It is important to note that Abdullayev and colleagues used nine types of tissue, and that potentially skin-specific transcripts were not distinguished in their paper.

Table 1

Basic Sequencing Statistics and Assembly Report for the CLC Genomics Workbench (GW) Version 10 and Lasergene SeqMan NGen (SMN) Version 14.

Sequencing parameters	Assembler	
	CLC	Lasergene SMN
Total reads	107,721,896	107,703,709
Contig (n)	176,940	18,387
Average contig size (nts*)	400	559
Median contig size (nts)	298	435
N50 (nts)	393	429
Min contig length	69	101
Max contig length	64,804	15,378
Contigs >1Kb	7,394	1,978
Total length of contig	48,415,145	10,297,868

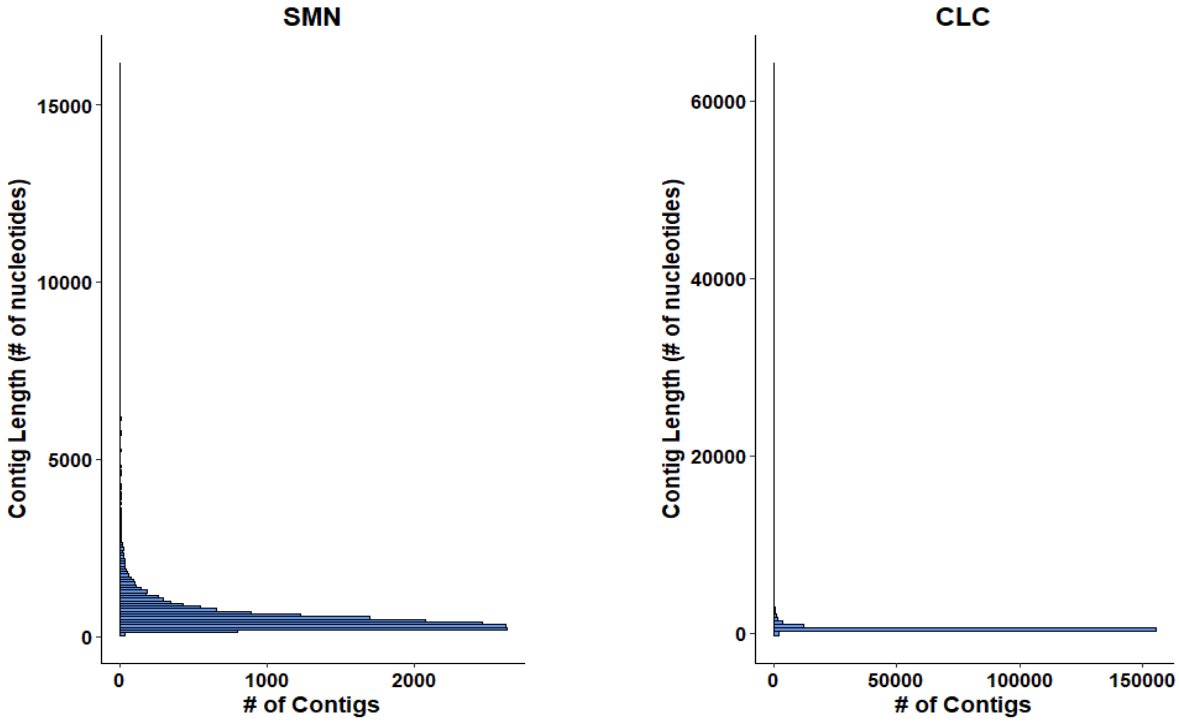
*nts = nucleotides

We used a contig-centric comparison of the CLC GW transcriptome with that of partially annotated *N. viridescens* transcriptome (Abdullayev et al. 2013), since the CLC GW platform allows for adjustable mismatch, insertion, and deletion penalties when mapping reads to contigs,

and populates a contig report (neither of which is available for Lasergene SMN's project wizard) at the end of each transcriptome assembly. Meanwhile, CLC GW had assembled over nine times the amount of contigs, which spanned at a length of 48,415,145 nucleotides – a full 370% increase from the combined length of contigs generated using Lasergene SMN (Table 1). Even after applying a filter which restricted the output to contigs that were larger than one kilobyte, contigs assembled by CLC GW outnumbered the Lasergene SMN transcripts by a factor of three. Despite the extensive discrepancies in contig abundance, Lasergene SMN's narrow total contig length produced an average contig size of 559 nucleotides that was greater than that of the CLC GW output (Table 1). Interestingly, mean-median comparison of contig lengths within each assembler revealed greater disparity within the CLC output, suggesting that CLC's GW conservative approach towards read handling allows for the generation of selectively lengthy contigs that are clear outliers. This bias was also observed when comparing the relative degree of skewness about the distribution of contig lengths between both assemblers (Figure 1).

Figure 1

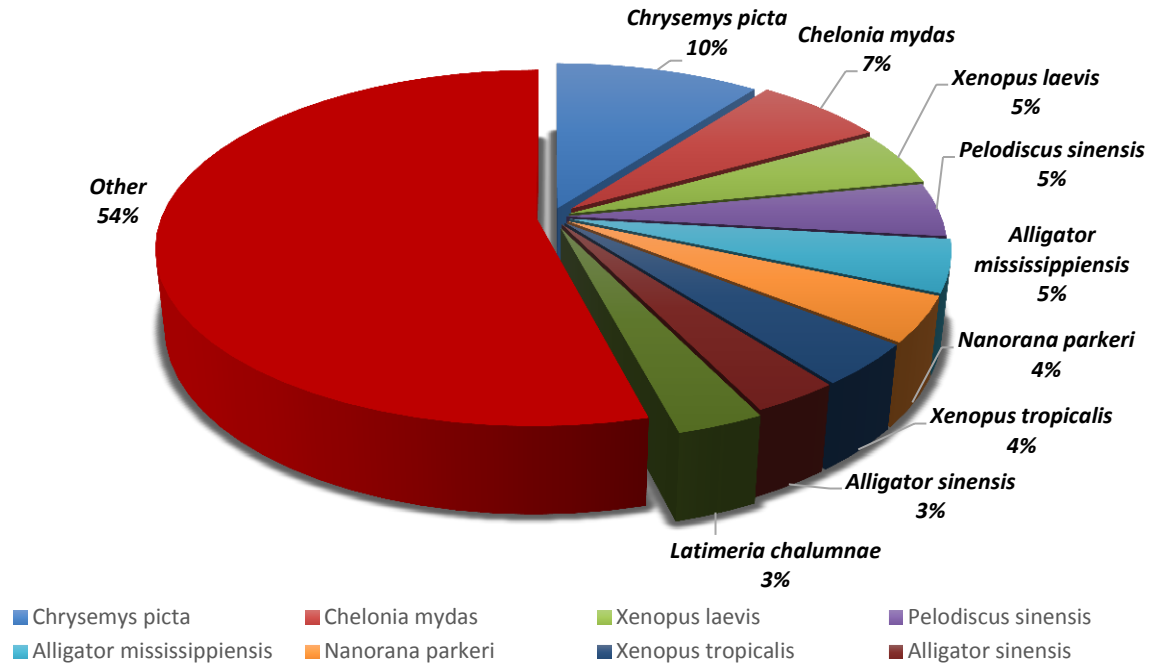
Histogram Distribution of Transcript Sequence Assembly



Contig Length distribution for both Lasergene SMN and CLC GW assemblies.

Figure 2

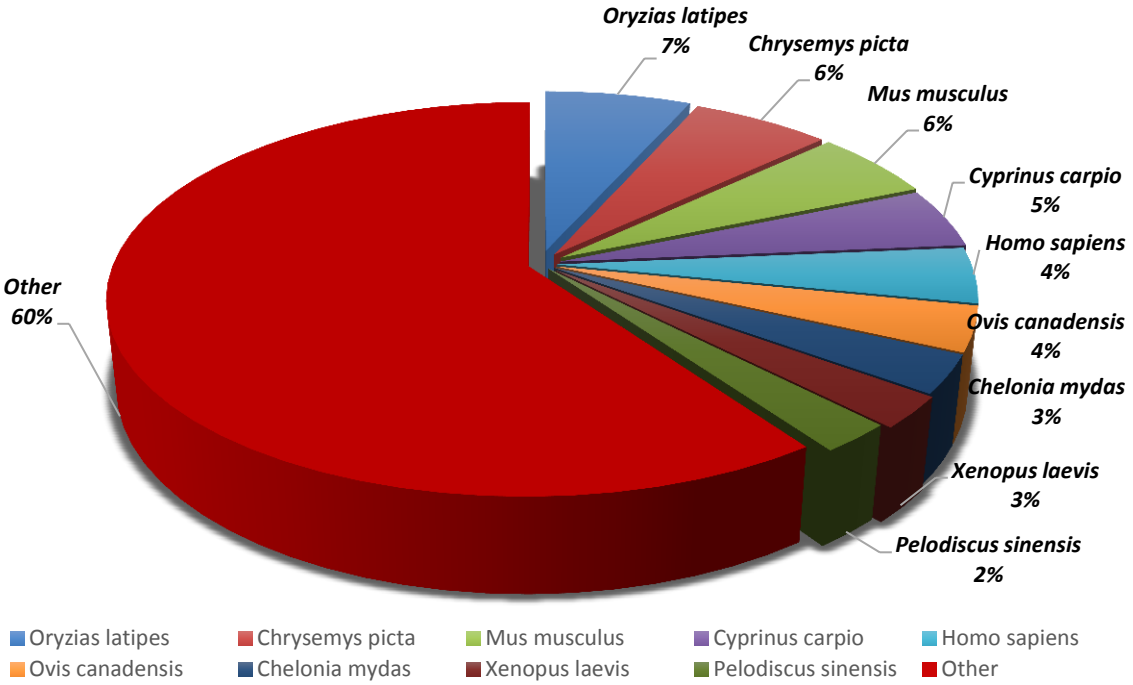
Schematic Diagram of Transcript Annotations for Lasergene SMN



The pie chart depicts the percentage of assembled transcriptomic sequences that matched or aligned from existing genomes of several organisms available from the public database.

Figure 3

Schematic Diagram of Transcript Annotations for CLC GW



Similar to Figure 2, the pie chart shows the percentage of assembled transcript sequences that mapped or aligned from existing genomes of various organisms.

Transcript Annotation

Lasergene SMN's narrowly-defined use of a RefSeq reference queried a specified non-mammalian vertebrate database. Specifically, we used the non-mammalian "Other vertebrate" (release 79) database option for Lasergene SMN transcript annotation, since no amphibian databases were available for input at the time of assembly (data not shown). The Lasergene SMN annotation approach differed with that of CLC GW, in which CLC GW utilized annotation data from redundant and non-redundant databases (NCBI's RefSeq and GenBank). The disparity with respect to primary transcript annotation approaches for each assembler managed to produce a condensed list of annotated transcripts from a diverse set of organisms (Figures 2 and 3). On the other hand, Lasergene SMN was unable to match any sequenced transcripts using a partially annotated *N. viridescens* genomic data available from NCBI. The lack of relatively stringent use of annotated references allowed CLC GW's "Transcript Detection" tool to identify two transcripts from the partially annotated NCBI database (E-value < 10^{-8}). Finally, we were able to obtain a GO report after each Lasergene SMN assembly, while CLC GW outputs rely on the use of third-party platforms to access GO terms such as BLAST2GO (data not shown).

Discussion

We show here that short sequencing data generated by Illumina instruments could be used for de novo transcriptome assembly using two well-known commercial workflows. The cross-assembler contig comparison revealed that the Lasergene SMN Trace Evidence consensus-calling algorithm generated longer contigs on average. Although the initial CLC GW Newt transcriptome surpasses that of Lasergene SMN's total contig lengths, there is no immediate information on intentionally excluded reads. Lasergene SMN, however, clearly defines excluded reads in its project report (e.g. 43,412,171 reads were excluded from the analysis done by Lasergene). The lack of information on excluded reads from CLC GW analysis could suggest that the output generated by this software may contain oversampling of the reads, which reduces some precision in the assembled transcriptome.

Lasergene SMN limits its annotation to a user-specified RefSeq database. The use of a single non-redundant reference database, although faster with respect to run time, appears to impose some limitations for Lasergene SMN's output since it did not identify any transcripts from the partially annotated *N. viridescens* sequence data. Nevertheless, Lasergene SMN and CLC GW performed satisfactorily in terms of producing annotated transcripts. The overall distribution of species-calling approach during the annotation process seems similar between both assembler, such that neither appeared to rely too heavily on sequences from a particular species.

Limitations and Conclusion

The CLC GW and Lasergene SMN workflows evaluated in the study have distinctive features to de novo assembly process. Although the observed disparity between the two software packages may be subtle, our observation is currently restricted to de novo assembly. Thus, comparison of the two workflows in assembling transcripts

from sequencing data of model organisms that have existing reference genomes may provide a completely different outcome.

Despite the convenience and ease of use with streamlined workflows in de novo assembly and some similarity between data analysis approach, the major differences between the two software packages suggest that researchers need to be aware of their limitations on data output (i.e. black box effect) and the reliance of certain platforms on the use of third-party software or additional plugin. It is thus advisable for researchers to take advantage of limited trials offered by the software company to determine its appropriateness to analyze their data. Researchers should also consider using publicly accessible open-source program as alternative means if a commercial software route is not a financially viable option.

Acknowledgements

We are grateful for the technical assistance provided by Denise Stephens (NuGen Technologies) for the creation of primer sequences to deplete rRNA contaminants. We also thank Dr. Christopher Tracy (CSU Fullerton) for providing the Newt skin tissue. Research reported in this publication was partially supported by the National Institutes of Health NINDS R15NS101594 and NIMHD R25MD01397 grants.

References

Abdullayev, I., et al. (2013). A reference transcriptome and inferred proteome for the salamander Notophthalmus viridescens. *Experimental Cell Research, 319*, 1187-1197. doi: 10.1016/j.yexcr.2013.02.013.

Bruckskotten, M., et al. (2012). Newt-omics: a comprehensive repository for omics data from the newt Notophthalmus viridescens. *Nucleic Acids Research, 40*, D895-D900. doi: 10.1093/nar/gkr873.

Kumar, S. and Blaxter, M.L. (2010). Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics*, *11*, 571. doi: 10.1186/1471-2164-11-571

Laube, F., et al. (2006). Re-programming of newt cardiomyocytes is induced by tissue regeneration. *Journal of Cell Science, 119*, 4719-4729. doi: 10.1242/jcs.03252.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26, 493–500. doi: 10.1093/bioinformatics/btp692

Lin, L., et al. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedical Biotechnology, 2012*, 1-11. Article ID 251364. doi:10.1155/2012/251364

Wall, P.K., et al. (2009). Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics, 10*, 347. doi: 10.1186/1471-2164-10-347

Zhao, Q.Y., et al. (2011). Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics, 12*, S2. doi: 10.1186/1471-2105-12-S14-S2

Author Information
Math P. Cuajungco, PhD,
Dept. of Biological Science
California State University Fullerton
800 N State College Blvd, Fullerton, CA 92831.
Tel: 657-278-8522
Fax: 657-278-3426
E-mail: mcuajungco@fullerton.edu.