12-17-2015

Finding Function in the Unknown

Kelly Boyd
Loyola University Chicago, kboyd3@luc.edu

Emma Highland
Loyola University Chicago, ehighland@luc.edu

Amanda Misch
Loyola University Chicago, amisch@luc.edu

Amber Hu
Stevenson High School, ahu7@students.d125.org

Sushma Reddy
Loyola University Chicago

See next page for additional authors

Authors
Kelly Boyd, Emma Highland, Amanda Misch, Amber Hu, Sushma Reddy, and Catherine Putonti

Finding Function in the Unknown

Methods for Examining Transcriptomes of Non-Model Organisms

Kelly Boyd
Department of Biology
Bioinformatics Program
Loyola University Chicago
Chicago, IL 60660

Emma Highland
Department of Biology
Bioinformatics Program
Loyola University Chicago
Chicago, IL 60660

Amanda Misch
Department of Biology
Bioinformatics Program
Loyola University Chicago
Chicago, IL 60660

Amber Hu
Stevenson High School
Lincolnshire, IL 60069

Sushma Reddy
Department of Biology
Bioinformatics Program
Loyola University Chicago
Chicago, IL 60660

Catherine Putonti
Departments of Biology and Computer Science
Bioinformatics Program
Loyola University Chicago
Chicago, IL 60660
Email: cputonti@luc.edu

***Abstract*— Through high-throughput RNA sequencing (RNA-seq), transcriptomes for a single cell, tissue, or organism(s) can be ascertained at a high resolution. While a number of bioinformatic tools have been developed for transcriptome analyses, significant challenges exist for studies of non-model organisms. Without a reference sequence available, raw reads must first be assembled *de novo* followed by the tedious task of BLAST searches and data mining for functional information. We have created a pipeline, PyRanger, to automate this process. The pipeline includes functionality to assess a single transcriptome and also facilitate comparative transcriptomic studies.**

Keywords—de novo transcriptomics, RNA-seq, comparative gene expression

I. Introduction

Sequencing technologies continue to increase their throughput, decrease in cost, and increase the length of reads produced. As a result genomic, metagenomic, transcriptomic, and metabolomic sequencing projects of new ecological niches and non-model organisms is now attainable. Precisely defining the genes present and expressed (the transcriptome) of a single organism or a complex community is critical for our understanding of life. A number of tools [1-4] and complete pipelines [5-8] have been explicitly developed for transcriptome studies. Based upon mapping raw reads to a reference genome, these tools are able to quantify expression and compare expression profiles between samples. When no reference genome is available, however, transcriptome analysis can be computationally challenging in addition to being labor intensive. The standard practice is to perform *de novo* sequencing followed by homology searches via BLAST and this methodology has been employed for the examination of numerous species and niches, e.g. [9-12]. To streamline this process, we have developed PyRanger for the analysis of non-model species, from raw sequencing reads to comparisons between samples. While our motivation for development was for investigation of transcriptomic datasets (as exemplified in our proof-of-concept presented here), the pipeline can be applied to metagenomic and genomic studies as well.

K.B., E.H, and A.M. are funded by CRA-W's CREU program. A.H. was supported through the S.P.A.R.K immersive summer internship program. S.R. and C.P. are funded by the Loyola University Chicago Provost's Summer Research Fellowship Program.

II. Implementation

PyRanger (www.putonti-lab.com/software.html) integrates existing tools and custom parsers to expedite the evaluation of one to many individual samples through four steps:

1. Assembly Optimization: Fastq files are first examined using the Perl script VelvetOptimiser (http://bioinformatics.net.au/software.velvetoptimiser.shtml); this tool identifies the optimal parameters for assembly.
2. *De novo* Assembly: The fastq files are assembled using the Velvet [13] *de novo* assembler (written in C) with the optimal *k*-mer value identified in the prior step. The Velvet assembler includes two executables: velveth and velvetg; the former prepares the dataset for velvetg, the assembler.
3. Identifying Functionality: The assembled sequences (contigs) are next compared against the reference protein database (determined by the user). This database can be a near-neighbor or more distant, well-annotated species or a conglomeration of species. Functionality is assessed via a local blastx search through the BLAST+ suite (written in C++) [14].
4. Analysis: Functionality was developed in C++ to automate parsing of the BLAST results and facilitate the comparing multiple samples. Thus metagenomes or transcriptomes from different locations, time points, environmental conditions, etc. can be evaluated.

Each stage in the pipeline and its associated software is called through a single Python module; this provides flexibility as the user can substitute/add tools to meet his/her preferences and update components as new versions become available. The computational resources required of the pipeline are practical and expect minimal scripting expertise. Users can specify individual parameters through the use of a single configuration file. The pipeline is designed to be run on a UNIX machine with a minimum of 16GB of RAM. Run-time is dependent upon the assembly of the reads (Velvet) and BLASTing the contigs.

III. Proof-of-Concept

The transcriptomes of three avian species were used as a test for our pipeline. Extant avian species emerged over 100 million years ago (MYA) [15] giving rise to a broad array of diverse species. Only two complete bird genomes have been sequenced to date – chicken (*Gallus gallus*) and zebra finch (*Taeniopygia guttata*). These two are quite distant relatives to many bird species (e.g. the molecular divergence between the ostrich and the chicken is approximated at 80-90 MYA [16]). As such, it is not surprising that there is significant variation amongst extant bird species; in fact genome sequencing and transcriptomic studies of some species cannot rely on the available genomes as a reference as they are too divergent.

Thus, to test our pipeline three avian transcriptomic data sets were selected: the kiwi (*Apteryx australis*) SRX026541 [10], the duck (*Anas platyrhynchos*) SRX026109 [9], and the rock pigeon (*Columba livia*) SRX159811 [17]. The size of the transcriptomic data (in FASTQ format) ranged from ~41 Mb to 10 Gb. Each sample was processed by the pipeline three times, for three different databases – chicken, zebra finch and mouse (*Mus musculus*). As expected, more BLAST hits were found when comparing the kiwi, duck and rock pigeon contigs to chicken and zebra finch. Significantly more hits were found to chicken (a likely factor of the more complete annotation available rather than kinship). Comparing the transcriptomes reveals a number of genes expressed in a single species, pair of species or all three species (Fig. 1). It is of note, that these transcriptomes are from different tissues, further contributing to the variation in the genes identified within the three species. For each hit, protein information is provided for further downstream analysis.

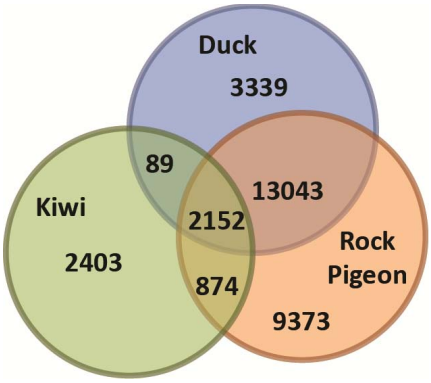

Fig. 1. Comparison of the BLAST hits to the three transcriptomes

IV. Conclusions and Future Directions

The PyRanger pipeline presented here expedites functional analysis when investigating species or mixed communities, particularly when near-neighbor reference sequences are not available. This approach can be applied for studies across all domains of life, from viral species and/or communities to eukaryotic genomes or transcriptomes, the latter being demonstrated in our proof-of-concept study. Future developments include the incorporation of additional assembly methods as well as downstream analysis tools.

Acknowledgment

The authors would like Mr. Steven Reisman for computational support.

References

[1] F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch, "Optimal spliced alignments of short sequence reads," Bioinformatics, vol. 24, pp. i174-180, 2008.

[2] C. Trapnell *et al*. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," Nat. Biotechnol., vol. 28, pp. 511-515, 2010.

[3] T. Bischler, M. Kopf, and B. Voss, "Transcript mapping based on dRNA-seq data," BMC Bioinformatics, vol. 15, 122, 2014.

[4] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," Nat. Methods, vol. 8, pp. 469-477, 2011.

[5] U. Hilgert, S. McKay, M. Khalfan, J. Williams, C. Ghiban, and D. Micklos, "DNA Subway: Making Genome Analysis Egalitarian." Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment - XSEDE '14, 70, 2014.

[6] J. Goecks *et al.* "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," Genome Biol., vol. 11, R86, 2010.

[7] D. Blankenberg *et al.* "Galaxy: a web-based genome analysis tool for experimentalists," Curr. Protoc. Mol. Biol., vol. 10, pp. 1-21, 2010.

[8] B. Giardine *et al.* "Galaxy: a platform for interactive large-scale genome analysis," Genome Res., vol. 15, pp. 1451-1455, 2005.

[9] Y. Huang *et al.* "The duck genome and transcriptome provide insight into an avian influenza virus reservoir species," Nat. Genet., vol. 45, pp. 776-783, 2013.

[10] S. Subramanian, L. Huynen, C. D. Millar, and D. M. Lambert, "Next generation sequencing and analysis of a conserved transcriptome of New Zealand's kiwi," BMC Evol. Biol., vol. 10, 387, 2010.

[11] E. D. Cooper, B. Bentlage, T. R. Gibbons, T. R. Bachvaroff, and C. F. Delwiche. "Metatranscriptome profiling of a harmful algal bloom," Harmful Algae, vol. 37, pp. 75-83, 2014.

[12] V. Cahais *et al.* "Reference-free transcriptome assembly in non-model animals from next-generation sequencing data," Mol. Ecol. Resour., vol. 12, pp. 834-845, 2012.

[13] D. R. Zerbino, and E. Birney, "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs," Genome Res., vol. 18, pp. 821-829, 2008.

[14] C. Camacho *et al.* "BLAST+: architecture and applications," BMC Bioinformatics, vol. 10, 421, 2009.

[15] E. D. Jarvis *et al.* "Whole-genome analyses resolve early brances in the tree of life of modern birds," Science, vol. 346, pp. 1320-1331, 2014.

[16] A. Härlid, A. Janke, and U. Arnason, "The mtDNA sequence of the ostrich and the divergence between paleognathous and neognathous birds," Mol. Biol. Evol., vol. 14, pp. 754-761, 1997.

[17] M. D. Shapiro *et al.* "Genomic diversity and evolution of the head crest in the rock pigeon," Science, vol. 339, pp. 1063-1067, 2013.