Evaluating quantification and expression methods with *Drosophila Melanogaster* data

Jimenez-Ruiz, Ivan[1] , Ropelewski, Alexander[2] , Agosto Rivera, Jose[1], Ortiz-Zuazaga, Humberto[1]

[1] Computer Sciences Department, University of Puerto Rico, Rio Piedras Campus, San Juan, Puerto Rico

[2] Pittsburgh Supercomputing Center, Carnegie Mellon University

Abstract

Even with the abundance of data that has been obtained from bioinformatics, detection of differential expression in two samples is a recurring problem. This detection is affected by the accuracy of the quantification methods used in the process, which in turn depend on the techniques used for assembly. The quantification methods that were compared include Cufflinks, RSEM and eXpress. Serially diluted RNA-seq data from *D. melanogaster* containing aliquots from both External RNA Control Consortium and Schneider 2 (S2) cells line at different percentages were assembled. We used Sequence Read Archive (SRA) data from the National Center for Biotechnology Information (NCBI) database to retrieve four transcriptomic datasets containing a total of 2,049,901,812 nucleotides (nt). The average number of bases per read in each dataset was 36 nt. Quantification results obtained from the assembled transcripts produced by Trinity and RNA Sequencing by Expectation Maximization (RSEM) [de-novo] were compared with those from TopHat and Cufflinks [reference based] to determine the validity of these protocols. We hypothesize that de-novo assembly is equally as powerful as reference-based assembly in the detection of differential expression. TopHat and Trinity are bioinformatics programs commonly used in the process of RNA assembly; Cufflinks and RSEM are quantification tools that generate Fragments per Kilobase of transcript per Million mapped reads (FPKM) that are used in differential expression detection. Using a heatmap produced by running RSEM, several genes were identified as being differentially expressed at a p-value exceeding 1e-3. Annotation of these genes through the Basic Local Alignment Search Tool (BLAST by NCBI) database identified them as coming from viral sources, specifically *Drosophila birnavirus* and X virus. These genes were not identified using reference-based approach, as their source was different from the reference used. This approach provides a draft workflow to be used with data produced from de-novo RNA-seq experiments using non-models organisms.

Background

Samples of DNA/RNA known as **sequences** can be used to understand the information of nucleotides in biological structures. Small fragments known as **reads** are produced from DNA by a **DNA Sequencer**. In the process of **sequence assembly** these fragments can be joined in order to reconstruct a complete sequence of the organism's DNA. ***De-novo***, meaning "from the beginning", refers to sequence assembly done without a reference genome, and it is used to discover/reconstruct new genome sequences. A common problem in sequence assembly can occur from errors in the sequencing data used. Reads can contain one or more mismatches from the original genome and could lead to inaccurate sequence assemblies. In de-novo sequence assembly it becomes particularly challenging, due to not having any reference to compare with and verify the integrity of the reads.

Data for the experiment consists of
D. melanogaster S2 cell lines and ERCC:

Accession	Data file	Source	Input RNA *Drosophila melanogaster*	External RNA (Control)	Method
SRX018872	SRR039460	Drosophila S2 Cells	50ng	2.5% (1.75ng) ERCC phase V pool 15	B
SRX018870	SRR039458	Drosophila S2 Cells	100ng	2.5% (2.5ng) ERCC phase V pool 15	A
SRX019234	SRR039933	Drosophila S2 Cells	100ng	5ng ERCC phase V pool 15	Untreated S2
SRX019236	SRR039935	Drosophila S2 Cells	100ng	1ng ERCC phase V pool 15	Mof RNAi S2 treatment

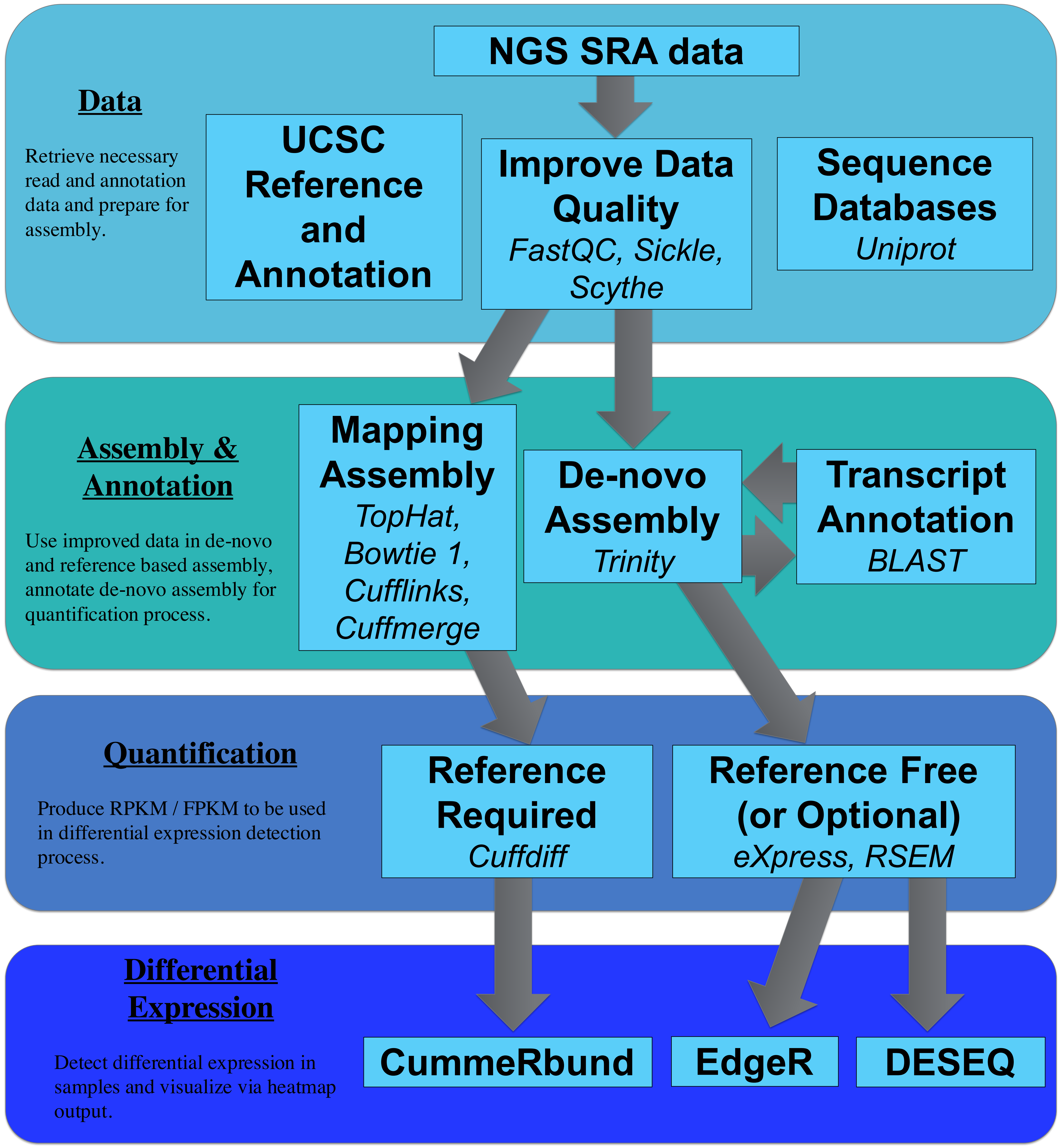
Differential expression test conditions:

Test Conditions	Datasets
One	SRR039460 + SRR039458 versus SRR039933
Two	SRR039460 + SRR039458 versus SRR039935

Aims

- Generate quantification workflow to be used for the assembly of RNA transcripts **both with a reference genome and without (de-novo)**
- Compare and contrast transcript quantification methods

Methods/Workflow



Data

Retrieve necessary read and annotation data and prepare for assembly.

NGS SRA data

UCSC Reference and Annotation

Improve Data Quality
FastQC, Sickle, Scythe

Sequence Databases
Uniprot

Assembly & Annotation

Use improved data in de-novo and reference based assembly, annotate de-novo assembly for quantification process.

Mapping Assembly
TopHat, Bowtie 1, Cufflinks, Cuffmerge

De-novo Assembly
Trinity

Transcript Annotation
BLAST

Quantification

Produce RPKM / FPKM to be used in differential expression detection process.

Reference Required
Cuffdiff

Reference Free (or Optional)
eXpress, RSEM

Differential Expression

Detect differential expression in samples and visualize via heatmap output.

CummeRbund

EdgeR

DESEQ

References

Forster, Samuel C. et al. "RNA-eXpress Annotates Novel Transcript Features in RNA-Seq Data." *Bioinformatics 29.6 (2013): 810–812. PMC. Web. 29 July 2015.*

Grabherr, Manfred G. et al. "Trinity: Reconstructing a Full-Length Transcriptome without a Genome from RNA-Seq Data." *Nature biotechnology 29.7 (2011): 644–652. PMC. Web. 29 July 2015.*

Jiang, Lichun et al. "Synthetic Spike-in Standards for RNA-Seq Experiments." *Genome Research 21.9 (2011): 1543–1551. PMC. Web. 29 July 2015.*

Langmead, Ben et al. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology 10.3 (2009): R25. PMC. Web. 29 July 2015.*

Li, Bo, and Colin N Dewey. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome." *BMC Bioinformatics 12 (2011): 323. PMC. Web. 29 July 2015.*

Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. "TopHat: Discovering Splice Junctions with RNA-Seq." *Bioinformatics 25.9 (2009): 1105–1111. PMC. Web. 29 July 2015.*

Trapnell, Cole et al. "Transcript Assembly and Abundance Estimation from RNA-Seq Reveals Thousands of New Transcripts and Switching among Isoforms." *Nature biotechnology 28.5 (2010): 511–515. PMC. Web. 29 July 2015.*

Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", *J Comput Biol 2000; 7(1-2):203-14. Web 29 July 2015.*

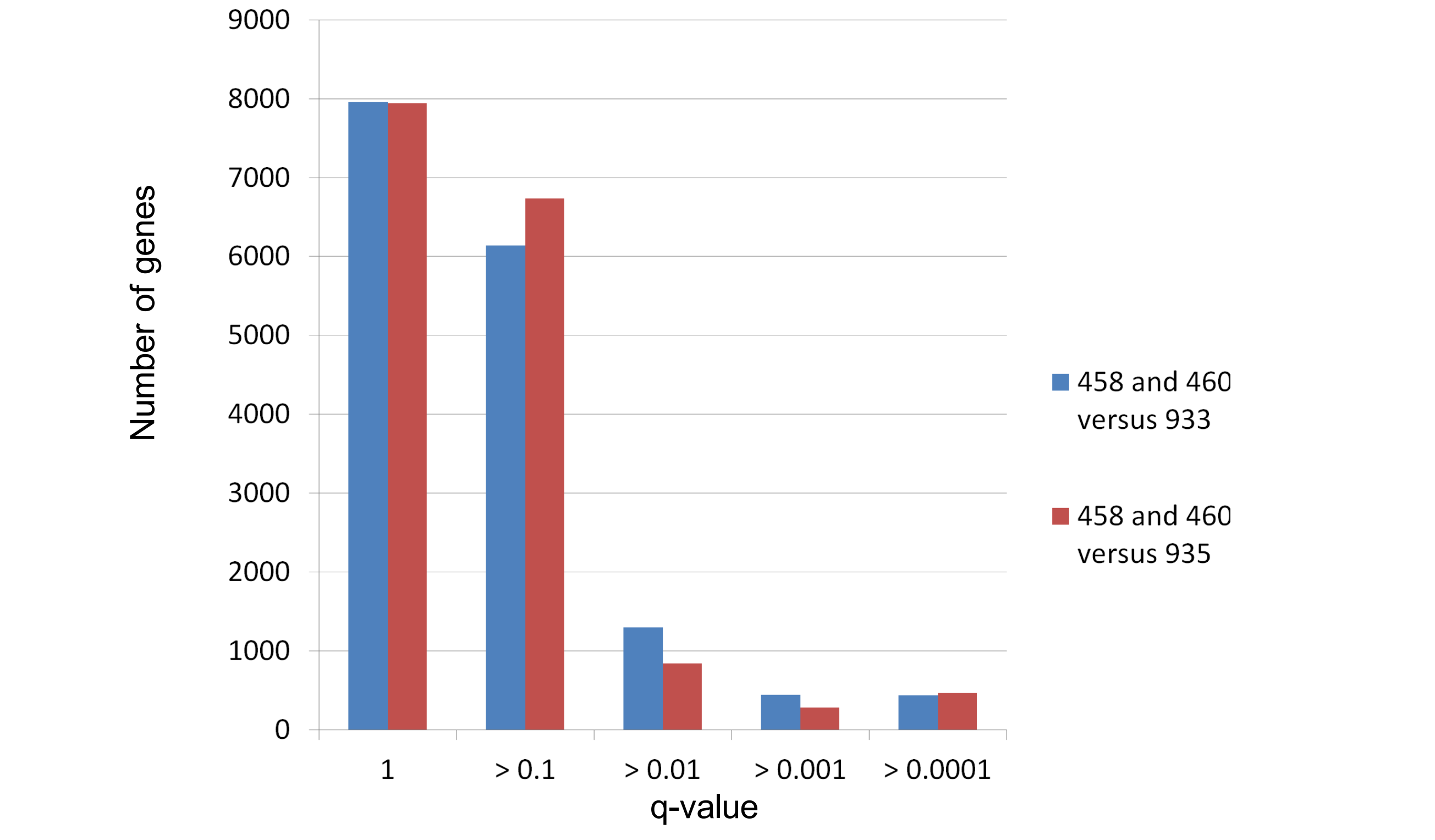
Aleksandr Morgulis, George Coulouris, Yan Raytselis, Thomas L. Madden, Richa Agarwala, Alejandro A. Schäffer (2008), "Database Indexing for Production MegaBLAST Searches", *Bioinformatics 24:1757-1764. Web. 29 July 2015.*

This work was supported by the National Institutes of Health Minority Access to Research Careers (MARC) grant T36-GM-095335 to the Pittsburgh Supercomputing Center. The computing resources that were used included BioU, which was made available by the National Institutes of Health grant T36-GM-008789 to the Pittsburgh Supercomputing Center. This work also used computational resources provided through the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant OCI-1053575. Specifically, it used the Blacklight supercomputer system at the Pittsburgh Supercomputing Center (PSC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Results

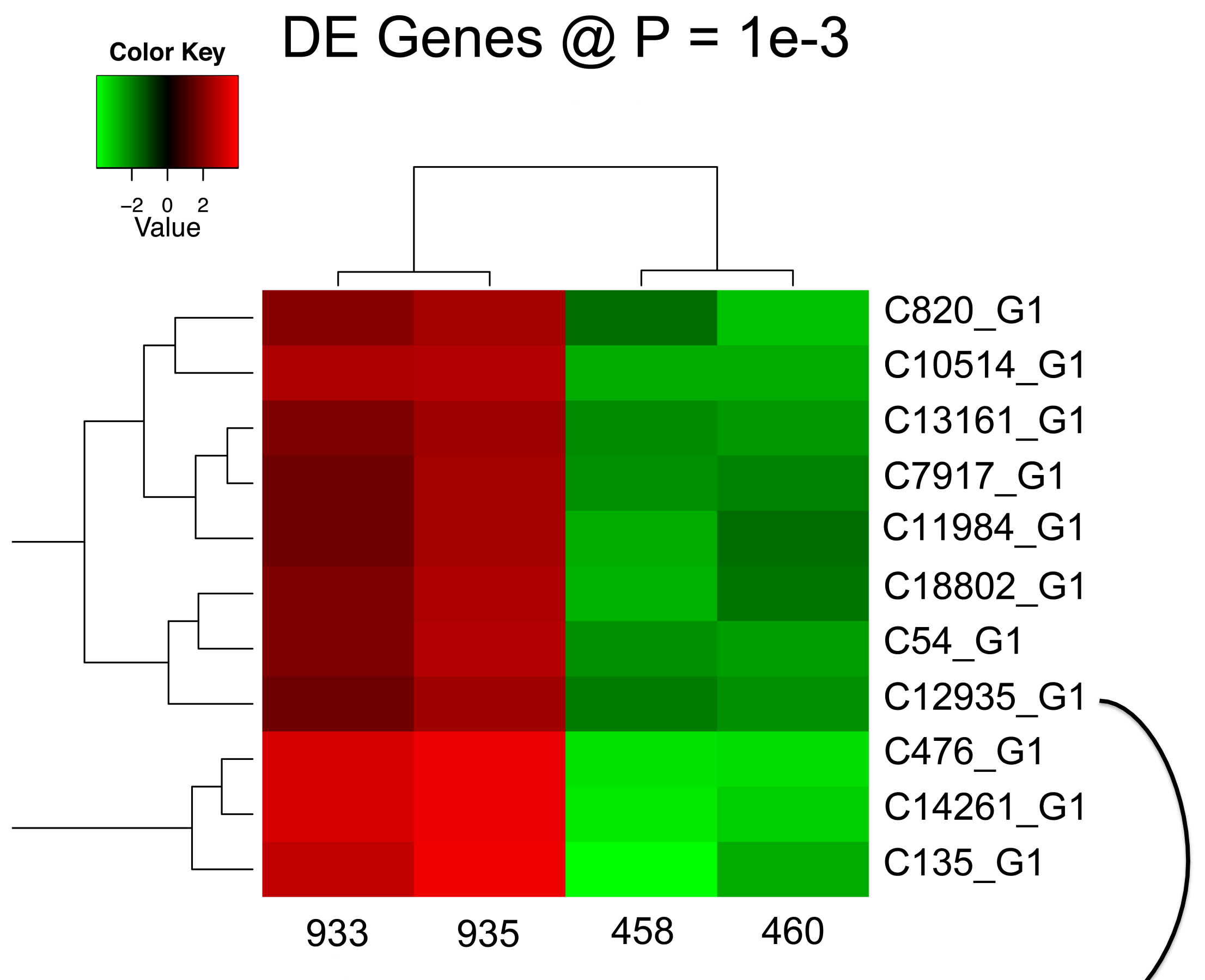
Mapping: TopHat/Cufflinks

This workflow identified hundreds of genes being differentially expressed at a p-value exceeding 1e-3.



De-Novo: Trinity/EdgeR

This workflow identified several genes being differentially expressed at a p-value exceeding 1e-3. Annotation of these genes identified them as coming from viral sources.



Example (BLAST): C12935_g1_i1 len = 1061 path = [1039:0-1060]

Top 5 sequences producing significant alignments ordered by coverage:	Max score	Total score	Query cover	E-value	Accession
Drosophila melanogaster birnavirus SW-2009a strain DBV segment A, complete sequence	**448**	**2506**	**99%**	**3e-160**	**GQ342962.1**
Nocardiopsis dassonvillei subsp. Dassonvillei DSM 43111 chromosome 1, complete sequence	42.3	77.3	35%	0.80	CP002040.1
Streptomyces nodosus strain ATCC 14899 genome	39.6	132	35%	7.2	CP009313.1
Actinoplanes sp. N902-109, complete sequence	38.6	74.2	28%	8.4	CP005929.1
PREDICTED: Dasypus novemcinctus ubiquitin specific peptidase 36 (USP36), transcript variant X1, mRNA	S35.0	64.5	27%	7.8	XM_004460442.2

Conclusion

Quantification methods could not be compared due to the presence of viral sequence data for Drosophila birnavirus and X virus. These genes were not identified using reference-based approach as their source was different from the reference used.

Future Work

- Remove virus transcripts at de-novo assembly step and repeat project workflow
- Apply draft methods to data from other non-model organism