Many-Sample de Novo Assembly and Analysis of Petromyzon marinus Transcriptome

Scott, Camille
camille.scott.w@gmail.com

Brown, C. Titus
ctbrown@ucdavis.edu

July 17, 2015

Background

Some biological background on sea lamprey, and why we care about them: ancestral jawed vertebrate, invasive species interest, regenerative capabilities, programmed genome rearrangement (PGR). This drives interest in a complete sea lamprey transcriptomic reference. [Cite: lamprey genome paper, more provided by weiming]

Background on NGS / RNAseq tech enabling deep mRNA sequencing. Lack of a complete reference due to PGR necessitates de novo assembly. De novo projects challenging because of difficulty in validation, data volume, data integrity. [Cite: ...]

Reproducibility crisis: many methods for pre-processing, assembly, and post-processing, but many difficult to replicate. Lack of documented tools, versions, parameters, source code for scripts. Demonstration of effective reproducible pipelines from start to finish. [Cite: assemblathon, mr-c, ...]

Goal: deeply characterize the sea lamprey transcriptome; produce a valuable resource for researchers; demonstrate efficacy of de novo approach for many-sample data; show open, reproducible pipelines.

Data Description

Section should include information on sea lamprey sample prep where available, table of sample descriptions including read lengths, insert sizes, sequencing technology, tissue type, and conditions. Also includes accessions.

[Table 1, samples]

Analyses

Assembly

We find 73.91% of annotations to be covered by a transcript from lamp10. Breaking down this percentage by feature type reveals that the results are biased by the inclusion of gene and transcripts features, both of which tend to contain large stretches of intronic sequence unlikely to be covered above our chosen cutoff by any transcript. When we consider only exons, 80.65% are covered, exons being a basic feature of mRNAs.

Table 1: Proportion of Annotations Covered by lamp00 and lamp10

	lamp00	lamp10
CDS	0.919196	0.814513
UTR	0.957400	0.862251
exon	0.896246	0.806504
gene	0.051476	0.147945
start_codon	0.960101	0.857505
stop_codon	0.636520	0.903922
transcript	0.048511	0.138392

Conversely, 29.72% of transcripts are covered by a single feature, while 99.89% of transcripts are covered in lamp00. We find the latter number encouraging; one would expect almost all transcripts in lamp00 to be covered by a single feature, as it was derived directly from the annotations, while previous evidence suggests that lamp10 is a superset of lamp00, thus explaining the disparity. Examining the extend to which our overlaps are a superset, we can break down transcript genome homologies by whether each transcript has only a homology, or both a homology and an annotation overlap, as follows.

Table 2: Proportion of Homologies and Annotation overlaps by Transcript in lamp00 and lamp10

	assembly	num	prop
presence			
+genome+ann	lamp00	11476	0.998868
+genome+ann	lamp10	212607	0.297209
+genome-ann	lamp00	0	0.000000
+genome-ann	lamp10	311948	0.436080

Futher, % of the genome is covered by annotations, while % is covered by alignments from lamp10; % of transcripts have any alignment to the genome.

We also find that % of transcript alignments entirely contain an annotation, increasing the annotation size by %. % of extensions are supported by homology to a known protein. % of transcript alignments are entirely contained by an annotation.

Pooled Assembly Discovers Novel Transcripts

Many-Sample Comparison

Include heatmaps, bar charts of unique gene content.

Discussion

Methods

In order to assess the completeness of our de Novo transcriptome assembly (lamp10), we have compared the alignment of the generated transcripts against the existing genome annotations released with Pmarinus v7.0.75. First, we use blastn to align transcripts to the genome, using parameters '-evalue 1e-6'. Then, we use the coordinates from the annotation and the corresponding coordinates from the alignments to calculate the proportion of annotated sequence overlapped, proportion of transcripts overlapped, and the respective proportions of non-overlapped sequence and transcripts. We consider an annotated region to be overlapped by a trancript if it is at least 90

We give particular attention to alignments which entirely contain annotated regions, as these suggest extensions to existing annotations. When these alignments are from transcripts with homology evidence from other species, we consider them to represent putative extensions [note: maybe not necessary to establish validity, instead just report the numbers]. Further, alignments which are entirely contained within an annotation suggest either an overly aggressive prediction in the genome, or an incompletely assembled transcript.

Pre-processing

Trinity Assembly

Post-processing