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

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

Brown, C. Titus
ctbrown@ucdavis.edu

September 4, 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

Many-sample de Novo Assembly

Here we describe basic assembly statistics:

1. Number of transcripts, coding regions.
2. Number of orthologies, homologies (broken down by genome support).
3. Accuracy, completeness, contiguity (with discussion of limitations).

de novo Assembly Improves Recall Over ab initio Predictions

This subsection has two goals: show that our assembly is reasonably valid, and show that it has better recall than lamp00. Though these two goals could ostensibly be split into two subsections, they have considerable overlap and mostly are shown by the same results. We make our case through the following points:

1. Our assembly has good recall of existing genes in lamp00. We show that we cover lamp00 by homology (blast lamp10 against lamp00), that we cover the gene products of lamp00 through annotation, and that we cover the GTF gene annotations on the genome.
2. Our assembly has good recall of core vertebrate orthologs. Here we present BUSCO results, both for lamp00 and lamp10.
3. Our assembly extends existing genes. Can be extracted from homology with lamp00 and TransDecoder results.
4. Our assembly includes novel genes. We show this by filtering out transcripts with homology to the lamprey genome or transcriptome, and finding orthologies (recipricol best hits) for those that remain. We do another level of filtering for false positives with the mygene API.
5. Our assembly has already proven useful to lamprey researchers **?** (perhaps this goes in background?)

Discussion point: do we split this into two sections?

We find 73.93% 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, 80.71% 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	0.896378	0.807120
gene	0.051323	0.147640
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	212606	0.297208
+genome-ann	lamp00	0	0.000000
+genome-ann	lamp10	311949	0.436082

With so many transcripts having alignments to the genome but no corresponding annotation, it would be valuable to further understand which of these transcripts have protein homologies to other databases. In particular, lamprey's uniquely valuable position in vertebrate evolution drives questions regarding loss and gain of genes within gnathostomes. To that end, we have subdivided these transcripts based on their homologies and orthologies with both zebrafish and amphioxus.

Table 3: BLAST Best Hits for Transcripts Filtered by Database Presence

braflo_best_hom	danrer_best_hom	no_ann	has_ann
True	True	36605	10293
True	False	847	873
False	True	7196	4445
False	False	167958	296338

Table 4: BLAST Orthologies for Transcripts Filtered by Database Presence

danrer_ortho	braflo_ortho	no_ann	has_ann
True	True	3664	833
True	False	2450	892
False	True	1255	437
False	False	205237	309787

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.

Improved recall discovers potential ancestral vertebrate genes

Here we talk about the genes we have shown to potentially be ancestral vertebrate orthologs. This is at least a useful result in its own right, but it would be nice to find something more compelling here. Immune-related genes might be a good starting point.

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% covered, with at least 98% identity [TODO: get better justification for these cutoffs other than "things Camille remembers reading"].

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

Describe pipeline: Trimmomatic PE or SE; digital normalization to C=20 on each sample (PE and orphans together for paired samples); pooled digital normalization C=20; filter-abund with variable coverage C=2 Z=20 using table output from pooled digital normalization run.

Trinity Assembly

Trinity assembly using all preprocessed reads. Final version will probably be with default settings.

Post-processing

cd-hit-est (or vsearch) used to remove redundancy. All transcripts aligned with BLASTX against zebrafish, amphioxus, mouse, lamprey, and human protein sequences downloaded from ensembl, and with BLASTN against lamprey version 7.0.75 genome, CDS, mRNA, and ncRNA. TransDecoder used to predict CDS, and hmmer used to make predictions against Pfam-A from predicted proteins. bowtie2 used to align all raw reads against assembly, and eXpress used for abundance estimation. Orthologies determined using recipricol best-hits (RBH). BUSCO ran to assess recall of core vertebrate orthologs.

Orthologs were filtered by whether they had any blastn hit to lamprey resources; protein IDs then queried with mygene to retreive gene symbols associated with each transcript, and symbols queried using the taxonomy tree option to determine gene membership in gnathostomata, cylcostomata, and cephalochordata lineages.