Extract Likely Coding Sequences from Trinity Transcripts

Likely coding regions can be extracted from Trinity transcripts using a set of utilities included in the Trinity software distribution. The system works as follows:

- the longest ORF is identified within the Trinity transcript (if strand-specific, this is restricted to the top strand).
- of all the longest ORFs extracted, a subset corresponding to the very longest ones (and most likely to be genuine) are identified and used to parameterize a Markov model based on hexamers. These likely coding sequences are randomized to provide a sequence composition corresponding to non-coding sequence.
- all longest ORFs found are scored according to the Markov Model (log likelihood ratio based on coding/noncoding) in each of the six possible reading frames. If the putative ORF proper coding frame scores positive and is highest of the other presumed wrong reading frames, then that ORF is reported.
- if a high-scoring ORF is eclipsed by (fully contained within the span of) a longer ORF in a different reading frame, it is excluded.

The scoring of ORFs is largely based on that described for GeneID, and the ORF selection process was added here.

2. Sequence Databases Required

Trinotate is built around specific releases of SwissProt and Pfam. You can also find these specific versions mirrored at the Trinotate ftp site:

Location on the HPC: **/home/mwamalwa/src/metagenome2014/transcriptome**

a) SwissProt Download SwissProt here

```
be sure the search database is uncompressed and properly formatted by
running the following (requires that blast+ is already installed as
indicated above) gunzip uniprot_sprot.fasta.gz makeblastdb -in
uniprot_sprot.fasta -dbtype prot
```

#Uniref90 Download Uniref90 here (Too big)

```
#gunzip uniref90.fasta.gz makeblastdb -in uniref90.fasta -dbtype prot
```

b) Pfam domains Download Pfam-A here

```
Uncompress and prepare it for use with 'hmmscan' like so: gunzip
Pfam-A.hmm.gz hmmpress Pfam-A.hmm
```

Running Sequence Analyses

Both the Trinity-assembled transcripts and any predicted protein-coding regions are subject to analysis using the various strategies below. Trinity transcripts are searched for sequence homologies using BLASTX, and RNAMMER is used to identify potential rRNA transcripts.

Transdecoder-predicted coding regions are also searched for sequence homologies using BLASTP, protein domains identified via a Pfam search, and additional properties such as signal peptides and likely transmembrane-spanning regions are explored.

1. Trinity files needed for execution

Trinity.fasta - Final product containing all the transcripts assembled by Trinity

Trinity.fasta.transdecoder.pep: Most likely Longest-ORF peptide candidates generated from the Trinity Assembly.

Transdecoder is included in Trinity at *$TRINITY_HOME/trinity-plugins/transdecoder/* ; Newer versions of TransDecoder should generate a file *Trinity.fasta.transdecoder.pep*. Earlier versions will generate an equivalent file called *transdecoder.pep*. These should be treated as equivalent outputs.

2. Extracting Best ORFs

Extracting likely coding regions from Trinity transcripts can be done using (beware, set up for strand-specific data by default, use -B for both strands):

NOTE: Use the perl script "run_Trinity_ORF.pl" to automate this task

```
% $TRINITY_HOME/trinity-
plugins/transdecoder/transcripts_to_best_scoring_ORFs.pl
with options:
############################## Options
################################################## # # Required: #
# -t transcripts.fasta # #  Optional: # # -m minimum protein length
(default: 100) # # -G genetic code (default: universal, options:
Euplotes, Tetrahymena, Candida, Acetabularia) # # -h print this
option menu and quit # -v verbose # # -C complete ORFs only ********
# -S strand-specific (only analyzes top strand) # -T top longest
ORFs to train Markov Model (hexamer stats) (default: 500) #
#####################################################################
#########################
```

Final output files include:

```
best_candidates.eclipsed_orfs_removed.cds
best_candidates.eclipsed_orfs_removed.pep
best_candidates.eclipsed_orfs_removed.gff3
best_candidates.eclipsed_orfs_removed.bed
```

The .bed file can be loaded into IGV for viewing along with the additional Trinity assembly and alignment data.

2. Capturing BLAST Homologies

BLAST information Instructions for installation of command line stand alone blast can be found here: http://www.ncbi.nlm.nih.gov/books/NBK52640/ NOTE: This step will undoubtedly take the longest,

for very large files execution on a multi-cpu server HPC environment is highly recommended, and your thread count should be equal to the number of CPU's present on the node the job is run on.

Blast Commands

search Trinity transcripts

blastx -query Trinity.fasta -db uniprot_sprot.fasta -num_threads 8 -max_target_seqs 1 -outfmt 6 > blastx.outfmt6

search Transdecoder-predicted proteins

blastp -query transdecoder.pep -db uniprot_sprot.fasta -num_threads 8 -max_target_seqs 1 -outfmt 6 > blastp.outfmt6

Optional: perform similar searches using uniref90 as the target database, rename output files accordingly.

blastp -query transdecoder.pep -db uniref90.fasta -num_threads 8 -max_target_seqs 1 -outfmt 6 > uniref90.blastp.outfmt6

blastx -query Trinity.fasta -db uniref90.fasta -num_threads 8 -max_target_seqs 1 -outfmt 6 > uniref90.blastx.outfmt6

3. Running HMMER to identify protein domains

hmmscan (HMMER) command:

hmmscan --cpu 8 --domtblout TrinotatePFAM.out Pfam-A.hmm transdecoder.pep > pfam.log