Transdecoder

TransDecoder identifies candidate coding regions within transcript sequences, such as those generated by de novo RNA-Seq transcript assembly using Trinity, or constructed based on RNA-Seq alignments to the genome using Tophat and Cufflinks.

TransDecoder identifies likely coding sequences based on the following criteria:

- a minimum length open reading frame (ORF) is found in a transcript sequence
- a log-likelihood score similar to what is computed by the GeneID software is > 0.
- the above coding score is greatest when the ORF is scored in the 1st reading frame as compared to scores in the other 2 forward reading frames.
- if a candidate ORF is found fully encapsulated by the coordinates of another candidate ORF, the longer one is reported. However, a single transcript can report multiple ORFs (allowing for operons, chimeras, etc).
- a PSSM is built/trained/used to refine the start codon prediction.
- **optional** the putative peptide has a match to a Pfam domain above the noise cutoff score.

Running Transdecoder

Running Transdecoder is done in two steps, first it takes the fasta File and extracts the longest ORFs to calculate a Markov model.

```
/Path/to/Transdecoder/TransDecoder.LongOrfs -t Transciptome.fasta
```

Then with its own Markov model it predicts the likely coding regions.

```
/Path/to/Transdecoder/TransDecoder.Predict -t Transcriptome.fasta [homology options]
```

Optionally you can identify ORFs with homology to known proteins via blast or pfam searches. If you do then you may add the homology options stated on their site. Link

Transdecoder Output

File	Content
longest_orf.pep	all ORFs meeting the minimum length criteria, regardless of coding potential.
longest_orfs.gff3	positions of all ORFs as found in the target transcripts
longest_orfs.cds	the nucleotide coding sequence for all detected ORFs
longest_orfs.cds.top_500_longest	the top 500 longest ORFs, used for training a Markov model for coding sequences.
hexamer.scores	log likelihood score for each k-mer (coding/random)
longest_orfs.cds.scores	the log likelihood sum scores for each ORF across each of the 6 reading frames

File	Content
longest_orfs.cds.scores.selected	the accessions of the ORFs that were selected based on the scoring criteria
transcripts.fasta.transdecoder.pep	peptide sequences for the final candidate ORFs; all shorter candidates within longer ORFs were removed
transcripts.fasta.transdecoder.cds	nucleotide sequences for coding regions of the final candidate ORFs
transcripts.fasta.transdecoder.gff3	positions within the target transcripts of the final selected ORFs
transcripts.fasta.transdecoder.bed	bed-formatted file describing ORF positions, best for viewing using GenomeView or IGV

The first seven files are generated from calculating the long ORFs while the last four are the actual ORFs that were predicted from the Algorithm. Load the reference Transcriptome with the transcripts.fasta.transdecoder.bed into IGV to visualize the results and for a more detailed information simply extract the sequences from the .cds or .pep file via the Terminal.