Unigene set selection

Basic problem

RNA seq is method to sequence mRNAs from any organisms. This method is so popular and being affordable to many research groups. It has important applications such as gene discovery, differential gene expression and SNP calling and Alternative splicing events etc. Unfortunately, the RNA seq assembly and analysis has many challenges. Because it produces multiple contigs per gene locus and they are not truly representative of isoforms. Therefore, I made a strategy to select one contigs per gene locus.

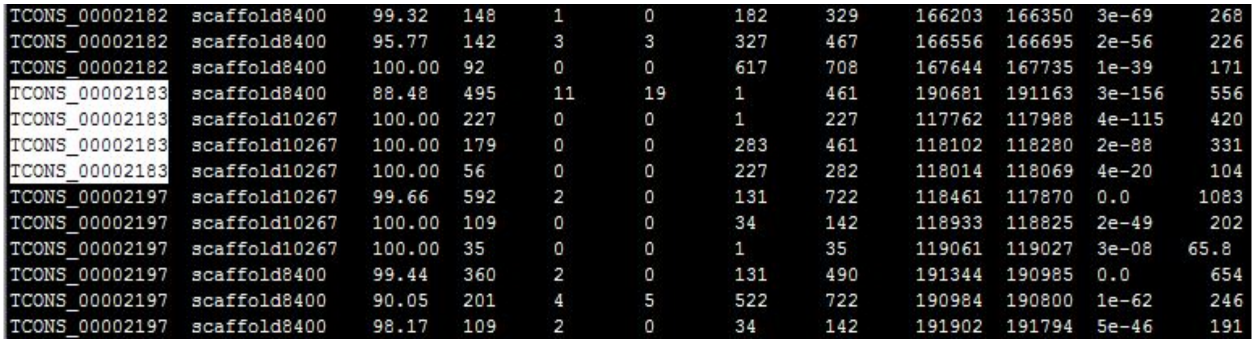
First, We align assembled transcriptome sequence to reference genome sequence using below blast command.

"blastn -db genome_v01 -query unigenes_seq.fa -num_threads 35 -evalue 1e-05 -outfmt 6 -max_target_seqs 2 -out genome_hits.txt"

All the steps are done on this output file (Pg_genome_hits.txt)

1_alignment_issue.py

This script resolves an error in the output file.



```
TCONS_00002182  scaffold8400    99.32   148     1       0       182     329     166203  166350  3e-69       268
TCONS_00002182  scaffold8400    95.77   142     3       3       327     467     166556  166695  2e-56       226
TCONS_00002182  scaffold8400    100.00  92      0       0       617     708     167644  167735  1e-39       171
TCONS_00002183  scaffold8400    88.48   495     11      19      1       461     190681  191163  3e-156      556
TCONS_00002183  scaffold10267   100.00  227     0       0       1       227     117762  117988  4e-115      420
TCONS_00002183  scaffold10267   100.00  179     0       0       283     461     118102  118280  2e-88       331
TCONS_00002183  scaffold10267   100.00  56      0       0       227     282     118014  118069  4e-20       104
TCONS_00002197  scaffold10267   99.66   592     2       0       131     722     118461  117870  0.0        1083
TCONS_00002197  scaffold10267   100.00  109     0       0       34      142     118933  118825  2e-49       202
TCONS_00002197  scaffold10267   100.00  35      0       0       1       35      119061  119027  3e-08      65.8
TCONS_00002197  scaffold8400    99.44   360     2       0       131     490     191344  190985  0.0         654
TCONS_00002197  scaffold8400    90.05   201     4       5       522     722     190984  190800  1e-62       246
TCONS_00002197  scaffold8400    98.17   109     2       0       34      142     191902  191794  5e-46       191
```

From the out result, we need to extract the top hits. But some hits can be first based on blast score but they may not be biologically from same locus. So these type of error can be detected by this script

2_Format_genome_hits.py

The above errors will be removed by this script and produce new output file "formatted_genome_hits.txt"

3_top_genome_alignment.py

From the above output file, we select only the top hit alignment and this will produce "new_top_hits_1e_5.txt "

4_mapped_region_extract.py

From the above top hits result, make each sequence ID’s start and end position on the genome sequence. This gives output of “T_mapped_regions.txt”

5_Gene_cluster_final.pl

Then each sequence IDs are clustered using the perl script. It generates “final_updated_cluster.txt” file.

6_mode_new_filter.py

Then we calculate mode value for each clusters based on alignment. It will give “Nr_mode_update.txt” file.

7_2_Final_transcript_cluster.py

This script will make consensus cluster based on the above mode output file. It gives “final_cds_update.txt”

7_1_RNA_contribution.py

This script calculates sequence size of each contigs and produce “seq_id_size.txt”

8_uni_merge.py

This script will merge “final_cds_update.txt” and “seq_id_size.txt” by ID wise and gives “formatted_cds_update.txt”

9_unigene_selector.py

Based one sequence length, it selects candidate set from each cluster and gives output as “candidate_unigenes_update.txt”.