CRSP: Comparative RNA-seq Pipeline

```
Python Requirement: Python version >= 3.0
You will need to have the following programs installed and in your PATH
 - cd-hit-est (https://github.com/weizhongli/cdhit)
 - NCBI-BLAST+ (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/)
 - RSEM (http://deweylab.github.io/RSEM/)
 - Bowtie (http://bowtie-bio.sourceforge.net/)
```

Workflow (Examples)

1. Merge multiple transcriptome assemblies, prepending a label to contigs from each assembly to avoid name collisions

```
./src/crsp_merge_assemblies.py assembly_1 ./User_Assembly_Files/Trinity_assembly_1.fasta \
                         assembly_2 ./User_Assembly_Files/Trinity_assembly_2.fasta \
                         > merged_assembly.fasta
```

2. Create a non-redundant transcriptome assembly using cd-hit-est

```
cd-hit-est -M 200000 -T 0 -i merged_assembly.fasta -o non_redundant_assembly.fasta -c 1 \
     > reduce_assembly.log
```

3. Create a BLAST database for a comparative reference protein set [please see the folder: Blastdb_Protein]

CRSP (default) has build the mouse protein blastdb (the users can build their own blastdb)

```
makeblastdb -in ./Blastdb_Protein/Screened_Mouse_Protein_with_UniqueID.fa \
             -dbtype prot \
         -out ./Blastdb_Protein/Mouse_protein_blast_index \
         -title mouse_protein \
             -logfile mouse_protein.makeblastdb.log
```

4. Run BLAST+ locally on the non-redundant assemlby and the comparative protein set

```
blastx \
    -num_threads 96 \
    -db ./Blastdb_Protein/Mouse_protein_blast_index \
    -outfmt 6 \
    < non_redundant_assembly.fasta \
    > mouse_protein.blast \
    2> mouse_protein.blast.log
```

5. Extract the best BLAST hit for each contig

```
./src/crsp_blast_tophit.py < mouse_protein.blast > mouse_protein.tophits
```

6. Create a mapping from contigs to comparative reference proteins with the given e-value threshold

```
./src/crsp_tophits_to_map.py -e 0.00001 < mouse_protein.tophits > contig_to_mouse_protein.map
```

7. Prepare an RSEM reference using the non-redundant assembly

```
rsem-prepare-reference \
    --bowtie \
    non_redundant_assembly.fasta \
    rsem_reference \
    &> rsem_prepare_reference.log
```

8. Compute contig expression levels using RSEM | The example is for paired end reads. For single end reads please see RSEM website: http://deweylab.github.io/RSEM/

```
rsem-calculate-expression \
    --bowtie-n 2 \
    --no-bam-output \
    --paired-end \
    ./User_RNASeq_Files/Sample_1.R1.fastq \
    ./User_RNASeq_Files/Sample_1.R2.fastq \
    rsem_reference \
    Sample_1 \
    &> rsem_calculate_expression.log
```

9. Map contig expression levels to comparative reference protein expresion levels

```
./src/crsp_map_abundance_estimates.py \
    contig_to_mouse_protein.map \
    < Sample_1.genes.results \
    > Sample_1.proteins.results
```

10. Map protein expression levels to gene symbol expresion levels

```
./src/crsp_map_abundance_estimates.py \
    ./Blastdb_Protein/Mouse_Protein_UniqueID_to_Symbol.map \
    < Sample_1.proteins.results \
    > Sample_1.gene_symbols.results
```

Output File:

Sample__1.gene__symbols.results

Citation

Bagheri A., Dewey C., Stewart R., Jiang P. CRSP: Comparative RNA-seq pipeline for species lacking both a reference genome and annotated transcriptome (Submitted)

Contact

Peng Jiang, Ph.D
Assistant Professor
Center for Gene Regulation in Health and Disease (GRHD)
Center for Applied Data Analysis and Modeling (ADAM)
Department of Biological, Geological and Environmental Sciences (BGES)
Cleveland State University, 2121 Euclid Ave, Cleveland, OH 44115
Lab Website: https://sites.google.com/view/jiang-lab/