1 Introduction

Welcome to the Bellerophon Pipeline. This pipeline is commissioned by the Institute for Biodiversity and Ecosystems Dynamics with the University of Amsterdam in cooperation with the department of Bio-informatics with the Applied University Leiden.
The goal of this pipeline is to assess the quality of a *de novo* assembled transcriptome. In short, the software uses different methods of filtering to optimize the transcriptome's quality.

2 Functionality

This pipeline filters a *de novo* transcriptome in either two or three ways, chosen by the user. The transcriptome is filtered using CDHIT-EST. This tool clusters highly similar transcripts and chooses one representative transcript per cluster. The second filtering is based on expression level. Using the RSEM tool, normalized expression values (TPM) are calculated. The user can execute a third filtering method if necessary. This filtering method is not used by default. This method filters the transcripts based on the length of predicted Open Reading Frames. The tool TransDecoder provides a set of ORFs that meet the cut-off length, regardless of coding potential.
To assess the quality of the transcriptome, the tool TransRate is used. This tool requires the paired end reads and the transcriptome. It calculates a quality score for the assembly and provides several statistics on the reads, the transcriptome and the quality of the transcriptome To assess the completeness of the transcriptome and the effect of the filtering on the completeness, the tool BUSCO is used. This tool searches for orthologs that are all supposed to be present single-copy in the transcriptome. The *Insecta* reference set is used by default.
The pipeline will execute in the current directory in a subdirectory with the name Run_(date)_(time)

3 Dependencies

This pipeline uses several tools. These tools can be divided in provided tools and dependent tools. The provided tools are:

1. TransRate
2. RSEM (as part of Trinity)
3. BUSCO
4. TransDecoder
5. faSomeRecords

These tools can be found in the utils directory.
The dependent tools are:

1. CDHIT-EST
2. Python
3. Unix environment
4. Bash

The dependent tools need to be installed on the user's system before the pipeline is able to execute properly. Along with these requirements, the pipeline needs at least 150 GB of free disk space. The alignment files created by TransRate are very large. These files are automatically removed by the pipeline (can be counteracted by a flag) but the space has to be available during the execution of the pipeline. Also, RSEM creates a large BAM file that is not removed by the pipeline. This file can safely be removed once RSEM is done running to free up the disk space. Multithreading is possible for most used tools. The more threads are used, the faster the pipeline will run. Some tools require a good amount of working memory. It is advised to have a minimum of 16 GB of RAM.

4 Usage information

To start the pipeline execute the script Bellerophon_startscript.sh with at least the following parameters:

- -a/–assembly Transcriptome assembly (fasta)
- -l/–left Left side reads (fastq)
- -r/–right Right side reads (fastq)

Example:

```
bash Bellerophon_startscript.sh --assembly assembly.fasta
        --left reads_R1.fastq --right reads_R2.fastq
```

Optional parameters are:

- -c/–cdhit_cutoff CDHIT-EST identity cutoff. 0.95 by default
- -t/–tpm_cutoff Expression cut-off. 1 By default.
- -o/–orf_cutoff ORF length cutoff. 50 aa by default
- -T/–threads number of threads to use. 4 By default
- -D/–debug run in Debug Mode (prints out parameter information). False by default

- -O/–order Filtering order. 1 By default. (Table 1)
- -k/–keep_bam the pipeline doesnt remove the large alignment files made by TransRate. False by default

Example:

```
bash Bellerophon_startscript.sh --assembly assembly.fasta
        --left reads_R1.fastq --right reads_R2.fastq
        --threads 12 -o 100 --debug -O 5
```

Filtering code	Filtering order
1	TPM, CDHIT-EST
2	CDHIT-EST, TPM
3	CDHIT-EST, ORF, TPM
4	CDHIT-EST, TPM, ORF
5	ORF, CDHIT-EST, TPM
6	ORF, TPM, CDHIT-EST
7	TPM, CDHIT-EST, ORF
8	TPM, ORF, CDHIT-EST

5 Output

The output files of the pipeline will be located in the current directory in a folder named Run_(date of execution)_(time of execution). In this directory, several subdirectories are found. /Bowtie2_files/ contains the index files Bowtie2 creates when used in RSEM. /Fasta_files/ contains all fasta files of the filtered assemblies. /Log/ contains all log and error files (plain text). RSEM_files/ contains all RSEM library files. /TPM_output/ contains the RSEM expression filter data. In this directory, a large BAM file can be found that can be removed to free disk space. /TransRate_output/ contains all TransRate subdirectories for all TransRate runs. The final assembly can be found in the Fasta files directory, indicatted with the suffix 'BEL'. /Misc_output/ contains several output files that are not to be placed elsewhere.

6 Disclaimers

Please note that this software is created for quality assessment and improvement of Illumina paired end reads, assembled with Trinity *de novo* Transcriptome assembler. Also, this pipeline is developed and only tested with insect data. The BUSCO part of the pipeline uses the Insecta dataset. This dataset can be changed to any BUSCO dataset, but the directory is hardcoded. The pipeline's code needs to be altered to find another BUSCO reference set.
Every bash subscript of the pipeline can be used stand-alone. However, the scripts are specifically created for this pipeline and there are no guarantees that the tools will operate correctly out of context. The R, Python and Tex scripts

are only to be used in context of this pipeline as they will not work without the pipeline's exact input.