dbHT-Trans (v1.0) Manual

Deng *et al.*

The dbHT-Trans is an efficient tool for filtering the assembled transcripts of RNA-Seq based on homologous protein sequence search against reference database. Two modules of Operator and Extractor are independently designed for data processing and filtering action, respectively. Due to the MySQL-based design, the key features of dbHT-Trans include the large reduction of storage requirement and flexible outputs.

Essential requirements

- **Operating system**

  The dbHT-Trans can only be operated on Linux systems at present.

- **MySQL**

  You can download the installation package from official website (http://dev.mysql.com/downloads/mysql/) or directly execute the terminal command as '*apt-get install mysql-server*'.

- **Python**

  The dbHT-Trans is written in python 2.7.3. Therefore, Python 2.x.x is required and can be downloaded from official website (www.python.org/downloads/). However, we recommend that you would prefer to install python 2.7.x. If you have correctly installed python, you can check the version information by executing terminal command as '*python --version*'.

- **Python package for MySQLdb**

  One additional third-party library of MySQLdb is required for operating dbHT-Trans (https://pypi.python.org/pypi/MySQL-python/). Please follow its official instructions to install this library.

Download and configure

- **Download**

  You can download dbHT-Trans from https://github.com/chengroup/dbHT-Trans and unzip the file of 'dbHT-Trans-master.zip'.

  ```
  $ unzip dbHT-Trans-master.zip
  $ cd dbHT-Trans-master
  ```

- **Configured by MySQL information**

You must first edit the file of ‘**config.txt**’ (./dbHT-Trans-master/config.txt) by updating the required MySQL information, which includes the following fields:

1) The field of ‘**host**’ (line 1) is the address of MySQL. In general, MySQL is locally installed and you just use the default value as ‘localhost’. Otherwise, you must replace this address by your MySQL server’s IP address.
2) The field of ‘**username**’ (line 2) is the username of your MySQL, which had been denoted when you installed MySQL database.
3) The field of ‘**password**’ (line 3) is the password of your MySQL, which had been denoted when you installed MySQL database.
4) The field of ‘**database**’ (line 4) is a customer name of database which you are going to create. All of your data in relation to dbHT-Trans will be stored and processed in this unique database. If you want to start a different and independent job, you must use one new name; otherwise the older data will be overwritten by the new data.
5) The field of ‘**port**’ (line 5) is the port of your MySQL. In general, you don’t need to change the default value (3306). The ‘port’ of database is an advanced setting.

After updating these fields, please enable the configuration by executing the following command:

```
$ chmod a+x dbHT-Trans-Extractor dbHT-Trans-Operator
$ chmod a+x model/usearch7.0.1090_i86linux32
$ python config.py
```

How to use dbHT-Trans

The dbHT-Trans is divided into two individual modules including (1) the processing system (**dbHT-Trans-Operator**) for ORF finding, protein translation and homologous search against reference database; (2) the outputting process for customer results (**dbHT-Trans-Extractor**). After completing the ‘dbHT-Trans-Operator’ process, the module of ‘dbHT-Trans-Extractor’ can be independently and repeatedly operated for outputting customer results.

- ### dbHT-Trans-Operator

Usage: dbHT-Trans-Operator [option(s)] -T <*.fasta> -d <*.fasta>

-h/--help
Note: this module will not output any result.

Required:

-T/--transcript <fasta> the inputted fasta file that contains all transcript sequences.
-d/--database <fasta> the reference protein database used for homologous search.

Optional:

-G/--genetic_code <string> genetic codes (default: universal, options: Euplotes, Tetrahymena, Candida).

-g/--gene_list <table file> a list for denoting the match between the transcript name and gene name. The first column is for transcript name and second column for gene name.

-L/--min_orfs <int> to only retain ORFs with nucleotide length equal to or longer than this value (default: 300 bp).

-p/--processes <int> to set the processing threads. default: 2

Advanced:

-a/--identity <float> the lowest threshold of sequence identity for USEARCH tool's homologous search (default: 0.5). Only the retained transcripts by USEARCH tool will be further subjected to the customer filtering of 'dbHT-Trans-Extractor'. **Note**, the customer filtering threshold of 'dbHT-Trans-Extractor' should be equal to or higher than this USEARCH tool's setting.

-q/--query_cov <float> the lowest threshold of query coverage (= alignment length / query length) for USEARCH tool's homologous search (default: 0.5). The explanation is same to 'identity' as above.

- ## dbHT-Trans-Extractor

Usage: dbHT-Trans-Extractor [options] -i <float> -c <float> -T <filename>

-h/--help

Required:

-i/--ident <float> the customer threshold of sequence identity for filtering transcripts. This value should be not lower than the lowest threshold of sequence identity for USEARCH tool's homologous search of 'dbHT-Trans-Operator'.

-c/--cov <float> the customer threshold of query coverage (=alignment length / query length) for filtering transcripts. This value should be not lower than the lowest threshold of query coverage for USEARCH tool's homologous search of 'dbHT-Trans-Operator'.

-T/--transcript_prefix <filename> to denote the file name's prefix for the two output files, which contain the retained transcripts and discarded transcripts, respectively. (default: transcript)

Optional:

-S/--stat <filename> to output the results of statistical comparison before and after dbHT-Trans filtering.

-M/--meta_table <filename> to output the detailed results of homologous sequence search for all ORF-positive transcripts (the transcripts with ORF(s) to be detected).

-C/--cds_file <filename> to output CDS sequences for these retained transcripts.

-p/--protein_file <filename> to output protein sequences fo these retained transcripts.

-F/--filtered_gene <filename> to output all genes which don't have any transcript finally retained by 'dbHT-Trans-Extractor'. This function is only available when the match list between transcript name and gene name was provided.

-t/--orf_style <string> to specify the ORF type for outputting results, which contains the "complete", "3primer_partial", "5primer_partial", "internal".More than one styles (splitted by comma) are supported. (default:complete,3primer_partial,5primer_partial, internal)

Details of dbHT-Trans output

By default, the module of '**dbHT-Trans-Extractor**' will only output two files (*<filename>.retained.fasta* and *<filename>.discarded.fasta*) in fasta format. Additional files would be outputted according to custom denotation as described above.

- **<filename>.retained.fasta**

This file is in fasta format and contains the retained transcripts after customer filtering of 'dbHT-Trans-Extractor'.

- **<filename>.discarded.fasta**

This file is in fasta format and contains the discarded transcripts after customer filtering of 'dbHT-Trans-Extractor'.

- **statistical results** (specified by argument of '-S/--stat')

  This file is in text format and contains the statistical information, including the length distribution of transcripts and gene counts.

- **Detailed results of homologous sequence search** (specified by argument of '-M/--meta_table')

  This file is in tab-delimited text format and contains 11 columns for describing the detailed results of homologous sequence search of these ORF-positive transcripts:

  Column 1: transcript name
  Column 2: the corresponding gene name
  Column 3: transcript length
  Column 4: start position of ORF
  Column 5: end position of ORF
  Column 6: start position of its parent ORF (The 'parent ORF' is referred to the longest ORF by fully encompassing this one.)
  Column 7: end position of its parent ORF
  Column 8: the type of ORF. This includes four types of 'complete' (having both start codon and end codon), 'internal' (having neither start codon nor end codon), '5primer_paritial' (having end codon but not start codon), and '3primer_paritial' (having start codon but not end codon).
  Column 9: strand type
  Column 10: the number of target sequences by homologous search
  Column 11: the detailed description for all homologous alignments of each query transcript. The description for each alignment is arranged in one bracket and contains four fields of 'target sequence name', 'sequence identity', 'query coverage', and 'target coverage'.

- **CDS sequences** (specified by argument of '-C/--cds_file')

  This file is in fasta format and contains the CDS sequences which have the homologous sequences in database.

- **Protein sequences** (specified by argument of '-p/--protein_file')

  This file is in fasta format and contains the protein sequences which have the homologous sequences in database.

- **Gene list** (specified by argument of '-F/--filtered_gene')

  This file is in text format and records the gene name in each line which doesn't have any transcript finally retained.

Example data and pipeline

We provide one example data and corresponding pipeline (both in the folder of ‘**example**’), which include 10, 000 assembled transcripts of mouse RNA-Seq. You can operate the tool of dbHT-Trans by this pipeline step by step.