TnClone manual for linux users

Contents

1. System requirements

- Windows 7

- Python 2.7

- PyQt4 library

- Python HTSeq library

- Python MatplotLib library

- Java 7 or higher

- Memory at lease 8GB

Python dependency check

Windows is simple. Just download python.

Java installation check

Open command line prompt and type things described below

```
C:\Users\허성훈>java -version
java version "9.0.4"
Java(TM) SE Runtime Environment (build 9.0.4+11)
Java HotSpot(TM) 64-Bit Server VM (build 9.0.4+11, mixed mode)

C:\Users\허성훈>
```

If one is missing any of these dependencies, please download dependent module via library supplier's description

2. Download

Download of TnClone from URL below

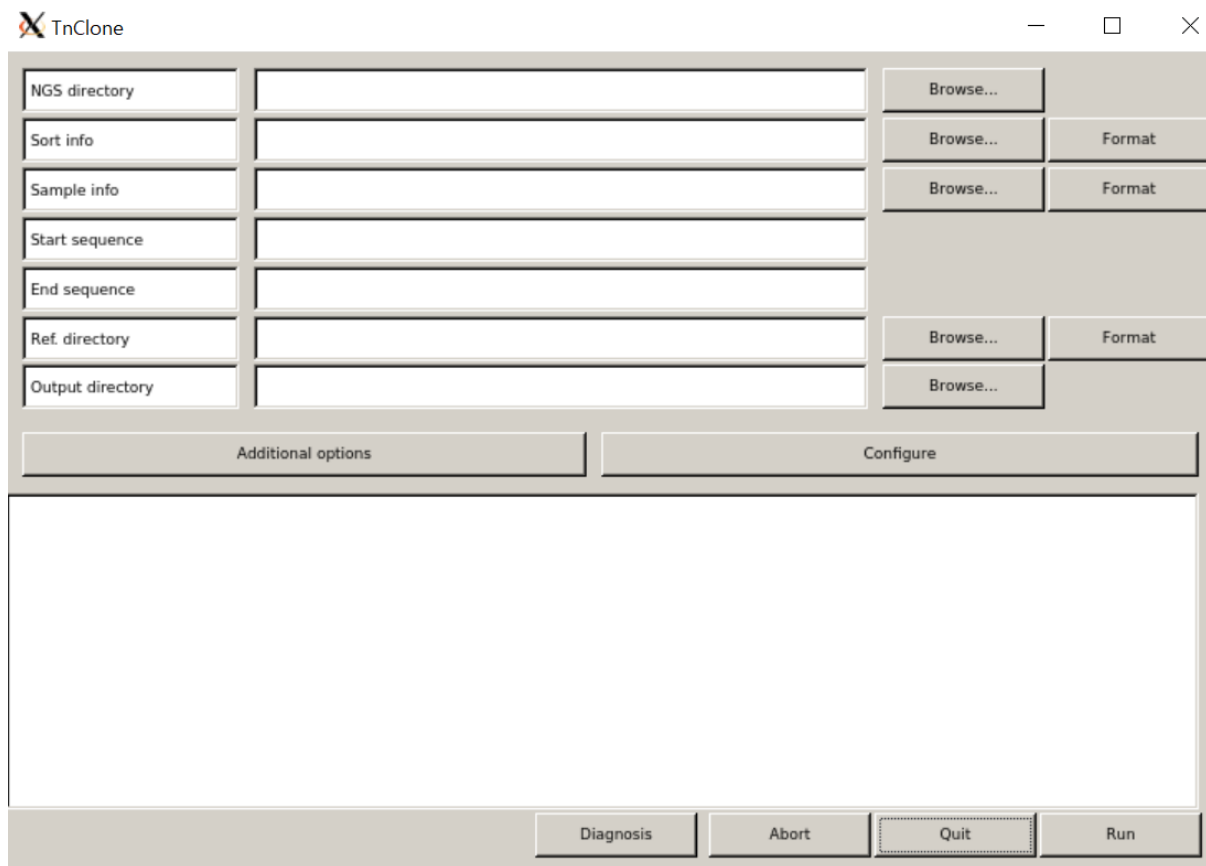https://drive.google.com/drive/u/4/folders/10x1B7rVWPt0rPMyT2ZIMdAXSn-2VuhVu

Double click tnclone icon. Now you are ready to run TnClone

3. Quick start guide

Double click downloaded TnClone

Then you can see the main window of TnClone

Give TnClone about information of samples

Then fill the fields for TnClone to be execute with input files specified in section 5

Execution of TnClone

Press "Configure" button and then press "Run" button

If user do not press "Configure" button TnClone do not know what to do and fails to execution.

4. UI description

(1) NGS directory

This field will make computer to recognize your raw NGS sequencing file location

Click Browse button next to the field and search for your NGS directory

(2) Sort info

This field tell program how to sort your files under Tn5 ME sequence combinations

Only reverse barcode is acceptable

See File format section for the file format

(3) Sample info

This field will tell program how many samples you required.

This field is a required field for trimming and assembly and downstream analysis step.

(4)   Start sequence

The assembly start sequence. Must be the same length as k-mer's length used for graph construction (length default : 63)

If Ref. directory (6) is set and BED file is supported(see 8), then it is no longer required. Program will automatically select start sequence

(5)   End sequence

Assembly end sequence. Must be the same length as k-mer's length used for graph construction (length default : 63)

If Ref. directory (6) is set and BED file is supported(see 8), then it is no longer required. Program will automatically select start sequence

(6)   Ref. directory

Reference directory required for downstream analysis or automatic start, end sequence detection. Choose a directory where your reference are located. Reference file must end with 'gene.fa' (i.e. extension if gene.fa)

(7)   Output directory

The parent directory for output

(8)   Additional options

This field will tell you more detailed options for analysis

Almost every option formats or descriptions can be obtained by pression Description button at relevant location. Simple description is listed below

K-mer size : The size of k-mer

Minimum k-mer occurrence : Cutoff value of k-mer abundance. If k-mer's abundance is smaller than this value, those k-mers will be dropped

Analysis steps : TnClone's analysis steps. If one wants to use de novo assembly mode, set "assembly" box as checked. If one wants to reference mapping based analysis then uncheck "assembly" box

See difference below

De novo assembly mode

Reference alignment mode

BED file : A region description file that user wants to assemble. See detailed format at https://genome.ucsc.edu/FAQ/FAQformat.html#format1

Down sampling : Tell tnclone to downsample if file size is too large

Down sampling ratio : The ratio if TnClone downsamples data

Seed Mismatch Correction : Correction method for seed search. One can choose either "Depth" based method or "Mismatch" based method. Default is "Depth" method

Num mismatch : Number of allowed mismatches for seed when user selects "Mismatch" based method for seed correction

Sequence match score : The match score when perform Smith-Waterman alignment. Must be positive integer number

Sequence mismatch score : The penalty score when perform Smith-Waterman

alignment. Must be positive integer number

Gap open penalty : The gap open penalty with positive integer.

Gap extension penalty : The gap extension penalty with positive integer

(9)　Configure

Configure program options. Must be clicked before use

(10)　Abort

Abort program while running if user wants to stop.

(11)　Quit

Quit program. Difference between abort and quit is abort will just stop analysis and leave screen opened but quit button will close the window and shutdown program.
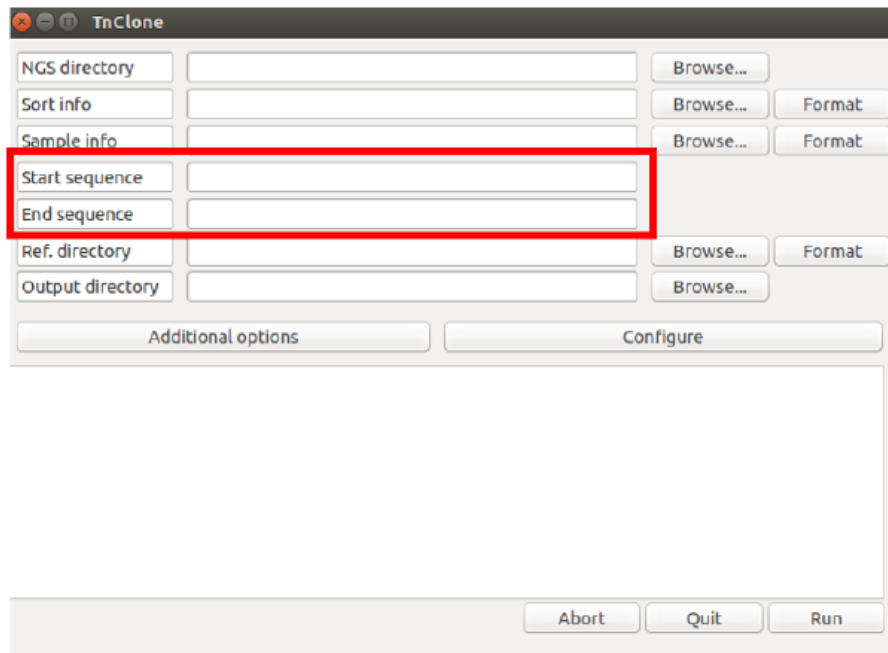
(12)　Run

Run the program

(13)　Browse buttons

Browse for specific files or paths

(14)　Format buttons

Will show description for formats

Optional : Start and End sequence

The program originally designed for analyzing insert DNA sequence after cloning so one must specify the start and end sequence of assembly. If one left reference blank and BED file field blank (will be discussed soon), then must fill these fields. Those can be flanking sequences at the end of insert. By choosing these sequences carefully, one can assemble insert or some part of backbone DNA sequence.

5. Input formats

(a) The sort information file

Sort information contains five columns. Each columns are separated by tab.

Those columns are arranged as described below

SAMPLE_NAME<TAB>NUMBER_OF_SAMPLES<TAB>ME_SEQUENCE1<TAB> ME_SEQUENCE2<TAB>FWD_FASTQ<TAB>REV_FASTQ

Field description

SAMPLE_NAME : The name of sample for your insert (Like Cas9, CRISPR, Myc, TP53 etc)

NUMBER_OF_SAMPLES : Number of samples (colonies) you have run NGS

ME_SEQUENCE1 : Mosaic End sequence for Tn5 barcoding step (F1~F8 in paper)

ME_SEQUENCE2 : Mosaic End sequence for Tn5 barcoding step (R1~R8 in paper)

FWD_FASTQ: The RAW FASTQ files for this sample.

REV_FASTQ : The RAW FASTQ files for this sample. Other pairs of files listed in FWD_FASTQ

(b) The sample information file

Sample information file contains two columns. Each columns are separated by tab.

Those two columns are exactly same as the first two columns of sort file

See example at https://github.com/tahuh/tnclone/tree/master/test_data

(c) Reference format

All reference file must be FASTA format

Also reference file must start with the gene/sample name specified in sample information or sort information and end with "gene.fa"

(d) BED file

BED file contains at least three fields (see details at https://genome.ucsc.edu/FAQ/FAQformat#format1) . But for our analysis, we will use this as region to assembly. The first column is not chromosome but the sample name used for sort file or sample information file. And chromStart column will denote the start position of assembly. chromEnd column will denote end position of assembly. Both coordinates depend on the reference sequence. If reference changes then one must carefully lookup bed file.

Example below

```
D1VXP4   0    3727
D3FJ35   0    3878
E1Z024   0    4625
F0RSV0   0    3593
G2Z1C1   0    4484
Q7P7J1   0    4186
```

CAUTION

The coordinates are 0-based coordinate not 1-based. So the 0 denotes the very first base of insert sequence.

6. Output format

TnClone will output number of assembled clones information, error contigs description as written in README file

(https://github.com/tahuh/tnclone/blob/master/README.md).

The other valuable information that TnClone generates are under directory of "sam" and "vcf" of output directory

Sam directory contains alignment result by aligner

See details about sam file format at

https://samtools.github.io/hts-specs/SAMv1.pdf

VCF directory contains the variant information of assembled contigs /mapped reads. If one performed de novo assembly mode, then this directory will

contain assembled contig variant information. If one used reference mapping mode, then this directory will contain variant information generated by reads for this sample

Analysis result

The analysis result will locate at 'report' folder of under your output directory. This file contains brief information of analysis. File will contain

Number of contigs assembled.

DNA error/match contig (compared to reference)

Protein error/match contig (compared to reference)

In 'umber of contigs' section in analysis.report , C,I/D,NC are followings

C : Confident Contig

I/D : InDel contig

NC : Non-confident contig

Other files such as dna_free_contig_info.txt and protein_free_contig_info.txt will contain information about the assembled contig's identifier for each field

Others

Assem folder : Assembly contig sequence will locate at this folder. The final contig file will end with 'final.fa'

Kfs folder : The k-mer frequency spectrum folder. This will tell you k-mer frequency spectrum

Tmp folder : This folder will contain graph's connection information

Cfastq folder : The artificial fastq file for alignment

Sorted : Sorted result containing folder

Trimmed : Trimmed result containing folder
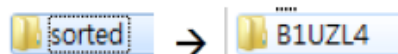
7. FAQ

Q . What if I don't want to sort?

    A. Please click additional options and <u>uncheck the sort box.</u>

    After uncheck, you have to do some manual task.

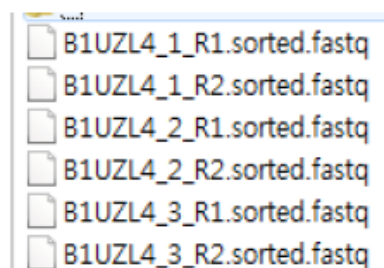**First create <u>sorted</u> folder under your desired output directory**

Then create folders with respect to each of your sample names

[EX]

📁 sorted → 📁 B1UZL4

Then make all your fastq file name as SAMPLENAME_SAMPLENUMBER_R1.sorted.fastq (for fwd files) and SAMPLENAME_SAMPLENUMBER_R2.sorted.fastq(for rev files)

[EX]

📄 B1UZL4_1_R1.sorted.fastq
📄 B1UZL4_1_R2.sorted.fastq
📄 B1UZL4_2_R1.sorted.fastq
📄 B1UZL4_2_R2.sorted.fastq
📄 B1UZL4_3_R1.sorted.fastq
📄 B1UZL4_3_R2.sorted.fastq

Q. What if I don't want to trim because I already trimmed?
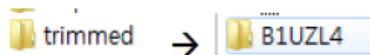
A. Please click additional options and <u>uncheck the trim box.</u>

After uncheck, you have to do some manual task.

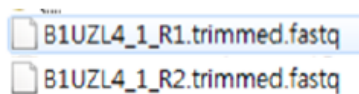First create <u>trimmed</u> folder under your desired output directory

Then create folders with respect to each of your sample names

[EX]

Then make all your fastq file name as SAMPLENAME_SAMPLENUMBER_R1.trimmed.fastq (for fwd files) and SAMPLENAME_SAMPLENUMBER_R2.trimmed.fastq(for rev files)

[EX]

Q. What if I don't know assembly start and end sequence?

A. Very simple. Just enter junk sequence to assembly start and end sequence. Then open 'additional option' and check Depth checkbox of Seed Mistmatch Correction field. Program will heuristically select contig

Q. Is it possible to assemble whole plasmid?

A. In theory this program can. If you specify start and end sequence for assembly correctly,program will do the task (see example below).
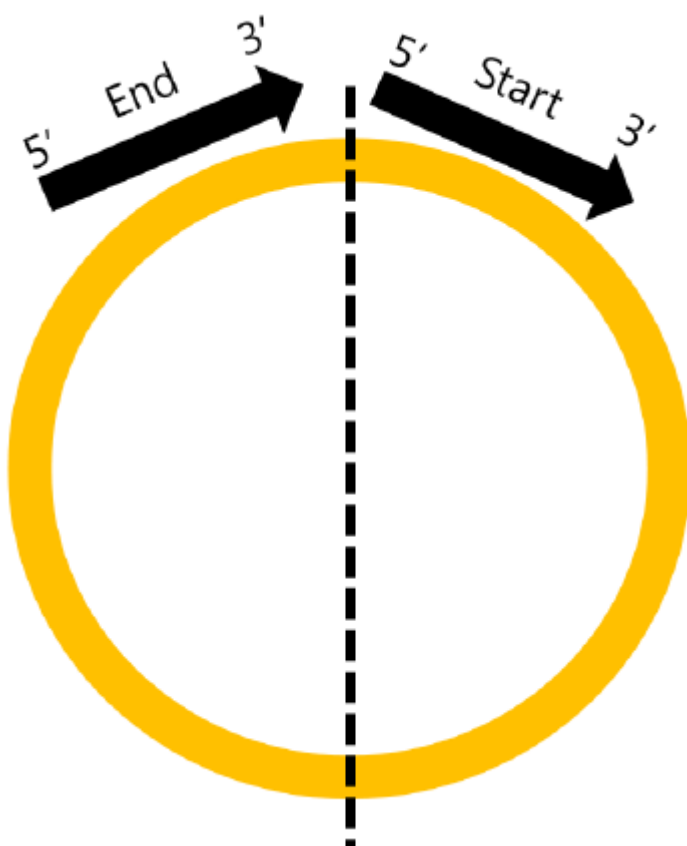
Q. Is it possible to assemble PCR amplicon?

A. Yes. Specify the start and end sequence for your PCR amplicon's $\pm$50bp or primer locations or other sequence that can figure it out.