

TnClone User Guide

System requirements

- Ubuntu 14.04 LTS (other systems are not currently tested)
- Python 2.7
- PyQt4 support
- Python HTSeq library
- Java 7 or higher
- At least 8Gb RAM

NOTE

If one will use this program from remote server, please install SSH capable software to use.

1 Check if your system meets requirements

A. Check your Ubuntu system version

Follow instruction below

<https://help.ubuntu.com/community/CheckingYourUbuntuVersion>

B. Check your system's python version

Open terminal by pressing Ctrl + Alt + T

Then type python and press enter button

Then you jump into python's idle as shown below

```
~$ python
Python 2.7.6 (default, Jun 22 2015, 17:58:13)
[GCC 4.8.2] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> 
```

As shown in yellow line you can see the version of python.

If something starts with 2.7 then you passed first step.

(Don't worry. If you are Ubuntu user then python is already installed)

If you want to exit from this screen then just type exit() and type enter key as shown below

```
~$ python
Python 2.7.6 (default, Jun 22 2015, 17:58:13)
[GCC 4.8.2] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> exit()
~$ 
```

C. Check if PyQt4 is installed

On your terminal screen, open python idle as described in section A above and type import PyQt4 as below

```
~$ python
Python 2.7.6 (default, Jun 22 2015, 17:58:13)
[GCC 4.8.2] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> import PyQt4
>>> 
```

If your system raises some error that sentence begin with "ImportError" , then you don't have library for it

Follow instruction guide at link below which is official PyQt homepage

<http://pyqt.sourceforge.net/Docs/PyQt4/installation.html>

D. Check if HTSeq is installed

To install HTSeq please go link below and download tar.gz file

<https://pypi.python.org/pypi/HTSeq> (Current version : 0.6.1)

Assume that your download location is ~/Downloads

Please follow instruction below

(Installation guide was adapted from <http://www-huber.embl.de/HTSeq/doc/install.html#install>)

Type the command below

(tar -zxvf HTSeq-0.6.1.tar.gz)

```
~/Downloads$ tar -zxvf HTSeq-0.6.1.tar.gz
```

After decompression, you will get screen like below

```
HTSeq-0.6.1/HTSeq/scripts/qa.py
HTSeq-0.6.1/HTSeq/scripts/__init__.py
HTSeq-0.6.1/HTSeq/StepVector.py
HTSeq-0.6.1/HTSeq/__init__.py
~/Downloads$
```

Then type the command below

(cd HTSeq-0.6.1)

```
~/Downloads$ cd HTSeq-0.6.1/
```

Now it is time to build your HTSeq package

```
~/Downloads$ cd HTSeq-0.6.1/
~/Downloads/HTSeq-0.6.1$ python setup.py build
```

You will get some compiler talking screen and build will be finished soon (< 1 min)

Then we have to install this package for all users

Type the last line of example below

```
copying and adjusting scripts/htseq-qa -> build/scripts-2.7
copying and adjusting scripts/htseq-count -> build/scripts-2.7
changing mode of build/scripts-2.7/htseq-qa from 664 to 775
changing mode of build/scripts-2.7/htseq-count from 664 to 775
~/Downloads/HTSeq-0.6.1$ sudo python setup.py install
```

After installation you might have similar screen as described below

```
Installed /usr/local/lib/python2.7/dist-packages/HTSeq-0.6.1-py2.7-linux-x86_64.egg
Processing dependencies for HTSeq==0.6.1
Finished processing dependencies for HTSeq==0.6.1
~/Downloads/HTSeq-0.6.1$
```

Now check for HTSeq is properly installed

```
~/Downloads/HTSeq-0.6.1$ python
Python 2.7.6 (default, Jun 22 2015, 17:58:13)
[GCC 4.8.2] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> import HTSeq
>>>
```

If no error occurred then you got HTSeq library

If you got error like 'ImportError: Cannot import 'HTSeq' when working directory is HTSeq's own build directory' then open a new terminal screen and try import HTSeq as described above.

E. Check your system's Java version

From your terminal screen, just simply type `java -version` as shown below

```
~/Downloads/HTSeq-0.6.1$ java -version
java version "1.8.0_101"
Java(TM) SE Runtime Environment (build 1.8.0_101-b13)
Java HotSpot(TM) 64-Bit Server VM (build 25.101-b13, mixed mode)
~/Downloads/HTSeq-0.6.1$
```

If you see java version "1.7.x_xxx" or like in this manual, "1.8.x_xxx", then you don't have to worry about your java system

(If error occurs then type 'sudo apt-get install openjdk-7-jdk')

2 Getting started with program

A. Download program from github(currently not avail, It will be updated)

On your terminal system please type command below

We will assume your download directory(path) is **TNCLONE**

Follow instructions below (all lines below are linux command. Just follow instruction)

mkdir TNCLONE

cd TNCLONE

git clone <https://github.com/tahuh/TnSuite.git>

```
:~$ git clone https://github.com/tahuh/TnSite.git
```

B. Launching program

Now you are ready to launch the program. Please follow instruction below

cd TNCLONE

python tnclone.py

```
TNCLONE $ python tnclone.py
```

This will give you screen below

TnClone

NGS directory

Browse...

Sort info

Browse...

Format

Sample info

Browse...

Format

Start sequence

End sequence

Ref. directory

Browse...

Format

Output directory

Browse...

Additional options

Configure

Abort

Quit

Run

3 UI description

The screenshot shows the TnClone application window. It features a series of input fields on the left, each with a corresponding 'Browse...' button on the right. The fields are: NGS directory (1), Sort info (2), Sample info (3), Start sequence (4), End sequence (5), Ref. directory (6), and Output directory (7). Below these is an 'Additional options' field (8) and a 'Configure' button (9). A large dark oval with the number 13 is centered in the main area. At the bottom right are three buttons: 'Abort' (10), 'Quit' (11), and 'Run' (12).

(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

Options

Additional de novo assembly options|

1 k-mer size 63 Description

2 Minimum K-mer occurrence 3 Description

3 Analysis Steps ☒ Sort ☒ Trim ☒ Assembly ☒ Analysis Description

4 BED file Browse... Description

5 Down Sampling ☒ On/Off

6 Down sampling ratio 0.3

7 Seed Mismatch Correction ☐ Depth ☐ Mismatch 8 Num mismatch 3

9 GPU available ☐ On/Off

Smith-Waterman alignment options

Sequence match score 1 10 Description

Sequence mismatch score 3 11 Description

Gap open penalty 5 12 Description

Gap extension penalty 3 13 Description

Set default Save changes

(1) K-mer length

(2) Cut off for k-mer abundance. If a k-mer's abundance is smaller than this value, then those k-mers will be discarded

(3) Analysis steps. If check box is off, then program will not run the step.

(4) BED file is a region file. See file format for detail

(5) Down sampling action (if unchecked, action will be turned off)

(6) If one wants to down sample, then give the ratio

(7) If user wants to introduce error in start sequence of assembly, choose method(depth for depth-based, mismatch for mismatch allowing)

(8) If one selects mismatch method then please specify the number of mismatches that user will allow for start sequence error

(9) Select if you have CUDA capable GPU (this will accelerate computation to some extent , see <https://developer.nvidia.com/cuda-gpus> for the list)

(10) Gap open penalty for alignment

(11) Gap extension penalty for alignment

(12) Restore to default values

(13) Set options and close

Browse Buttons : Browse where the specific files are

Description : The description for each field. As described here or detailed compared to here

(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

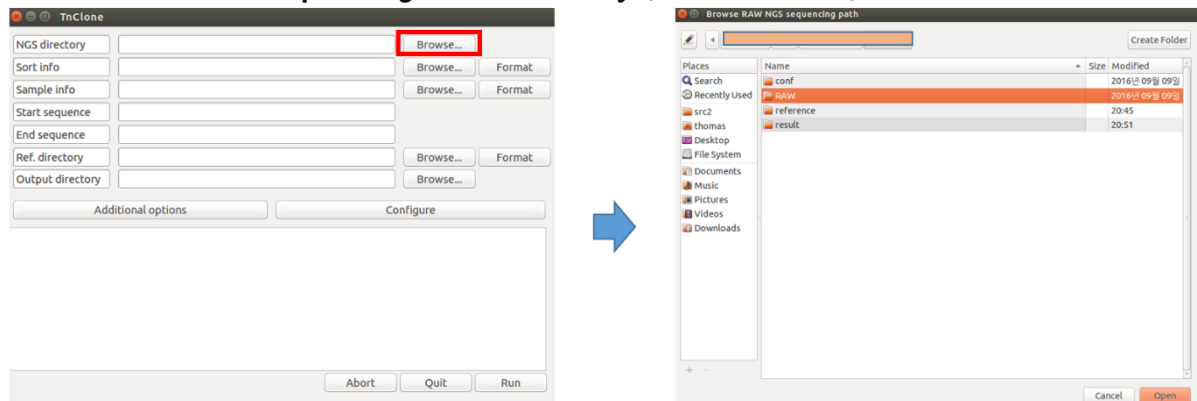
Will show description for formats

4 Instruction to the program

First initialize program as described in section 2-B

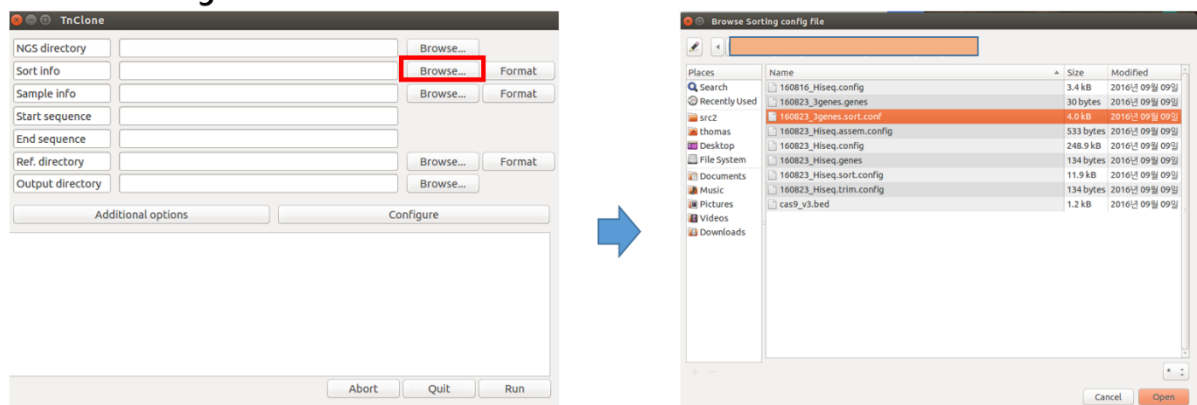
To run the program please follow instruction below

Browse RAW NGS sequencing result directory (NOT FILES !!!)



Click the browse button then you will see a pop up screen that will lead you to your NGS directory. Choose your directory (NOT FILES) and click open or OK button (depend on your system) on the right lower corner of popup screen.

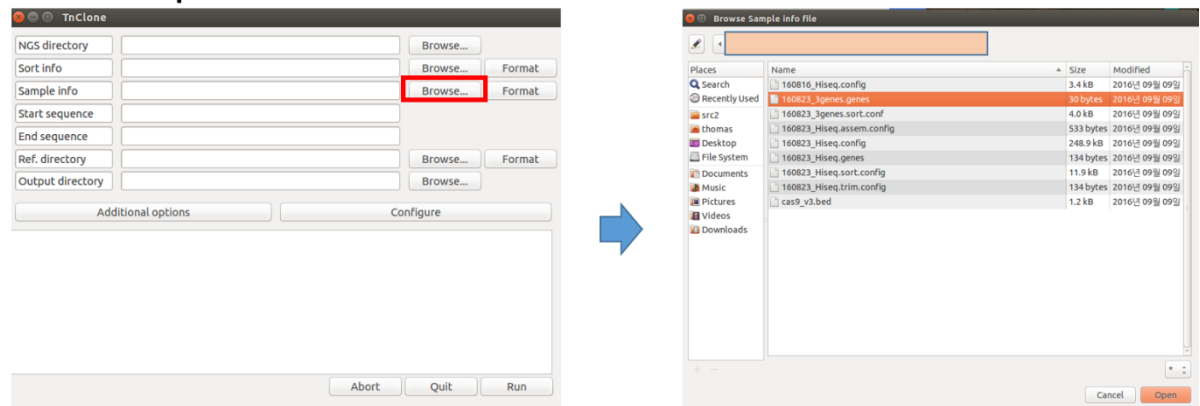
Browse sorting information file



By clicking browse button next to sort info field, you can load sort information file

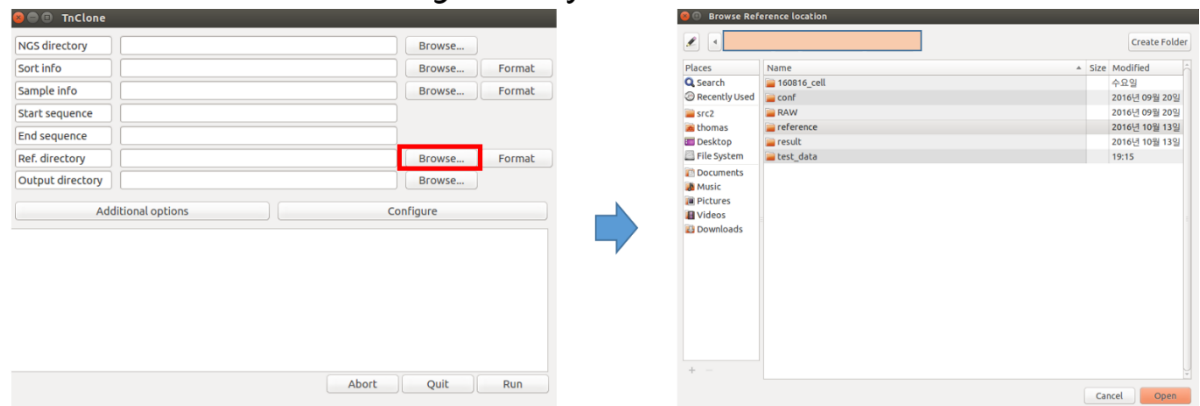
For detailed description for file, please see section '5 File formats'

Browse sample information file



Now you have to load sample information file(see section 5 for details for the file). This is important since except sorting step, all other steps for analysis rely on this file information. Please load this file. If not, program will crash.

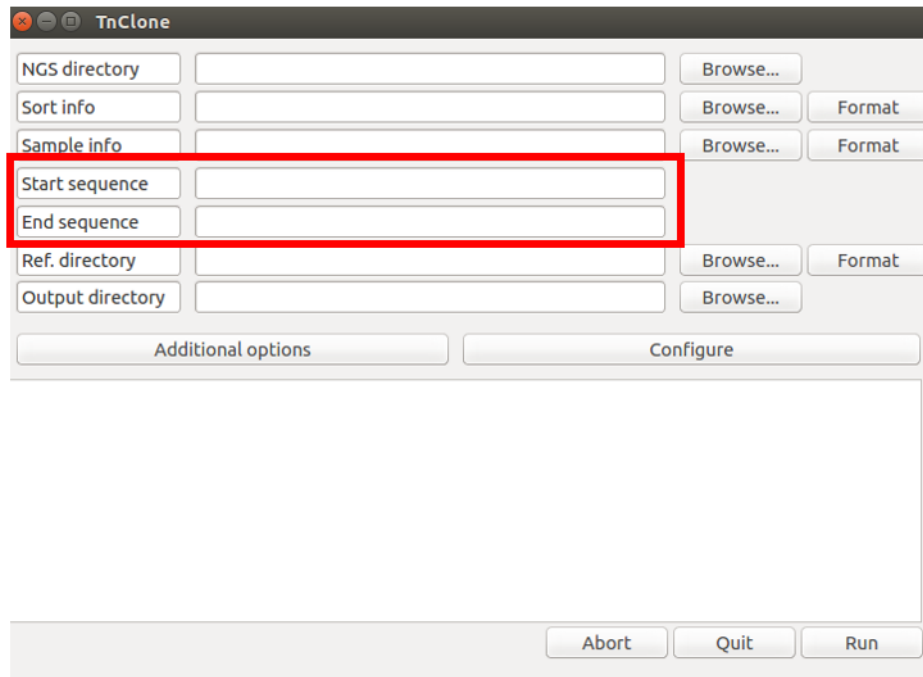
Browse reference file containing directory



Browse your reference directory. In reference directory, it is okay to have many reference files if those file formats meet the criteria as described in section 5.

If you don't have reference and only wants to assemble, then leave this field blank.

Optional : Start and End sequence



The screenshot shows the TnClone application window. It features a series of input fields for configuration: 'NGS directory', 'Sort info', 'Sample info', 'Start sequence', 'End sequence', 'Ref. directory', and 'Output directory'. Each field has a corresponding 'Browse...' button. The 'Start sequence' and 'End sequence' fields are highlighted with a red rectangle. To the right of the 'Sample info' field are 'Browse...' and 'Format' buttons. To the right of the 'Ref. directory' field are 'Browse...' and 'Format' buttons. To the right of the 'Output directory' field is a 'Browse...' button. Below these fields are two buttons: 'Additional options' and 'Configure'. At the bottom right are three buttons: 'Abort', 'Quit', and 'Run'.

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.

Other options

Options

Additional de novo assembly options

k-mer size: 63 [Description](#)

Minimum K-mer occurrence: 3 [Description](#)

Analysis Steps: ☒ Sort ☒ Trim ☒ Assembly ☒ Analysis [Description](#)

BED file: [Browse...](#) [Description](#)

Down Sampling: ☒ On/Off

Down sampling ratio: 0.3

Seed Mismatch Correction: ☐ Depth ☐ Mismatch Num mismatch: 3

GPU available: ☐ On/Off

Smith-Waterman alignment options

Sequence match score: 1 [Description](#)

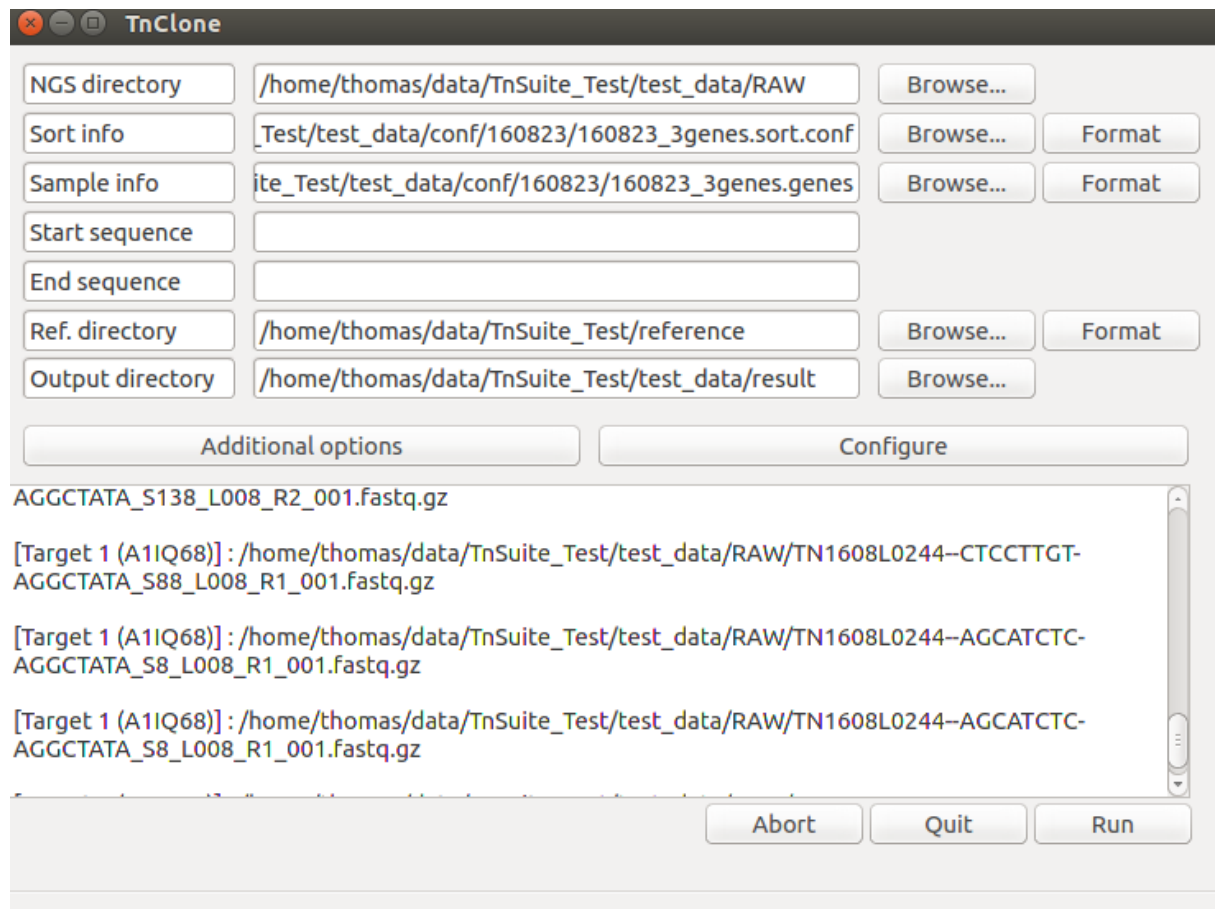
Sequence mismatch score: 3 [Description](#)

Gap open penalty: 5 [Description](#)

Gap extension penalty: 3 [Description](#)

[Set default](#) [Save changes](#)

You can choose detailed options for this program. Such as selecting steps, choosing k-mer length, down sampling, seed (start sequence) correction method, etc.. For the BED file(See section 5 for detail) one do not have to specify this field if he/she selected start and end sequence manually. If not then please specify this file and reference path. Program will automatically generate assembly configuration file based on these information. Click save changes after you changed options.



If you filled up all entry, then click 'configure' button then launch program by clicking 'Run' button

Then you will see logs on the log board as described above.

5 File 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

CAUTION

SAMPLE_NAME field must NOT contain underscore ('_') character (i.e. TP53_1 or TP53_colon, etc are not allowed. Use TP53-1 or TP53-colon instead)

ME SEQUENCE : For Illumina sequencing, if user use Tn5 ME sequence as barcode described in our paper, there are two different sequences we call forward barcode and reverse barcode. While tagmentation those barcodes will be inserted to your samples.

Use reverse barcode(see our paper).

FWD_FASTQ / REV_FASTQ : All files must be 'COMMA (,)' separated and list sequence must be the same(i.e. if FWD fast are listed 1_fwd.fq,2_fwd.fq, ... then reverse side must be 1_rev.fq,2_rev.fq,...)

Example below

A1IQ68	12	GTCCATCA	TN1608L0244--GCATAGTG-AGGCTATA_S131_L008_
C4ZA16	12	CCGTGTTA	TN1608L0244--GCATAGTG-AGGCTATA_S131_L008_
E1LBR5	7	TTGCCAAC	TN1608L0244--GCATAGTG-AGGCTATA_S131_L008_
E3ELL7	12	ACAACCGA	TN1608L0244--GCATAGTG-AGGCTATA_S131_L008_
B1UZL4	12	TGGTTGGC	TN1608L0244--GCATAGTG-AGGCTATA_S131_L008_
C0FXH5	12	ATGTGCTG	TN1608L0244--GCATAGTG-AGGCTATA_S131_L008_
D1VXP4	11	GTCCATCA	TN1608L0244--GCATAGTG-GCCTCTAT_S134_L008_
D3FJ35	6	CCGTGTTA	TN1608L0244--GCATAGTG-GCCTCTAT_S134_L008_
E1Z024	1	TTGCCAAC	TN1608L0244--GCATAGTG-GCCTCTAT_S134_L008_
G2Z1C1	4	ACAACCGA	TN1608L0244--GCATAGTG-GCCTCTAT_S134_L008_
Q7P7J1	2	TGGTTGGC	TN1608L0244--GCATAGTG-GCCTCTAT_S134_L008_
Q9CLT2	7	ATGTGCTG	TN1608L0244--GCATAGTG-GCCTCTAT_S134_L008_
Q73QW6	5	CTGCCGTA	TN1608L0244--GCATAGTG-GCCTCTAT_S134_L008_

(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

Example below

A1IQ68	12
C4ZA16	12
E1LBR5	7
E3ELL7	12
B1UZL4	12
C0FXH5	12
D1VXP4	11
D3FJ35	6
E1Z024	1
G2Z1C1	4
Q7P7J1	2
Q9CLT2	7
Q73QW6	5

(c) Reference file

Reference file requires strict file format.

First of all, it must be FASTA format(see details at https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYP E=BlastHelp)

Second, The file extension must end with 'gene.fa' (To avoid program for selecting different reference for analysis).

Third, file name (except extension) must be the same compared with SAMPLE_NAME field in sort information file.

Fourth the header must also same as file name.

Example below

File name :

It's header line :

(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.

(e) SAM format

This is usual alignment file format(see <https://samtools.github.io/hts-specs/SAMv1.pdf> for details) . If analysis finishes, then under your output directory, there will be a directory called 'sam'. Please look into files for detailed alignment result. DNA/protein alignment both are provided.

(f) VCF format

This is variant call format. If one specifies reference, then assembler will assemble contig and compare those contigs with reference and calls substitution/indels according to alignment options. Please see details for vcf at <https://samtools.github.io/hts-specs/VCFv4.2.pdf>

But our program will give you small information compared to the document above. We will only have chrom, pos, id, ref, alt (5 first columns).

Files will be located under 'vcf' folder of your output directory.

(g) 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

(h) 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

6 FAQ

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

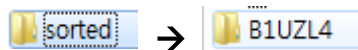
A. Please click additional options and uncheck the sort box.

After uncheck, you have to do some manual task.

First create sorted folder under your desired output directory

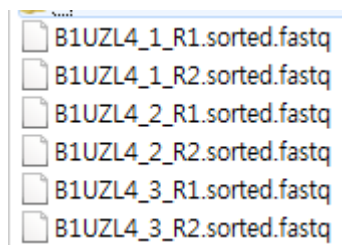
Then create folders with respect to each of your sample names

[EX]



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

[EX]



Q. What if I don't want to trim because I already trimmed?

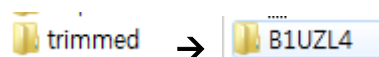
A. Please click additional options and uncheck the trim box.

After uncheck, you have to do some manual task.

First create trimmed folder under your desired output directory

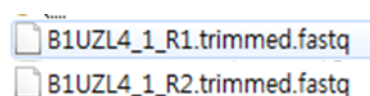
Then create folders with respect to each of your sample names

[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]



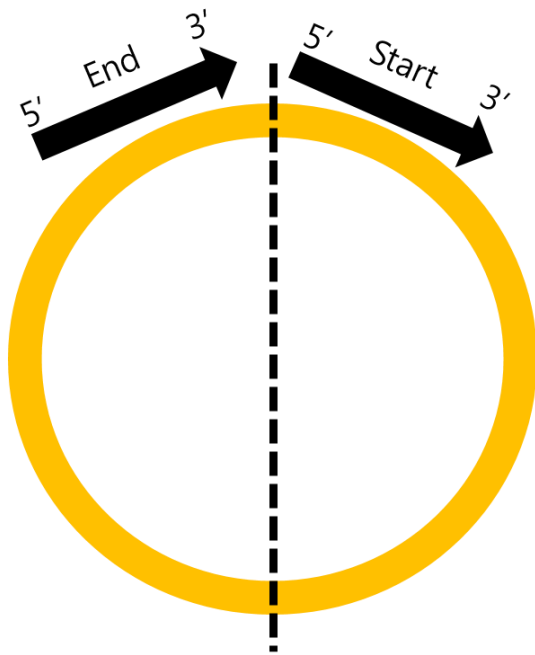
Q. What if I don't know assembly start and end sequence?

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

.

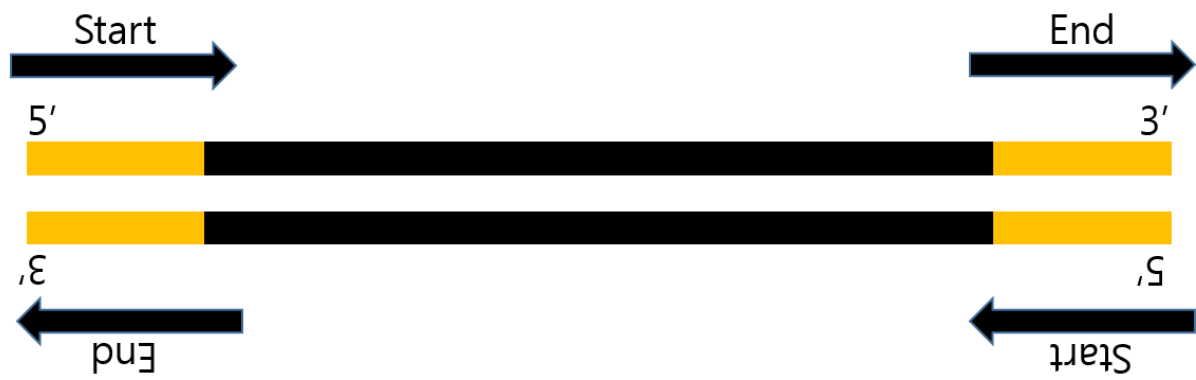
Q. Is it possible to assemble whole plasmid?

A. In theory this program can. If you specify start and end sequence for assembly correctly, program will do the task (see example below), but currently not tested



Q. Is it possible to assemble PCR amplicon?

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



To report BUG or other issues, please contact developer via e-mail

team.tnclone@ gmail.com