Manual of NCLscan pipeline

The package of NCLscan and the preparation of reference sequences can be downloaded from our FTP site: ftp://treeslab1.genomics.sinica.edu.tw/NCLsacn or GitHub: https://github.com/TreesLab/NCLscan

1. System Requirement

The NCLscan pipeline is executed on the 64-bit Linux operation system (e.g., Bio-Linux 6; see also http://nebc.nerc.ac.uk/ for further details). The three used aligners (i.e., BWA, Novoalign, and BLAT) were involved in the package of NCLscan, and can be downloaded from http://bio-bwa.sourceforge.net/, http://www.novocraft.com, and https://genome.ucsc.edu, respectively.

2. Preparation

2.1 Reference sequences

2.1.1 Reference sequences for NCLscan

The genomic sequences (FASTA files) were downloaded from the GENCODE website at http://www.gencodegenes.org/. Taking the human reference genome (GRCh37.p13) as an example, the following 4 reference files should be downloaded (http://www.gencodegenes.org/releases/19.html):

(1) Genome sequence FASTA file in GRCh37.p13 assembly (sequence region names are the same as in the GTFs): GRCh37.p13.genome.fa.gz

(2) Protein-coding transcript sequences: gencode.v19.pc_transcripts.fa.gz

(3) Long non-coding RNAs: gencode.v19.lncRNA_transcripts.fa.gz

(4) Gene annotation: gencode.v19.annotation.gtf.gz

2.1.2 Reference sequences for BWA

Before executing BWA alignment, the users need to create mapping index files for the reference genome. For the human reference genome (GRCh37.p13), the index files ("bwa_AllRef.fa*") has been created and accessible at our FTP site (ftp://treeslab1.genomics.sinica.edu.tw/NCLsacn). The users can also perform the following command to get the index files ("bwa_AllRef.fa*"):

```
>bwa index bwa_AllRef.fa
```

2.2 Configuration

The 4 reference files stated in 2.1.1 and 2 parameters are listed in order in “config.txt” as follows:

```
====================
GRCh37.p13.genome.fa
gencode.v19.pc_transcripts.fa
gencode.v19.lncRNA_transcripts.fa
gencode.v19.annotation.gtf
151
500
=====================
```

Note: The final two parameters, 151 and 500, indicate the maximal read length and fragment size of the used paired-ended RNA-seq data (FASTQ files).
If the read length (L) <= 151 bp and fragment size <= 500 bp, the users don't need to change these two parameters. If L > 151, the users should change these two parameters to (L, L*2 + insert size).

2.3 Setting

All the 4 reference files stated in 2.1.1, mapping index files stated in 2.1.2, (i.e., “bwa_AllRef.fa*”), the config.txt and the used paired-end RNA-seq data should be put in the same sub-directory where NCLscan executes.

3. NCLscan pipeline

The NCLscan pipeline includes six steps, which are all involved in the batch file: NCLscan.sh

Usage:

```
>./NCLscan.sh 01.fastq 02.fastq MyProject 20
```

Note: 01.fastq and 02.fastq are the two files of a set of paired-end RNA-seq data. NCLscan outputs two main results: "MyProject.result.gene_name" and "MyProject.result.sam". MyProject is the prefix of all output filenames. The third parameter (e.g., 20) means the cutoff of sequence quality score of the considered RNA-seq reads. The users can changes this cutoff quality score depends on their cases. If 0 is given, it means that all reads are considered in NCLscan.

The six steps and their usages are explained as follows:

Step 1: NCL_Scan0

Usage:

```
>./NCL_Scan0.sh 01.fastq 02.fastq MyProject
```

Note: NCL_Scan0 aligns the reads against the reference genome and the annotated transcripts using BWA.

Step 2: NCL_Scan1:

Usage:

```
> NCL_Scan1 MyProject.bwa.unmapped_1.fastq MyProject.bwa.unmapped_2.fastq
MyProject
```

Note: NCL_Scan1 aligns the BWA-unmapped reads against the reference genome and the annotated transcripts by using Novoalign.

Step 3: NCL_Scan2

Usage:

```
> NCL_Scan2 20
```

Note: NCL_Scan2 includes three steps: concatenating the two ends of each unmapped read; BLAT-aligning the concatenated sequences against the reference genome; and removing the concatenated sequences with an alternative co-linear explanation. A cutoff of sequence quality score of the considered reads is given.

Step 4: NCL_Scan3

Usage:

```
> NCL_Scan3
```

Note: NCL_Scan3 makes putative NCL references with putative NCL junction sites for each retained concatenated-sequence.

Step 5: NCL_Scan4

Usage:

```
> NCL_Scan4
```

Note: NCL_Scan4 includes the following steps:

- aligning unmapped reads against all putative NCL references by Novoalign
- retaining the putative NCL references that satisfy all of the three rules: (1) all

mapped reads must not have an alternative co-linear explanation; (2) at least one read confirms the junction site; and (3) the collection of the supported reads must span the NCL junction boundary by ≥ 50 bp on both sides of the junction site

- aligning the retrieved putative NCL references against the reference genome with a different set of BLAT parameters
- removing candidates with an alternative co-linear explanation within a single gene or between-in two close genes

Step 6: Add_gene_name

Usage:

```
>./get_GeneName_1.1.py
```

Note: This program appends gene name in the final result (i.e., "MyProject.result.gene_name") according to the gene annotation.

4. NCLscan Outputs

Two output files are generated after executing the pipeline: "MyProject.result.gene_name" (tab-delimited text file) and “MyProject.result.sam” (SAM-format file). The former is the list of the identified NCL events, and the latter records the supporting reads of the identified NCL events. For the SAM format, users can see http://samtools.github.io/hts-specs/SAMv1.pdf for further details.

The column format of "MyProject.result.gene_name" is described as follows:

(1) System generated ID for indicating the supporting reads in “MyProject.result.sam”
(2) Chromosome name of the donor side (5'ss)
(3) Junction coordinate of the donor side
(4) Strand of the donor side
(5) Chromosome name of the acceptor side (3'ss)
(6) Junction coordinate of the acceptor side
(7) Strand of the acceptor side
(8) (Reserved column, don't care)
(9) Distance to the far left-side end of alignments from the junction site
(10) Distance to the far right-side end of alignments from the junction site
(11) Total number of the supporting reads
(12) (Reserved column, don't care)
(13) Gene name of the donor side
(14) Gene name of the acceptor side