NextClip

Richard Leggett
richard.leggett@tgac.ac.uk
https://github.com/richardmleggett/nextclip

August 27, 2013

1 Introduction

In this manual, we describe the installation and use of NextClip, a tool for quality control and read preparation of data sequenced from Nextera Long Mate Pair (LMP) libraries. We describe both the core NextClip tool and the NextClip pipeline which can be used for further analysis in situations where a reference or partial assembly is available. This manual is best read after understanding the concepts described in the NextClip paper (currently under review).

2 Build and installation

2.1 NextClip tool

To compile NextClip, change into the directory containing the `Makefile` and type:

```
make all
```

A `nextclip` file should appear in the bin directory. To check it is working, type:

```
bin/nextclip -h
```

You should see a help message describing the program options.

2.2 NextClip pipeline

The NextClip pipeline is a set of shell scripts and Perl scripts that run NextClip in association with the BWA aligner in order to produce a more comprehensive PDF report that includes insert size information. As well as having BWA installed, the pipeline requires a LaTeX installation which includes the `pdflatex` command. To install, carry out the following steps.

1. Ensure that BWA and LaTeX are installed in your environment.
2. Copy the scripts directory to a suitable location - this can be anywhere you like.
3. Open `nextclip_lmp_analysis.sh` and:
    - Change the line beginning `scriptdir=` to point to the location of the scripts directory.

- Change the line beginning `nextclip=` to point to the location of the NextClip tool.
- Change the line beginning `numthreads=` to set the number of threads to use for BWA. If not running in a cluster environment, or unsure, this can be set to 1.

4. Open `nextclip_plot_graphs.sh` and:
   - Change the line beginning `scriptdir=` to point to the location of the scripts directory.

3 Running NextClip

NextClip can be run as follows:

```
nextclip --input_one R1.fastq --input_two R2.fastq
         --output_prefix output/out --log log.txt --min_length 25
         --number_of_reads 10000000 --trim_ends 0
         --remove_duplicates
```

The options have the following meanings:

- the `input_one` and `input_two` options specify the filenames of the input R1 and R2 files.
- the `output_prefix` option specifies the output prefix to use for the output FASTQ files and txt files.
- the `log` option specifies the name of an optional output log which will contain details of all adaptor alignments.
- the `min_length` parameter specifies the minimum read length after adaptor clipping. By default, this is 25.
- the `number_of_reads` parameter specifies the approximate number of read pairs. This is used to work out the size of hash table needed for PCR duplicate analysis.
- the `trim_ends` parameter specifies how many bases to trim off the ends of reads which do not have an adaptor match.
- the `remove_duplicates` parameter instructs NextClip to deduplicate the output files.

NextClip provides some other options which you are less likely to use:

- the `adaptor_sequence` option allows you to specify the adaptor sequence to search for - by default this is the Nextera sequence `CTGTCTCTTATACACATCT`.
- the `category_e` option instructs NextClip to produce category E reads, as well as A, B, C and D. These are reads where the normal adaptor alignment criteria only produce a match to one of the pair, but by relaxing the match criteria, we can get a second match.

- the `strict_match` option allows you to specify the adaptor matching criteria, in the form `X,Y` - where `X` is the number of bases to match for two adaptors (one forward, one reverse) and `Y` is the number of bases to match for a single adaptor. The default is `34,18`.
- the `relaxed_match` option allows you to specify the relaxed adaptor matching criteria. This is used if a strict match is found on one read, but not on it's pair. The default is `32,17`.

4 Understanding NextClip output

A typical NextClip report will look something like this:

```
SUMMARY

     Strict match parameters: 34, 18
           Minimum read size: 25
                   Trim ends: 0

         Number of read pairs: 516923
   Number of duplicate pairs: 293 0.06 %
Number of pairs containing N: 1802 0.35 %
   R1 Num reads with adaptor: 251263 48.61 %
       R1 long adaptor reads: 182481 35.30 %
          R1 reads too short: 68782 13.31 %
     R1 Num reads no adaptor: 265660 51.39 %
   R2 Num reads with adaptor: 231003 44.69 %
       R2 long adaptor reads: 164784 31.88 %
          R2 reads too short: 66219 12.81 %
     R2 Num reads no adaptor: 285920 55.31 %
   Total pairs in category A: 72146 13.96 %
         A pairs long enough: 53149 10.28 %
           A pairs too short: 18997 3.68 %
   Total pairs in category B: 158857 30.73 %
         B pairs long enough: 103204 19.97 %
           B pairs too short: 55653 10.77 %
   Total pairs in category C: 179117 34.65 %
         C pairs long enough: 119646 23.15 %
           C pairs too short: 59471 11.50 %
   Total pairs in category D: 106803 20.66 %
         D pairs long enough: 106803 20.66 %
           D pairs too short: 0 0.00 %
          Total usable pairs: 275999 53.39 %
             All long enough: 382802 74.05 %
    All categories too short: 134121 25.95 %
      Duplicates not written: 0 0.00 %
          Overall GC content: 66.04 %
```

The top three lines summarise input options. After that, there are three lines of information on numbers of pairs:

- A line giving the number of pairs of reads.
- A line giving the number of PCR duplicates as an absolute number and as a percentage of the read pairs.
- A line giving the number of pairs containing ambiguous bases as an absolute number and a percentage.

Next are four lines for each read pair:

- The number of reads containing the adaptor, as a number and as a percentage of total reads.
- The number of reads which, after clipping the adaptor, are greater than or equal to the minimum read length specified.
- The number of reads which, after clipping the adaptor, are less than the minimum read length specified.
- The number of reads not containing the adaptor.

Then, there are some lines of output for each of the categories - A, B, C, D and optional category E:

- Number of pairs in category.
- Number of pairs where both reads are greater than or equal to the minimum read length.
- Number of pairs where one or both reads are less than the minimum read length.

Finally, some more overall information:

- The total number of usable pairs - this is the number of category A, B, C and E pairs where both reads have a length greater than or equal to the minimum read length.
- The number of reads not written because they are PCR duplicates.
- The overall GC content.

5 NextClip output files

A number of files are output which are suitable for plotting as graphs. These will begin with the specified output prefix:

- `outputprefix_R1_gc.txt` and `outputprefix_R2_gc.txt` give GC content data for read 1 and read 2. This is a two column file, the first column being percentage GC and the second number of reads with that GC content.
- `outputprefix_duplicates.txt` gives PCR duplication information. The first column is n, the number of times a read is seen, the second column is the number of reads which are in duplicates of n and the final column gives this as a percentage. As an example, if the count of $n = 3$ is 27, this means there are 27 reads that are of a read that appears 3 times, or in other words there are 9 reads which are duplicated 3 times each.

- `outputprefix_A_pair_hist.txt` and equivalents for each category give clipped read length for pairs - ie. this reflects the shortest length read in a pair. The first column gives the read length, the second the number of pairs with this length (or for which the shortest of the pair has this length) and the last column gives a cumulative total - ie. number of pairs with at least this length.
- `outputprefix_A_R1_hist.txt` and equivalents for each category give individual read lengths after clipping. The first column is the read length, the second the number of reads with this length.

6 Running the NextClip pipeline

6.1 Variables

Before invoking the pipeline, it is necessary to set some temporary environment variables. This would typically be done in a simple script as follows:

```
export ref_min_size=0
export lib=LIB1468
export read_one=LIB1468_ATCACG_L001_R1_001.fastq
export read_two=LIB1468_ATCACG_L001_R2_001.fastq
export reference=Reference/AL645882.fasta
export organism="Streptomyces coelicolor"
nextclip_lmp_analysis.sh
```

The variables have the following meaning:

- `ref_min_size` defines the minimum reference contig size for a matching alignment to be included in the insert size histogram. This is useful if the reference is incomplete and we wish to avoid skewing results through the inclusion of contigs that are not substantially larger than the expected insert size.
- `lib` gives the library name, which will also be the directory name where intermediate files and the final report will be created. If the directory does not exist, it will be created, along with subdirectories called `bwa`, `reads`, `logs`, `graphs`, `analysis` and `latex`.
- `read_one` and `read_two` provide the filenames of the input R1 and R2 FASTQ files.
- `reference` provides the name of an indexed reference. See below for indexing.
- `organism` provides the name of the organism involved, for inclusion in the report.

The final command invokes the pipeline - ensure you specify the full path, or that your `PATH` variable points to the directory containing `nextclip_lmp_analysis.sh`.

6.2 Indexing the reference

Each time a new reference is used, it must be indexed with BWA and with NextClip. To index with BWA, type:

```
bwa index -a bwtsw reference.fasta
```

To index with NextClip, type:

```
nextclip_index_reference.pl reference.fasta
```

7 Understanding NextClip pipeline reports

The output of the NextClip pipeline is a thee page PDF report. A description of each section of the report follows.

7.1 Overall

This section reports number of read pairs with and without the adaptor sequence. The number of pairs in each category are reported, as well as those pairs that are too short (as defined by the user specified minimum length) and those that are long enough. Figures are expressed as absolute numbers of reads and as percentages of total number of pairs.

7.2 Category reports

There then follows a mapping report for each category of read, with a table for pairs producing good BWA mappings and those producing bad mappings. By default, a good mapping is defined as one with a score of 10 or more, but this can be changed by adjusting `minmapq` towards the top of `nextclip_lmp_analysis.sh`. Percentages expressed here are out of the number of reads in the given category.

The tables give figures for pairs that are 'in range' and 'out of range'. Mate pair oriented reads are considered to be in range if they are have an insert less than 25 kb, paired end oriented reads if they have an insert size less than 1 kb and tandems if they are less than 10 kb. These figures are also given in the Notes section at the end of the document.

Insert size distributions are plotted for those reads that are in mate pair orientation, paired end orientation and tandem orientation (both reads pointing in the same direction). Reads are included in these distributions only if they are good mapping and they map to contigs at least as long as the minimum contig size specified.

7.3 GC content

This section gives overall GC content of all reads, as well as graphs of GC content for reads 1 and 2. Reads with ambiguous bases are ignored.

7.4 Shortest pair length

This section gives cumulative plots of shortest pair length. Given two clipped reads, R1 and R2, the shortest pair length is whichever of the two reads is smallest. Therefore, these plots show for each length represented on the x axis, how many pairs have both reads of at least this length.

7.5 Clipped read lengths

This graph shows the length of clipped reads for categories A, B and C. In the case of B and C, only one read is shown, as in each case its pair is unclipped.

7.6 Duplication

This section reports the number of duplicate reads, as well as plotting a graph showing the percentage of reads at each duplication level. For illustration, a reading of 40 % at duplication level 2 means 40 % of reads are of sequences that appear exactly twice.

8 Comments, bugs and problems

Please report to `richard.leggett@tgac.ac.uk`.

9 Acknowledgements

The hash table code in NextClip is derived from that in the assembler Cortex_con - thanks are due to Mario Caccamo and Zamin Iqbal for their initial work on this.