NextClip Manual

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

March 5, 2014

Contents

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 (Figure 1). The manual is best read after understanding the concepts described in the NextClip paper (available from `http://bioinformatics.oxfordjournals.org/content/early/2013/12/02/bioinformatics.btt702.abstract`) and Illumina's technical note 'Data Processing of Nextera Mate Pair Reads on Illumina Sequencing Platforms'.

In outline, NextClip takes a set of read files as an input and classifies each pair of reads according to the presence of the junction adaptor in one or both reads:

- Category A pairs contain the adaptor in both reads.
- Category B reads contain the adaptor in only read 2.
- Category C reads contain the adaptor in only read 1.
- Category D reads do not contain the adaptor in either read.

NextClip will separate the input FASTQ files into separate files representing each category, with reads trimmed up to the adaptor starting point. Reads will only be written if the length of the trimmed read exceeds a user-configurable minimum read length (default 25bp). NextClip will also look for the presence of the external adaptor, clipping as necessary, but this typically occurs after the junction adaptor. Most users will want to use only category A, B and C reads in resulting downstream analysis - for example, assembly - and to treat category D as unreliable, as there is no way of telling if they are true mate pair reads or not.

In the sections that follow, we describe the installation of the tool and of the pipeline. We then describe use of the tool, followed by use of the pipeline. If you have a reference available, you may wish to skip the description of the tool and go straight to the use of the pipeline.

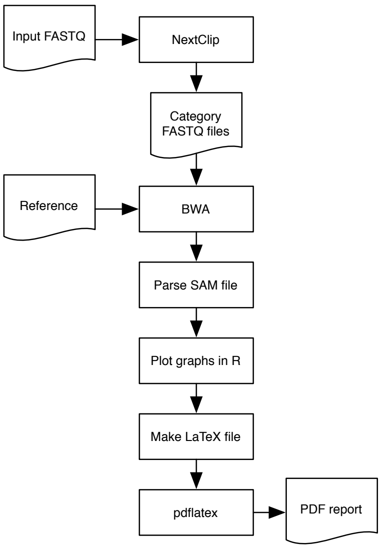

Figure 1: The NextClip pipeline is an enhancement to the basic tool: reads are processed with the NextClip tool, then aligned against a reference or assembly using BWA. Scripts parse the results, plot graphs with R and generate a PDF report via LaTeX .

2 Build and installation

2.1 NextClip tool

NextClip has been designed for compilation with GCC and should work on any platform that GCC is available for. To compile, change into the directory containing the `Makefile` and type:

```
make all
```

The situation with Mac OS X is slightly complicated. In version 4.x of Xcode, GCC is supplied as one of the command line tools. From version 5 onwards, Apple have removed GCC in favour of their own LLVM compiler. The easy way to find out what you have is to try to compile NextClip:

```
make all MAC=1
```

If this works without error, then you have GCC. If not, you will need to download GCC. Binaries of GCC for Mac OS can be downloaded from various sites - for example `http://hpc.sourceforge.net` - and need simply be copied to a suitable location. Once you have installed GCC, you need to change the line beginning `CC=` in the Makefile to point to GCC, for example:

```
ifdef MAC
CC=~/gcc/bin/gcc
endif
```

In the above instance, the gcc binaries have been installed into `~\gcc`.
Once you have compiled NextClip, you can check all is working by typing:

```
bin/nextclip -h
```

You should see a help message describing the program options.

2.2 NextClip pipeline

The NextClip pipeline is a set of Perl and shell scripts that run NextClip in association with the BWA aligner in order to produce a more comprehensive PDF report that includes insert size information. The pipeline has a number of dependencies which must be installed before running:

- BWA - version 0.6.1 or greater.
- R - version 2.12.2 or greater.
- LaTeX - should be TexLive 2012 or later.

In order to install, carry out the following:

1. Ensure that BWA, R and LaTeX are installed in your environment.
2. Copy the scripts directory to a suitable location - this can be anywhere that suits your software installation practices.
3. Open `nextclip_lmp_analysis.pl` and:
    - Change the line beginning `script_dir=` to point to the location of the scripts directory.
    - Change the line beginning `nextclip_tool=` to point to the location of the NextClip tool.

2.3 Job schedulers

The NextClip pipeline currently supports LSF, PBS or no job scheduler. In order to add support for an alternative scheduler, users can alter the function `submit_job` in `nextclip_lmp_analysis.pl`. This function takes the following parameters:

- a command to run.
- an ID for the job.
- the ID (or wild carded ID) of the dependency.
- the number of threads required.
- the amount of memory, in Mb, required.

These tend to be standard attributes across most job schedulers.

3 Running the NextClip tool

The `examples` directory contains some example data and a `ReadMe.md` file that describes how to run the examples. 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 junction 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 junction 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 junction adaptor matching criteria, in the form `X,Y` - where `X` is the number of bases to match for two junction adaptors (one forward, one reverse) and `Y` is the number of bases to match for a single junction 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`.

3.1 Guidance on match parameters

Our experience is that the default `strict_match` and `relaxed_match` parameters should be suitable for most situations and we do not recommend changing them unless your results lead you to believe that they may have a significant effect.

3.2 Error messages

- `too much rehashing` - this means that the hash table used to store PCR duplicate signatures is full. To remedy this, specify a more realistic value to the `number_of_reads` parameter.

4 Understanding NextClip output

4.1 A typical report

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: 251244 48.60 %
   R1 Num with external also: 9704 1.88 %
       R1 long adaptor reads: 182299 35.27 %
          R1 reads too short: 68945 13.34 %
     R1 Num reads no adaptor: 265679 51.40 %
  R1 no adaptor but external: 9702 1.88 %

   R2 Num reads with adaptor: 230971 44.68 %
   R2 Num with external also: 8973 1.74 %
       R2 long adaptor reads: 164662 31.85 %
          R2 reads too short: 66309 12.83 %
     R2 Num reads no adaptor: 285952 55.32 %
  R2 no adaptor but external: 9760 1.89 %

   Total pairs in category A: 72133 13.95 %
         A pairs long enough: 53136 10.28 %
           A pairs too short: 18997 3.68 %
A external clip in 1 or both: 11 0.00 %

   Total pairs in category B: 158838 30.73 %
         B pairs long enough: 103087 19.94 %
           B pairs too short: 55751 10.79 %
B external clip in 1 or both: 132 0.03 %

   Total pairs in category C: 179111 34.65 %
         C pairs long enough: 119480 23.11 %
           C pairs too short: 59631 11.54 %
C external clip in 1 or both: 198 0.04 %

   Total pairs in category D: 106841 20.67 %
         D pairs long enough: 97336 18.83 %
           D pairs too short: 9505 1.84 %
D external clip in 1 or both: 10187 1.97 %
```

```
        Total usable pairs: 275703 53.34 %
           All long enough: 373039 72.17 %
  All categories too short: 143884 27.83 %
    Duplicates not written: 0 0.00 %
        Overall GC content: 66.04 %

Done. Completed in 96 seconds.
```

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 six lines for each read pair:

- The number of reads containing the junction adaptor, as a number and as a percentage of total reads.
- The number of reads containing the junction adaptor and also the external adaptor.
- The number of reads which, after clipping the junction adaptor, are greater than or equal to the minimum read length specified.
- The number of reads which, after clipping the junction adaptor, are less than the minimum read length specified.
- The number of reads not containing the junction adaptor.
- The number of reads not containing the junction adaptor, but including the external 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.
- Number of pairs where one or both reads were clipped for the external adaptor.

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.
- Number of pairs in all categories which are long enough.
- Number of pairs in all categories which are too short.
- The number of reads not written because they are PCR duplicates.
- The overall GC content.

4.2 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.

5 Running the NextClip pipeline

5.1 Parameters

The pipeline is invoked by passing it a configuration file. This file is of the form `parameter:value` (no spaces) and will include the following:

```
library_name:LIB1468
organism:Streptomyces coelicolor
output_dir:/data/workarea/leggettr/matepairs/LIB1468b
read_one:/data/workarea/LIB1468/reads/LIB1468_ATCACG_L001_R1_001.fastq
read_two:/data/workarea/LIB1468/reads/LIB1468_ATCACG_L001_R2_001.fastq
reference:/data/workarea/leggettr/matepairs/Reference/AL645882.fasta
minimum_contig_alignment_size:0
number_of_pairs:125000
```

The parameters have the following meaning:

- `library_name` gives the library name.
- `organism` provides the name of the organism involved, for inclusion in the report.
- `output_dir` defines a working directory for the analysis. 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 full pathnames of the input R1 and R2 FASTQ files.
- `reference` provides the pathname of an indexed reference. See below for indexing.
- `minimum_contig_alignment_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. As a rule of thumb, try specifying a value that is double the expected insert size.
- the `number_of_pairs` is an optional parameter to specify the approximate number of pairs of reads in the input. Without it, the pipeline will perform a line count using the `wc` command.

Once the configuration file has been created, the pipeline can be invoked as follows:

```
nextclip_lmp_analysis.pl -config LIB1468.txt -scheduler LSF -bwathreads 8
```

The parameters have the following meanings:

- the `config` parameter specifies the filename of the configuration file.
- the `scheduler` parameter specifies the name of the job scheduler - currently `LSF`, `PBS`, or `NONE`.
- the `bwathreads` parameter specifies how many threads to use for BWA. Typically 1 when not using a scheduler.

5.2 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
```

5.3 Outputs

A PDF file is created inside the `latex` subdirectory of the output directory. Inside the analysis subdirectory, you will find a file called `libraryname_analysis.txt` which contains a plain-text, easily parsable version of the information contained in the PDF report. The file consists of multiple `parameter:value` lines, for example:

```
R1NumWithJunctionAdaptor:61020
R1PcWithJunctionAdaptor:48.82
R1NumWithJunctionAdaptorAndExternalAdaptor:2395
R1PcWithJunctionAdaptorAndExternalAdaptor:1.92
...
```

5.4 Logs

Each of the stages of the pipeline will write log files to the logs subdirectory. There is also a master log file called `nextclip_analysis.log` which contains selected output, warnings and errors from all scripts. This is the easiest place to go to diagnose any failures. Note, if running with a job scheduler, the order of items in this file can become mixed up. However, warnings and errors should still be clear to see - these are usually lines beginning with the words ”Error” or ”Warning”.

5.5 Diagnosing problems

If the pipeline fails to produce a result, check the following:

1. Check you have an up-to-date version of BWA installed (0.6.1 or better).
2. Check you have an up-to-date version of R installed (2.12.2 or better).
3. Check you have an up-to-date version of LaTeX installed (TexLive 2012).
4. Check that the `nextclip_tool` and `script_dir` variables are set correctly at the top of `nextclip_lmp_analysis.pl`.
5. Check you have indexed the reference - both with BWA and with NextClip (see above).
6. Check the input files - reads, reference - exist.
7. Examine the `nextclip_analysis.log` file inside the logs subdirectory. Look for the first occurrence of ”Error” or ”Warning” in this file - later steps may continue to run (and fail), but the first occurrence of ”Error” is probably the cause.

8. Examine the other logs inside the logs subdirectory.

9. If the .tex file is produced, but no .pdf file, check the .log file inside the latex directory.

If you are running with a job scheduler, try running without (`-scheduler none`) in order to more easily diagnose the problem.

6 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.

6.1 Overall

This section reports number of read pairs with and without the junction 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.

6.2 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.

6.3 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.

6.4 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.

6.5 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.

6.6 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.

6.7 Ambiguous bases

Currently, this section simply reports the number of read pairs which contain Ns.

6.8 External adaptor

Sometimes reads can extend beyond the junction adaptor sequence and extend into the external adaptor. NextClip will search for this too, trimming as necessary, and reports occurrences of it in this section.

7 Acknowledgements

The hash table code in NextClip is derived from that in the assembler Cortex_con, for which we acknowledge the work of Mario Caccamo and Zamin Iqbal.

8 Comments

Please pass comments or bug reports to `richard.leggett@tgac.ac.uk`.