ngskit4b Reads Filter Application Note

Release 0.4.6

CAUTION – this document has not been updated to include new functionality added since original cloning of the BioKanga 4.2.0 documentation

Overview

There are a number of processing contexts in which filtering of NGS readsets prior to downstream applied analytics is desirable. The 'ngskit4b filter' is targeted at filtering out reads with likely library and/or sequencing induced artefact bases from input into downstream processing such as alignment based SNP calling or de Novo assembly. The approach taken is essentially that any given read is assumed to be error free if that read can be overlapped by 2 other supporting reads with no substations required in the overlaps. One of the overlapping supporting reads must overlap onto the 5' end of the given read and the other supporting read must overlap onto the 3' end of the given read; the overlap lengths required are parameterised and default to be 70% of the mean read length. Other overlap requirements are that the overlapping reads must extend past the ends of the given read (contained or duplicate reads do not count as overlapping) and the user may request that overlaps be strand dependent (default is to process for both sense and antisense overlaps) plus repeat the overlapping for a number of iterations (default is for 2 iterations). End result is that the remaining reads are unlikely to contain many library preparation or sequencing artefacts, thus may be treated as though error free in downstream processing. Note that the number of reads remaining after filtering may be significantly reduced, necessitating relaxing of certain thresholds in the downstream analytics – i.e if SNP calling then the minimum coverage required to call a SNP may need to be reduced.

Filter Parameterisation

For the 'ngskit4b filter' processing module, the primary inputs are expected to be the NGS readsets which may specified as either single ended or paired ended. These readsets would normally be raw fastq, but may be supplied as pre-filtered or multifasta if the experimenter wishes to apply additional filtering.

- -m, mode=<int>
    - Specify the overall processing mode, by default this is 0 which is filtering mode with accepted reads output as multifasta. The experimenter can request that the filtering mode output the accepted reads as binary packed sequences which can be used as input into the 'biokanga assemb' module. There is little advantage to using this binary output mode other than spacing saving in disk resources. The 'mode' can be specified as 2, which can be used to convert a previously generated 'mode' 1 back into multifasta compatible with the same file as if generated in 'mode' 0.
    -
- -p, --minphred=<int>
    - Only accept reads for filtering if their mean (over complete read sequence bases) Phred scores are at least this threshold (default 20, 0 to disable, range 15..40). Additionally, if any single base in the read sequence has a Phred score of less than 10 then that read is not accepted even if the mean Phred score is at least that requested.

- -P, --iterativepasses=<int>
    - Iterative passes repeating overlap processing (default 2, range 1..5). Reads in earlier passes which are not supported by overlaps are discarded and not processed in subsequent passes.

- -S, --strand
    - Strand specific filtering - filter reads with overlaps in read orientation, sense only – the default is for non-strand specific with overlaps attempted in both sense and antisense orientation. This parameter also applies to the duplicate read detection processing.

- -n, --indeterminates=<int>
    - Filter out input sequences having higher than this number of indeterminate bases (default is 0, range 0..5). Reads which do contain indeterminate bases after this parameter threshold is applied will have the indeterminate bases substituted with a randomly selected canonical base prior to any duplicate or overlap processing.

- -x, --trim5=<int>
    - Trim this number of 5' bases from each input sequence (default is 0, range 0..20bp) when loading reads. After any trimming reads are checked to ensure conformance with the minimum length specified with ‘minlen’.

- -X, --trim3=<int>
    - Trim this number of 3' bases from each input sequence (default is 0, range 0..20bp) when loading reads. After any trimming reads are checked to ensure conformance with the minimum length specified with ‘minlen’.

- -l, --minlen=<int>
    - Filter out input sequences (after any trimming with ‘trim5’ and/or ‘trim3’) which are less than this length (default is 50bp, range 20..5000bp)

- -L, --trimseqlen=<int>
    - Sequences which are longer than this length will be 3’ trimmed down to this length. Defaults to 0 for no trimming and can be specified to be in the range of ‘MinSeqLen’ to a maximum of 1000bp.

- -y, --minoverlap=<int>
    - When overlapping other reads on to a given read then the overlap must be of at least this length. By default the required overlap is at least 70% of the mean read length. If specified then the required overlap must be in the range of 25 to 150bp, but if -1 is specified as the overlap then overlap processing is disabled.

- -D, --nodedupe
    - The default is to dedupe all exactly matching input sequences with only a single copy of the duplicated sequence retained for further processing. Use this parameter to disable duplicate processing. If paired ends are being processed then by default these

will be checked for corresponding mate end overlaps before calling as duplicate (this can be overridden with the ‘dedupepe’ parameter).

- -d, --dedupepe
    - By default, paired ends are processed as dependent ends when deduping. For a given paired end, if another paired end is an exact copy of the given end then it is only deemed to be a duplicate if the corresponding mate ends are also exact copies. Use the ‘dedupepe’ parameter if the paired end ends are to be treated as if single end reads with no mate end dependencies.

- -c, --contaminants=<file>
    - Putative contaminant sequences are contained in this multifasta file. The fasta descriptor lines contain a specification of how the contaminant sequence is to be applied. For instance, only process if sense to the 3’ end of a paired end or perhaps antisense to the 5’ end. Contaminate sequences must in the size range of 4 to 128bp. A following section in this application note describes the accepted descriptor line convention by which the user can specify the context in which putative overlaps are to be explored. If a given read is overlapped by a contaminate then that overlap will be trimmed from the read. After any trimming reads are checked to ensure conformance with the minimum length specified with ‘minlen’.

- -i, --inpe1=<file>
    - Load single ended readsets, or 5' end if paired end, from fasta or fastq file(s); if single ended then wildcards allowed. Gzip’d readsets can be directly specified, no need to unzip. There is a hardcoded limit of 500bp for read sequence lengths; reads longer than this length will cause process termination. Read sequences less than 16bp will not be accepted for further processing. If single ended reads then wildcards may be used. Max of 100 files may be specified

- -I, --inpe2=<file>
    - Load 3' ends if paired end reads from fasta or fastq file(s). Note that ordering is important, it is assumed that the 5’ ends specified to ‘inpe1’ are in the same corresponding order as the 3’ ends specified to ‘inpe2’. There is a hardcoded limit of 500bp for read lengths; reads longer than this length will cause process termination. Read sequences less than 16bp will not be accepted for processing.

- -o, --out=<file>
    - Accepted filtered reads are output to either one (single end reads input readsets) or two (paired end input readsets) multifasta file(s). These are multifasta fasta files, not fastq files and hence contain no quality scores. If the processing ‘mode’ was specified to be 1 then the output format is binary, and this format can be directly loaded by the ‘biokanga asemb’ de Novo assembly module. If processing ‘mode’ was specified to be 2 then a previously generated binary format ‘mode’ 1 file can be specified as input to ‘inpe1’ and this will now be converted/output as multifasta.

- -O, --dupdist=<file>

- Use to specify the name of a file to which the duplicate read distributions are to be written.

ngskit4b Filter Internal Processing

The following describes the internal processing flow assuming a single paired end NGS readset is being characterised, and there was also a contaminate containing sequences file specified by the experimenter. If multiple NGS paired end NGS readsets are being characterised then the processing flow is essentially the same as all input reads are pooled retaining the mate end relationship if paired ended.

All required output files are created (truncated to 0 length if already existing) with the naming incorporating the source readset file name plus a characteristic specific file name suffix.

Input readsets are checked to ensure they are accessible and an estimate of the number of contained reads and their mean length obtained allowing for preallocation of the majority of required memory structures, these will be reallocated later on demand if the initial estimate was incorrect.

Contaminate file is loaded and parsed using contaminate naming conventions as described in a following section.

The specified NGS readsets are then progressively loaded, and each parsed read pair is processed in the following processing rule order –

a) Process for putative contaminate sequence overlaps allowing at most 1 mismatch base per 25bp of overlap out past the initial 10bp of the overlap. Note that the contaminate sequences are ordered with higher ordered contaminates having processing priority over lower ordered contaminates. Ordering priority is determined by the order of contaminate sequences in the contaminate file; earlier occurring sequences have higher priority than later occurring sequences.
    a. If paired end processing then the pair of reads are independently processed for contaminate sequence overlaps.
    b. Contaminate sequences which qualify for processing (have matching 5’ or 3’, sense orientation, paired end, etc.) are progressively checked for overlapping onto the current read starting from a maximal overlap.
    c. If the current putative overlap O would be at least 10bp then allowed substitutions are calculated as:
        i. AllowedSubs = (‘—maxcontamsubrate’ * O) + 15) /25
    d. If the length of any overlap is less or equal to the ‘--trim5’ or ‘—trim3’ experimenter set threshold then that read is assumed to be contaminate free

b) Ends are trimmed according to the ‘—trim5’ and ‘—trim3’ parameters. If the remaining sequence post-trimming is less than ‘--minlen ‘ or the read contains contaminates then that read (read pair if pair ended) is discarded.
c) If Phred scores are associated with the (end trimmed) read and the experimenter has specified a minimum required Phred ‘--minphred ‘ then if any base has a Phred less than the minimum then that read (read pair if pair ended) will be discarded.
d) The number of indeterminate ‘N’ bases in the (end trimmed) read is counted and normalised to 100bp. If the rate per 100bp is higher than that specified with ‘--indeterminates’ then that

read (read pair if pair ended) is discarded. If the accepted read contains 'N's fewer than '--indeterminates' then these are substituted with random canonical bases.

e) A read (read pair) remaining after the previous filtering steps is then packed at 15bp per 32bit word (2bits per base plus a 2bit header) into memory resident structures.

When all input readsets have been filtered with accepted sequences packed into memory then a sparse suffix array index is constructed on the initial 15bp (32bit word) of each packed sequence. If the '—nodedupe' has not been requested by the experimenter then this sparse index is used to efficiently identify all exactly matching duplicate sequences (sense, and antisense according to experimenter set '—strand' parameter) and mark duplicates for removal with only one instance of any duplicated sequence retained. If paired end readsets, and '—dedupepe' not requested, then both ends a read pair must exactly match both ends of another read pair before classified as being a duplicate.

Remaining non-duplicate, or retained single instances of duplicates, reads are then processed for 5' and 3' end overlaps by two other reads. A sparse index is constructed at 15bp intervals (32bit word boundaries) over all read sequences. Read overlap processing is treating paired ends as if single ended reads and only retained the paired end linkage for when reporting and generating the output filtered multifasta files. The processing for overlaps is iterative, the number of iterations is specified with '--iterativepasses' which defaults to 2. Each iteration follows these processing phases -

a) A given read is checked for an exact match putative overlap of it's 3' onto all other reads at their 5' ends; sense overlap sense. The given read must have at least 1bp 5' which does not overlap. If there is an exact match, and the match overlap is at least the minimum K-mer as specified with '—minoverlap', then the given read is marked as having a 3' overlap and the overlapped read marked as having a 5' overlap. Note that in the forgoing the given read will be checked for overlaps against all other reads, and as a result there may be multiple overlapped reads marked as being 5' overlapped.

b) If the '—strand' has not been specified by the user then the above is repeated but this time the given read is reverse complemented so the matches are antisense onto sense. In this instance the marking is reversed for the given read; it will be marked as having a 3' overlap, with the overlapped read(s) still being marked as 5' overlapped.

c) In the final overlap phase, if '—strand' not specified, all reads are reverse complemented and reindexed. A given read in it's original sense orientation is processed for sense overlap onto antisense using the same required overlap criteria as in the earlier sense overlap sense phase. If there is an accepted overlap then the given read is marked as 3' overlap, and overlapped reads are also marked as 3' overlap.

d) When all overlap phases have completed then any reads which are not marked as having both the 5' and 3' markers set are removed. If either end of a paired end read do not have both 5' and 3' markers set then both ends are removed.

Reads remaining are now written out to one (single ended) or two (paired ended) multifasta files. The naming convention is to use the '--out' parameter name specified as the name prefix, which will be for single end readsets be suffixed with '.SE.fasta'; if paired end processing then the 5' end file name will be suffixed with '.R1.fasta' and the 3' end file name will be suffixed with '.R2.fasta'.

Contaminate Sequence File Format

The contaminate sequence file is multifasta and contains those sequences for which the experimenter is interested in counts of the number of times contaminates are observed as overlapping read

sequences in NGS readsets. Contaminants may be adaptor sequences or any other artefact sequences, the presence of which may impact on the degree of filtering required in subsequent processing of the effected NGS readsets. Contaminate sequences must be in the range of 4 to 200bp in length.

Ordering priority is determined by the order of contaminate sequences in the contaminate file; earlier occurring sequences have higher priority than later occurring sequences. Higher priority contaminate sequences are processed for putative overlaps on to read sequences before lower priority contaminate sequences with overlap counts accrued to the first contaminate with matching overlap.

The fasta descriptor names are expected to conform to the following convention and will be parsed as such, the convention allows the experimenter to specify the read sequence context for which putative overlaps by that contaminate sequence may be explored. An example fasta contaminate sequence may be:

>contamABC@12

ACGATAGGATTTTTT

The above descriptor line will be parsed and interpreted as meaning that the sequence identified as 'contamABC' is only to be processed for sense overlaps onto the 5' end of a single end read, or the 5' ends of either end of a paired end pair.

In addition to the standard canonical 'A','C','G','T' bases, contaminate sequences may contain wildcard bases designated as being indeterminate 'N's. These wildcard bases will match any base ('A','C','G','T','N') at the offset being processed for a match in the targeted sequence.

The contaminate sequence naming convention is that sequence names are suffixed with a '@' character followed by contaminant overlay codes. If not present, then the default is to process for 5' end sense overlaps only as in the forgoing example. If the sequence name is suffixed with a '&' instead of the usual '@' character then the check will be for reads which are completely contained within the sequence such as would be the case if BAC clones are being sequenced with the sequence being the BAC vector. These vector sequences can be up to 16Mbp in length and a maximum of 10 vector sequences may be specified.

Contaminate overlay codes are one or more combinations of the following numeric characters and must be prefixed by the '@' or '&' character:

Numeric Code	Accepting Putative Overlaps
1	Sense 5' end of SE, or 5' PE1 of PE
2	Sense 5' PE2 of PE
3	Sense 3' end of SE, or 3' PE1 of PE
4	Sense 3' PE2 of PE
5	Antisense 5' end of SE, or 5' PE1 of PE
6	Antisense 5' PE2 of PE
7	Antisense 3' end of SE, or 3' PE1 of PE
8	Antisense 3' PE2 of PE