ngskit4b Reads Quality Check 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

ngskit4b include a processing module 'ngskit4b ngsqc' by which NGS readsets can be efficiently quality checked for a number of quantitative characteristics generally accepted as being representative of datasets suitable for the targeted analytics. Most existing NGS quality checking and assurance processing toolkits (Fastqc etc.,) provide a summary of the Phred scores, length, duplicate instance counts, and nucleotide composition distributions. The 'ngskit4b ngsqc' also provides the forgoing summaries, but significantly scope extended; i.e. duplicate instance counts include paired end duplicates and by are by default generated over the first 5 million unique sequences, with nucleotide composition distributions by default generated up to pentamers. Additional quantitative characteristics are generated for Pearson K-mer concordance distributions and Phred base score derived error free reads probability distributions. Furthermore, an estimate of contaminate sequence containment can be obtained. All the quantitate summaries are generated both as CSV tables allowing for ease of downstream post-processing analytics, and also as graphs in SVG format for ease of visualisation.
Examination of the generated summary quantitative characteristics should help inform the experimenter as to the appropriate parameterisation to be used in the subsequent processing of the NGS datasets – e.g. is trimming required, number of substitutions to allow etc.

NGS Quality Check Parameterisation

For the 'ngskit4b ngsqc' 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, but may be supplied as pre-filtered if the experimenter is interested in quantifying the efficacy of the filtering process utilised.

- -m, --mode=<int>
    - Processing mode: 0 - independent processing of single/paired readsets, 1 - pooled processing of single/paired readsets. If processing mode 0 specified (the default) then each single or paired readset is processed independently of other readsets resulting in separate sets of characteristics summary and graphical SVG files for the individual readsets. In processing mode 1, the readsets are all pooled into one and a single set of characteristics summary and graphical SVG files will be generated.

- –S, --strand
    - This allows the experimenter to specify that the readsets are strand specific, thus there will be no antisense processing when quantifying the readset characterisations.

- -y, --trim5=<int>

    - The specified number (defaults to 0, range 0..50) of bases will be 5' end trimmed from each read when loading prior to any characterisations.

- -Y, --trim3=<int>
    - The specified number (defaults to 0, range 0..50) of bases will be 3' end trimmed from reads when loading prior to any characterisations.

- -k, --maxkmerlen=<int>
    - Many characterisations are K-mer related – i.e. compositional. Use this parameter to specify the maximum length K-mer of experimental interest. The default maximum K-mer is 6bp and may be specified to be in the range 3..12

- -K, --kmerccc=<int>
    - Use to specify the maximum Pearson concordance correlation coefficient measure KMer length, defaults to 6 and may be specified to be in the range of 1 .. 'maxkmerlen'.

- -s, --seeds=<int>
    - When characterising readsets for duplicate instance counts the first unique <seeds> reads will be used as seed reads against which other reads will be checked and counted towards read duplicate counts. If processing paired end readsets then both ends are used as a single seed and to count as a duplicate of that seed then other read pairs must exactly match on both ends. By default the first 5 million unique reads (or 5 million read pairs) are used as duplicate seeds, the experimenter may specify the number of seeds to be in the range of 100K to 25M seeds. Be aware that specifying 25M seeds will significantly increase memory and runtime requirements.

- -p, --minphred=<int>
    - The experimenter can request that only reads with a minimum mean Phred score (average of all Phred scores along length of read) contribute to duplicate instance and K-mer count dependent characterisations. By default the 'minphred' is 0 (disables Phred filtering), but may be specified in the range 10..40.

- -z, --maxcontamsubrate=<int>
    - The maximum allowed contaminant sequence (adaptors etc.) substitution rate (bases per 25bp of contaminant overlap, 1st 10bp of overlap no subs allowed) only applies if a contaminate sequences file is also specified with the '-c<file>' parameter. All contaminate sequences are checked for putative 5' overlaps onto each read. If the overlap is at most 10bp then the overlap must be an exact match, if would be longer then up to 'maxcontamsubrate' substitutions for each 25bp of overlap extension is allowed. By default 1 sub per 25bp extension is allowed, but may be specified to be in the range 0..3

- -Z, --mincontamlen=<int>
    - Putative contaminates must overlap by at least this many bases (default is 5, range 1..100) before they are accepted as being contaminant overlaps.
    -
- -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.

- -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. Additionally, alignments in SAM or BAM format containing the aligned sequences will be accepted as input – but because the contained sequences may be reverse complemented, or may contain both 5’ and 3’ ends, then the resultant distributions generated will be for merged combinations.
  - 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.

- -u, --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>
  - This specifies the output naming prefix to use when generating the output summary characteristics files. In general the appended file name suffix will contain the applicable readset file name appended with the characteristic name followed by ‘.csv’ if summary detail and ‘svg’ if graphical.
    - .contaminates.<svg|csv> Contaminates distributions
    - .duplicatesdist.<svg|csv> Duplicated read distributions
    - .kmerdist.<svg|csv> K-mer distributions
    - .pearsondist.<svg|csv> Pearson concordance correlations
    - .qscoredist.<svg|csv> Phred score distributions
    - .errfreedist.<svg|csv> Phred score derived error free reads
    - .readlendist.<svg|csv> Read length distributions

NGS Quality Check 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 with separate threads processing each readset independently in parallel.

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 then processed in the following order –

a) Ends are trimmed according to the '—trim5' and '—trim3' parameters. If the remaining sequence post-trimming is less than 16bp then that read (or read pair if pair ended) is not processed further.
b) If Phred scores are associated with the read then these scores will be accumulated at each base position allowing for summary Phred score distributions to be calculated when all reads have been processed.
c) The mean Phred score is then calculated over the length of the read, and if any base position has a Phred score of less than 10 or if the mean Phred score is less than that specified with the '—minphred' parameter then no additional characterisations are processed for that read.
d) Reads passing the Phred score filter step c) are then processed for their duplicate counts
    a. For paired ends, the 5' end is concatenated with the 3' end
    b. A 20bit hash is generated over the read sequence (concatenation if paired end) with another 20bit hash antisense. If the read sequence contains any indeterminate bases then that read is not processed further for duplicate counts.
    c. If a previously hashed seed sequence exists with the same hash as the current sequence then a check is made for an exact match and if matching the duplicate instance count is incremented. A check is made for sense and antisense (if '–strand' specified) matching.
    d. If no existing exact match can be found then subject to the limit specified with '—seeds' a new seed sequence is added with count 1 and either it's hashes added or linked with existing seed hashes. If already at the maximum seed limit then this read will not be contributing towards the duplicate counts.
e) Reads accepted as contributing towards the duplicate counts are then processed for their K-mer counts.
    a. A sliding window is utilised; the size K of window is iterated from 1 to the maximum specified length 'maxkmerlen' and counts accrued for each K-mer of size K over the full read length. K-mer counts are accrued sense strand only.
f) Reads which were accepted as contributing towards the duplicate counts are then processed for putative contaminate sequence overlaps according to the experimenter specified thresholds. Counts are accrued to the highest priority contaminate sequence for which an overlap match is detected. 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

When all reads in the readset have been processed with Phred, lengths, duplicates, K-mers, and contaminate counts accrued then the summary CSV and SVG graph results are generated.

Contaminates Distributions

Summarised contaminate distribution CSV file contains a header line followed by rows for each contaminate sequence; there may be multiple rows for a given contaminate sequence if of a compound overlap type, i.e perhaps the sequence was processed for overlaps on both 5' and 3' ends of reads. Column 1 contains the sequence name prefix stripped of any suffixed context numeric codes, column 2 the overlap type, column 3 the contaminate sequence length in bp, column 4 number of times a putative check for overlap was made, and column 5 the number of times a putative overlap was accepted. Subsequent columns contain the number of instances an overlap was of the column specific length. Thus a count of say 213 in column labelled 'Overlap:22' represents that for the given contaminant sequence there were 213 accepted overlap instances where the overlap was 22bp.

If checking for vector contamination then the number of reads which are completely contained within the vector sequences will be reported as additional appended rows in the contaminate distribution file.

Duplicate Read Distributions

The generated CSV summary file contains two columns (labelled as 'Instances' and 'Counts') with 2000 (excluding the 1st header row) rows of counts; counts are reported for instances 1 to 2000+. The 'Instances' column contains the number of copies with the number of times there were that many copies in the 'Count' column. An 'Instances' of one will have the number of read sequences which were unique (no duplicate copies) in 'Counts' column. 'Instances' of 2 will have the number of times there were read sequences duplicate 2x in the 'Counts' column, through to 'Instances' of 2000+ with 'Counts' containing the number of sequences which were duplicated >= 2000 fold.

K-mer Distributions

The generated CSV summary file contains as many columns plus one as there are in the maximum read length processed from the input NGS readsets. The first column identifies the K-mer instance over which the count accrual at each base position along the length of the read is being reported. For example the dimers are identified in column 1 as 'aa'…'tt'. Following columns after the first column are reporting the total number of times a given K-mer instance (column 1) was observed in the NGS readset at the base offset (column 2 has counts for the K-mer instance starting at read base offset 0, column 3 for base offset 1 …) within all reads.

Pearson Concordance Correlations

A window starting at 1/3rd of the sequence length is used as the reference, and accrued counts in the K-mer (length 1 .. 'kmerccc') instances starting at the left edge of the reference window are then

iteratively compared to all other windows of the same length starting from the first 5' base position through to the maximal 3' base position using the Pearson concordance measure as the quantitate comparison. A concordance of near 1.0 would be expected if the distribution of K-mer counts at the reference window were in concordance with little variation relative to the window read offset at which the K-mer count distributions were being compared. Assuming that the reads were being uniformly sampled with very low biases in start sites and little preferential library or sequencing artefacts then readsets should evidence uniform Pearsons for all K-mers across the full length of the readset which is near 1.0 with the exception of the 3' read end where the Pearson will be set to be 0 for windows in which the K-mer being processed would extend past the end of reads. The summary CSV file contains a header row labelling the K-mer window starting base at which the Pearson has been calculated, and this is followed by 'kmerccc' rows containing the Pearsons for each base offset in the read.

Phred Score Distributions

The Phred score distributions CSV summary file contains counts for each time a specific Phred score was observed at base positions along the length of read sequences. The 1st row is a header row with labels corresponding to the base position, and the 1st column identifies the Phred score for which that row contains the observed counts.

Phred Score Derived Error Free Reads Distribution

Of interest to the experimenter is a summary as to the discrete probabilities of reads being error free as derived from the individual base Phred scores. The processing is to calculate for each read the probability of that read being error free and then summarise over all reads in histogram form (100 bins) whereby each bin contains the number of reads observed with the bin range probability of being error free according to Phred scores.

Read Length Distributions

Read lengths are summarised as the read lengths in column 1 with the number of read instances of that length in column 2.

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