Biokanga Scaffolder Application Note

Covers release 3.7.6

11th May 2015

Overview

After assembly of very large numbers of NGS reads into smaller numbers of maximally contiguous sequences (contigs) by a de Novo assembler, perhaps with the 'biokanga assemb' subprocess, then the next workflow task is to attempt the ordering and relative orientation of these contigs using one or more paired end readsets. These scaffolding readsets normally will insert sizes such that the 5' end of a given pair could be expected to overlay a different contig to that which the 3' end of the same read pair overlays thus establishing the order and relative sense of the two overlaid contigs. A series of contigs linked by multiple scaffolding read pairs form a scaffold set, and any given contig can be a member of at most one scaffold set. By convention, if a contig can't be scaffolded then that contig will be the only member of a singular scaffold set.

If scaffolding read sets are available with different ranges of insert sizes then it is recommended that there be multiple invocations of the scaffolder subprocess with the initial invocation scaffolding with the shortest sized insert readset. The scaffolded set from the initial invocation are then used as input to the next scaffolder invocation together with the next larger insert sized scaffolding readset, with invocations repeated until all scaffolding readsets have been utilised.

Dependent on the estimated scaffolding readset coverage and actual readset end sequence lengths, the readsets used to scaffold may be filtered with the 'biokanga filter' subprocess; but it must be noted that with many scaffolding readsets there may be very little end overlapping resulting in too many reads being filtered out.

The scaffolding readsets may have been generated by one of four basic library sequencing protocols which will determine the relative orientation of the 5' PE1 read sequence relative to the 3' PE2 read sequence in any given readset. By default 'biokanga scaffold' will be expecting sense/antisense aka Illumina paired end short insert readsets, but parameterisation is provided by which the researcher can specify mate pair Roche 454 (sense/sense) or mate pair Illumina circularised (antisense/sense) or mate pair SOLiD (antisense/antisense).

The 'biokanga scaffolder' subprocess constructs a graph in which the contigs are nodes and the scaffolding read pair end alignments are represented as undirected edges annotated with sense orientations of the alignment of the respective read pair end to the contig node. Contig nodes are iterated and the most parsimonious edge linkage between the contig nodes is used to establish scaffold set membership.

Biokanga Scaffolder Parameterisation

The 'biokanga scaffolder' processing module requires, as a minimum, the researcher to specify the input file containing contigs (or the generated scaffolds from a previous invocation) and the scaffolding paired end readset. A number of internal thresholds, most having reasonable empirically derived defaults, are exposed through command line parameters by which the researcher can specify optimal parameterisations by which specific experimental objectives can be best realised.

- -m, --mode=<int>

- Currently there is only one processing mode which is simply referenced as being the default scaffolding mode. Future releases may allow for additional processing modes

- -M, --orientatepe=<int>
  - Use to specify the orientation of the paired end reads; individual end sequences may be either sense or antisense relative to each other dependent on the library preparation and sequencing protocols
  - Preparation and sequencing protocol used can be specified as one of
    - 0: PE1/PE2 is sense/antisense (PE short insert)
    - 1: PE1/PE2 is sense/sense (MP Roche 454)
    - 2: PE1/PE2 is antisense/sense (MP Illumina circularised)
    - 3: PE1/PE2 is antisense/antisense (MP SOLiD)
  - The default is for PE short insert

- -s, --maxsubs100bp=<int>
  - Allow max induced substitutions per 100bp overlapping paired end sequence onto a contig sequence
  - Can be specified to be in the range of 0 to 5
  - Defaults to 1

- -E, --maxendsubs=<int>
  - Allow max induced substitutions in overlap 12bp ends
  - Intended to compensate for any hexamer mispriming, especially evident with many RNA-seq datasets
  - These overlap end 12bp allowed substitutions are not combined with any allowed substitutions per 100bp overlapping end sequence as may be specified with the '—maxsubs100bp' parameterisation.
  - Can be specified to be in the range 0 to 6
  - Defaults to 0 whereby no substitution processing specific to the 12bp fragment ends is allowed.

- -p, --minpeinsert=<int>
  - Minimum PE insert size
  - Must be in the range of 100 to 50000
  - Defaults to 300
  - Note that the expected insert size should be treated as nominal only, and the minimal and maximum insert sizes will need to be set accordingly. As an example, using a PE readset with a nominal insert size of 300bp then a reasonable minimum PE insert size could be in the range 120 to 150, and a reasonable maximum PE insert size could be in the range 800 to 1200. The foregoing needs to take into account the read sizes used during assembly and the sequencing fragment size selection protocol – for instance gel-free fragment sizing protocols generally result in fragment sizes with much broader sizing ranges than gel-based fragment sizing selection protocols.

- -P, --maxpeinsert=<int>

- Maximum PE insert size
- Must be in the range of '—minpeinsert' to 50000
- Defaults to 10000
- Note: See earlier note relating to user specification of the minimum PE insert size parameter; the PE insert size is nominal only and the min/max limits should be set taking into account the original read sizes used in assembly and the sequencing fragment size selection protocol.

- -l, --minscafflen=<int>
  - Only report scaffold sets which contain sequences totalling at least this length
  - Must be in the range 100 to 5000
  - Defaults to 300

- -a, --inpe1=<file>
  - Load PE1 5' paired end scaffolding readset which is expected to be in fasta or fastq format May have been pre-processed with 'biokanga filter', or some other external process, to remove most duplicate and error containing reads.
- -A, --inpe2=<file>
  - Load PE2 3' paired end scaffolding readset which is expected to be in fasta or fastq format.
- -c, --contigsfile=<file>
  - Load SE readset which must be in fasta or fastq format:

- -o, --out=<file>
  - Output scaffolded contigs to this file

- -T, --threads=<int>
  - Number of processing threads 0..n (defaults to 0 which sets threads to number of CPU cores, max 64)

- -a, --inpe1=<file>
  - Load PE1 5' paired end readset which is expected to be in fasta or fastq format:
  - Expected to have been pre-processed with 'biokanga filter', or some other external process, to remove most duplicate and error containing reads.
  - If only SE readsets ('—seedcontigsfile') are to be assembled then this parameter is optional.

- -A, --inpe2=<file>
    - Load 3' paired end readset which is expected to be in fasta or fastq format:
        - The PE1 readset must have been specified with '--inpe1'.
    - Expected to have been pre-processed to remove most duplicate and error containing reads.
    - If only SE readsets ('—seedcontigsfile') are to be assembled then this parameter is optional.

- -c, --seedcontigsfile=<file>
    - Load SE readset which must be in fasta or fastq format:
        - Readset could be NGS readset
        - Readset could be high confidence fasta seed contigs or a previously assembled fasta SE sequences file
    - Optional and need not be specified if only PE assembly processing

- -i, --inartreducfile=<file>
    - Load PE or SE readset from a previously generated 'biokanga filter' artefact reduced packed reads file created with the

- -o, --out=<file>
    - Prefix name of files to which assembled contigs and remaining PE sequences are to be written in multifasta format.
    - The output file containing SE assembled contigs will have been named as the prefix name followed by '.SE.fasta'
    - If PE assembling, then the output file containing the 5' PE1 sequences (may have been extended) will have been named as the prefix name followed by '.PE1.fasta', whilst the 3' PE2 sequences will be written to output file having a name suffix of '.PE2.fasta'.
        - For example if the researcher used the parameter '—out=assembled' with a PE readset then 3 output files will be created named as 'assembled.SE.fasta', 'assembled.PE1.fasta' and 'assembled.PE2.fasta'.

- -T, --threads=<int>
    - Number of processing threads 0..n
    - Can be specified to be in the range 0 (uses all CPU cores limited to max 64) to 64
    - Defaults to 0 which sets threads to number of CPU cores but will use no more than 64