-=- dnAQET- README -=-

-- DESCRIPTION –

The *de novo* Assembly Quality Evaluation Tool (dnAQET) is a framework designed to evaluate a *de novo* assembly along with its scaffolds/contigs against a trusted reference genome. It first generates a quality score for each scaffold of an assembly based on its length, the misassemblies observed in the alignments and the total alignment percentage of a scaffold. Based on the quality scores assigned to the contigs/scaffolds, it computes an overall quality score for the whole assembly.

-- INSTALLATION & REQUIREMENTS--

The dnAQET is a Java package designed to be used in a Unix based operating system (such as Linux, MacOS, etc.) and it requires Java 1.7 Runtime Environment installed. It depends on Minimap2 tool and MUMmer4 package, which requires the following to run successfully.

- perl5 (5.6.0)
- sh
- sed
- awk

On the other hand, the dnAQET package does not require any installation. It automatically compiles MUMmer or Minimap2 in the first use and the main jar file does not require any compilation or installation since it is in byte code designed to work in a Java RTE.

-- RUNNING dnAQET --

The dnAQET provides three commands: '*evaluate*', '*reevaluate*', '*model'*. To do the evaluation of a de novo assembly for the first time, the user should run the '*evaluate*' command. For re-evaluating an existing evaluation with different parameters or a model, the user can run '*reevaluate*' command. The '*reevaluate*' command cannot be run on a **not previously evaluated** data. The '*model*' command generates a multiple linear regression model, which is used to predict the total number of expected artifacts for a given reference genome. Note that if the user does not specify a model when running the '*evaluate*' command, a model is automatically computed and saved as the result of this command.

The general usage of dnAQET is as follows:

'java -jar dnAQET.jar <command> [options]'

Below we explain each of the listed commands one by one in detail along with the required parameters.

1. ** evaluate **

DESCRIPTION:

The '*evaluate*' command takes a de novo assembly, which is a set of contigs/scaffolds in the fasta format as input and a trusted reference genome along with several other parameters and then generates a quality score in the [0, 1] interval for each scaffold/contig and a quality score for the whole assembly. The results are written into a user specified folder as multiple reports. This command is designed in such a way that the reference genome and the assembly file are partitioned into multiple files so that the alignment can be done in a parallel way. Although the user can determine the total number of partitions, the number genome partitions are optimized by dnAQET according to the selected alignment tool. If the user is running dnAQET in a High Performance Computing (HPC) system and he/she can choose to do the alignment in the processing nodes. Otherwise, the alignment processing will be done in the same node in multiple threads by default.

USAGE:

java -jar dnAQET.jar evaluate [options/parameters]

OPTIONS/PARAMETERS:

* -f, --file

  Input contig file path (this is a required parameter, not optional!)

* -r, --ref

  A trusted reference genome file. (this is a required parameter, not optional!)

-q, --alignmentTool

Tool which will be used for the alignment of the contigs/scaffolds back to reference

Default: minimap2

Possible Values: [minimap2, nucmer]

-d, --evaldestfolder

Evaluation destination directory where all the results are saved. If this folder is not specified by the user, a RESULT folder is created and the results are stored in that folder

Default: RESULT

-j, --jobcommand

Alignment job submission command. This command defines how the alignment jobs are processed. If the user is in a HPC environment and selects qsub the alignment processing is done in the processing nodes of the cluster. Otherwise he/she needs to use multithread (which is the default) option

Default: multithread

Possible Values: [sbatch, qsub, multithread]

-p, --jobcommandparams

Alignment job submission command parameters in quotes. The user can determine what parameters he/she wants to use for the job submission command. Example: "-pe multicore 4 -cwd"

Default: <empty string>

-l, --lengthscalingfactor

The length scaling factor constant that is used to scale the quality value of a scaffold/contig based on its size. The default value is the size of the smallest chromosome of the reference genome which is used to compute the regression model

-n, --npartition

Number of partitions for the input assembly file. The user can choose to partition the assembly into multiple files to speed up the alignment process using parallel processing

Default: 1

-m, --ov_dist_threshold

Distance (or overlap) threshold for the alignments of consecutive parts of a contig to be considered a RELOCATION type misassembly

Default: 1000

-t, --threadno

Number of threads for alignment (in case -j is selected to be 'multithread') step and for evaluation step.

Default: 4

-k, --model

A binary file that contains the multiple linear regression model coefficients to decide the penalty given to a misassembly. This file could have been produced by executing the *model* command separately. If this parameter is not specified, the evaluation process automatically computes a model from the given reference genome.

OUTPUT:

This command creates the folder specified with the '-d' option (in case it is specified. Otherwise RESULT folder is created) and this folder contains three sub folders, namely **bin**, **charts** and **reports** as well as a log file with '.params' extension. This log file contains the actual parameters which are used to run this command and a record of the run time of the individual steps of this command. The **reports** folder contains the following report files:

assembly.stat This file contains the basic statistics such as total number of contigs/scaffolds, total assembly lengths, N50 and L50 values etc. The quality score that dnAQET computes for this assembly is also reported in this file.

scaffolds.stat This file contains the actual statistics such as scaffold length, aligned scaffold length, alignment percentage, the quality score, number of misassemblies etc. for the scaffolds.

scaffolds.alignment This file contains the alignments for each scaffold produced by the alignment tool and then filtered and trimmed by dnAQET.

misassembly.report This file contains the details of the misassemblies detected in the *de novo* assembly. The columns are the Scaffold ID, Misassembly Type, ScaffoldCoordinates1, ScaffoldCoordinates2 (i.e., the coordinates of the two flanking segments in a scaffold which happened to have the specified type of misassembly), AlignmentCoords1 and AlignmentCoords2 (i.e., the alignment coordinates of the two specified adjacent scaffold segments on the reference genome, respectively).

reference_scaffoldinggaps.report This file contains the coordinates of the gaps (ambiguous segments) in the reference genome.

The **charts** folder contains the following files:

qualityScoreHistogram.jpeg This is a chart showing the distribution of the quality scores of the contigs using a histogram.

qualityScoreHistogram.txt This is a table displaying the histogram in textual format.

qualityScoreHistogram_RelativeFreq.jpeg This is a chart displaying the relative frequency distribution of the quality scores of the contigs using a histogram.

qualityScoreHistogram_RelativeFreq.txt This is a table displaying the relative frequency histogram in textual format.

cumsizeratio@Qthresholds.jpeg This is the chart demonstrating how the ratio of the cumulative size of contigs/scaffolds (whose quality scores exceed the quality threshold specified on the x-axis) to the whole assembly size behave as the quality thresholds change.

cumsizeratio@Qthresholds.txt This is a table displaying the data in cumsizeratio@Qthresholds.jpeg file in textual format.

genomecoverage@Qthresholds.jpeg This is a chart demonstrating how the reference genome coverage behave as the quality thresholds change (x-axis). To calculate the values on the y-axis, only the contigs/scaffolds whose quality values exceed the corresponding threshold value on the x-axis are considered for coverage computation.

genomecoverage@Qthresholds.txt This is the table displaying the data in genomecoverage@Qthresholds.jpeg file in textual format.

The **bin** folder contains the following files:

scaffold_stats.bin This file contains binary data about the contigs/scaffolds and their alignments.

references.bin This file contains binary data about the reference genome used.

model.bin This file contains binary data about the linear regression model which is computed for the given reference genome. If the user wants to use the same model for other assembly evaluations, he/she can directly use this file.

2. ** reevaluate **

DESCRIPTION:

The '*reevaluate*' command takes either the results of a previous '*evaluate*' or a '*reevaluate*' command. It then re-computes these with the given parameters to obtain the same type of results as the '*evaluate*' command.

USAGE:

java -jar dnAQET.jar reevaluate [options/parameters]

OPTIONS/PARAMETERS:

* -s, --sourcefolder

This is the source folder where a previous result of either an 'reevaluate' or 'evaluate' command is stored. (this is a required parameter, not optional!)

-d, --evaldestfolder

Evaluation destination directory where all the results are saved. If this folder is not specified by the user, the results are saved to the folder spcified by -s -parameter by overriding its contents.

-l, --lengthscalingfactor

The length scaling factor constant that is used to scale the quality value of a scaffold/contig based on its size. The default value is the size of the smallest chromosome of the reference genome which is used to compute the regression model

-m, --ov_dist_threshold

Distance (or overlap) threshold for the alignments of consecutive parts of a contig to be considered a RELOCATION type misassembly

Default: 1000

-t, --threadno

Number of threads for alignment (in case -j is selected to be 'multithread') step and for evaluation step

Default: 4

OUTPUT:

This command creates the folder specified with the '-d' option (in case it is specified. Otherwise the folder specified with the -s option is overridden) with exactly the types of files and folders produced by the '*evaluate*' command.

3. ** model **

DESCRIPTION:

The '*model*' command generates a binary file which contains the linear regression model coefficients for computing the expected number of misassemblies/artifacts for a reference genome. Note that although a model is always needed for running both '*evaluate*' or a '*reevaluate*' commands, they are not required. In case of '*evaluate*' command, if a pre-computed model is not specified then the dnAQET automatically computes a model for the given reference. In case of the '*reevaluate*' command, if the user does not specify a new model, then the model used to generate the current results will be used

USAGE:

java -jar dnAQET.jar model [options/parameters]

OPTIONS/PARAMETERS:

* -r, --ref

Input reference genome file upon which the model will be built. (this is a required parameter, not optional!)

-q, --alignmentTool

Tool which will be used for the alignment of the contigs/scaffolds back to reference

Default: minimap2

Possible Values: [minimap2, nucmer]

-t, --threadno

Number of threads for alignment (in case -j is selected to be 'multithread') step and for evaluation step.

-j, --jobcommand

Alignment job submission command. This command defines how the alignment jobs are processed. If the user is in a HPC environment and selects qsub the alignment processing is done in the processing nodes of the cluster. Otherwise he/she needs to use multithread (which is the default) option

Default: multithread

Possible Values: [sbatch, qsub, multithread]

-p, --jobcommandparams

Alignment job submission command parameters in quotes. The user can determine what parameters he/she wants to use for the job submission command. Example: "-pe multicore 4 -cwd"

Default: <empty string>

-d, --modeldestfolder

Model file destination directory