Serial NGS Reads Classification

The workflow sample, described below, takes FASTQ files with metagenomic NGS reads as input and process them as follows:

- Improve reads quality with Trimmomatic
- Provide FastQC reads quality reports
- Classification:
  Classify the pre-processed reads with Kraken
  Get reads that were not classified by Kraken
  Classify these reads with CLARK
  Get reads that were not classified (in case of SE reads)
  Classify these reads with DIAMOND (in case of SE reads)
  Provide general classification reports

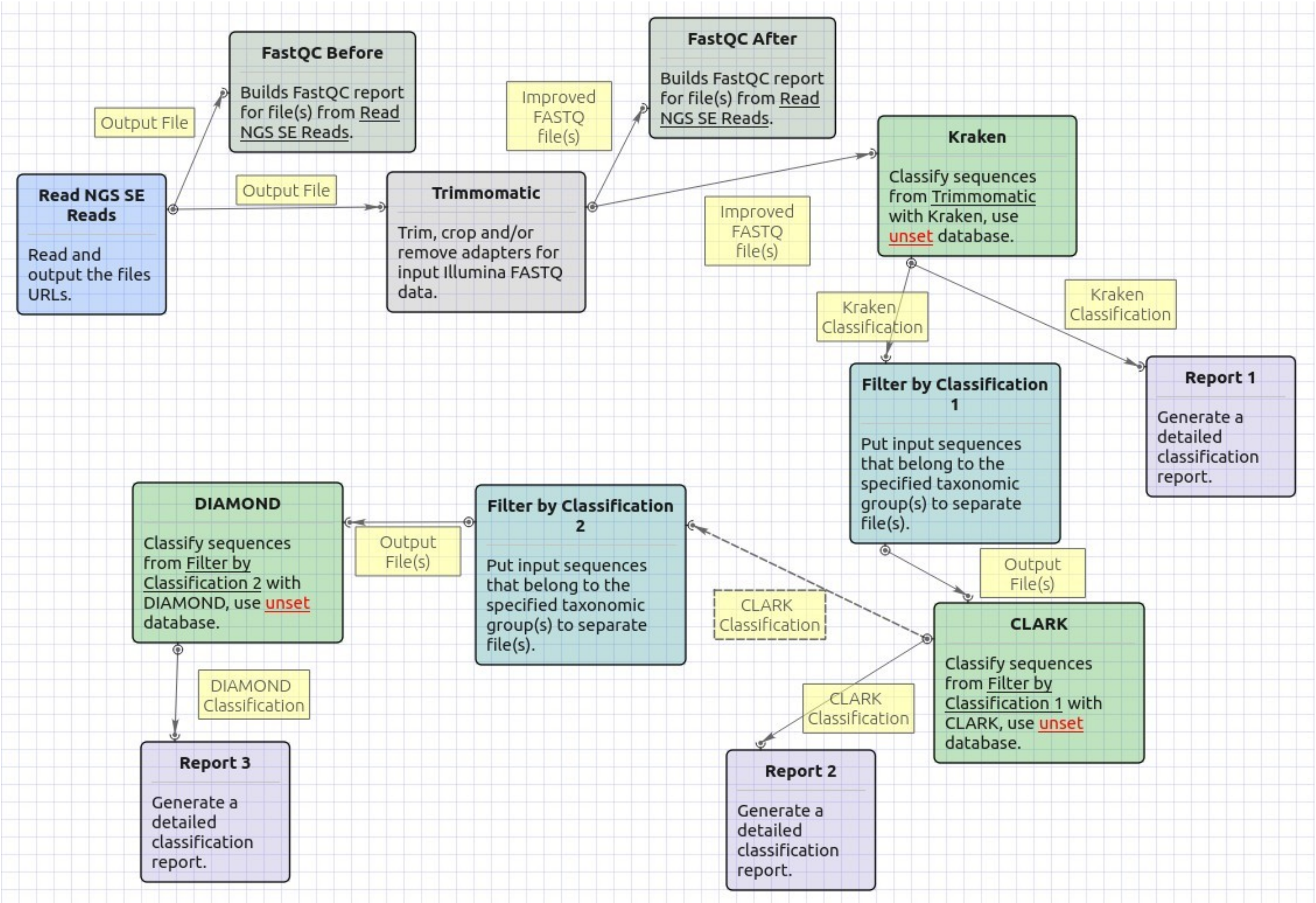
> **How to Use This Sample**
>
> If you haven't used the workflow samples in UGENEbefore, look at the "How to Use Sample Workflows" section of the documentation.

Workflow Sample Location

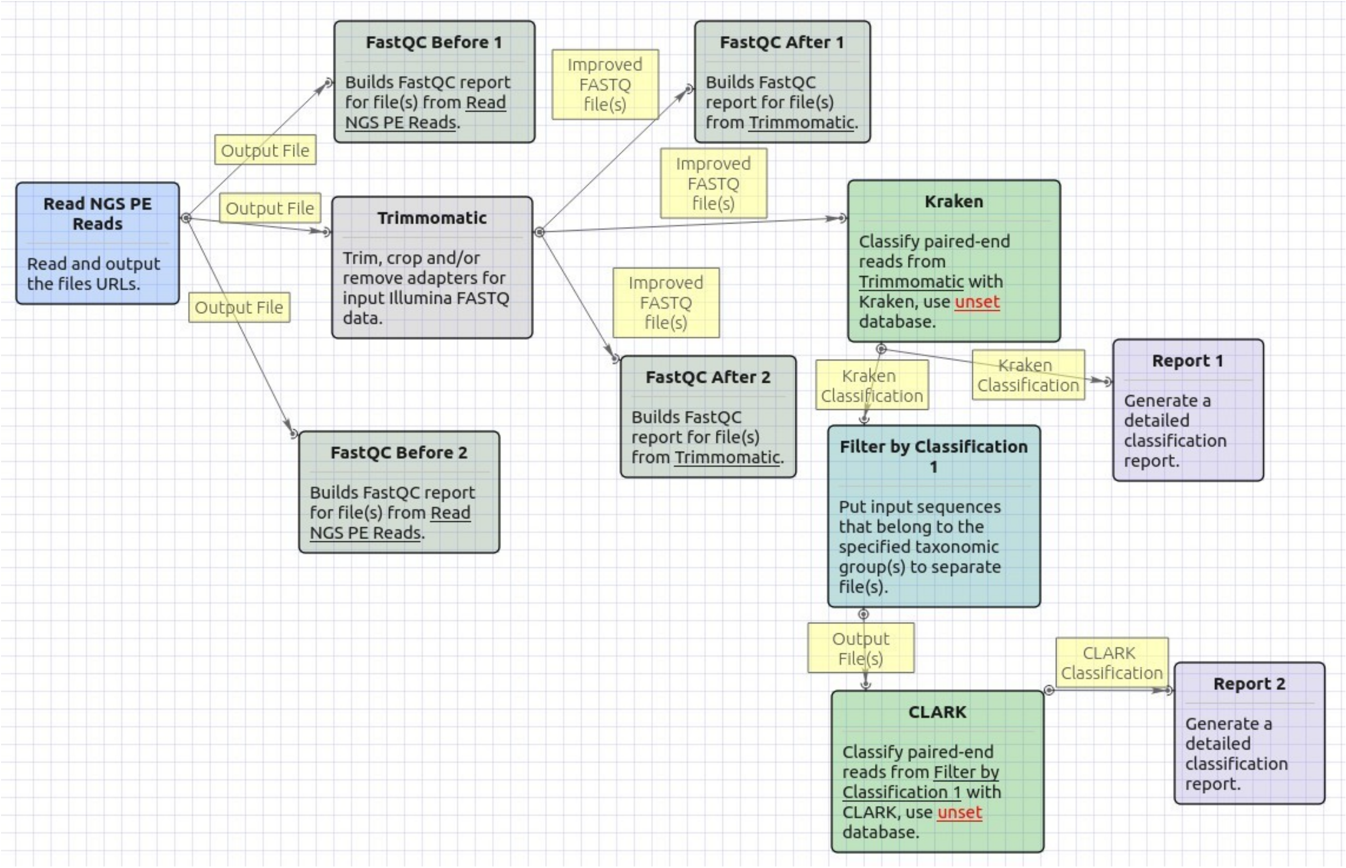
The workflow sample "Serial NGS Reads Classification" can be found in the "NGS" section of the Workflow Designer samples.

Workflow Image

The opened workflow for single-end reads looks as follows:


FastQC Before
Builds FastQC report for file(s) from Read NGS SE Reads.
Output File
Improved FASTQ file(s)
FastQC After
Builds FastQC report for file(s) from Read NGS SE Reads.
Read NGS SE Reads
Read and output the files URLs.
Output File
Trimmomatic
Trim, crop and/or remove adapters for input Illumina FASTQ data.
Improved FASTQ file(s)
Kraken
Classify sequences from Trimmomatic with Kraken, use unset database.
Kraken Classification
Kraken Classification
Filter by Classification 1
Put input sequences that belong to the specified taxonomic group(s) to separate file(s).
Report 1
Generate a detailed classification report.
DIAMOND
Classify sequences from Filter by Classification 2 with DIAMOND, use unset database.
Output File(s)
Filter by Classification 2
Put input sequences that belong to the specified taxonomic group(s) to separate file(s).
CLARK Classification
Output File(s)
CLARK
Classify sequences from Filter by Classification 1 with CLARK, use unset database.
DIAMOND Classification
CLARK Classification
Report 3
Generate a detailed classification report.
Report 2
Generate a detailed classification report.

The opened workflow for paired-end reads looks as follows:



Read NGS PE Reads
Read and output the files URLs.

FastQC Before 1
Builds FastQC report for file(s) from Read NGS PE Reads.

Trimmomatic
Trim, crop and/or remove adapters for input Illumina FASTQ data.

FastQC Before 2
Builds FastQC report for file(s) from Read NGS PE Reads.

FastQC After 1
Builds FastQC report for file(s) from Trimmomatic.

FastQC After 2
Builds FastQC report for file(s) from Trimmomatic.

Kraken
Classify paired-end reads from Trimmomatic with Kraken, use unset database.

Report 1
Generate a detailed classification report.

Filter by Classification 1
Put input sequences that belong to the specified taxonomic group(s) to separate file(s).

CLARK
Classify paired-end reads from Filter by Classification 1 with CLARK, use unset database.

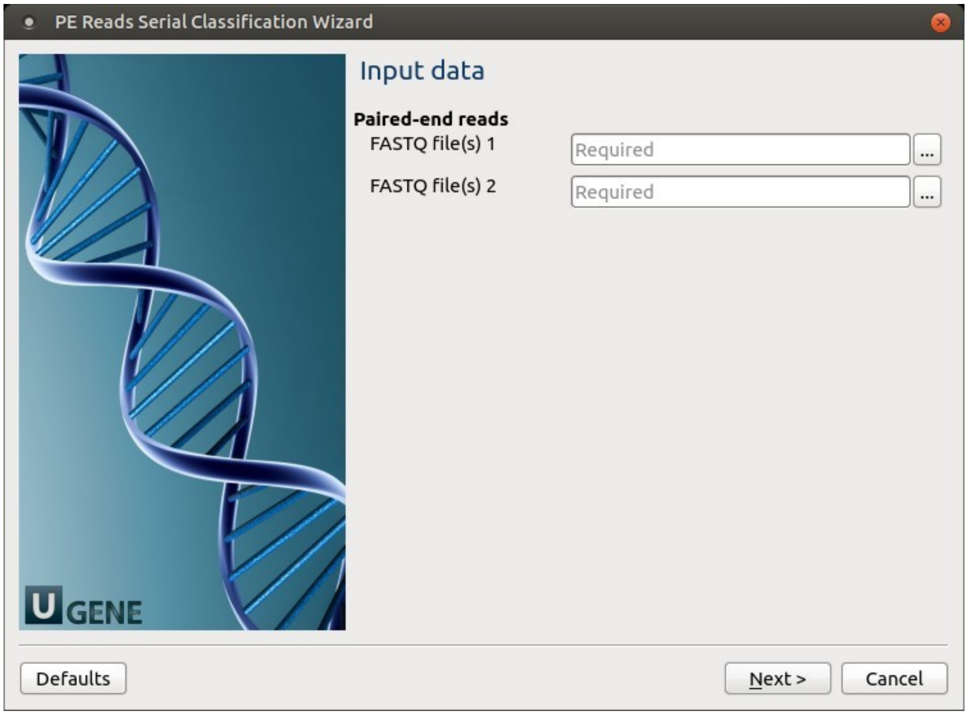
Report 2
Generate a detailed classification report.

Output File
Output File
Output File
Improved FASTQ file(s)
Improved FASTQ file(s)
Improved FASTQ file(s)
Kraken Classification
Kraken Classification
Output File(s)
CLARK Classification

Workflow Wizard

The wizard has 5 pages.

1. Input data: On this page, input files must be set.



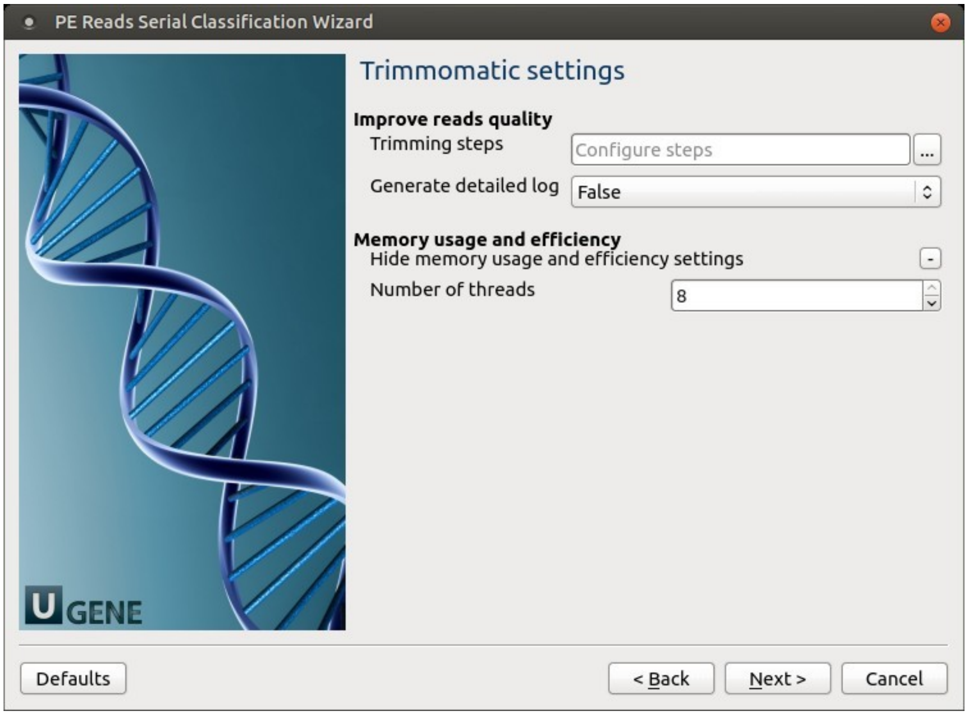
PE Reads Serial Classification Wizard

Input data

Paired-end reads
FASTQ file(s) 1: Required
FASTQ file(s) 2: Required

...
...
UGENE
Defaults
Next >
Cancel

2. Trimmomatic settings: The Trimmomatic parameters can be changed here.



PE Reads Serial Classification Wizard

Trimmomatic settings

Improve reads quality

Trimming steps: Configure steps

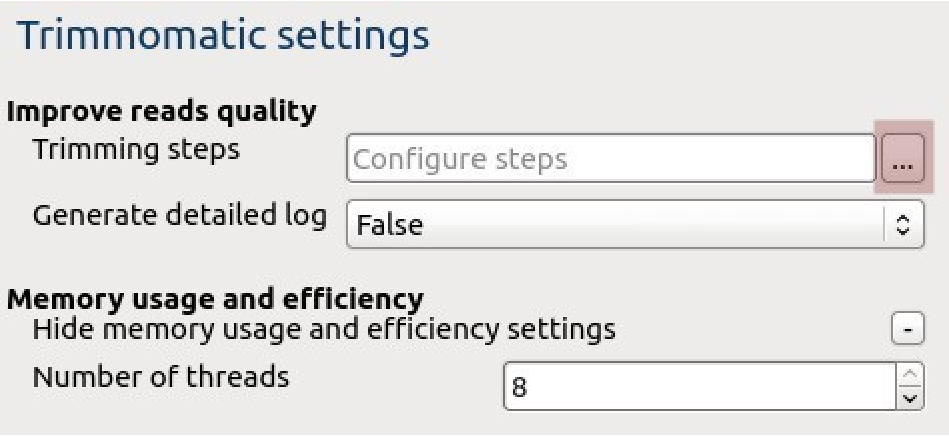
Generate detailed log: False

Memory usage and efficiency

Hide memory usage and efficiency settings

Number of threads: 8

To configure trimming steps use the following button:



Trimmomatic settings

Improve reads quality

Trimming steps: Configure steps

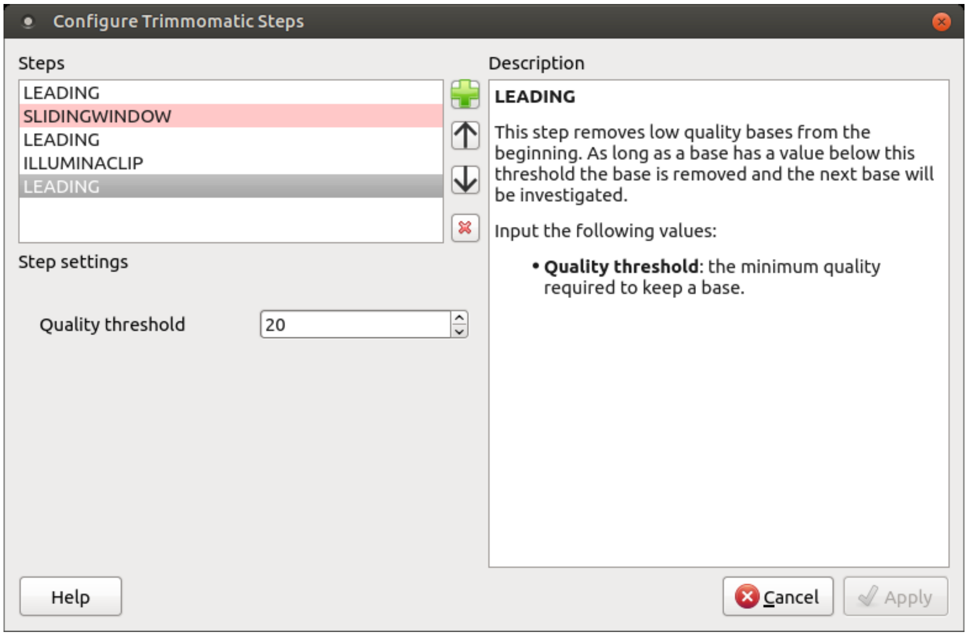
Generate detailed log: False

Memory usage and efficiency

Hide memory usage and efficiency settings

Number of threads: 8

The following dialog will appear:



LEADING
SLIDINGWINDOW
LEADING
ILLUMINACLIP
LEADING

Step settings

Quality threshold 20

LEADING

This step removes low quality bases from the beginning. As long as a base has a value below this threshold the base is removed and the next base will be investigated.

Input the following values:

- **Quality threshold**: the minimum quality required to keep a base.

Click the *Add new* step button and select a step. The following options are available:

- ILLUMINACLIP: Cutadapterandotherillumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality.
- TRAILING: Cut bases off the end of a read, if below a threshold quality.
- CROP: Cut the read to a specified length.
- HEADCROP: Cut the specified number of bases from the start of the read.
- MINLEN: Drop the read if it is below a specified length.
- AVGQUAL: Drop the read if the average quality is below the specified level.
- TOPHRED33: Convert quality scores to Phred-33.
- TOPHRED64: Convert quality scores to Phred-64.

Each step has the own parameters:

AVGQUAL

This step drops a read if the average quality is below the specified level.

Input the following values:

- Quality threshold: the minimum average quality required to keep a read.

CROP

This step removes bases regardless of quality from the end of thread, so that the readhas maximally the specified length after this step has been performed. Steps performed after CROP might of course further shorten the read.

Input the following values:

- Length: the number of bases to keep, from the start of the read.

HEADCROP

This step removes the specified number of bases, regardless of quality, from the beginning of the read.

Input the following values:

- Length: the number of bases to remove from the start of the read.

ILLUMINACLIP

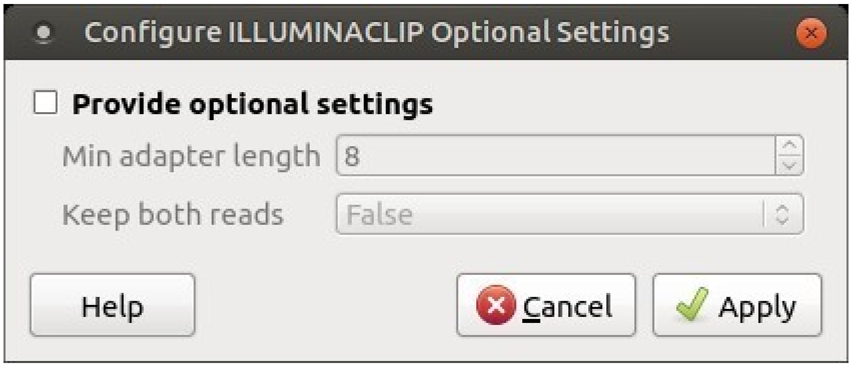
This step is used to find and remove Illumina adapters.

Trimmomatic first compares short sections of an adapter and a read. If they match enough, the entire alignment between the read and adapter is scored. For paired-end reads, the "palindrome" approach is also used to improve the result. See Trimmomatic manual for details.

Input the following values:

- Adapter sequences: a FASTA file with the adapter sequences. Files for TruSeq2 (GAII machines), TruSeq3 (HiSeq and MiSeq machines) and Nextera kits for SE and PE reads are now available by default. The naming of the various sequences within the specified file determines how they are used.
- Seed mismatches: the maximum mismatch count in short sections which will still allow a full match to be performed.
- Simple clip threshold: a threshold for simple alignment mode. Values between 7 and 15 are recommended. A perfect match of a 12 base sequence will score just over 7, while 25 bases are needed to score 15.
- Palindrome clip threshold: a threshold for palindrome alignment mode. For palindromic matches, a longer alignment is possible. Therefore the threshold can be in the range of 30. Even though this threshold is very high (requiring a match of almost 50 bases) Trimmomatic is still able to identify very, very short adapter fragments.

There are also two optional parameters for palindrome mode: Min adapter length and Keep both reads. Use the following dialog. To call the dialog press the *Optional* button.


Configure ILLUMINACLIP Optional Settings

☐ Provide optional settings

Min adapter length: 8

Keep both reads: False

LEADING

This step removes low-quality bases from the beginning. As long as a base has a value below this threshold the base is removed and the next base will be investigated.

Input the following values:

- Quality threshold: the minimum quality required to keep a base.

MAXINFO

This step performs an adaptive quality trim, balancing the benefits of retaining longer reads against the costs of retaining bases with errors. See Trimmomatic manual for details.

Input the following values:

- Target length: the read length which is likely to allow the location of the read within the target sequence. Extremely short reads, which can be placed into many different locations, provide little value. Typically, the length would be in the order of 40 bases, however, the value also depends on the size and complexity of the target sequence.
- Strictness: the balance between preserving as much read length as possible vs. removal of incorrect bases. A low value of this parameter (0.8) favours read correctness.

MINLEN

This step removes reads that fall below the specified minimum length. If required, it should normally be after all other processing steps. Reads removed by this step will be counted and included in the "dropped reads" count.

Input the following values:

- Length: the minimum length of reads to be kept.

SLIDINGWINDOW

This step performs a sliding window trimming, cutting once the average quality within the window falls below a threshold. By considering multiple bases, a single poor quality base will not cause the removal of high-quality data later in the read.

Input the following values:

- Window size: the number of bases to an average across.
- Quality threshold: the average quality required.

TOPHRED33

This step (re)encodes the quality part of the FASTQ file to base 33.

TOPHRED64

This step (re)encodes the quality part of the FASTQ file to base 64.

TRAILING

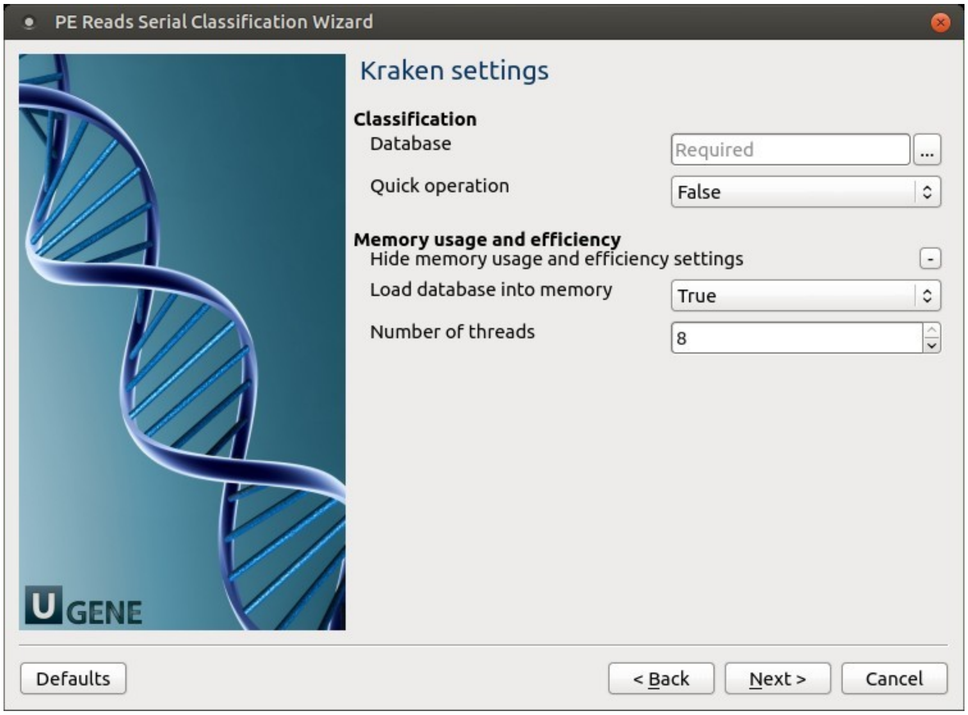
This step removes low-quality bases from the end. As long as a base has a value below this threshold the base is removed and the next base (i. e. the preceding one) will be investigated. This approach can be used removing the special Illumina " low-quality segment" regions (which are marked with a quality score of 2), but SLIDINGWINDOW or MAXINFO are recommended instead.

Input the following values:

- Quality threshold: the minimum quality required to keep a base.

To remove a step use the *Remove selected step* button. The pink highlighting means the required parameter has not been set.

3. Kraken settings: Default Kraken parameters can be changed here.



PE Reads Serial Classification Wizard

Kraken settings

Classification

Database: Required

Quick operation: False

Memory usage and efficiency

Hide memory usage and efficiency settings

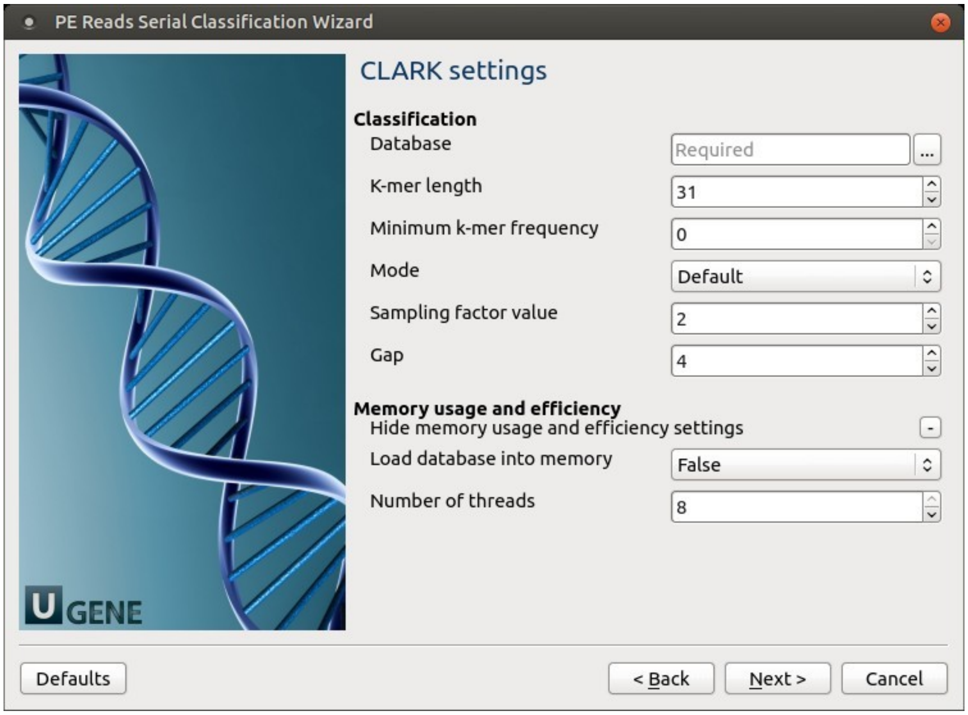
Load database into memory: True

Number of threads: 8

The following parameters are available:

Database	A path to the folder with the Kraken database files.
Quick operation	Stop classification of an input read after the certain number of hits. The value can be specified in the "Minimum number of hits" parameter.

4. CLARK settings: Default CLARK parameters can be changed here.



CLARK settings

Classification

Database: Required

K-mer length: 31

Minimum k-mer frequency: 0

Mode: Default

Sampling factor value: 2

Gap: 4

Memory usage and efficiency

Hide memory usage and efficiency settings -

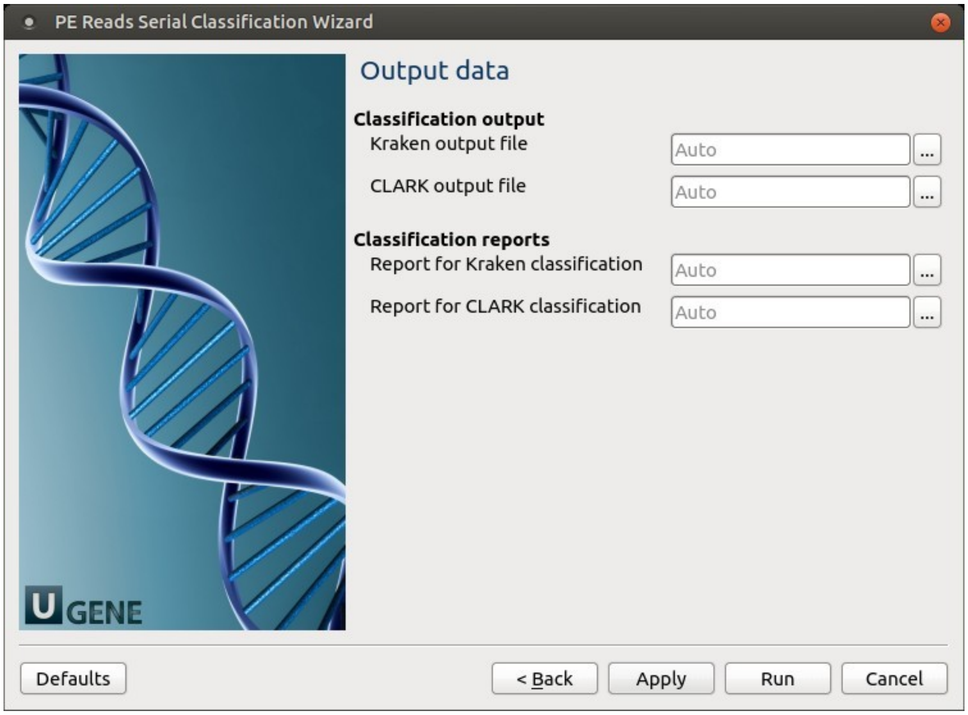
Load database into memory: False

Number of threads: 8

The following parameters are available:

Database	A folder that should be used to store the database files.
K-mer length	This value is critical for the classification accuracy and speed. For high sensitivity, it is recommended to set this value to 20 or 21 (along with the "Full" mode). However, if the precision and the speed are the main concern, use any value between 26 and 32. Note that the higher the value, the higher is the RAM usage. So, as a good tradeoff between speed, precision, and RAM usage, it is recommended to set this value to 31 (along with the "Default" or "Express" mode).
Minimum k-mer frequency	Minimum of k-mer frequency/occurrence for the discriminative k-mers(-t). For example, for 1 (or, 2), the program will discard any discriminative k-mer that appear only once (or, less than twice).
Mode	Set the mode of the execution (-m): • "Full" to get detailed results, confidence scores, and other statistics. • "Default" to get results summary and perform the best trade-off between classification speed, accuracy and RAM usage. • "Express" to get results summary with the highest speed possible.
Sampling factor value	
Gap	"Gap" or number of non-overlapping k-mers to pass when creating the database (-). Increase the value if it is required to reduce the RAM usage. Note that this will degrade the sensitivity.

5. Output Files Page: On this page, you can select an output directory:


PE Reads Serial Classification Wizard

Output data

Classification output

Kraken output file: Auto

CLARK output file: Auto

Classification reports

Report for Kraken classification: Auto

Report for CLARK classification: Auto

UGENE
Defaults
< Back
Apply
Run
Cancel