Raw ChIP-Seq Data Processing

Download and install the UGENE NGS package to use this pipeline.

Use this workflow sample to process raw ChIP-seq next-generation sequencing (NGS) data from the Illumina platform. The processing includes:

- *Filtration:*
  - Filtering of the NGS short reads by the CASAVA 1.8 header;
  - Trimming of the short reads by quality;
- *Mapping:*
  - Mapping of the short reads to the specified reference sequence (the BWA-MEM tool is used in the sample);
- *Post-filtration:*
  - Filtering of the aligned short reads by SAMtools to remove reads with low mapping quality, unpaired/unaligned reads;
  - Removing of duplicated short reads.

The result of the data processing is provided in the BED format. Intermediate data files from the filtration and mapping steps are also available in the output.

How to Use This Sample

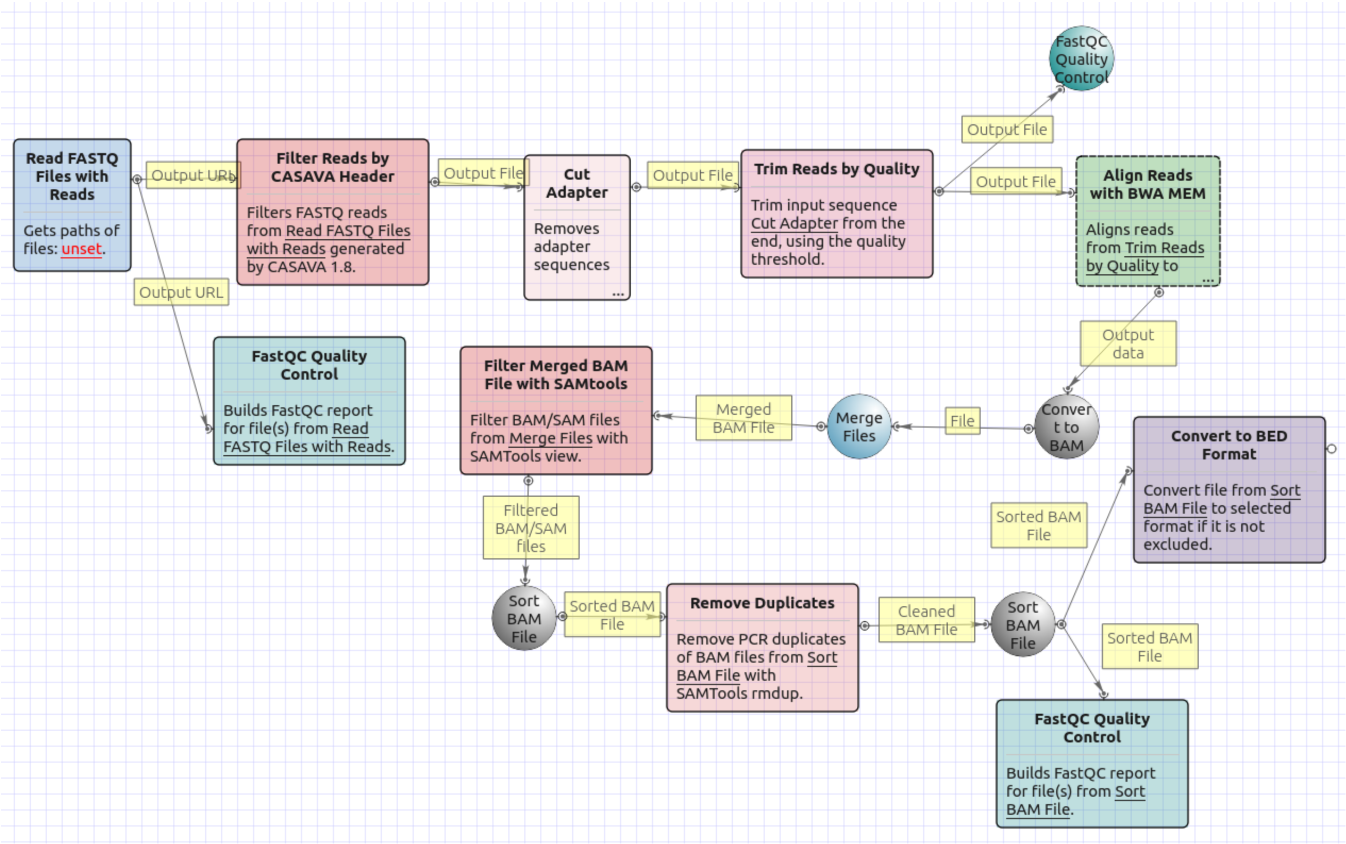
If you haven't used the workflow samples in UGENE before, look at the "How to Use Sample Workflows" section of the documentation.

Workflow Sample Location

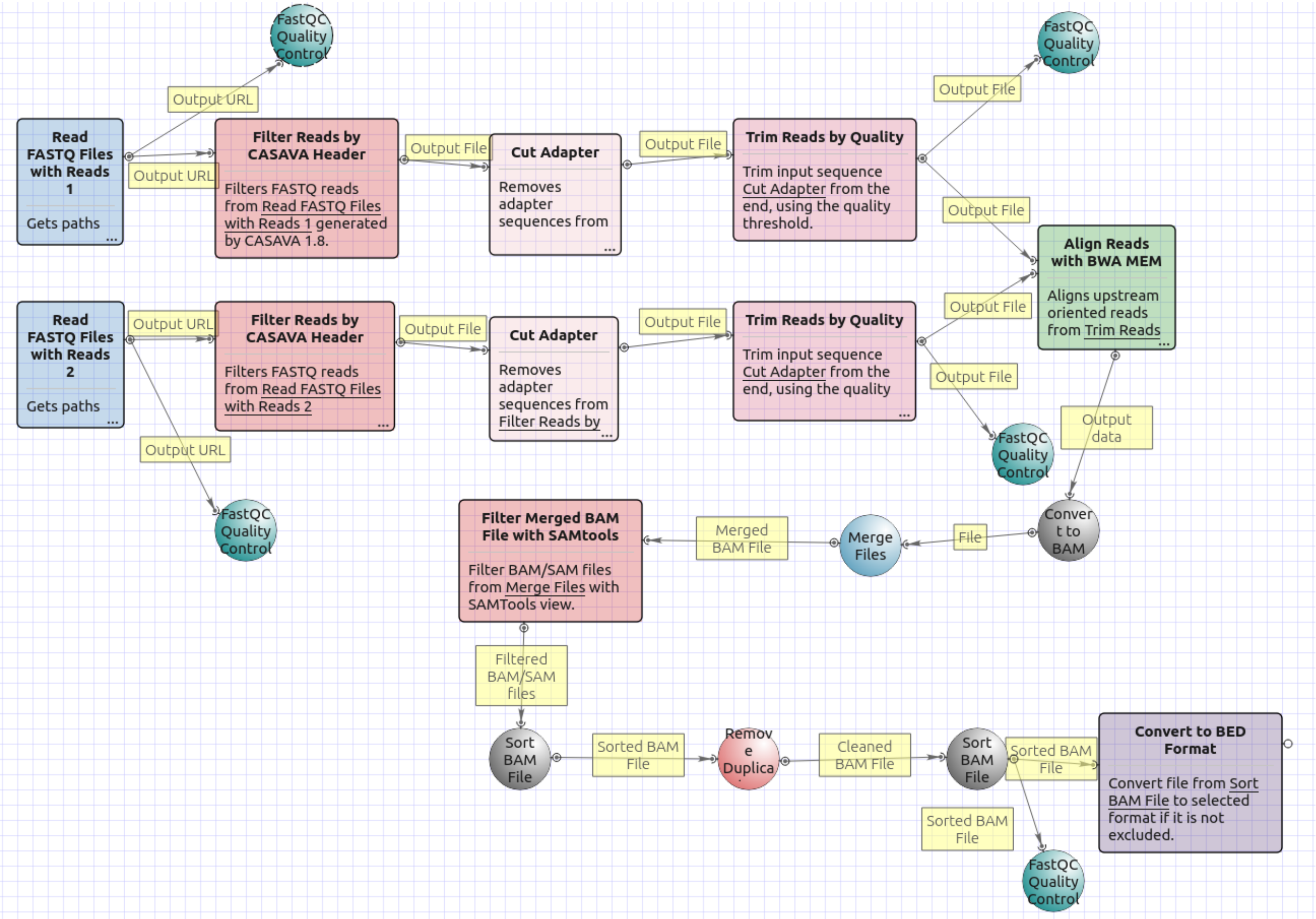
The workflow sample "Raw ChIP-Seq processing" can be found in the "NGS" section of the Workflow Designer samples.

Workflow Image

There are two versions of the workflow available. The workflow for single-end reads looks as follows:


FastQC Quality Control
Output File
Read FASTQ Files with Reads
Gets paths of files: unset.
Output URL
Filter Reads by CASAVA Header
Filters FASTQ reads from Read FASTQ Files with Reads generated by CASAVA 1.8.
Output File
Cut Adapter
Removes adapter sequences ...
Output File
Trim Reads by Quality
Trim input sequence Cut Adapter from the end, using the quality threshold.
Output File
Align Reads with BWA MEM
Aligns reads from Trim Reads by Quality to ...
Output URL
FastQC Quality Control
Builds FastQC report for file(s) from Read FASTQ Files with Reads.
Filter Merged BAM File with SAMtools
Filter BAM/SAM files from Merge Files with SAMTools view.
Merged BAM File
Merge Files
File
Output data
Convert to BAM
Convert to BED Format
Convert file from Sort BAM File to selected format if it is not excluded.
Filtered BAM/SAM files
Sorted BAM File
Sort BAM File
Sorted BAM File
Remove Duplicates
Remove PCR duplicates of BAM files from Sort BAM File with SAMTools rmdup.
Cleaned BAM File
Sort BAM File
Sorted BAM File
FastQC Quality Control
Builds FastQC report for file(s) from Sort BAM File.

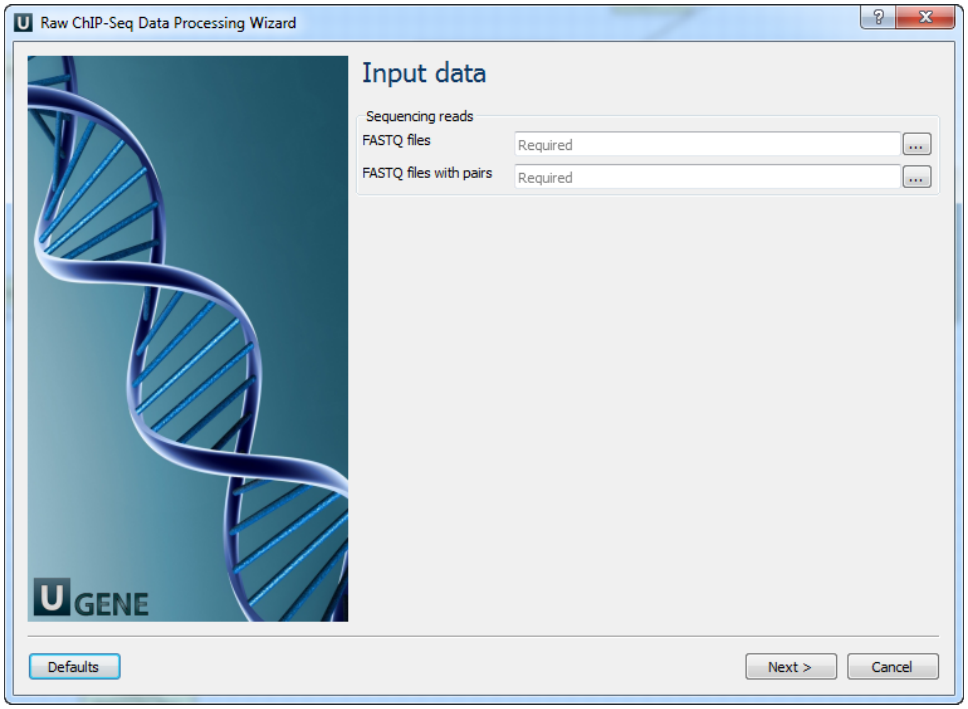
The workflow for paired-end short appearance is the following:


FastQC Quality Control
Output URL
Read FASTQ Files with Reads 1
Gets paths
Output URL
Filter Reads by CASAVA Header
Filters FASTQ reads from Read FASTQ Files with Reads 1 generated by CASAVA 1.8.
Output File
Cut Adapter
Removes adapter sequences from
Output File
Trim Reads by Quality
Trim input sequence Cut Adapter from the end, using the quality threshold.
Output File
FastQC Quality Control
Output File
Align Reads with BWA MEM
Aligns upstream oriented reads from Trim Reads
Read FASTQ Files with Reads 2
Gets paths
Output URL
Filter Reads by CASAVA Header
Filters FASTQ reads from Read FASTQ Files with Reads 2
Output File
Cut Adapter
Removes adapter sequences from Filter Reads by
Output File
Trim Reads by Quality
Trim input sequence Cut Adapter from the end, using the quality
Output File
Output File
FastQC Quality Control
Output data
Output URL
FastQC Quality Control
Filter Merged BAM File with SAMtools
Filter BAM/SAM files from Merge Files with SAMTools view.
Merged BAM File
Merge Files
File
Convert to BAM
Filtered BAM/SAM files
Sort BAM File
Sorted BAM File
Remove Duplica
Cleaned BAM File
Sort BAM File
Sorted BAM File
Convert to BED Format
Convert file from Sort BAM File to selected format if it is not excluded.
Sorted BAM File
FastQC Quality Control

Workflow Wizard

The workflows have the similar wizards. The wizard for paired-end reads has 5 pages.

1. Input data: On this page you must input FASTQ file(s).


Raw ChIP-Seq Data Processing Wizard
UGENE
Defaults
Next >
Cancel

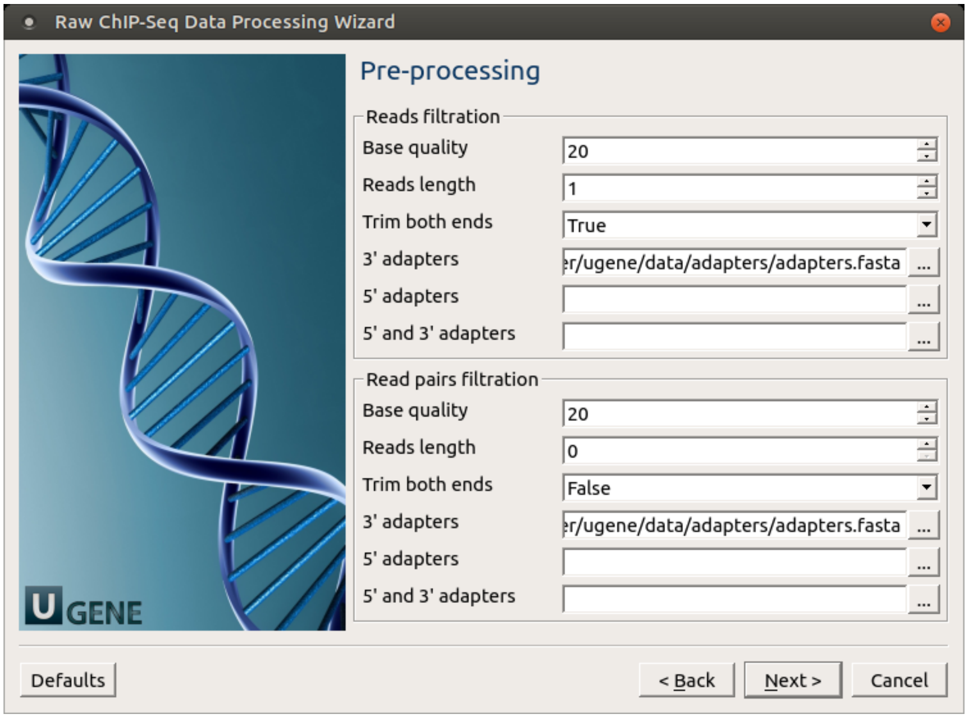
Input data

Sequencing reads

FASTQ files: Required

FASTQ files with pairs: Required

2. Pre-processing: On this page you can modify filtration parameters.


Raw ChIP-Seq Data Processing Wizard
UGENE
Defaults
< Back
Next >
Cancel

Pre-processing

Reads filtration

Base quality: 20

Reads length: 1

Trim both ends: True

3' adapters: r/ugene/data/adapters/adapters.fasta

5' adapters:

5' and 3' adapters:

Read pairs filtration

Base quality: 20

Reads length: 0

Trim both ends: False

3' adapters: r/ugene/data/adapters/adapters.fasta

5' adapters:

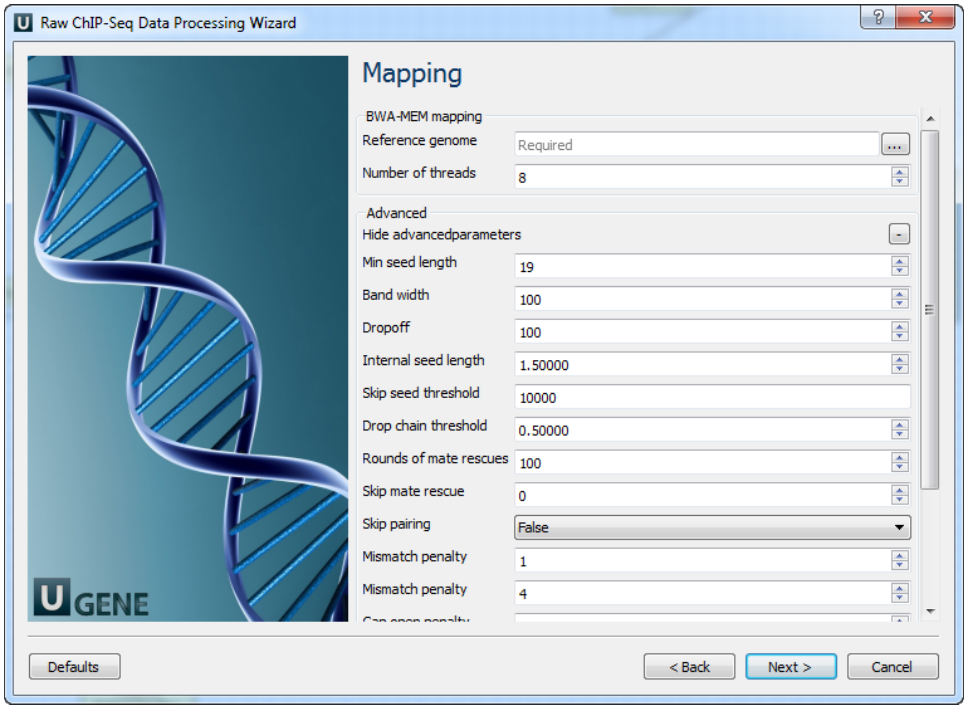
5' and 3' adapters:

The following parameters are available for reads and reads pairs filtration:

Base quality	Quality threshold for trimming.
Reads length	Too short reads are discarded by the filter.
Trim both ends	Trim the both ends of a read or not. Usually, you need to set True for Sanger sequencing and False for NGS
3'	

adapters	A FASTA file with one or multiple sequences of adapter that were ligated to the 3' end. The adapter itself and anything that follows is trimmed. If the adapter sequence ends with the '$ character, the adapter is anchored to the end of the read and only found if it is a suffix of the read.
5' adapters	A FASTA file with one or multiple sequences of adapters that were ligated to the 5' end. If the adapter sequence starts with the character ^, the adapter is 'anchored'. An anchored adapter must appear in its entirety at the 5' end of the read (it is a prefix of the read). A non-anchored adapter may appear partially at the 5' end, or it may occur within the read. If it is found within a read, the sequence preceding the adapter is also trimmed. In all cases, the adapter itself is trimmed.
5' and 3' adapters	A FASTA file with one or multiple sequences of adapter that were ligated to the 5' end or 3' end.

3. Mapping: On this page you must input reference and optionally modify advanced parameters.



Mapping

BWA-MEM mapping

Reference genome: Required

Number of threads: 8

Advanced

Hide advancedparameters

Min seed length: 19

Band width: 100

Dropoff: 100

Internal seed length: 1.50000

Skip seed threshold: 10000

Drop chain threshold: 0.50000

Rounds of mate rescues: 100

Skip mate rescue: 0

Skip pairing: False

Mismatch penalty: 1

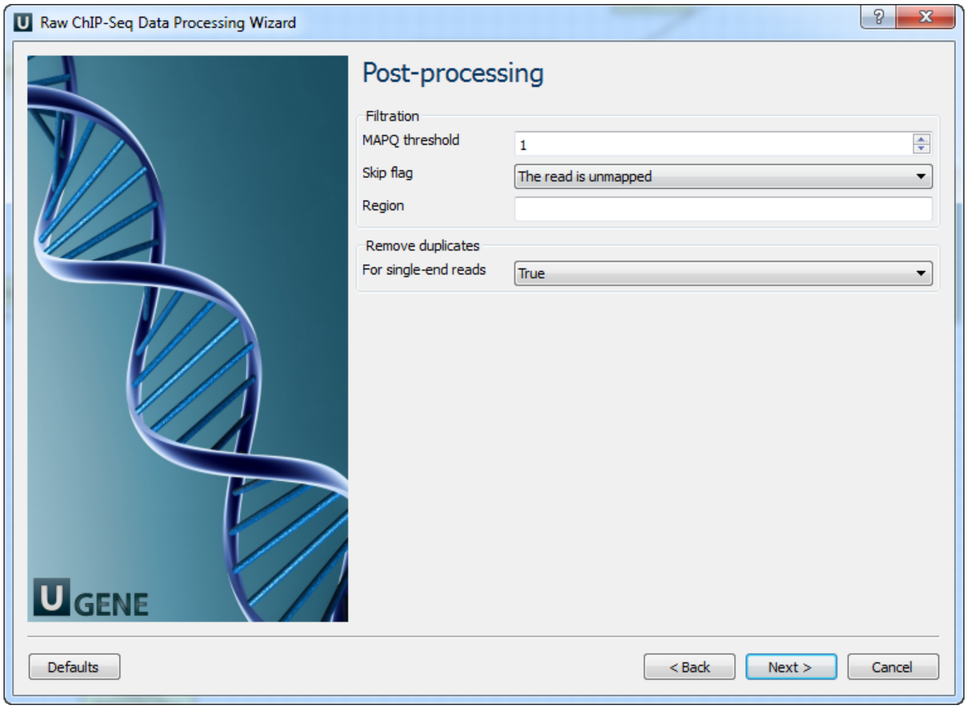
Mismatch penalty: 4

The following parameters are available:

Reference genome	Path to indexed reference genome.
Number of threads	Number of threads (-t).
Min seed length	Path to indexed reference genome (-k).
Band width	Band width for banded alignment (-w).
Dropoff	Off-diagonal X-dropoff (-d).
Internal seed length	Look for internal seeds inside a seed longer than {-k} (-r).
Skip seed threshold	Skip seeds with more than INT occurrences (-c).
Drop chain threshold	Drop chains shorter than FLOAT fraction of the longest overlapping chain (-D).
Rounds of mate rescues	Perform at most INT rounds of mate rescues for each read (-m).
Skip mate rescue	Skip mate rescue (-S).
Skip pairing	Skip pairing; mate rescue performed unless -S also in use (-P).
Mismatch penalty	Score for a sequence match (-A).
Mismatch penalty	Penalty for a mismatch (-B).

Gap open penalty	Gap open penalty (-O).
Gap extention penalty	Gap extension penalty; a gap of size k cost {-O} (-E).
Penalty for clipping	Penalty for clipping (-L).
Penalty unpaired	Penalty for an unpaired read pair (-U).
Score threshold	Minimum score to output (-T).

4. Post-processing: On this page you can modify post-processing parameters.



Raw ChIP-Seq Data Processing Wizard

Post-processing

Filtration

MAPQ threshold: 1

Skip flag: The read is unmapped

Region:

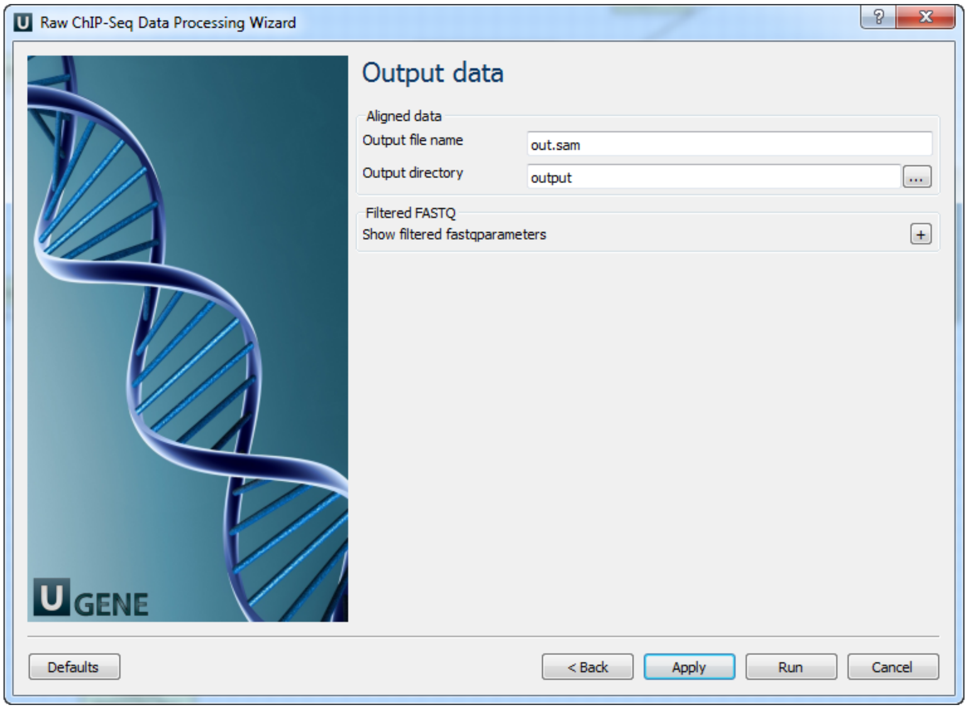
Remove duplicates

For single-end reads: True

The following parameters are available:

MAPQ threshold	Minimum MAPQ quality score.
Skip flag	Skip alignment with the selected items. Select the items in the combobox to configure bit flag. Do not select the items to avoid filtration by this parameter.
Region	Regions to filter. For BAM output only. chr2 to output the whole chr2. chr2:1000 to output regions of chr 2 starting from 1000. chr2:1000-2000 to ouput regions of chr2 between 1000 and 2000 including the end point. To input multiple regions use the space seprator (e.g. chr1 chr2 chr3:1000-2000).
For single-end reads	Remove duplicates for single-end reads.

5. Output data: On this page you must input output parameters.



Raw ChIP-Seq Data Processing Wizard

Output data

Aligned data

Output file name: out.sam

Output directory: output

Filtered FASTQ

Show filtered fastqparameters

UGENE
Defaults
< Back
Apply
Run
Cancel