
Qualimap Documentation

Release 1.0

Fernando Garcia-Alcalde, et al

March 27, 2012

CONTENTS

1	Introduction	1
1.1	What is Qualimap?	1
1.2	Installation	1
1.3	Requirements	1
1.4	Installing Qualimap on Ubuntu	2
2	Workflow	5
2.1	Starting a new analysis	5
2.2	Viewing the results of the analysis	6
2.3	Exporting results	6
2.4	Using tools	6
3	Analysis types	7
3.1	Genomic	7
3.2	RNA-seq	8
3.3	Epigenomic (Clustering)	9
3.4	Compute counts	10
4	Command line interface	13
4.1	General description	13
4.2	Genomic	13
4.3	RNA-seq	14
4.4	Epigenomic	14
4.5	Compute counts	15

INTRODUCTION

1.1 What is Qualimap?

Qualimap is a platform-independent application written in Java and R that provides both a Graphical User Interface (GUI) and a command-line interface to facilitate the quality control of alignment sequencing data. The aim of Qualimap is to provide an easy way for the quality control of mapping data, considering the sequence features and their genomic properties.

Basically, the application accepts and examines sequence alignment data, summarizing some interesting issues found in such data. The main features offered by Qualimap are: fast analysis across the reference genome of mapping coverage and nucleotide relative content; easy-to-interpret summary of the main properties of the alignment data; analysis of the reads mapped inside/outside of the regions defined in an annotation reference; analysis of the adequacy of the sequencing depth in RNA-seq experiments and clustering of !HERE!.

1.2 Installation

Download the ZIP archive with Qualimap from the [Qualimap web page](#).

Unpack it to desired directory.

Now you can run Qualimap from this directory using the prebuilt script:

```
‘./qualimap’
```

Qualimap was tested on GNU Linux, MacOS and MS Windows. !Revise Windows!

Note: On MS Windows use script qualimap.bat to launch Qualimap.

1.3 Requirements

Qualimap requires

- JAVA runtime (6 or above)
- R enviroment (2.14 or above)

The JAVA runtime can be downloaded from the [official web-site](#). There are prebuilt binaries available for many platforms.

R enviroment can be downloaded from [R project web-site](#).

Several Qualimap features are implemented in R, using a number of external packages.

Note: If R enviroment is not availble, “Epigenetics” and “RNA-seq” features will be disabled.

Currently Qualimap requires the following R-packages:

- `optparse` (available from [CRAN](#))
- `Repitools`, `Rsamtools`, `GenomicFeatures`, `rtracklayer` (available from [Bioconductor](#))

One can install these packages manually or use the script from Qualimap distribution.

Once R environment is available the installation script can be invoked from Qualimap installation folder:

```
'Rscript scripts/installDependencies.r'
```

Note: In general the installation of R packages is platform-specific and may require additional effort.

1.4 Installing Qualimap on Ubuntu

This manual is specific for Ubuntu(Debian) Linux distributive, however with slight differences this can be applied for others Unix systems.

1.4.1 Install JAVA

It is possible to use `openjdk`:

```
'sudo apt-get install openjdk-6-jre'
```

1.4.2 Install R

!Modify! The R latest version can be installed from public repos.

However, the repos must be added to the sources. Open `sources.list`:

```
'sudo gedit /etc/apt/sources.list'
```

Add the following lines:

```
'deb http://<my.favorite.cran.mirror>/bin/linux/ubuntu <name.of.your.distribution>/'
```

Then install R:

```
'sudo apt-get update'
```

```
'sudo apt-get install r-base-core'
```

If you don't have the public key for the mirror add it:

```
'gpg --keyserver subkeys.pgp.net --recv-key <required.key>'
```

```
'gpg -a --export <required.key> | sudo apt-key add -'
```

More details available here:

<https://stat.ethz.ch/pipermail/r-help/2009-February/187644.html>

<http://cran.r-project.org/bin/linux/ubuntu/README>

Note: Alternatively it is possible to build R environment directly from sources downloaded from r-project.org.

1.4.3 Install required R-packages

Use special script from Qualimap package:

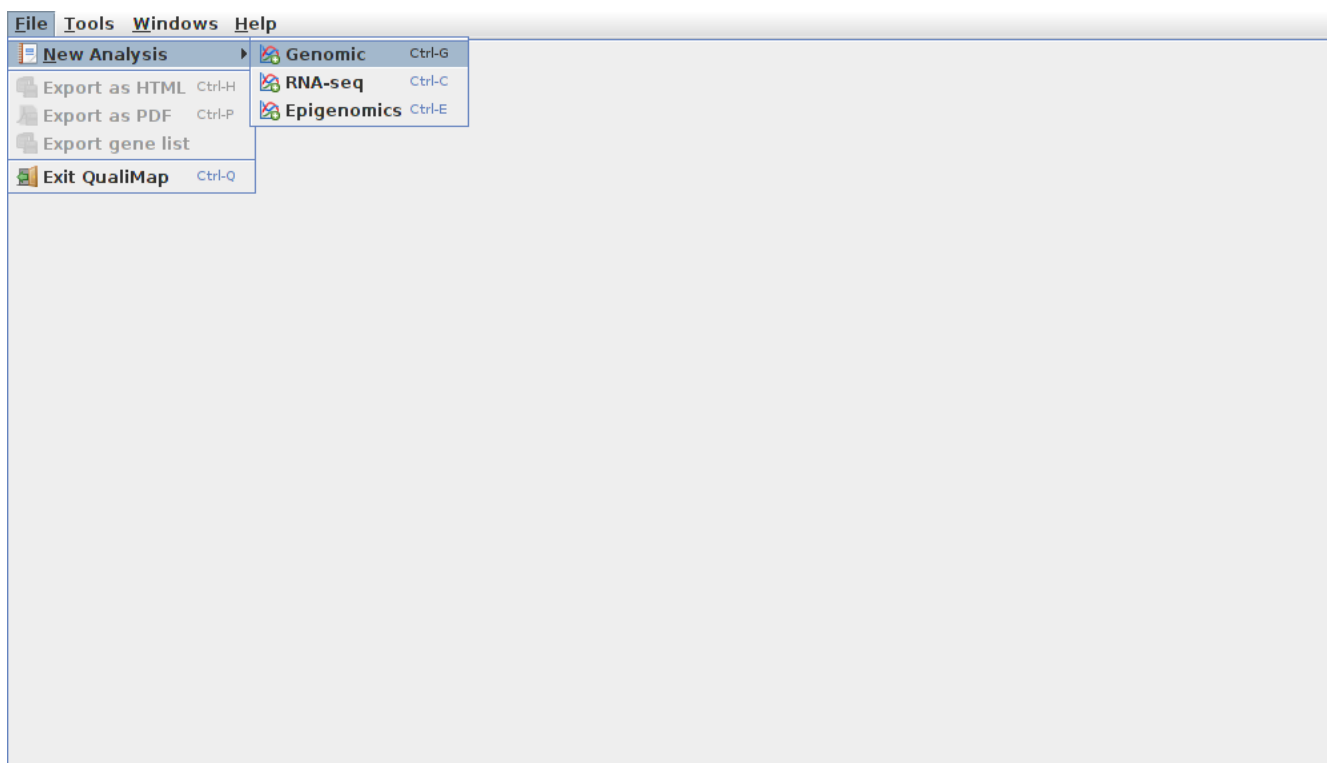
```
'Rscript $QUALIMAP_HOME/scripts/installDependencies.r'
```

where '\$QUALIMAP_HOME' is the full path to the Qualimap installation folder.

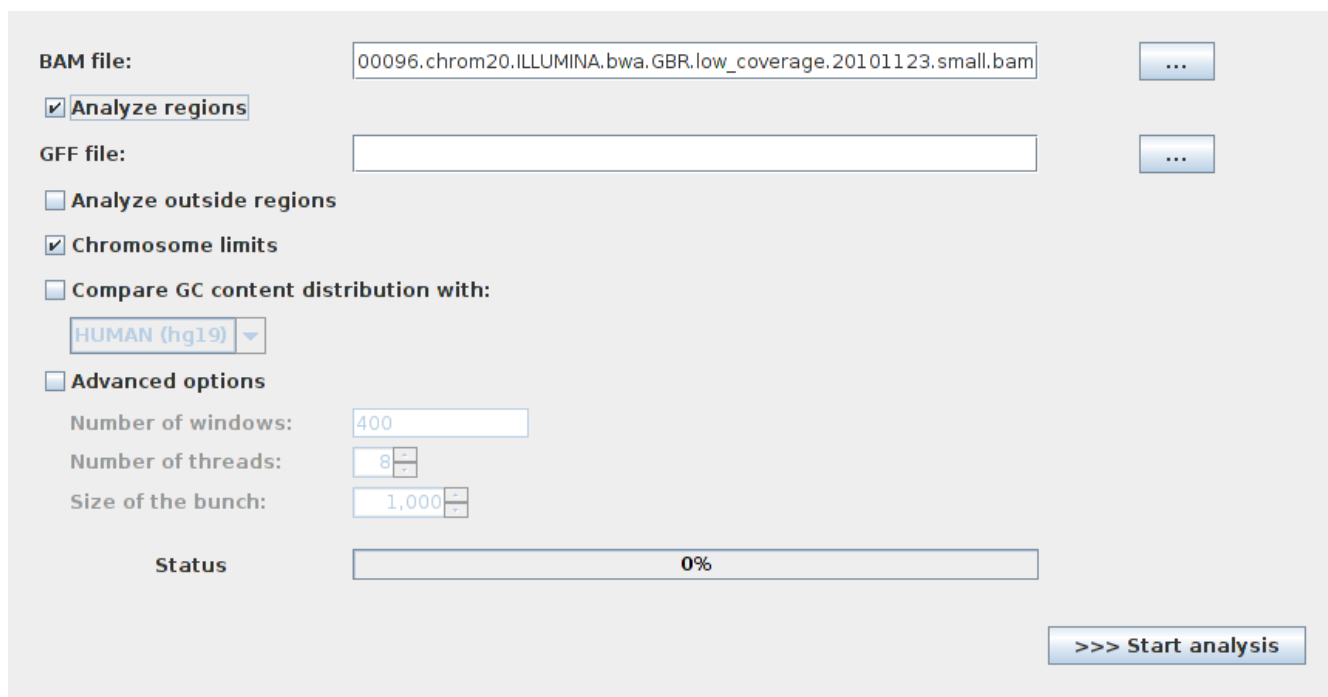
WORKFLOW

2.1 Starting a new analysis

To start new analysis activate main menu item *File* → *New Analysis* and select the desired type of analysis. Read more about different types of analysis [here](#).



After the corresponding item is selected a dialog will appear that allows customizing analysis options (input files, algorithm parameters, etc.).

The image shows a software interface for configuring a genomic analysis. It includes fields for a BAM file (00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20101123.small.bam) and a GFF file. There are checkboxes for 'Analyze regions' (checked), 'Analyze outside regions' (unchecked), 'Chromosome limits' (checked), and 'Compare GC content distribution with:' (unchecked). A dropdown menu shows 'HUMAN (hg19)'. Under 'Advanced options', there are input fields for 'Number of windows' (400), 'Number of threads' (8), and 'Size of the bunch' (1,000). A 'Status' bar shows '0%' progress. A '>>> Start analysis' button is at the bottom right.

BAM file: 00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20101123.small.bam ...

☒ Analyze regions

GFF file: ...

☐ Analyze outside regions

☒ Chromosome limits

☐ Compare GC content distribution with:

HUMAN (hg19) ▼

☐ Advanced options

Number of windows: 400

Number of threads: 8

Size of the bunch: 1,000

Status 0%

>>> Start analysis

To run the analysis click the *Start analysis* button.

During the computation a status message and a graphic bar will indicate the progress of the computation.

2.2 Viewing the results of the analysis

After the selected analysis is finished the results are shown as an interactive report in the Qualimap main window. Several reports can be opened at the same time in different tabs.

In the left part of the report window one can find a list containing available result items. Clicking on an item will automatically show the corresponding information report or graph. Some report items are common for different types of analysis.

For example, *Summary* section provides a short summary of performed quality control checks, while *Input* section lists all the input parameters. Further information about each specific result is provided [here](#).

2.3 Exporting results

The resulting report can be saved to HTML page or PDF document.

To export results to HTML use a main menu item *File* → *Export to HTML*. In the appeared window one can select the path to the output folder. After clicking *OK* button the web-page, containing analysis results will be saved to specified directory.

Similarly one can save the report to a PDF document by using a main menu item *File* → *Export to PDF*.

2.4 Using tools

Qualimap provides also additional functionality other than quality control checks. The *Tools* main menu item allows to access this functionality. Currently there is only one tool available – *Compute Counts*.

In future we plan to add more tools.

ANALYSIS TYPES

3.1 Genomic

Genomic analysis provides information for quality evaluation of the alignment data: base statistics summary showing mean GC-content, coverage, mapping quality and also a number of useful graphs, that allow to evaluate the quality of the sequencing data.

3.1.1 Input

BAM file Path to the sequence alignment file in BAM format. Note, that the BAM file has to be sorted. Sorting can be performed with [satmttools](#).

Draw chromosome limits If selected, vertical dotted lines will be placed at the beginning of each chromosome according to the information found in the header of the BAM file.

Analyze regions Activating this option results in analysis of the alignment data only in specified region.

GFF file The path to the annotation file that specifies regions of interest.

Analyze outside regions This option allows to provide the information about the alignment data outside of the regions of interest. Useful for comparison purposes.

Compare GC content distribution with... Option allows to compare calculated distribution with selected precalculated genome distribution. Currently two genome distributions are available: human (hg19) and mouse (mm9)

Advanced parameters

Number of windows Number of windows used to split the selected genome reference. Default is 400.

Number of threads The genomic analysis computation can be performed in parallel on a multicore system using the given number of threads. The default number of threads equals number of available processors.

Reads per chunk To speed up the computation reads are analyzed in chunks. Each chunk is analyzed by single thread. This option controls the number of reads in a chunk. Smaller number may result in lower performance, but also the memory consumption will be reduced.

3.1.2 Output

Summary Basic information and statistics for the alignment sequencing input. For example, number of reads, number of mapped reads, mean coverage, chromosome-based statistics, etc.

Input data & parameters Section provides the information about selected input parameters.

Coverage across reference This plot consists of two figures. The upper figure provides the coverage (red color) and coverage deviation across the reference sequence. The lower figure shows gc-content across reference (black color) and its deviation.

Coverage Histogram Frequency histogram of the coverage.

Coverage histogram (0-50X) There is often big picks of coverage across the reference and the scale of the Coverage Histogram graph scale may not be adequate. In order to solve this, in this graph genome locations with a coverage greater than 50X are grouped into the last bin.

Coverage quota Provides an easy way of viewing how much reference has been sequenced with a coverage higher than a selected level.

Mapped Reads Nucleotide Content This plot demonstrates the nucleotide content per read position.

Mapped Reads GC Content Distribution This graph shows the distribution of GC content per read. When compared with genome distribution this plot allows to check if there is a shift in the GC content.

Duplication Rate Histogram The histogram shows how many reads start at unique position. This plot is helpful to see if the fragment distribution across genome.

Mapping quality across reference Plot provides mapping quality across reference as indicated by the parameters.

Mapping quality histogram Histogram of the mapping quality frequency

3.2 RNA-seq

In RNA-seq experiments, the reads are mapped to a reference genome. If the total amount of sequencing reads is enough, the number of those reads mapping to a certain biological feature of interest (gene, transcript, exon, ...) is an estimation of the abundance of that feature in the sample and can be used as the quantification of its expression level.

These count data are usually utilized to assess differential expression between two or more experimental conditions. But before computing differential expression, users should be aware of some potential limitations of the RNA-seq data they are working with, as for example: has the saturation been reached or could more features be detected by increasing sequencing depth? Which kind of features are being detected in the experiment? how good is the quantification of expression in the sample? All of these questions are easily answered by looking at the plots generated by Qualimap.

To study the quality of a sample from the count data in a RNA-seq experiment, please use the RNA-seq option from the Analysis menu.

For this option to work, the R language must be installed in the user's computer along with the R library "optparse" (both are freely available from <http://cran.r-project.org/>).

3.2.1 Input

First sample (counts) File containing the count data from sample. Count data must be provided in a tab-delimited txt file, with the features names or IDs in the first column and counts in the second column. This file must not contain any header or column names. One can calculate the counts from a GFF file and a BAM file using option *Tools* → *Compute Counts*

First sample name Name for sample 1 to appear in plots legends

Second sample (counts) Optional. If a second sample is available, this file should contain the same information as in *First sample*, but for the second sample. Mark the *Compare with other sample* box to enable this option.

Second sample name Name for second sample to appear in plots legends.

Count threshold A feature is considered as detected if the corresponding number of counts is greater than this count threshold. By default, the threshold value is set to 5 counts.

Info File Optional. File containing the biological classification of features in the count files. The info file must be a tab-delimited txt file, with the features names or IDs in the first column and the biological group (e.g. the biotype field from Biomart in the Ensembl database) in the second column. Again, the file must not contain any header or column names. If this file is provided, further exploratory plots can be generated to evaluate characteristics of the sample such as the kind of features being detected or the counts distribution for detected features. Please, make sure that the features IDs are the same in the *Info file* and in the *Count file*.

Species Optional. If the Info File is not given by the user, Qualimap provides the Ensembl biotype classification for certain species (human and mouse in Qualimap version 1.0), whenever the features names in the counts file are the Ensembl gene or transcripts IDs (e.g. ENSG00000251282 or ENST00000508921). If so, mark the box to enable this option and select the species.

3.2.2 Output

Global Saturation

This plot provides information about the level of saturation in the sample, so it helps the user to decide if more sequencing is needed or if no many more features will detected when increasing the number of reads. These are some tips for the interpretation of the plot:

- The increasing sequencing depth of the sample is represented at the X-axis. The maximum value is the real sequencing depth of the sample(s). Smaller sequencing depths correspond to samples randomly generated from the original sample(s).
- The curve(s) is associated to the left Y-axis and represents the number of detected features when working with each of the sequencing depths in the X-axis. “Detected features” mean features with more than k counts, where k is the Count threshold chosen by the user.
- The bars are associated to the right Y-axis and they represent the number of newly detected features when increasing the sequencing depth in one million reads at each sequencing depth value.

When an *Info File* is provided by the user or chosen from the ones supplied by Qualimap, a series of plots are additionally generated that are described next.

Detection per group This barplot allows the user to know which kind of features are being detected in their sample(s). The X-axis shows all the biological groups included in the Info file (or the biotypes supplied by Qualimap). The grey bar is the percentage of features in each biological group within the reference genome (or transcriptome, etc.). The striped color bar is the percentage detected in the sample with regard to the genome. The solid color bar is the percentage that the group (or biotype) represents in the total detected features in the sample.

Counts per group A boxplot per each group (or biotype) describes the counts distribution for the detected features in that group.

Saturation per group For each group (or biotype), a saturation plot is generated like the one described above.

Counts & Sequencing Depth For each group (or biotype), a plot is generated containing a boxplot with the distribution of counts at each sequencing depth. X-axis shows the increasing sequencing depths of randomly generated samples from the original one till the true sequencing depth is reached. This plot allows the user to see how the increase of sequencing depth is changing the expression level quantification.

3.3 Epigenomic (Clustering)

This analysis type allows to cluster the regions of interest in the input alignment based on the coverage. Typical usecase for such clustering can be evaluation of epigenomics experiment. For example due to epigenetics changes When analyzing promoter regions one should expect some genes to be activated and some silenced.

To perform this evaluation you need to provide alignment data (both methylated and control) and list of transcript ids as BED annotation file.

3.3.1 Input

Experiment ID The experiment name

Alignment data Here you can provide your replicates to analyze. Each replicate includes sample file and a control file. For example, in an epigenomics experiment, the sample file could be the MeDIP-seq data and the control the non-enriched data (the so-called INPUT data). Thus, for each replicate the following information has to be provided:

Replicate name Name of the replicate

Sample file Path to sample BAM file

Control file Path to control BAM file

To add a replicate click *Add* button. To remove a replicate select it and click *Remove* button. You can modify replicate by using *Edit* button.

Regions of interest Path to an annotation file in BED format which contains regions of interest. The BED file should be a tab delimited text file with exactly 6 fields per line:

- chromosome
- start
- end
- name
- score
- strand (+ or -)

Location Relative location to analyze

Left offset Offset in bp upstream the selected regions

Right offset Offset in bp downstream the selected regions

Bin size Can be thought as the resolution of the plot. Bins of the desired size will be computed and the information falling on each bin will be aggregated

Number of clusters Number of groups that you the user wants to divide the data. Several values can be used by separating them with commas

Fragment length Length of the fragments that were initially sequenced. All reads will be enlarged to this length.

Visualization type You can visualize cluster using heatmaps or line-based graphs

3.3.2 Output

After the analysis is performed, the regions of interest are clustered in groups based on the coverage pattern. The output graph shows the coverage pattern for each cluster either as a heatmap or a line graph. There can be multiple graphs based on the number of clusters provided as input. The name of each graph consists of the experiment name and the number of clusters.

It is possible to export list of features belonging to the particular cluster. To do this use main menu item *File* → *Export gene list* or context menu item *Export gene list*. After activating the item a dialog will appear where you can choose some specific cluster. One can either copy the list of features belonging to this cluster in the clipboard or export it to a text file.

3.4 Compute counts

This tool allows to calculate how many reads belong to each region of interest in the alignment. To access the tool use menu item *Tools* → *Compute counts*.

3.4.1 Input

BAM file Path to BAM alignment file

Annotation file Path to GTF file containing regions of interest

Protocol Three options are avalalbe:

non-strand-specific Feature is counted independent of strand

forward-stranded Feature is counted only if it has the same strand as the read

reverse-strand Feature is counted only if the it has the strand reverse to the one of the read

Feature type Third column of the GTF file. Only features of this particular type are counted.

Feature name The name of the feature to be counted.

Output Path to the file which will contain output.

Save computation summary This option controls whether to save overall computation statistics.

COMMAND LINE INTERFACE

4.1 General description

Each analysis type presented in QualiMap GUI is also available as command line tool. The common pattern to launch the tool is the following:

```
qualimap <tool_name> <tool_options>
```

<tool_name> is the name of the desired analysis. This could be: *genomic*, *rna-seq*, *epigenomic* or *counts*.

<tool_options> are specific to each type analysis. If not option is provided for the specific tool a full list of available options will be shown

To show available tools use command:

```
qualimap --help
```

4.2 Genomic

The following command allows to perform genomic analysis:

```
qualimap genomic -bam <arg> [-c] [-gff <arg>] [-home <arg>] [-nr <arg>] [-nt <arg>] [-nw <arg>] [-o,--outside-stats] [-outdir <arg>] [-outformat <arg>]
-bam <arg>                input mapping file
-c,--paint-chromosome-limits  paint chromosome limits inside charts
-gff <arg>                region file (gff format)
-home <arg>                home folder of Qualimap
-nr <arg>                  number of reads in the bunch (advanced)
-nt <arg>                  number of threads (advanced)
-nw <arg>                  number of windows (advanced)
-o,--outside-stats          compute region outside stats (only with -gff option)
-outdir <arg>              output folder
-outformat <arg>           output report format (PDF or HTML, default is HTML)
```

The only required parameter is *bam* – the input mapping file.

If *outdir* is not provided, it will be created automatically in the same folder where BAM file is located.

Detailed explanation of available options can be found [here](#).

Example:

```
./qualimap genomic -bam ~/sample_data/pl.bam -gff ~/sample_data/pl_anns.gff --outside-stats
```

4.3 RNA-seq

To perform RNA-seq analysis use the following command:

```
qualimap rna-seq -d1 <arg> [-d2 <arg>] [-home <arg>] [-i <arg>] [-k <arg>] [-n1 <arg>] [-n2 <arg>]
[-outdir <arg>] [-outformat <arg>] [-s <arg>]
-d1,--data1 <arg>          first file with counts
-d2,--data2 <arg>          second file with counts
-home <arg>                 home folder of Qualimap
-i,--info <arg>            info file
-k,--threshold <arg>       threshold for the number of counts
-n1,--name1 <arg>          name for the first sample
-n2,--name2 <arg>          name for second sample
-outdir <arg>              output folder
-outformat <arg>           output report format (PDF or HTML, default is HTML)
-s,--species <arg>         use default file for the given species [human | mouse]
```

Detailed explanation of available options can be found [here](#).

Example:

```
./qualimap rna-seq -d1 ~/sample_data/counts-kidney.txt -d2 ~/sample_data/counts-liver.txt -s human
```

4.4 Epigenomic

To perform epigenomic analysis use the following command:

```
qualimap epigenomic [-b <arg>] [-c <arg>] -control <arg> [-expr <arg>] [-f <arg>] [-home <arg>]
[-l <arg>] [-name <arg>] [-outdir <arg>] [-outformat <arg>] [-r <arg>] -regions <arg> -sample <arg>
-b,--bin-size <arg>          size of the bin (default is 100)
-c,--clusters <arg>          comma-separated list of cluster sizes
-control <arg>               path to control BAM file
-expr <arg>                  name of the experiment
-f,--fragment-length <arg>   smoothing length of a fragment
-home <arg>                  home folder of Qualimap
-l <arg>                     left offset (default is 2000)
-name <arg>                  name of the replicate
-outdir <arg>                output folder
-outformat <arg>             output report format (PDF or HTML, default is HTML)
-r <arg>                     right offset (default is 500)
-regions <arg>               path to regions file
-sample <arg>                path to sample BAM file
-viz <arg>                   visualization type: heatmap or line
```

Detailed explanation of available options can be found [here](#).

Example:

```
./qualimap epigenomic -sample ~/sample_data/24h-i-medip.bam -control ~/sample_data/24h-i-control.bam
```

4.5 Compute counts

To compute counts from mapping data use the following command:

```
qualimap counts -bam <arg> [-f <arg>] -gff <arg> [-home <arg>] [-p <arg>]
-bam <arg>                mapping file in BAM format)
-f,--output <arg>        path to output file
-gff <arg>                region file in GFF format
-home <arg>              home folder of Qualimap
-p,--protocol <arg>      forward-stranded, reverse-stranded or non-strand-specific
```

Detailed explanation of available options can be found [here](#).

Example:

```
./qualimap counts -bam ~/sample_data/pl.bam -gff ~/sample_data/pl_anns.bam
```