Qualimap Documentation

Release 2.2-dev

F. Garcia-Alcalde, K. Okonechnikov, et al

January 19, 2016

CONTENTS

CHAPTER

ONE

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. Shortly, Qualimap:

1. Examines sequencing **alignment data** according to the features of the mapped reads and their **genomic properties**
2. Povides an **overall view** of the data that helps to to the **detect biases** in the sequencing and/or mapping of the data and eases **decision-making** for further analysis.

The main features offered by Qualimap are:

- fast analysis across the reference genome of mapping coverage and nucleotide distribution;
- 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;
- computation and analysis of read counts obtained from intersting of read alignments with genomic features;
- analysis of the adequacy of the sequencing depth in RNA-seq experiments;
- support for multi-sample comparison for alignment data and counts data;
- clustering of epigenomic profiles.

1.2 Installation

Download the ZIP file from the Qualimap web page.

Unpack it to desired directory.

Run Qualimap from this directory using the prebuilt script:

```
./qualimap
```

Qualimap was tested on GNU Linux and MacOS.

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

1.3 Requirements

Qualimap requires:

- JAVA runtime version 6 or above.
- R enviroment version 3.1 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.

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

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

Note: If R environment is not available or required R-packages are missing, "Counts QC" and "Clustering" features will be disabled.

Currently Qualimap requires the following R-packages:

- optparse (available from CRAN)
- NOISeq, Repitools, Rsamtools, GenomicFeatures, rtracklayer (available from Bioconductor)

One can install these packages manually or by executing the script found in the installation folder:

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

1.4 Installing Qualimap on Ubuntu

This manual is specific for Ubuntu(Debian) Linux distribution, however with slight differences this can be applied for other GNU Linux systems.

1.4.1 Install JAVA

It is possible to use openjdk:

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

1.4.2 Install R

The R latest version can be installed from public repos.

The repos must be added to the sources file. Open sources.list:

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

Add the following line:

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

List of cran mirrors can be found here

Here is an example for Ubuntu 10.04 (Lucid):

```
deb http://cran.stat.ucla.edu/bin/linux/ubuntu lucid/
```

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:

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

Qualimap needs R version 3.1 or above. This can be checked with the following command:

```
Rscript --version
```

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

1.4.3 Install required R-packages

Some packages depend on external libraries, so you might need to install them either:

```
sudo apt-get install libxml2-dev
sudo apt-get install libcurl4-openssl-dev
```

You can install required packages manually or use special script from Qualimap installation folder:

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

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

1.5 Citing Qualimap

If you use Qualimap 2 for your research, please cite the following:

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data." Bioinformatics, btv566

The first version of the tool was described in the following manuscript:

García-Alcalde, et al. "Qualimap: evaluating next generation sequencing alignment data." Bioinformatics(2012) 28 (20): 2678-2679

CHAPTER

TWO

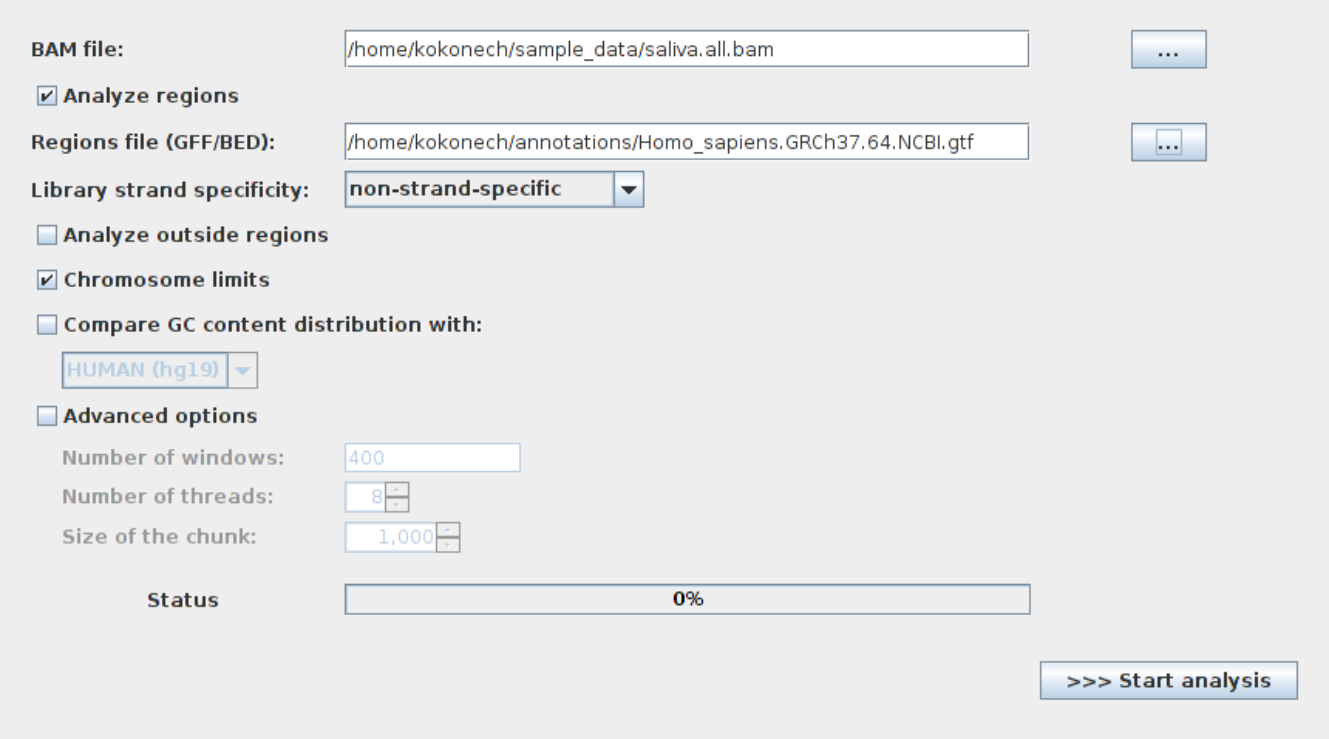
WORKFLOW

This chapter describes how to perform QC analysis of alignment data with graphical user interface of Qualimap. To run analysis with command line interface please refer to the *corresponding chapter*.

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*.

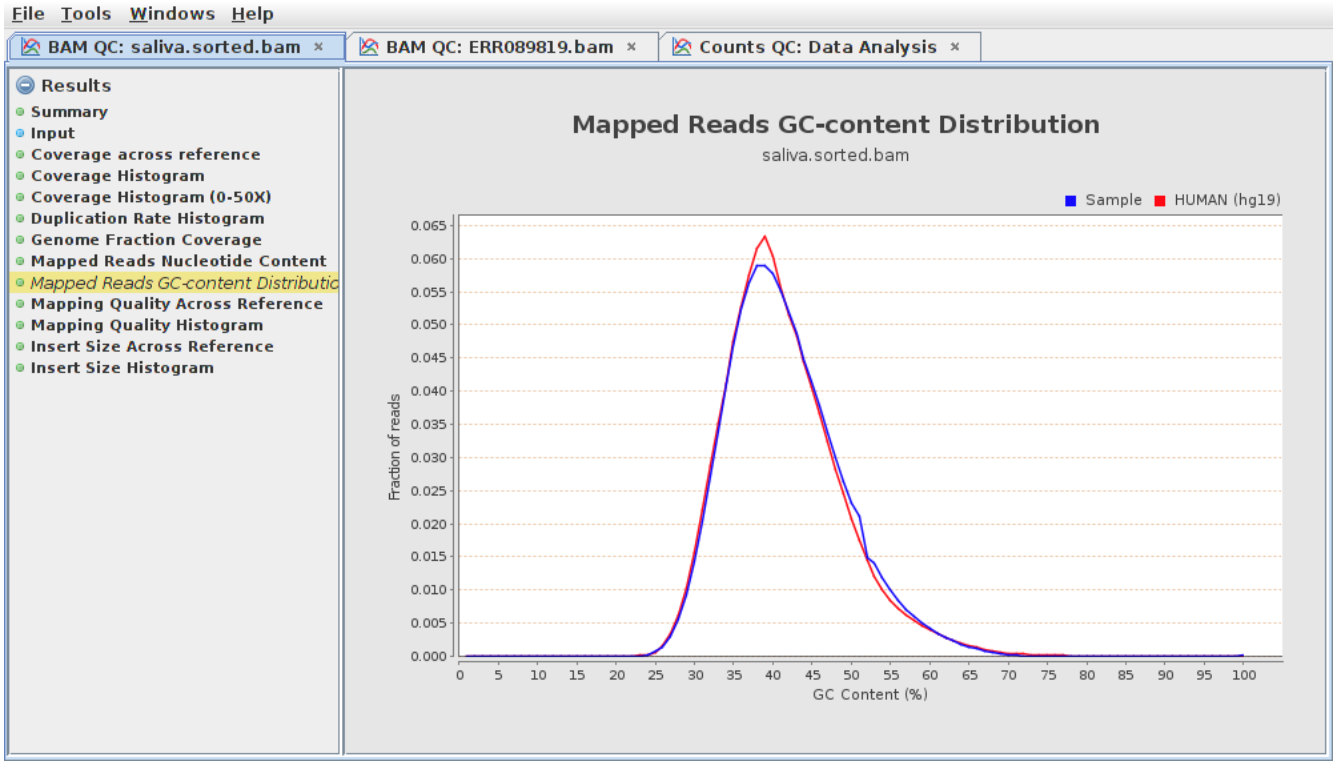
- After the corresponding item is selected a dialog will appear that allows customizing analysis options (input files, algorithm parameters, etc.).



- 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, the *Summary* section provides a short summary of performed quality control checks, while the *Input* section lists all the input parameters. Further information about each specific result is provided *here*.

2.3 Exporting results

- The resulting report along with raw statistics data 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 dialog window one can select the output folder. After clicking *OK* button the web-page, containing analysis results along with raw statistics data will be saved to the specified directory.
- Similarly one can save the report to a PDF document by using a main menu item *File* → *Export to PDF*.
- Note that for plots in *BAM QC* and *Counts QC* it is also possible to export the underlying raw data using the context menu, with appears by clicking the right mouse button in the corresponding plot. In addition, when the report is exported to HTML, the raw data for all plots can be found in the output folder.

2.4 Using tools

- Qualimap is desgined to provide NGS-related tools that can be used aside from the quality control analysis. Currently two tools are available (more are planned to be added in the future):
  1. *Compute Counts* for counting how many reads are mapped to each region of interest at the desired level (genes, transcripts, etc.)
  2. *Clustering* for obtaining groups of genomic features that share similar coverage profiles

CHAPTER

THREE

ANALYSIS TYPES

3.1 BAM QC

BAM QC reports information for the evaluation of the quality of the provided alignment data (a BAM file). In short, the basic statistics of the alignment (number of reads, coverage, GC-content, etc.) are summarized and a number of useful graphs are produced. This analysis can be performed with any kind of sequencing data, e.g. whole-genome sequencing, exome sequencing, RNA-seq, ChIP-seq, etc.

In addition, it is possible to provide an annotation file so the results are computed for the reads mapping inside (and optionally outside) of the corresponding genomic regions, which can be especially useful for evaluating *target-enrichment* sequencing studies.

To start a new BAM QC analysis activate main menu item *File* → *New Analysis* → *BAM QC*.

3.1.1 Examples

- Whole-genome sequencing: HG00096.chrom20.bam. HTML report for sample alignment file from 1000 Genomes project.
- Whole-genome sequencing: ERRR089819.bam. PDF report created using the whole-genome sequencing data of *Caenorhabditis elegans* from the following study.
- See the *Sample data* section for more details about the data used in the examples.

3.1.2 Input Parameters

BAM file Path to the sequence alignment file in **BAM format**. Note, that the BAM file has to be **sorted by chromosomal coordinates**. Sorting can be performed with samtools sort.

Analyze regions Activating this option allows the analysis of the alignment data for the **regions of interest**.

Regions file(GFF/BED file) The path to the annotation file that defines the regions of interest. The file must be **tab-separated** and have GFF/GTF or BED format.

Note: A typical problem when working with human genome annotations is the inconsistency between chromosome names due to "chr" prefix. For example, Ensemble annotations do not include this prefix, while UCSC annotations do. This can become a problem when asscociating regions file with the BAM alignment. Qualimap handles this problem: if the reference sequence of a region has "chr" prefix, it tries to search for sequence name with prefix and without prefix.

Library strand specificity

The sequencing protocol strand specificity: *non-strand-specific*, *forward-stranded* or *reverse-stranded*. This information is required to calculate the number of **correct strand** reads.

Analyze outside regions If checked, the information about the **reads** that are **mapped outside** of the regions of interest will be also computed and shown in a separate section.

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.

Compare GC content distribution with This allows to **compare** the **GC distribution** of the sample with the selected pre-calculated **genome** GC distribution. Currently two genome distributions are available: human (hg19) and mouse (mm9). More species will be included in future releases.

Detect overlapping paired-end reads In case of small insert size the paired-end read alignmetns might overlap in high proportion. Using this option detection of overlapping pairs can be activated. Additionally, adapted mean coverage is calcualted based on extraction of pair overlap-region.

Skip duplicates This option allows to skip duplicate alignments from analysis. There are three modes of this option. By default, the duplicates are skipped only if they are flagged in BAM file and remaining alignments are futher analyzed by Qualimap. Additionally it is possible to skip only the duplicates detected by Qualimap method (based on duplication rate estimation) or aplly both approaches. Number of skipped duplicates will be shown in the report.

Advanced parameters

Number of windows Number of **windows** used to **split** the reference **genome**. This value is used for computing the graphs that plot information across the reference. Basically, reads falling in the same window are aggregated in the same bin. The higher the number, the bigger the resolution of the plots but also longer time will be used to process the data. By default 400 windows are used.

Homopolymer size Only homopolymers of this size or larger will be considered when estimating homopolymer indels count.

Number of threads In order to speed up the computation, the BAM QC analysis **computation** can be performed **in parallel** on a multicore system using the given number of threads. More information on the parallelization of qualimap can be found in *FAQ*. The default number of threads equals number of available processors.

Size of the chunk In order to **reduce the load of I/O**, reads are analyzed in chunks. Each chunk contains the selected number of reads which will be loaded into memory and analyzed by a single thread. Smaller numbers may result in lower performance, but also the memory consumption will be reduced. The default value is 1000 reads.

3.1.3 Output

Summary

Basic information and statistics for the alignment data. The following sections are available:

Globals

This section contains information about the total number of reads, number of mapped reads, paired-end mapping performance, read length distribution, number of clipped reads and duplication rate (estimated from the start positions of read alignments).

ACGT Content

Nucleotide content and GC percentage in the mapped reads.

Coverage

Mean and standard deviation of the coverage depth.

Mapping quality

Mean mapping quality of the mapped reads.

Insert size

Mean, standard deviation and percentiles of the insert size distribution if applicable. The features are computed based on the TLEN field of the SAM file.

Mismatches and indels

The section reports general alignment error rate (computed as a ratio of total collected edit distance to the number of mapped bases), total number of mismatches and total number of indels (computed from the CIGAR values). Additionally fraction of the homopolymer indels among total indels is provided. Note, the error rate and mismatches metrics are based on optional fields of a SAM record (**NM** for edit distance, **MD** for mismatches). The features are not reported if these fields are missing in the SAM file.

Chromosome stats

Number of mapped bases, mean and standard deviation of the coverage depth for each chromosome as defined by the header of the SAM file.

For region-based analysis the information is given inside of regions, including some additional information like, for example, number of correct strand reads.

Input

Here one can check the **input data** and the **parameters** used for the analysis.

Coverage Across Reference

This plot consists of two figures. The upper figure provides the **coverage distribution** (red line) and coverage deviation across the reference sequence. The coverage is measured in X [1]. The lower figure shows **GC content** across reference (black line) together with its average value (red dotted line).

Coverage Histogram

Histogram of the number of **genomic locations** having a given **coverage rate**. The bins of the x-axis are conveniently scaled by aggregating some coverage values in order to produce a representative histogram also in presence of the usual NGS peaks of coverage.

Coverage Histogram (0-50X)

Histogram of the number of **genomic locations** having a given **coverage rate**. In this graph genome locations with a coverage greater than **50X** are grouped into the last bin. By doing so a higher resolution of the most common values for the coverage rate is obtained.

Genome Fraction Coverage

Provides a visual way of knowing how much **reference** has been **sequenced** with **at least** a given **coverage rate**. This graph should be interpreted as in this example:

If one aims a coverage rate of **at least 25X** (x-axis), how much of reference (y-axis) will be considered? The answer to this question in the case of the whole-genome sequencing provided example is **~83%**.

Duplication Rate Histogram

This plot shows the **distribution** of **duplicated** read **starts**. Due to several factors (e.g. amount of starting material, sample preparation, etc) it is possible that the same **fragments** are **sequenced several times**. For some experiments where enrichment is used (e.g. ChIP-seq) this is expected at some *low* rate. If most of the reads share the exact same genomic positions there is very likely an associated bias.

Mapped Reads Nucleotide Content

This plot shows the **nucleotide content per position** of the **mapped reads**.

Mapped Reads GC Content Distribution

This graph shows the distribution of **GC content** per **mapped read**. If compared with a precomputed *genome distribution*, this plot allows to check if there is a shift in the GC content.

[1] Example for the meaning of X: If one genomic region has a coverage of 10X, it means that, on average, 10 different reads are mapped to each nucleotide of the region.

Mapped Reads Clipping Profile

Represents the percentage of clipped bases across the reads. The clipping is detected via SAM format CIGAR codes 'H' (hard clipping) and 'S' (soft clipping). In addition, the total number of clipped reads can be found in the report Summary. The plot is not shown if there are no clipped-reads are found. Total number of clipped reads can be found in *Summary*. Example.

Homopolymer Indels

This bar plot shows separately the number of indels that are within a **homopolymer** of A's, C's, G's or T's together with the number of **indels** that are not within a homopolymer. Large numbers of homopolymer indels may indicate a problem in a sequencing process. An indel is considered homopolymeric if it is found within a homopolymer (defined as at least 5 equal consecutive bases). Owing to the fact that Qualimap works directly from BAM files (and not from reference genomes), we make use of the CIGAR code from the corresponding read for this task. Indel statistics cam be found in a dedicated section of the Summary report.

This chart is not shown if the sample doesn't contain any indels.

Mapping Quality Across Reference

This plot provides the **mapping quality** distribution **across the reference**. To construct the plot mean mapping quality is computed for each window.

Mapping Quality Histogram

Histogram of the number of **genomic locations** having a given **mapping quality**. To construct the histogram mean mapping quality is computed at each genome position with non-zero coverage and collected. According to Specification of the SAM format the range for the mapping quality is [0-255].

Insert Size Across Reference

This plot provides the **insert size** distribution **across the reference**. Insert size is collected from the SAM alignment field TLEN. Only positive values are taken into account. To construct the plot mean insert size is computed for each window.

Insert Size Histogram

Histogram of **insert size** distribution. To construct the histogram all collected insert size values are applied.

3.2 RNA-seq QC

RNA-seq QC reports quality control metrics and bias estimations which are specific for whole transcriptome sequencing, including reads genomic origin, junction analysis, transcript coverage and 5'-3' bias computation. This analysis could be applied as a complementary tool together with *BAM QC* and additionally to produce gene counts for further analysis with *Counts QC*.

To start a new RNA-seq QC analysis activate main menu item *File → New Analysis → RNA-seq QC*.

3.2.1 Examples

- RNA-seq QC report. This report was produced using the RNA-seq alignment of *Homo sapiens* kidney sample [Marioni] and Ensembl v.64 GTF file.
- These data can be downloaded from *here*.

3.2.2 Input parameters

BAM file Path to the sequence alignment file in **BAM** format, produced by a splicing-aware aligner similar to Tophat.

GTF file Genomic annotations in Ensembl **GTF** format. The corresponding annotations can be downloaded from the Ensembl website.

Note: Only annotations in GTF format are supported for this analysis mode. GTF annotations allow to reconstruct the exon structure of transcripts to compute the coverage. For simple region-based analysis please use BAM QC.

Library protocol The strand-specficity of the sequencing library. By default non-strand specific library is assumed.

Paired-end analysis This option activates counting of pair fragments instead of counting of single reads. Only valid for paired-end sequencing experiments.

Alignment sorted by name The paired-end analysis requires the BAM file to be sorted by name. If the BAM file is already sorted by name, then this option should be check, otherwise temporary BAM sorted by name will be created.

Output counts If checked, the gene counts will be saved to a specified file.

Path to counts Path to the output file with the computed counts.

Advanced parameters

Multi-mapped reads Select method to count reads that are mapped to several genome locations. By default only **uniquely-mapped-reads** are used to compute counts. However, it is possible to include multimapped reads by activating **proprtional** method. More details *here*.

3.2.3 Output

Summary

The summary contains the following sections:

Reads alignment

The assignment of read counts per-category:

- total number of mapped reads
- total number of alignments
- number of secondary alignments (duplicates are marked as SAM flag)
- number of non-unique alignments (SAM format "NH" tag of a read is more than one)
- number of reads aligned to genes
- number of ambiguous alignments (belong to several genes, ignored during counting procedure)
- number of alignments without any feature (intronic and intergenic)
- number of unmapped reads.

Transcript coverage profile

The profile provides ratios between mean coverage at the 5' region, the 3' region and the whole transcript. The 5' bias is the ratio between mean coverage at the 5' region and the whole transcript, while the 3' bias is the ratio between mean coverage at the 3' region and the whole transcript. 5'-3' bias is the ratio between both biases.

To compute these values for each transcript mean coverage along with mean coverage in first 100 bp (5' region) and last 100 bp (3'region) are calculated and collected. Afterwards, the collected values are sorted and median is selected from each array to compute the ratios.

Reads genomic origin

Shows how many alignments fall into exonic, intronic and intergenic regions along with a number of intronic/intergenic alignments overlapping exons. Exonic region includes 5'UTR,protein coding region and 3'UTR region.

Junction analysis

Total number of reads with splice junctions and 10 most frequent junction rates.

Input

Here one can check the **input data** and the **parameters** used for the analysis.

Reads Genomic Origin

Pie chart showing how many of read alignments fall into exonic, intronic and intergenic regions.

Coverage Profile (Total)

The plot shows mean coverage profile of the transcripts. All transcripts with non-zero coverage are used to calculate this plot.

Coverage Profile (Low)

The plot shows mean coverage profile of 500 lowest-expressed genes.

Coverage Profile (Total)

The plot shows mean coverage profile of 500 highest-expressed genes.

Coverage Histogram (0-50x)

Coverage of transcripts from 0 to 50X. If certain genes have higher coverage level they are added to the last column (50X).

Junction Analysis

This pie chart shows analysis of junction positions in spliced alignments. **Known** category represents percentage of alignments where both junction sides are known. **Partly known** represents alignments where only one junction side is known. All other alignments with junctions are marked as **Novel**.

3.3 Counts QC

In **RNA-seq** experiments, the reads are usually **first mapped** to a reference genome. It is assumed that if the **number of reads** mapping to a certain biological feature of interest (gene, transcript, exon, ...) is sufficient, it can be used as an **estimation** of the **abundance** of that feature in the sample and interpreted as the quantification of the **expression level** of the corresponding region.

These **count data** can be utilized for example to assess differential expression between two or more experimental conditions. Before assessing differential expression analysis, researchers should be aware of some potential **limitations** of RNA-seq data, as for example: Has the **saturation** been reached or more features could be detected by increasing the sequencing depth? Which **type of features** are being detected in the experiment? How good is the **quantification** of expression in the sample? All of these questions are answered by interpreting the plots generated by Counts QC.

Starting from **version 2.0** Counts QC module has been redisigned to work with **multiple samples** under different conditions. The new functionality is based on NOISeq package, therefore to use Counts QC it is required to have **R** language along with **NOISeq** and **optparse** packages installed.

To run this analysis activate from the main menu *File → New Analysis → Counts QC*.

Note: If count data need to be generated, one can use the provided tool *Compute counts*.

3.3.1 Example

- RNA-seq counts analysis from 2 experiments can be found *here*
- Sample counts data can be downloaded from *here*.

3.3.2 Input Parameters

Samples

The input samples can be added using button *Add*.

For each input sample it is required to provide the following information:

- **Sample name**. Name of the analyzed sample as it will be used as a legend in the plots.
- **Path** to the input file containing the counts data for the sample. This must be a **tab-delimited** file with at least **two columns**. First column of the file must contain feature IDs, while other columns should contain counts for features. Rows starting with # symbol and empty lines are ignored.
- **Data column index**. By default it is assumed that the counts are contained in the second column of the input file. However if the input file contains counts for multiple samples it is possible to define the column corresponding for the sample.
- **Condition index**. If comparison of conditions is activated, this index defines under which condition was the input sample.

Each added sample will be shown in **Samples** table. One can edit samples using button *Edit* and remove them using button *Remove*.

Counts threshold

In order to **remove** the influence of **spurious reads**, a feature is considered as detected if its corresponding number of counts is **greater than this threshold**. By default, the threshold value is set to 5 counts, meaning that features having less than 5 counts will not be taken into account.

Compare conditions

This option allows to compare groups of samples under different conditions. The name of a specific condition can be given using field *Condition name*.

Note: Currently Qualimap allows to compare samples under two conditions. More conditions will be supported in future versions.

Include feature classification

Optional. This option enables analysis of distribution of counts among feature groups defined by the biotype. In addition GC-content and length bias will be estimated.

Species

For convinience, Qualimap provides the Ensembl annotations for certain species (currently *Human* and *Mouse*). In order to use these annotations, **Ensembl Gene IDs** should be used as the feature IDs on the **count files** (e.g. ENSG00000251282). If this is true, mark the box to enable this option and select the corresponding species. More annotations and species will be made available in future releases.

Info File

File containing annotations of the features of the count files. It must be a **four-column tab-delimited** text file, with the features names or IDs in the first column, the group (e.g. the biotype from Ensembl database) in the second column, feature length in the third column and feature GC-content in the last column (see human.ens68.txt for an example). Please, make sure that the **features IDs** on this file are the same in the **count files**.

Note: To generate info file based on an arbitrary GTF annotations and genome FASTA file, one can use the following Python script available from Qualimap repo.

3.3.3 Output

Many of plots in Counts QC mode are created using NOISeq package. The NOISeq vignette contains a lot of useful information about the plots and how to interpret them. Here we provide short explanation of the plots.

Global Plots

Plots from this report present a global overview of the counts data and include all samples.

Counts Density

This plot shows density of counts computed from the histogram of log-transformed counts. In order to avoid infinite values in case of zero counts the transformation *log2(expr + 0.5)* is applied, where *expr* is a number of read counts for a given feature. Only log-transformed counts having value greater than 1 are plotted.

Scatterplot Matrix

The panel shows a scatterplot along with smoothed line (lower panel) and Pearson correlation coefficients (upper panel) for each pair of samples. Plots are generated using log-transformed counts.

Saturation

This plot provides information about the level of saturation in the samples, so it helps the user to decide if more sequencing is needed and more features could be 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 curves are associated to the left *y*-axis. They represent the number of detected features at each of the sequencing depths in the *x*-axis. By "detected features" we refer to features with more than *k* counts, where *k* is the *Count threshold* selected by the user.
- The bars are associated to the right *y*-axis. They represent the number of newly detected features when increasing the sequencing depth in one million reads at each sequencing depth value.

Counts Distribution

This box plot shows the global distribution of counts in each sample.

Features With Low Counts

This plot shows the proportion of features with low counts in the samples. Such features are usually less reliable and could be filtered out. In this plot, the bars show the percentage of features within each sample having more than 0 counts per million (CPM), or more than 1, 2, 5 and 10 CPM.

Individual Sample Plots

Apart from global overview there are plots generated individually for each sample.

Saturation

For each sample, a saturation plot is generated like the one described in *Global Saturation*.

When a **Info File** is provided by the user or annotations are chosen from those supplied by Qualimap, additional series of plots are generated:

Bio Detection

This barplot allows the user to know which kind of features are being detected his sample(s). The *x*-axis shows all the groups included in the annotations file. The gray bars are the percentage of features of each group within the reference genome (or transcriptome, etc.). The striped color bars are the percentages of features of each group detected in the sample with regard to the genome. The solid color bars are the percentages that each group represents in the total detected features in the sample.

Counts Per Biotype

A boxplot per each group describes the counts distribution in the given biotype.

Length Bias

The plot describes the relationship between the length of the features and the expression values. The length is divided into bins. Mean expression of features falling into a particular length interval is computed and plotted. A cubic spline regression model is fitted to explain the relation between length and expression. Coefficient of determination R^2 and p-value are shown together with regression curve.

GC Bias

The plot describes the relantionship between the GC-content of the features and the expression values. The data for the plot is generated similar to *Length Bias* plot. The GC content divided into beans and then mean expression of features corresponding to given GC interval are computed. The relation between GC-content and expression is investigated using cubic spline regression model.

Comparison Plots

When **Compare conditions** option is selected, additional plots comparing data in groups of samples having the same biological condition or treatment are available.

Counts Distribution

The plot is similar to the one in *Global* report. It compares distributions of **mean** counts across conditions.

Features With Low Counts

The plot is similar to the one in *Global* report. It compares proportions of features with low counts using **mean** counts across conditions.

Bio Detection

The plot is similar to the one in *Indvidual Sample Plots* report. It compares distribution of the detected features for the given biotype for **mean** counts across conditions.

Length Bias

The plot is similar to the one in *Individual Sample Plots* report. It analyzes relation between feature length and expression across conditions.

GC Bias

The plot is similar to the one in *Individual Sample Plots* report. It analyzes realtion between GC-content and expression across conditions.

3.4 Multi-sample BAM QC

Very often in genomics one has to work with multiple samples, which could represent sequencing results from either biological replicates or different conditions. For example, to reliably detect significant mutations from sequencing data in cancer it is required to analyze tens or even hundreds of samples from matched normal-tumor

data. When performing such large scale experiments it is always important to know if all samples belonging to a specific group pass the quality controls. To detect possible outliers one can compare results of *BAM QC analysis* performed on each individual sample.

QualiMap provides an automated solution for this task. Basically, the QC metrics computed in *BAM QC analysis* are combined together for all samples. Additionally **Principal Component Analysis** is performed to analyze variability and detect outliers.

Note: Starting from version 2.2 it is possible to assign groups marking biolobical or technical conditions of the samples.

One can apply multi-sample analysis for precomputed results of QualiMap BAM QC or directly for raw BAM files. In latter case firstly BAM QC analysis will be performed for each input file and then multi-sample analysis will be executed.

To start a new multi-sample BAM QC analysis activate main menu item *File* → *New Analysis* → *Multisample BAM QC*.

3.4.1 Examples

- gh2ax chip-seq data: 12 samples: example report for a ChIP-seq experiment having 12 samples.
- gh2ax chip-seq data: 4 conditions, 3 replicates per condition: example report for the same ChIP-seq experiment with 4 biological conditions marked. Each condition group includes 3 samples.

See the *Sample data* section for more details about the data used in the example.

3.4.2 Input Parameters

There are 2 types of input data that are accepted by *Multi-sample BAM QC*:

1. By default directory with the summary statistics and plot data produced by BAM QC analysis is expected as input data for multi-sample comparison.
2. If a special **"raw data" mode** is activated, then BAM files can be provided as input. In this case Qualimap will first run the *BAM QC analysis* on each indvidual BAM file, and then multi-sample report will be computed.

The input samples can be added using button *Add*. For each sample one has to provide the following information:

1. **Name** of the sample as it will be used in legend.
2. **Path** to the folder with which contains results of BAM QC analysis performed on the sample. The folder must include file **genome_results.txt** and subfolder **raw_data_qualimapReport** containing data of BAM QC plots. If **"Raw data" mode** is activated then the path to the BAM file should be provided.
3. **Group** of the sample. This option allows to combine the samples of the same condition. After the group is assigned, the samples in the plots belonging to the group will have the same color. Importantly, if the groups are avaialble, they should be provided for **each sample**. Empty value will mean no group.

Note: In QualiMap version <= 2.0 directory with raw data of BAM QC analysis was called **raw_data**. This name is also supported.

Each added sample will be shown in **Samples** table. One can edit samples using button *Edit* and remove them using button *Remove*.

Additionally it is possible to import configuration file, that is applied for command line interface using button *Import configuration...*. The configuration file is explained in *the overview of the command line mode*.

"Raw data" mode: run BAM QC on input samples

Activate this checkbox to analyze BAM files directly. A selected set of options is available to customize *BAM QC* process. One can read detailed explantion of these options in a *corresponding section* of the manual.

To start the analysis click button *Run analysis*.

3.4.3 Output

Summary

The summary table contains comparison of selected critical alignment metrics for all samples. The metrics include mean and standard deviation of coverage, mean GC content, mean insert size and mean mapping qualities. If the sample groups are provided, they are also shown for each sample.

Input

Here one can check the **input data** and the **parameters** used for the analysis.

PCA

The alignment features presented in the *Summary* section undergo Principal Component Analysis. Afterwards the biplot presenting first and second principal component is constructed. The plot shows how much variability demonstarte the analyzed samples. It allows to detect if any samples group together and if there are any outliers among analyzed samples.

Coverage Across Reference, *Coverage Histogram (0-50X)* , *Genome Fraction Coverage*, *Duplication Rate Histogram*, *Mapped Reads GC Content*, *Mapped Reads GC Content Distribution*, *Mapped Reads Clipping Profile*, *Mapping Quality Across Reference*, *Mapping Quality Histogram*, *Insert Size Across Reference*, *Insert Size Histogram*

The following plots demonstrate the comparison of samples using data from corresponding plots computed during BAM QC analysis. Each curve on a plot represents a single sample.

Please refer to documentation of *BAM QC* for detailed information about the plots.

***

CHAPTER

FOUR

TOOLS

4.1 Compute counts

- Given a BAM file and an annotation (GTF file), this tool calculates how many reads are mapped to each region of interest.
- The user can decide:
  - At which level wants to perform the counting (genes, transcripts...).
  - What to do whith reads mapped to multiple locations.
  - Paired-end reads status and strand-specifity.
  - When a transcriptome GTF file is provided the tool allows to calculate 5' and 3' prime coverage bias.

To access the tool use *Tools* → *Compute counts*.

4.1.1 Example

- Input data:
  - BAM file: liver.bam. RNA-seq of liver tissue from Marioni JC et al
  - GTF file: human.64.gtf . Human annotation from Ensembl (v. 64)
  - Parameters:
    - Feature ID: gene_id (to count at the level of genes)
    - Feature type: exon (to ignore other features like start/end codons)
    - Paired-end reads counts computation and strand-specificity
    - Multimapped reads: uniquely-mapped-reads (to ignore not unique alignments)
- Output:
  - liver.counts. Two-column tab-delimited text file, with the feature IDs in the first column and the number of counts in the second column.

4.1.2 Input

BAM file Path to the BAM alignment file.

Annotation file Path to the GTF or BED file containing regions of interest.

Protocol

Controls when to consider reads and features to be overlapping:

non-strand-specific Reads overlap features if they share genomic regions regardless of the strand.

forward-stranded For single-end reads, the read and the feature must have the same strand to be overlapping. For paired-end reads, the first read of the pair must be mapped to the same strand as the feature, while the second read must be mapped to the opposite strand.

reverse-strand For single-end reads, the read and the feature must have the opposite strand. For paired-end reads, the first read of pair must be mapped to the opposite strand of the feature, while the second read of the pair must be on the same strand as the feature.

Feature ID The user can select the attribute of the GTF file to be used as the feature ID. Regions with the same ID will be aggregated as part of the same feature. The application preload the first 1000 lines of the file so a list with possible feature IDs is conveniently provided.

Feature type The user can select the feature type (value of the third column of the GTF) considered for counting. Other types will be ignored. The application preload the first 1000 lines of the file so a list with possible feature IDs is conveniently provided.

Paired-end reads This option allows to activate counting of pairs of reads instead of single reads

Alignment sorted by name For correct analysis of paired-end reads alignment should be sorted by name. If this operation is already performed, sorting can be skipped.

Output Path to the ouput file.

Save computation summary This option controls whether to save overall computation statistics. If selected, the statistics will be saved in a file named *$INPUT_BAM*.counts

Multi-mapped reads This option controls what to do whith reads mapped to multiple location:

uniquely-mapped-reads Reads mapped to multiple locations will be ignored.

proportional Multi-mapped reads are detected based on "NH" tag from SAM format. Each read is weighted according to the number of mapped locations. For example, a read mapped to 4 different locations will add 0.25 to the counts of each location. After analysis is finished the value will converted to intger value.

Calculate 5' and 3' coverage bias If a **GTF file** is provided, the user has the possibility of computing **5' - 3' bias**. The application automatically constructs the 5' and 3' UTR (100 bp) from the gene definitions of the GTF file and determines the coverage rate of the 1000 most highly expressed transcripts in the UTR regions. This information is then stored in the *computation summary* file, together with the statistics of the counting procedure.

Note: This option requires a standard gene model definition. The UTRs are computed for the first and last exons of each transcript. Therefore, *exon* is the feature of interest (third field of the GTF) and *gene_id*, *transcript_id* should be attributes (ninth field of the GTF).

4.1.3 Output

A two-column tab-delimited text file, with the feature IDs in the first column and the number of counts in the second column, and overall calculation stats.

The calculation stats include:

Feature counts Number of reads assigned to various features

No feature Number of reads not aligned to any feature

Not unique alignment Number of reads with non-unique alignment

Ambiguous Number of reads that align to features ambigously

The following stats are calculate only if option *Calulate 5' and 3' bias* was set:

Median 5' bias For 1000 most expressed genes the ratio between coverage of 100 leftmost bases and mean coverage is calcualted and median value is provided.

Median 3' bias For 1000 most expressed gene the ratio between coverage of 100 rightmost bases and mean coverage is calculated and median value is provided.

Median 5' to 3 For 1000 most expressed genes the ratio between coverag of 100 leftmost and 100 rightmost bases is calculated and median value is provided.

4.2 Clustering

- Qualimap provides the possibility of clustering genomic features according to their surrounding coverage profiles. This is particulary interesting in epigenomic studies (e.g. methylation). The user can import a set of features (e.g. TSSs or CpG Islands) together with the BAM file. Then the application preprocess the data and clusters the profiles using the Repitools package (Statham et al). The obtained groups of features are displayed as a heatmap or as line graphs and can be exported for further analysis (e.g. for measuring the correlation between promoter methylation and gene expression).
- Summary of the process:
  - filter out the non-uniquely-mapped reads
  - compute the smoothed coverages values of the samples at the desired locations
  - apply k-means on the smoothed coverage values for the desired values of k
- To perform this analysis the user needs to provide at least two BAM files – one for the sample (enriched) and other for the control (input) – and a list of features as BED file.
- Clustering analysis can be accesed using the menu item *File* → *Tools* → *Clustering*.

Note: Clustering coverage profiles is not a straightforward task and it may be necessary to perform a number of empirical filter steps. In order to correctly interpret the approach the results we encourage the users to read Repitools User Manual.

4.2.1 Input Parameters

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 or GFF format, which contains regions of interest.

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.

4.2.2 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 beloning 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.

CHAPTER

FIVE

COMMAND LINE INTERFACE

5.1 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: *bamqc*, *rnaseq*, *multi-bamqc*, *counts*, *clustering* or *comp-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

---

Note: If you are using Qualimap on Unix server without X11 system, make sure that the DISPLAY environment variable is unset. Otherwise this might result in problems when running Qualimap. *Here* is an instruction how to solve this issue.

To show available tools use command:

```
qualimap --help
```

There are certain options that are common to most of the command line tools:

```
-outdir <arg>                          Output folder for HTML report and raw
                                        data.
-outfile <arg>                         Output file for PDF report (default value
                                       is report.pdf).
-outformat <arg>                       Format of the ouput report (PDF or HTML,
                                        default is HTML).
```

These options allow to confugre output of Qualimap. *-outdir* option sets the output folder for HTML report and raw data:

```
qualimap bamqc -bam file.bam -outdir qualimap_results
```

If the *-outfile* option is given then the output will be produced in PDF format. In this case *-outdir* option controls only the path to raw data.

Example:

```
qualimap bamqc -bam file.bam -outfile result.pdf
```

It is also possible to explictily set output format by using option *-outformat*. In this case report will be saved in the output dir under default name.

Example:

```
qualimap bamqc -bam file.bam -outdir qualimap_results -outformat pdf
```

Additionally each tool has its own defaults for output directory name. Check tools' description for details.

5.2 BAM QC

The following command allows to perform BAM QC analysis:

```
usage: qualimap bamqc -bam <arg> [-c] [-gd <arg>] [-gff <arg>] [-hm <arg>] [-nr
   <arg>] [-nt <arg>] [-nw <arg>] [-oc <arg>] [-os] [-outdir <arg>]
   [-outfile <arg>] [-outformat <arg>] [-p <arg>]
-bam <arg>                              Input mapping file in BAM format
-c,--paint-chromosome-limits            Paint chromosome limits inside charts
-gd,--genome-gc-distr <arg>             Species to compare with genome GC
                                       distribution. Possible values: HUMAN or
                                       MOUSE.
-gff,--feature-file <arg>               Feature file with regions of interest in
                                       GFF/GTF or BED format
-hm <arg>                               Minimum size for a homopolymer to be
                                       considered in indel analysis (default is
                                       3)
-ip,--collect-overlap-pairs             Activate this option to collect statistics
                                       of overlapping paired-end reads
-nr <arg>                               Number of reads analyzed in a chunk
                                       (default is 1000)
-nt <arg>                               Number of threads (default is 8)
-nw <arg>                               Number of windows (default is 400)
-oc,--output-genome-coverage <arg>      File to save per base non-zero coverage.
                                       Warning: large files are  expected for large
                                       genomes
-os,--outside-stats                     Report information for the regions outside
                                       those defined by feature-file  (ignored
                                       when -gff option is not set)
-outdir <arg>                           Output folder for HTML report and raw
                                       data.
-outfile <arg>                          Output file for PDF report (default value
                                       is report.pdf).
-outformat <arg>                        Format of the ouput report (PDF or HTML,
                                       default is HTML).
-p,--sequencing-protocol <arg>          Sequencing library protocol:
                                       strand-specific-forward,
                                       strand-specific-reverse or
                                       non-strand-specific (default)
-sd,--skip-duplicated                   Activate this option to skip duplicate
                                       alignments from the analysis. If the
                                       duplicates are not flagged in BAM file,
                                       then they will be detected by Qualimap.
-sdmode,--skip-dup-mode <arg>           Specific type of duplicated alignments to
                                       skip (if this option is activated).
                                       0 : only flagged duplicates (default)
                                       1 : only estimated by Qualimap
                                       2 : both flagged and estimated
```

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 (data available *here*):

```
qualimap bamqc -bam ERR089819.bam -c
```

5.3 RNA-seq QC

To perform RNA-seq QC analysis use the following command:

```
usage: qualimap rnaseq [-a <arg>] -bam <arg> -gtf <arg> [-oc <arg>] [-outdir
      <arg>] [-outfile <arg>] [-outformat <arg>] [-p <arg>]
-a,--algorithm <arg>             Counting algorithm:
                                 uniquely-mapped-reads(default) or
                                 proportional.
-bam <arg>                       Input mapping file in BAM format.
-gtf <arg>                       Annotations file in Ensembl GTF format.
-oc <arg>                        Path to output computed counts.
-outdir <arg>                    Output folder for HTML report and raw data.
-outfile <arg>                   Output file for PDF report (default value is
                                 report.pdf).
-outformat <arg>                 Format of the ouput report (PDF or HTML,
                                 default is HTML).
-p,--sequencing-protocol <arg>   Sequencing library protocol:
                                 strand-specific-forward,
                                 strand-specific-reverse or non-strand-specific
                                 (default)
-pe,--paired                     Setting this flag for paired-end experiments
                                 will result in counting fragments instead of
                                 reads.
-s,--sorted                      This flag indicates that the input file is
                                 already sorted by name. If not set, additional
                                 sorting by name will be performed. Only
                                 required for paired-end analysis.
```

The required parameteres for this type of analysis are the spliced-alignment file in BAM format and annotations in GTF format.

Detailed explanation of available options can be found *here*.

Example (data available *here*):

```
qualimap rnaseq -bam kidney.bam -gtf human.64.gtf -outdir rnaseq_qc_results
```

5.4 Multi-sample BAM QC

To perform multi-sample BAM QC use the following command:

```
usage: qualimap multi-bamqc [-c] -d <arg> [-gff <arg>] [-hm <arg>] [-nr <arg>]
       [-nw <arg>] [-outdir <arg>] [-outfile <arg>] [-outformat <arg>] [-r]
-c,--paint-chromosome-limits   Only for -r mode. Paint chromosome limits inside
                               charts
-d,--data <arg>                File describing the input data. Format of the
                               file is a 2-column tab-delimited table.
                               Column 1: sample name
                               Column 2: either path to the BAM QC result or
                               path to BAM file (-r mode)
-gff,--feature-file <arg>      Only for -r mode. Feature file with regions of
                               interest in GFF/GTF or BED format
-hm <arg>                      Only for -r mode. Minimum size for a homopolymer
                               to be considered in indel analysis (default is
                               3)
-nr <arg>                      Only for -r mode. Number of reads analyzed in a
                               chunk (default is 1000)
-nw <arg>                      Only for -r mode. Number of windows (default is
                               400)
-outdir <arg>                  Output folder for HTML report and raw data.
-outfile <arg>                 Output file for PDF report (default value is
                               report.pdf).
-outformat <arg>               Format of the ouput report (PDF or HTML, default
                               is HTML).
-r,--run-bamqc                 Raw BAM files are provided as input. If this
                               option is activated BAM QC process first will be
                               run for each sample, then multi-sample analysis
                               will be performed.
```

The main argument for this command is the configuration file describing input data (-d). This has to be a 2- or 3-column tab-delimted file. The first column should contain the sample name and the second column should contain either path to the results of BAM QC analysis or path to the BAM file (if -r mode is activated). The path for the data could be absolute or relative to the location of the configuration file. Additionally the third optional column can provide the condition of the sample. This is an optional column. However, if conditions are available they should be provided for each sample.

Here's an example of configuration file describing samples with conditions:

```
sample_1 sample_1_stats group_1
sample_2 sample_2_stats group_1
sample_3 sample_3_stats group_1
sample_4 sample_4_stats group_2
sample_5 sample_5_stats group_2
sample_6 sample_6_stats group_2
```

Detailed explanation of the analysis can be found here *here*.

Example (data available *here*):

```
unzip gh2ax_chip_seq.zip
cd gh2ax_chip_seq.txt
qualimap multi-bamqc -i gh2ax_chip_seq.txt -outdir gh2ax_multibamqc
```

5.5 Counts QC

To perform counts QC analysis use the following command:

```
usage: qualimap counts [-c] -d <arg> [-i <arg>] [-k <arg>] [-outdir <arg>]
      [-outfile <arg>] [-outformat <arg>] [-R <arg>] [-s <arg>]
-c,--compare             Perform comparison of conditions. Currently 2 maximum
                         is possible.
-d,--data <arg>          File describing the input data. Format of the file is
                         a 4 column tab-delimited table.
                         Column 1: sample name
                         Column 2: condition of the sample
                         Column 3: path to the counts data for the sample
                         Column 4: index of the column with counts
-i,--info <arg>          Path to info file containing genes GC-content, length
                         and type.
-k,--threshold <arg>     Threshold for the number of counts
-outdir <arg>            Output folder for HTML report and raw data.
-outfile <arg>           Output file for PDF report (default value is
                         report.pdf).
-outformat <arg>         Format of the ouput report (PDF or HTML, default is
                         HTML).
-R,--rscriptpath <arg>   Path to Rscript executable (by default it is assumed
                         to be available from system $PATH)
-s,--species <arg>       Use built-in info file for the given species: HUMAN or
                         MOUSE.
```

The main argument for this command is the configuration file describing the input samples (-d). This has to be a 4-column tab-delimited file. The first column should contain the name of the sample, the second - name of the biological condition (e.g treated or untreated), the third - path to the file containing counts data for the sample and the fourth - the index of the column in the data file which contains counts. This is useful when counts for all samples are contained in the one file, but in different columns.

Detailed explanation of the analysis can be found *here*.

Example. Note: requires counts file mouse_counts_ensembl.txt (data available *here*):

```
qualimap counts -d GlcN_countsqc_input.txt -c -s mouse -outdir glcn_mice_counts
```

5.6 Clustering

To perform clustering of epigenomic signals use the following command:

```
usage: qualimap clustering [-b <arg>] [-c <arg>] -control <arg> [-expr <arg>]
       [-f <arg>] [-l <arg>] [-name <arg>] [-outdir <arg>] [-outformat <arg>]
       [-r <arg>] -regions <arg> -sample <arg> [-viz <arg>]
 -b,--bin-size <arg>          size of the bin (default is 100)
 -c,--clusters <arg>          comma-separated list of cluster sizes
 -control <arg>               comma-separated list of control BAM files
 -expr <arg>                  name of the experiment
 -f,--fragment-length <arg>   smoothing length of a fragment
 -l <arg>                     upstream offset (default is 2000)
 -name <arg>                  comma-separated names of the replicates
 -outdir <arg>                output folder
 -outformat <arg>             output report format (PDF or HTML, default is
                              HTML)
```

```
 -r <arg>                      downstream offset (default is 500)
 -regions <arg>                path to regions file
 -sample <arg>                 comma-separated list of sample BAM files
 -viz <arg>                    visualization type: heatmap or line
```

Detailed explanation of available options can be found *here*.

Example (data available *here*):

```
qualimap clustering -sample clustering/hmeDIP.bam -control clustering/input.bam -regions annotati
```

5.7 Compute counts

To compute counts from mapping data use the following command:

```
usage: qualimap comp-counts [-a <arg>] -bam <arg> -gtf <arg> [-id <arg>] [-out
       <arg>] [-p <arg>] [-pe] [-s <arg>] [-type <arg>]
-a,--algorithm <arg>              Counting algorithm:
                                  uniquely-mapped-reads(default) or proportional
-bam <arg>                        Mapping file in BAM format
-gtf <arg>                        Region file in GTF, GFF or BED format. If GTF
                                  format is provided, counting is based on
                                  attributes, otherwise based on feature name
-id <arg>                         GTF-specific. Attribute of the GTF to be used
                                  as feature ID. Regions with the same ID will
                                  be aggregated as part of the same feature.
                                  Default: gene_id.
-out <arg>                        Path to output file
-p,--sequencing-protocol <arg>    Sequencing library protocol:
                                  strand-specific-forward,
                                  strand-specific-reverse or non-strand-specific
                                  (default)
-pe,--paired                      Setting this flag for paired-end experiments
                                  will result in counting fragments instead of
                                  reads
-s,--sorted <arg>                 This flag indicates that the input file is
                                  already sorted by name. If not set, additional
                                  sorting by name will be performed. Only
                                  required for paired-end analysis.
-type <arg>                       GTF-specific. Value of the third column of the
                                  GTF considered for counting. Other types will
                                  be ignored. Default: exon
```

Detailed explanation of available options can be found *here*.

Example (data available *here*):

```
qualimap comp-counts -bam kidney.bam -gtf ../annotations/human.64.gtf  -out kidney.counts
```

CHAPTER

SIX

EXAMPLES

6.1 Sample Data

6.1.1 Alignments

- **ERR089819.bam (2.6 GB)** Whole genome sequencing data of C. elegans from the following study.
- **HG00096.chrom20.bam (278 MB)** Sequencing of the chromosome 20 from a H. sapiens sample from 1000 Genomes project. The header of the BAM file was changed in order to contain only chromosome 20. Original file can be found here.
- **kidney.bam (386 MB) and liver.bam (412 MB)** Human RNA-seq sequencing data from from the paper of Marioni JC et al

6.1.2 Annotations

- **human.64.gtf** Human genome annotations from Ensembl database (v. 64).
- **transcripts.human.64.bed** Human transcripts in BED format from Ensembl database (v. 64).

6.1.3 Multisample BAM QC

- gh2ax_chip_seq.zip

  Example dataset from an unpublished ChiP-seq experiment with 4 condtions, each having 3 replicates (12 sampels in total). The archive contains BAM QC results for each sample and input configurations (normal: *input.txt*, with marked conditions: *input.with_groups.txt*) for command line version of Multisample BAM QC.

6.1.4 Counts QC

- **mouse_counts_ensembl.txt** Mouse counts data from a study investigating effects of D-Glucosamine:
- **GlcN_countsqc_input.txt** Command line input configuration for the counts data above.
- **kidney.counts and liver.counts** Counts data from the paper by Marioni JC et al.
- **marioini_countsqc_input.txt** Command line input configuration for the counts data above.

6.1.5 Clustering

- **hmeDIP.bam (988M)** MeDIP-seq of human embryonic stem cells from the study of Stroud H et al.
- **input.bam (1.8G)** Input data of the same study

6.2 Sample Output

6.2.1 BAM QC

Analysis of the WG-seq data (HG00096.chrom20.bam): QualiMap HTML report.

Analysis of the WG-seq data (ERR089819.bam): QualiMap PDF report.

6.2.2 RNA-seq QC

Analysis of RNA-seq data (kidney.bam, human.64.gtf): QualiMap HTML report.

6.2.3 Multisample BAM QC

Multisample analysis of 12 gH2AX ChiP-seq alignments: Qualimap HTML report.

Multisample analysis of the same data with condition assigned for each sample: Qualimap HTML report.

6.2.4 Counts QC

Counts QC HTML reports computed from RNA-seq experiment analyzing influence of D-Glucosamine on mice. The analysis was performed for 6 samples in 2 conditions - GlcN positive and negative (mouse_counts_ensembl.txt):

- Global report
- Comparison of conditions
- Sample 01 (GlcN negative)
- Sample 02 (GlcN negative)
- Sample 03 (GlcN negative)
- Sample 04 (GlcN positive)
- Sample 05 (GlcN positive)
- Sample 06 (GlcN positive)

Counts QC HTML reports from human RNA-seq data from study by Marioni JC et al (kidney.counts, liver.counts):

- Global report
- Comparison of conditions
- Sample 01 (Kidney)
- Sample 02 (Liver)

6.2.5 Clustering

Analysis of MeDIP-seq data: QualiMap HTML report.

CHAPTER

SEVEN

FREQUENTLY ASKED QUESTIONS

7.1 General

Q: *How to cite Qualimap?*

A: If you use Qualimap 2 for your research, please cite the following:

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2015). "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data." Bioinformatics, btv566

Q: *How to increase maximum Java heap memory size?*

A: The Qualimap launching script allows to set desired memory size using special command line argument `--java-mem-size`. Here are some usage examples:

```
qualimap --java-mem-size=1200M
```

```
qualimap bamqc -bam very_large_alignment.bam --java-mem-size=4G
```

Note that there should be **no whitespace** between argument and its value.

Alternatively one can change default memory size parameter by modifying the following line in the launching script:

JAVA_MEM_DEFAULT_SIZE="1200M"

Also one can override this parameter by setting environment variable $JAVA_OPTS.

Q: *Does Qualimap run on MS Windows?*

A: Qualimap can be launched on Windows using script `qualimap.bat`. However, officially we do not support MS Windows.

Q: *Does Qualimap work with R version 3?*

A: Yes, Qualimap works with R v3. There was a bug with R-version recogintion GUI, but starting from version 0.8 the bug was fixed.

Q: *Counts QC* mode launched on MacOS doesn't produce any results. Some of the following errors are reported:

```
unable to load shared object ../libs/cairo.so
```

or

```
libpng warning:  Application built with libpng-1.5.18 but running
with 1.6.17
```

A: Issue is related to the update of R package structure on MacOS. The error can be fixed by updating R to version >= 3.2.2 and installation of a specific library XQuartz. More details can be found in the following discussion.

Q: *I always get a message "Out of Memory". What should I do?*

A: You can try decreasing the number of reads in chunk or increasing *maximum Java heap memory size*.

7.2 Command line

Q: *I launch Qualimap command-line tool on my big and powerful Linux server. However it doesn't finish properly and outputs some strange message like*:

Exception in thread "main" java.lang.InternalError: Can't connect to X11
window server using 'foo:42.0' as the value of the DISPLAY variable.

What is going on?

A: Java virtual machine uses DISPLAY environment variable to detect if the X11 system is available. Sometimes this variab

```
unset DISPLAY
```

or like this: `export DISPLAY=:0`

Additionally it's possible to use a special option of Java **-Djava.awt.headless=true** to disable display requirement.

Enabling this option can be performed by setting **JAVA_OPTS** variable in system or by changing **java_options** variable in the *qualimap* script:

```
java_options="-Djava.awt.headless=true -Xmx$JAVA_MEM_SIZE
-XX:MaxPermSize=1024m"
```

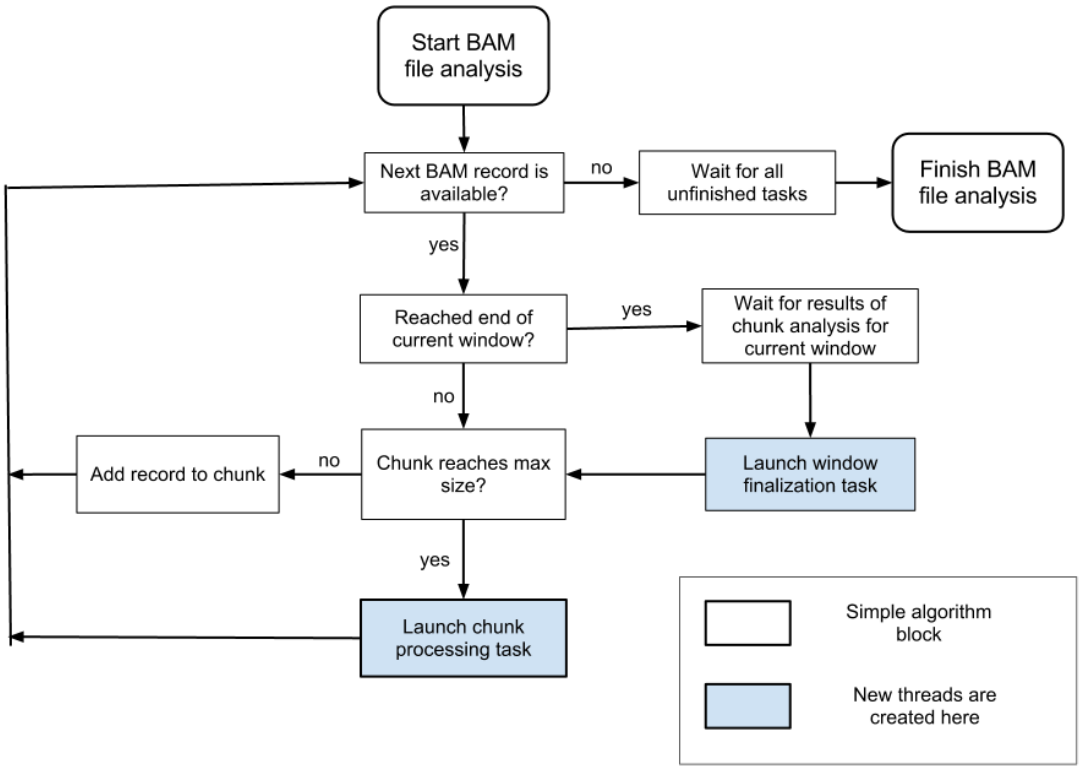
7.3 Performance

Q: *Does Qualimap make use of multicore systems to improve computation speed?*

A: Yes, Qualimap uses threads to perform BAM QC analysis.

In short, reads are processed in chunks and each chunk is analyzed in parallel.

Below you can find a schema, depicting the applied algorithm.



Here each block denotes a certain algorithm step. The analysis starts dividing the reference genome into windows. The first window is set to be the current one. Then the analysis continues processing BAM records belonging to the current window.

When all the reads belonging to the current window are processed, the window is finalized in a newly created thread.

The analysis is finished when all windows are processed.

Q: *What is the scalability of QualiMap? Can it run on a cluster?*

A: Currently qualimap is designed to run in a single multicore machine. In the future we plan to support cluster and computational cloud execution for BAM QC.

Q: *I have a powerful computer with a lot of memory. Can I make Qualimap run faster?*

A: Sure, just increase your *maximum JAVA heap size*.

BIBLIOGRAPHY

[Marioni] Marioni JC et al, “RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays”. Genome Res. 2008. 18: 1509-1517.