Qualimap Documentation

Release 1.0

Fernando Garcia-Alcalde, et al

May 29, 2014

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;
- analysis of the adequacy of the sequencing depth in RNA-seq experiments;
- 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 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.

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)
- 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 2.14 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 for your research, please cite the following:

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

CHAPTER

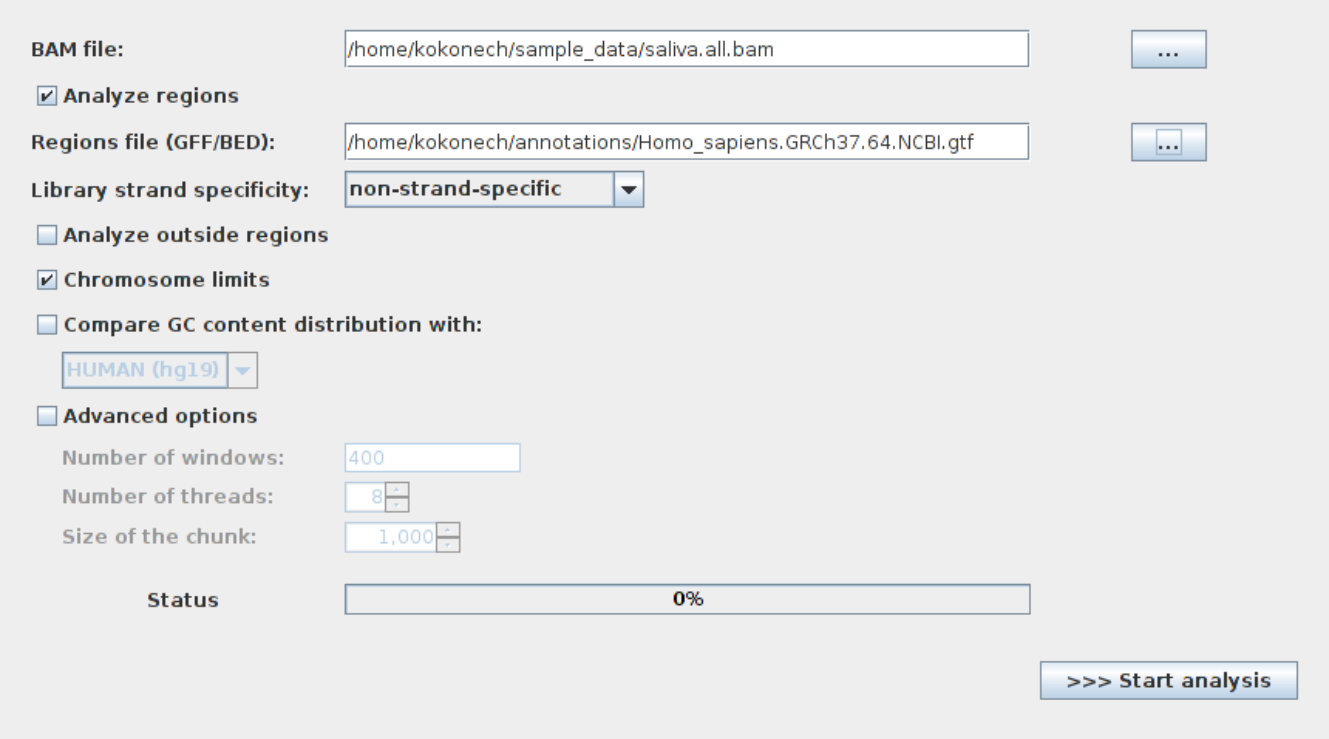
TWO

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*.

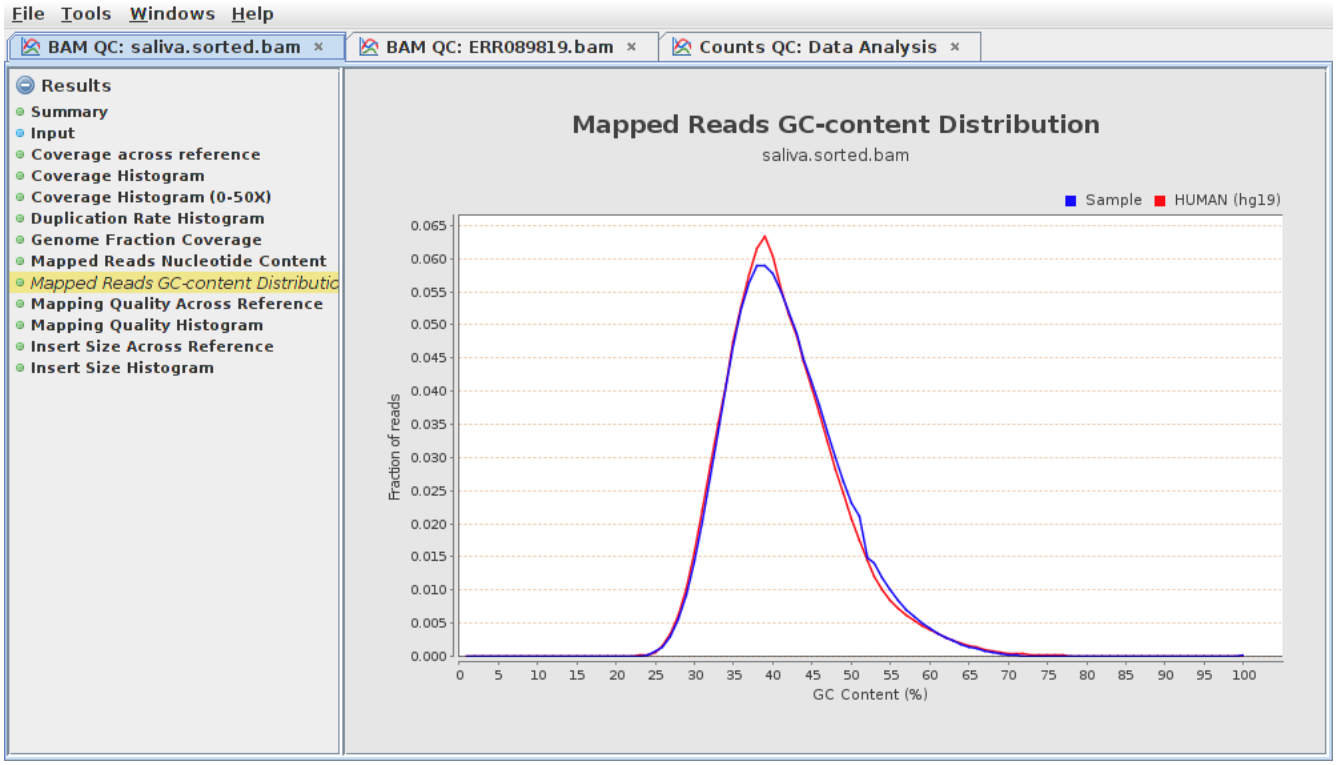
- After the corresponding item is selected a dialog will appear that allows customizing analysis options (input files, algorithm parameters, etc.).



- 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. Report for sample alignment file from 1000 Genomes project.
- Whole-genome sequencing: ERRR089819.bam. 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.

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 by 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%**.

Mapped Reads Nucleotide Content

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

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.

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.

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 report Summary.

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

[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.

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.

Mapping Quality Across Reference

This plot provides the **mapping quality** distribution **across the reference**.

Mapping Quality Histogram

Histogram of the number of **genomic locations** having a given **mapping quality**. According to Specification of the SAM format the range for the mapping quality is [0-255].

3.2 RNA-seq QC

RNA-seq QC reports quality control metrics and bias estimations which are specific for whole transcriptome sequencing, such as for example transcript coverage and 5'-3' bias. This analysis could be applied as 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.

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.

3.2.3 Output

Summary

The summary contains the following sections:

Reads alignment

The assignment of read counts per-category: mapped to genes, ambiguous, without any feature etc.

Transcript coverage profile

The ratios between mean coverage at the 5' region, 3' region and whole transcript.

Junction analysis

Total number of reads with splice junctions and 10 most frequest junctions rate.

Input

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

Coverage Profile

The plot shows mean coverage profile of the transcripts.

Coverage Histogram (0-50x)

Coverage of transcripts from 0 to 50X

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 assesing 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 Qualimap.

For assesing this analysis just 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*.

Note: For this option to work, the **R** language must be **installed** along with the R package **optparse** (both are freely available from http://cran.r-project.org/).

3.3.1 Example

- RNA-seq count data. This report was produced using the counts from the RNA-seq of *Homo sapiens* kidney and liver samples [Marioni].
- These counts can be downloaded from *here* or generated using the *Compute counts* tool.

3.3.2 Input Parameters

First sample (counts)

File containing the count data from the sample. This must be a **two-column tab-delimited** text file, with the feature IDs in the first column and the number of counts in the second column. This file must not contain header nor column names. See *Counts* for examples

First sample name

Name for the first sample that will be used as legend in the plots.

Second sample (counts)

Optional. If a second sample is available, this file should contain the same information as in *First sample* for the second sample, i.e. the same feature IDs (first column) and the corresponding number of counts (second column). Mark the *Compare with other sample* checkbox to enable this option.

Second sample name

Name for the second sample that will be used as legend in the plots.

Count 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 theshold value is set to 5 counts, meaning that features having less than 5 counts will not be taken into account.

Group File

Optional. File containing a classification of the features of the count files. It must be a **two columns tab-delimited** text file, with the features names or IDs in the first column and the group (e.g. the biotype from Ensembl database) in the second column (see human.64.genes.biotypes for an example). Again, the file must not contain any header or column names. If this file is provided, specific plots for each defined group are generated. Please, make sure that the **features IDs** on this file are the same in the **count files**.

Species

Optional. For convinience, Qualimap provides the Ensembl biotype classification [2] 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 so, mark the box to enable this option and select the corresponding species. More annotations and species will be made available in future releases.

3.3.3 Output

Global Plots

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 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.

An example for this plot can be seen here.

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

Samples Correlation

[2] Downloaded from Biomart v.61.

When two samples are provided, this plot determines the **correlation level** between both samples. Due to the often wide range of expression data (counts), a log2-transformation is applied in order to improve the graphical representation. Features not detected in any of the two samples are removed for this analysis. To avoid infinite values in the case of genes with 0 counts in one of the samples, log2(expression + 1) is used. Thus, sample 1 is depicted in X-axis and sample 2 in Y-axis. The colors of the plot should be interpreted as a map. The blue color is the level of the sea and the white color the top of the mountain. Hence, the higher you are over the sea level, the more genes you have in that range of X-Y values. In addition, the title of the plot includes the **Pearson's correlation coefficient**, which indicates if both samples present a linear relationship.

Detection Per Class

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 *Group File* (or the biotypes supplied by Qualimap). The grey 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 Class

A boxplot per each group describes the counts distribution for the detected features in that group.

Individual Group Plots

Saturation per group

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

Counts & Sequencing Depth

For each group, a plot is generated containing a boxplot with the distribution of counts at each sequencing depth. The *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.

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.
  - The 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*.

Note: For paired-end reads currently each mate of a pair is considered independently (taking into account the strand-specificity of the protocol). We will add full support for paired-end reads in future versions of Qualimap.

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)
    - 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.

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 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.

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*, *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
```

5.2 BAM QC

The following command allows to perform BAM QC analysis:

```
usage: qualimap bamqc -bam <arg> [-c] [-gd <arg>] [-gff <arg>] [-nr <arg>] [-nt
       <arg>] [-nw <arg>] [-os] [-outdir <arg>] [-outformat <arg>]
 -bam <arg>                       input mapping file
 -c,--paint-chromosome-limits     paint chromosome limits inside charts
 -gd <arg>                        compare with genome distribution (possible
                                  values: HUMAN or MOUSE)
 -gff <arg>                       region file (in GFF/GTF or BED format)
 -hm <arg>                        minimum size for a homopolymer to be considered
                                  in indel analysis (default is 3)
 -nr <arg>                        number of reads in the chunk (default is 500)
 -nt <arg>                        number of threads (default equals the number of cores)
 -nw <arg>                        number of windows (default is 400)
 -os,--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)
 -p <arg>                         specify protocol to calculate correct strand
                                  reads (works only with -gff option, possible
```

```
                                values are STRAND-SPECIFIC-FORWARD or
                                STRAND-SPECIFIC-REVERSE, default is
                                NON-STRAND-SPECIFIC)
```

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 [-algorithm <arg>] -bam <arg> [-counts <arg>] -gtf <arg>
   [-outdir <arg>] [-outfile <arg>] [-outformat <arg>] [-protocol <arg>] [-rscriptpath <arg>]
-algorithm <arg>     Counting algorithm: uniquely-mapped-reads(default) or proportional
-bam <arg>           Mapping file in BAM format
-counts <arg>        Path to output computed counts
-gtf <arg>           Annotations file in Ensembl GTF format.
-outdir <arg>        Output folder
-outfile <arg>       Output file for PDF report (default value is report.pdf)
-outformat <arg>     Output report format (PDF or HTML, default is HTML)
-protocol <arg>      Library protocol: strand-specific-forward,strand-specific-reverse or non-stra
```

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 Counts QC

To perform counts QC analysis use the following command:

```
usage: qualimap counts -d1 <arg> [-d2 <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
 -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 (data available *here*):

```
qualimap counts -d1 kidney.counts -d2 liver.counts -s human -outdir results
```

5.5 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
```

5.6 Compute counts

To compute counts from mapping data use the following command:

```
usage: qualimap comp-counts [-algorithm <arg>] -bam <arg> -gtf <arg> [-id <arg>]
       [-out <arg>] [-protocol <arg>] [-type <arg>]
 -algorithm <arg>   uniquely-mapped-reads(default) or proportional
 -b                 calculate 5' and 3' coverage bias
 -bam <arg>         mapping file in BAM format)
 -gtf <arg>         region file in GTF format
 -id <arg>          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
 -protocol <arg>    forward-stranded,reverse-stranded or non-strand-specific
 -type <arg>        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.

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 Counts

Human RNA-seq data from the paper of Marioni JC et al.

- Counts:

  kidney.counts and liver.counts
- BAM files used to produce the counts:

  kidney.bam and liver.bam
- Genes Biotypes:

  human.64.genes.biotypes.txt

6.1.4 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 (ERR089819.bam): QualiMap HTML report.

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

6.2.2 Counts QC

Analysis of RNA-seq counts data: QualiMap HTML report.

6.2.3 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 for your research, please cite the following:

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

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: *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`

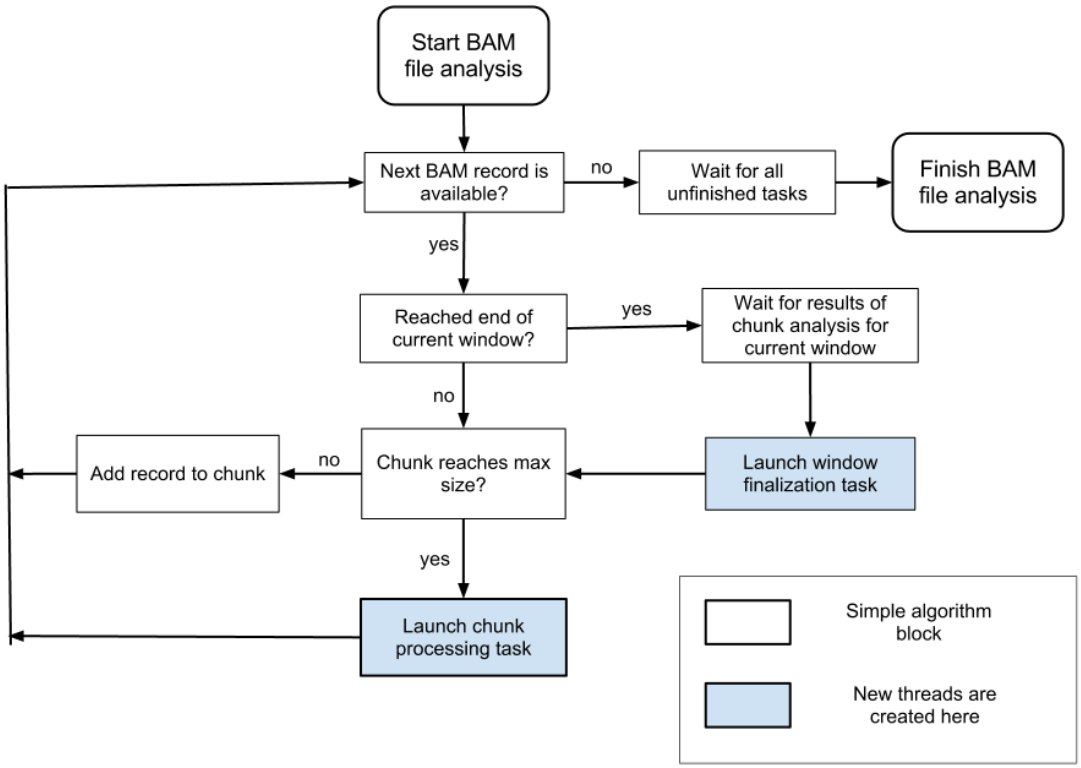
7.3 Performance

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.