Qualimap: evaluating next-generation sequencing alignment data

Fernando García-Alcalde¹,², Konstantin Okonechnikov², José Carbonell¹, Luis M. Cruz¹, Stefan Götz¹, Sonia Tarazona¹, Joaquín Dopazo¹, Thomas F. Meyer² and Ana Conesa¹,*

¹Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, 46012, Valencia, Spain and ²Department of Molecular Biology, Max-Planck Institute for Infection Biology, D-10117, Berlin, Germany

ABSTRACT

Motivation: The sequence alignment/map (SAM) and the binary alignment/map (BAM) formats have become the standard method of representation of nucleotide sequence alignments for next-generation sequencing data. SAM/BAM files usually contain information from tens to hundreds of millions of reads. Often, the sequencing technology, protocol and/or the selected mapping algorithm introduce some unwanted biases in these data. The systematic detection of such biases is a non-trivial task that is crucial to drive appropriate downstream analyses.

Results: We have developed Qualimap, a Java application that supports user-friendly quality control of mapping data, by considering sequence features and their genomic properties. Qualimap takes sequence alignment data and provides graphical and statistical analyses for the evaluation of data. Such quality-control data are vital for highlighting problems in the sequencing and/or mapping processes, which must be addressed prior to further analyses.


Contact: aconesa@cifp.es

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on May 17, 2012; revised on July 25, 2012; accepted on August 7, 2012

1 INTRODUCTION

With the advent of next-generation sequencing (NGS), a number of novel alignment methods have been developed (see Trapnell and Salzberg, 2009 for a review). The sequence alignment/map (SAM) and the binary alignment/map (BAM) formats have become the standards used for representation of nucleotide sequence alignments for these algorithms (Li et al., 2009). The results from such alignments can be used in subsequent analyses, such as genome-wide comparative studies, to drive conclusions concerning a variety of biological processes, such as gene expression and epigenomic modifications. SAM/BAM files usually contain the information from tens to hundreds of millions of reads, and the quantity of data contained in these files is continuously increasing. Unfortunately, SAM/BAM data files frequently contain biases that are introduced by sequencing technologies, during sample preparation, (Harismendy et al., 2009; Metzker, 2009) and/or the selected mapping algorithm (Flick and Birney, 2009). Therefore, one of the fundamental requirements during analysis of these data is to perform quality control, i.e. to get an idea of how reliable mapping data are and how well data fit with the expected outcome. To these ends, we have developed Qualimap, a Java application that aims to facilitate the quality-control analysis of mapping data. Recently, some efforts have been made to facilitate this task, for example see SAMStat (Lassmann et al., 2011), RNA-SeQC (DeLuca et al., 2012) or Picard; Qualimap advances this area by providing some additional features (Supplementary Table S1).

2 THE QUALIMAP APPLICATION

Qualimap is a multi-threaded application built in Java and R that provides a graphical user interface to perform the quality control of alignment sequencing data. A command-line interface has been also implemented, so Qualimap can be incorporated in particular analysis pipelines. The first step in the program is to select the type of analysis to be run, which can be: BAM QC for alignment data—with optionally a set of regions of interest—or Count QC for count data.

When dealing with alignment data, the main input for Qualimap is the BAM file to be analyzed. The application processes information by splitting the reference genome into a given number of windows (400 by default), collecting the information and parallelizing the process where appropriate. The results are summarized in a dedicated panel and graphically represented in different charts, which can be exported for further analysis. The user can also concentrate the analysis to specific regions of interest by including a general feature format (GFF) or a browser extensible data (BED) file. In this case, the information is shown separately for reads that are mapped inside or outside of the defined regions.

Qualimap also provides insights into mapping performance by studying the read counts overlaps with genomic features of interest. These read counts can be loaded into the program as a text file that can be directly computed by using a dedicated tool in Qualimap. For example, the saturation rate for the detected features can be analyzed with respect to the sequencing depth, unveiling whether more features could be detected by increasing the sequencing depth. This is of particular interest, for example, in RNA-seq assays. Likewise, it is possible to analyze the read counts separately in user-defined groups of features.

In addition, Qualimap is designed 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...
with default parameters, a reasonable number of mapped reads and level of mean coverage were obtained (data not shown); however, difficulties arose when performing the variant calling. By running Qualimap, it was found that there was a coverage bias for GC-rich areas (Fig. 1a). Further examples of BAM QC performance can be found in the Supplementary Figures S1–S3.

In order to show the performance of Qualimap with read-count data, we made use of the study by Marioni et al. (2008) in which the authors estimated differences in gene expression profiles between human liver and kidney RNA samples using multiple sequencing replicates. For the sake of clarity, in this work, we considered data from the kidney sample only. To begin, we mapped the reads using TopHat (Trapnell et al., 2009) with default parameters; next, we obtained the read counts for each gene using the tool provided in Qualimap, discarding non-uniquely mapped reads and using the Ensembl 64 annotation (Flicek et al., 2011). Figure 1b shows a per-biotype detection plot. There is an enrichment of the protein-coding biotype in the mapped reads, as expected, but also a significant number of other detected biotypes such as pseudogene, lincRNA and processed-transcript. Likewise, the saturation point was not reached, as more genes were detected when sequencing depth was increased (Supplementary Fig. S4).

Further examples and case studies, as well as updates of the software, can be found on the Qualimap web page (http://www.qualimap.org).

Conflict of Interest: none declared.

3 RESULTS

To demonstrate how Qualimap can be used in the analysis of genome-wide sequencing, we analyzed an in-house unpublished plant resequencing dataset. Using bowtie (Langmead el al., 2009)