BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 9 2011, pages 1316–1317
doi:10.1093/bioinformatics/btr122

Sequence analysis

Advance Access publication March 12, 2011

TEQC: an R package for quality control in target capture experiments

Manuela Hummel*, Sarah Bonnin, Ernesto Lowy and Guglielmo Roma
Core Facilities, Centre for Genomic Regulation (CRG), UPF, Barcelona, Spain

Associate Editor: Alex Bateman

ABSTRACT

Summary: *TEQC* is an R/Bioconductor package for quality assessment of target enrichment experiments. Quality measures comprise specificity and sensitivity of the capture, enrichment, per-target read coverage and its relation to hybridization probe characteristics, coverage uniformity and reproducibility, and read duplicate analysis. Several diagnostic plots allow visual inspection of the data quality.

Availability and implementation: *TEQC* is implemented in the R language (version >2.12.0) and is available as a Bioconductor package for Linux, Windows and MacOS from www.bioconductor.org.

Contact: manuela.hummel@crg.es

Received and revised on January 17, 2011; accepted on February 28, 2011

1 INTRODUCTION

Next-generation sequencing technologies have made it possible to sequence efficiently a whole genome. However, in population studies, the current cost for sequencing entire human genomes at sufficient read coverage is still prohibitively expensive. Alternatively, targeted sequencing is a promising approach to identify variants in large numbers of individuals focusing on specific regions of interest (e.g. linkage regions or the complete exome) that are selected and enriched previous to deep sequencing. A frequently used application for the enrichment of the target sequences is based on hybridization with pre-designed probes, either on microarrays or in solution. The hybridized molecules are captured (i.e. eluted from the microarrays or pulled-down from the solution, respectively), and are then amplified and sequenced (Bainbridge *et al*., 2010; Gnirke *et al*., 2009; Tewhey *et al*., 2009).

Many tools exist to check the quality of the sequenced data (Martínez-Alcántara, 2009; Morgan *et al*., 2009; Toedling *et al*., 2010). Quality criteria can be sequencing error rates, per-base Phred scores, total number of sequenced reads, and fraction of reads aligning to the respective reference genome. However, in target enrichment experiments, it is additionally crucial to assess whether the capture of the targeted regions was successful. Here, we have developed *TEQC,* an R/Bioconductor package (Gentleman *et al*., 2004; R Development Core Team, 2010) that enables users to leverage a multitude of methods to monitor the target enrichment process. Our tool allows to answer specific technical questions, among others whether most of the sequenced reads actually fall on the target and whether most of the targeted bases reach sufficient coverage for follow-up analyses like variant detection. Moreover, *TEQC* allows to evaluate the amount of read duplication and its effect on target coverage, as well as the relation between GC content of the capture hybridization probes and coverage for the respective bases.

*To whom correspondence should be addressed.

2 DESCRIPTION

2.1 Input files

All operations in *TEQC* are based on genomic positions of the sequencing reads previously aligned to the reference genome. The read positions are related to the genomic positions of the targeted regions. The input for the package consists basically of two files containing genomic positions of reads and targets. If the positions for each single hybridization probe are available, they can also be introduced into the analysis. The format of the input files is kept quite flexible, which makes *TEQC* analysis independent of sequencing platforms and alignment tools.

In case of paired-end sequencing data, the read pairs can be put together keeping the start position of the first read and the end position of the respective second read, which correspond to the genomic positions of the actual DNA molecule that had been sequenced from both ends. Some parts of the *TEQC* quality analysis can be applied to both, single reads or read pairs.

2.2 Specificity and enrichment

The target capture specificity is calculated as the fraction of the provided reads that overlap with any of the targets. The enrichment is defined as the fraction of on-target reads, divided by the fraction of targeted bases within the whole genome. Hence, it relates the capture specificity to the size of the target and reveals the magnitude of the sequence enrichment process.

2.3 Read coverage and reproducibility

The degree of coverage for the targeted region is going to condition all the downstream analyses. Because of this, *TEQC* contains several functions that provide coverage-based information. Specifically, the package provides the user with per-base coverage, average coverage per single target region and average coverage over all targeted bases. In some calculations and figures the normalized coverage is used, which is the per-base coverage divided by the average coverage over all targeted bases. Normalized coverage does not depend on the absolute quantity of sequenced reads and is hence better for comparing different samples or experiments. A variant of target coverage is given by the number of reads overlapping each target. The coverage histogram visualizes the

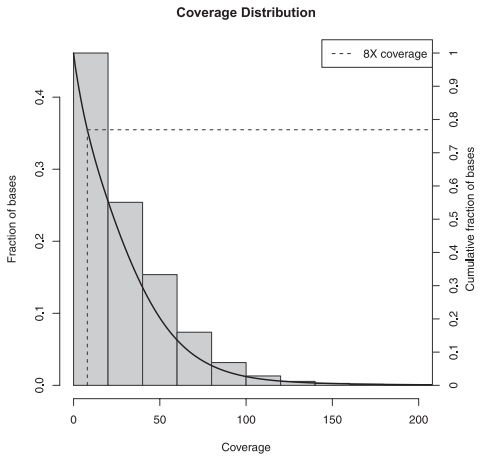

Fig. 1. On-target coverage histogram. The thick line shows the cumulative fraction of target bases (right *y*-axis) with a read coverage of at least *x*. Dashed lines highlight the fraction of target bases with $\geq 8\times$ coverage.

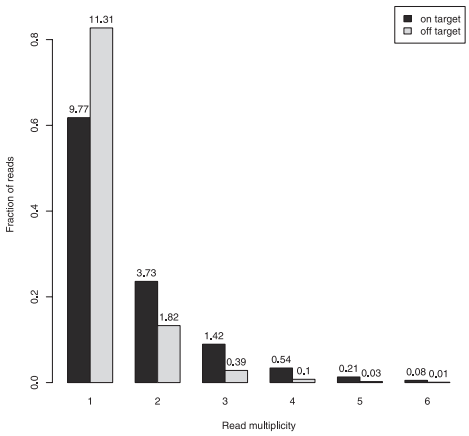

Fig. 2. Read duplicates bar plot. It shows which fraction of reads (*y*-axis) is present in the data in what number of copies (*x*-axis). The *x*-axis is truncated at a multiplicity of 6. Read multiplicity proportions are calculated and shown separately for on-target (dark gray) and off-target (light gray) reads. The numbers on top of the bars are absolute counts in millions.

coverage distribution for targeted bases; see Figure 1 for an example. The figure gives information about the fraction of target bases that have at least certain coverage (capture sensitivity). Further diagnostic plots regarding read coverage are: a coverage uniformity plot, a regression plot between target coverage and target length and a regression plot showing the relation between GC content of the hybridization probes and respective per-probe coverage. Also per-base coverage along chromosomal positions can be displayed, or alternatively exported as wiggle files for visualization in genome browsers.

The assessment of reproducibility, essential for any new technology, is based on read coverage as well. Coverage density and uniformity plots can be done for several replicate samples simultaneously. For pairs of replicates, correlation plots of per-base coverage values are provided. In these figures it is useful to consider the normalized instead of the actual coverage, since it assures comparability across samples.

2.4 PCR artifacts and read duplication

A crucial issue in target capture experiments is read duplication. Usually, read duplicates are removed before downstream analyses because they are supposed to be PCR artifacts. However, in a targeted resequencing experiment, one should expect a substantial amount of 'natural' read duplication due to the enrichment process. These 'natural' read duplicates would be derived from separate input DNA molecules that by chance were fragmented at the same position. Unfortunately, it is not possible to distinguish the PCR artifacts from the naturally occurring ones, so we recommend to apply *TEQC* to the data (also) after removing duplicates. In *TEQC*, two reads are considered read duplicates when they have exactly the same start and end positions. Figure 2 displays a duplicates bar plot generated by *TEQC* where we can see which fraction of the reads is present in the data in what number of copies. Read multiplicity proportions are calculated separately for on- (Fig. 2, dark gray bars) and off-target reads (light gray bars), giving information about the amount of 'natural' duplication (expected mostly in the target regions) versus read duplication originated by artifacts in the PCR (expected both on- and off-target).

3 CONCLUSIONS

Assessing the quality of the enrichment process is an important step in target capture experiments. At the time of writing, the R/Bioconductor package *TEQC* is one of the few tools available for the quality control of targeted resequencing experiments. The package imports genomic positions of sequencing reads and target regions, and allows to evaluate the whole enrichment process by estimating specificity and sensitivity of the capture, per-target read coverage and amount of read duplication.

ACKNOWLEDGEMENTS

We thank Mònica Bayés, Anna Ferrer and Heidi Mattlin for implementing the experimental part of target capture experiments, Heinz Himmelbauer, Debayan Datta and Matthew Ingham for performing sequencing and analysis and Xavier Estivill and Raquel Rabionet for providing an example dataset.

Funding: Spanish Ministry of Science with FEDER funds (European Regional Development Fund).

Conflict of Interest: none declared.

REFERENCES

Bainbridge,M.N. *et al.* (2010) Whole exome capture in solution with 3 Gbp of data. *Genome Biol.*, **11**, R62.

Gentleman,R. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Gnirke,A. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.

Martínez-Alcántara,A. (2009) PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics*, **25**, 2438–2439.

Morgan,M. *et al.* (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, **25**, 2607–2608.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tewhey,R. *et al.* (2009) Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol.*, **10**, R116.

Toedling,J. *et al.* (2010) Girafe–an R/Bioconductor package for functional exploration of aligned next-generation sequencing reads. *Bioinformatics*, **26**, 2902–2903.