BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 12 2014, pages 1767–1768
doi:10.1093/bioinformatics/btu108

Sequence analysis

Advance Access publication February 26, 2014

ngsCAT: a tool to assess the efficiency of targeted enrichment sequencing

Francisco J. López-Domingo[1], Javier P. Florido[1], Antonio Rueda[1], Joaquín Dopazo[1,2,3] and Javier Santoyo-Lopez[1,*]

[1]Bioinformatics Department, Genomics and Bioinformatics Platform of Andalusia (GBPA), 41092 Seville, [2]Computational Genomics Department, Centro de Investigación Príncipe Felipe, 46012 Valencia, Spain and [3]Functional Genomics Node (INB), Centro de Investigación Príncipe Felipe, 46012 Valencia, Spain

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Targeted enrichment sequencing by next-generation sequencing is a common approach to interrogate specific loci or the whole exome in the human genome. The efficiency and the lack of bias in the enrichment process need to be assessed as a quality control step before performing downstream analysis of the sequence data. Tools that can report on the sensitivity, specificity, uniformity and other enrichment-specific features are needed.

Results: We have implemented the next-generation sequencing data Capture Assessment Tool (ngsCAT), a tool that takes the information of the mapped reads and the coordinates of the targeted regions as input files, and generates a report with metrics and figures that allows the evaluation of the efficiency of the enrichment process. The tool can also take as input the information of two samples allowing the comparison of two different experiments.

Availability and implementation: Documentation and downloads for ngsCAT can be found at http://www.bioinfomgp.org/ngscat.

Contact: support@bioinfomgp.org

Supplementary information: Supplementary data is available at *Bioinformatics* online.

Received on December 3, 2013; revised on January 24, 2014; accepted on February 16, 2014

1 INTRODUCTION

Next-generation sequencing (NGS) technologies have opened new opportunities for inspecting and understanding genomic and transcriptomic sequences providing a wealth of new data. Whole-genome sequencing is now technically feasible; however, it is still expensive to run, and in many cases, only specific genomic regions are being sequenced. For instance, whole-exome sequencing is being extensively used to discover gene mutations for hereditary diseases, and high-throughput sequencing of gene panels is starting to be explored as a molecular diagnostic tool (Sikkema-Raddatz *et al.*, 2013). Thus, targeted NGS has become a common tool for interrogating at once several loci or all coding regions of the genome at a relatively low cost (Clark *et al.*, 2011).

Target enrichment for high-throughput sequencing includes a capture step where probes are hybridized to complementary fragments of genomic DNA. Successful sequencing is highly dependent on the efficiency of this target-enrichment procedure. The degree of enrichment can be estimated before sequencing by quantitative polymerase chain reaction of a few target regions. Only the analysis of the sequence data can reveal the degree of enrichment for all regions of interest (ROIs) and can provide information on the whole capture process showing if any bias is present.

Comparative studies for target-enrichment experiments have been carried out by a number of works that allow to identify some critical aspects to evaluate: (i) *sensitivity*, to assess the quality of the coverage on target regions; (ii) *specificity*, to measure the proportion of off-target reads; and (iii) *uniformity*, to detect sequencing biases in the targeted regions (Asan *et al.*, 2011; Clark *et al.*, 2011; Tewhey *et al.*, 2009). Systematic detection of biases owing to the capture step is critical to save time and to avoid drawing incorrect conclusions from downstream analysis.

Currently, the number of software packages that allow systematic assessment of the enrichment process is limited. For instance, there are some tools that may be used to get some general statistics for the coverage on/off-target reads and GC metrics like BEDTools (Quinlan and Hall, 2010) or Picard tools (http://picard.sourceforge.net). The functionality of these tools for capture efficiency assessment is limited and requires the implementation of complex scripts to produce informative results, limiting their use to researchers with knowledge on scripting languages. A more specific tool called TEQC has recently been implemented in *R*, but it has high requirements of RAM memory when target regions of the size of current exomes (~50 Mb) are analyzed (Hummel *et al.*, 2011). NGSrich is another recent tool, implemented in Java, intended to be used as part of a pipeline in a high-performance computing environment, but it lacks of some functionality and its reporting is limited (Frommolt *et al.*, 2012). A comparison of NGS data Capture Assessment Tool (ngsCAT), TEQC and NGSrich is provided in the Supplementary Data.

We here present ngsCAT, a Linux command line tool that allows a comprehensive evaluation of the performance of the capture step in terms of sensitivity, specificity and uniformity. ngsCAT can be run on a standard computer and integrates the functionalities of the aforementioned tools extending them to generate a detailed report with metrics, summary tables, figures and plots that evaluate the efficiency of the targeted enrichment process.

*To whom correspondence should be addressed.

2 APPLICATION DESCRIPTION

ngsCAT is a command-line application written in Python that only requires as input a binary alignment map (BAM) file and a browser extensible data (BED) file describing the ROIs. Therefore, the tool can handle data from all sequencing platforms and can be easily integrated as part of automated analysis pipelines. ngsCAT implements an efficient multi-threaded processing enabling even a full execution for human exome data in a standard computer (Intel Core 2 Duo, 4 GB RAM).

The software calculates a number of graphs and statistics that allow to assess the performance of the enrichment steps. Thus, ngsCAT reports the number and percentage of reads on/off target, the percentage of target bases covered at different coverage thresholds, the number of duplicated reads on/off target, bedgraph tracks of off-target regions with high coverage, the distribution of the coverage in the ROIs, the variability of the coverage within the ROIs and the distribution of the coverage as a function of GC content. The program can also draw a saturation curve of the coverage as a function of the number of reads that can serve to estimate whether sequencing a higher number of reads will produce a significant increase of the coverage in the ROIs. ngsCAT can also process two samples at a time, which allows a simple comparison of two samples. It can be used for the analysis of small target regions as well as for larger regions like whole exomes (see Supplementary Data).

The output is integrated in a concise and self-explanatory html report. For each report section, a minimum threshold value can be set to produce a warning if the values of the tested experiment are beyond the preset threshold. Further documentation, examples and updates can be found at http://www.bioinfomgp.org/ngscat.

3 RESULTS

To show how ngsCAT can assess the performance of the capture process, three in-house datasets targeting ~5.8 Mb of the human genome were used (see Supplementary Data for analysis pipeline details). ngsCAT was used taking as input the BAM file and the BED file that describes the targeted regions. First, we could check the percentage of target bases covered at different coverage thresholds (Fig. 1a and Supplementary Fig. S1). Interestingly, we found that ~80% of the targeted bases have at least a 10× coverage, which is a relatively low percentage, and that the percentage of reads on target is ~80%, which is a typical value for capture experiments, suggesting that reads were properly enriched. The 10× coverage saturation curve tends to plateau at ~80% of covered positions (Fig. 1b and Supplementary Fig. S2), indicating that no more target bases with a 10× coverage will be obtained by increasing sequencing depth. When ngsCAT was run with probe coordinates, instead of ROI coordinates, ~95% of the bases were covered at 10×. In addition, the inspection of the off-target information files showed the presence of a number of off-target regions with a coverage higher than 15×. Thus, ngsCAT results allowed us to hypothesize that the low percentage of covered target bases could be due to the capture probe tiling and to an off-target effect (see Supplementary Data).

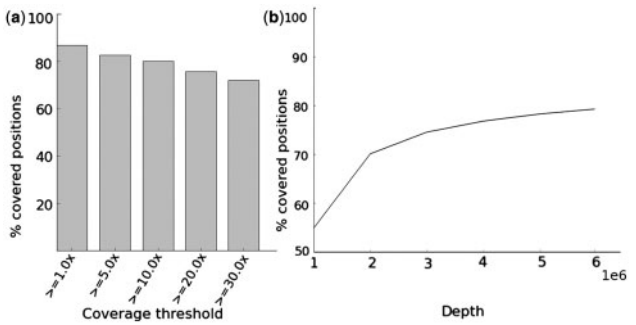

Fig. 1. Graphs generated by ngsCAT on an in-house dataset (sample 1). (**a**) Percentage of covered target bases at different coverage thresholds. (**b**) Coverage saturation curve. X axis represents the number of mapped reads. Y axis represents the percentage of target bases covered at ≥10×

4 CONCLUSIONS

Targeted NGS experiments are becoming a common way to interrogate genomic ROIs at a relatively low cost. Systematically assessing the performance of the capture step is critical before continuing with downstream analysis. We have implemented ngsCAT, a tool that allows to easily assess the enrichment efficiency with just one command-line run in a common computer. This tool has been used as a quality control for >600 exomes sequenced in our facility in the context of the Medical Genome Project (http://www.medicalgenomeproject.com), allowing us to detect samples not properly hybridized, optimize targeted enrichment protocols and adjust data analysis pipelines.

Funding: Ministerio de Economia y Competitividad (MINECO) (BIO2011-27069), the Conselleria de Educacio of the Valencia Community (PROMETEO/2010/001), the Regional Ministry of Health of the Andalusia Community (PI-0445-2013), ACTEPARQ (PCT-30000-2009-12), INNPLANTA (PCT-300000-2010-007) and Fondo Europeo de Desarrollo Regional (FEDER).

Conflict of Interest: none declared.

REFERENCES

Asan *et al.* (2011) Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol.*, **12**, R95.

Clark,M.J. *et al.* (2011) Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.*, **29**, 908–914.

Frommolt,P. *et al.* (2012) Assessing the enrichment performance in targeted resequencing experiments. *Hum. Mutat.*, **33**, 635–641.

Hummel,M. *et al.* (2011) TEQC: an R package for quality control in target capture experiments. *Bioinformatics*, **27**, 1316–1317.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Sikkema-Raddatz,B. *et al.* (2013) Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. *Hum. Mutat.*, **34**, 1035–1042.

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