TarSeqQC: Quality Control on Targeted Sequencing Experiments in R

Gabriela A. Merino[1, 2, *], Yanina A. Murua[3], Cristóbal Fresno[1], Juan M. Sendoya[3], Mariano Golubicki[4], Soledad Iseas[5], Mariana Coraglio[5], Osvaldo L Podhajcer[3], Andrea S Llera[3], Elmer A. Fernández[1, 2, *]

[1]UA AREA CS. AGR. ING. BIO. Y S, CONICET, Universidad Católica de Córdoba, Córdoba, Argentina.

[2]Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Córdoba, Argentina.

[3]Fundación Instituto Leloir and Instituto de Investigaciones Bioquímicas de Buenos Aires-CONICET, Buenos Aires, Argentina.

[4]Intergrupo Argentino para el Tratamiento de los Tumores Gastrointestinales, Buenos Aires, Argentina

[5]Hospital de Gastroenterología "Dr. Carlos Bonorino Udaondo", Buenos Aires, Argentina.

*To whom correspondence should be addressed.

Gabriela A. Merino

gmerino@bdmg.com.ar

And

Elmer A. Fernández

efernandez@bdmg.com.ar

ABSTRACT

Targeted sequencing is growing as a screening methodology used in research and medical genetics to identify genomic alterations causing human diseases. In general, a list of possible genomic variants is derived from mapped reads through a variant calling step. This processing step is usually based on variant coverage, although it may be affected by several factors. Therefore, under-covered relevant clinical variants may not be reported, impacting on pathology diagnosis or treatment. Thus,

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/humu.23204.

a prior quality control of the experiment is critical to determine variant detection accuracy and to avoid erroneous medical conclusions. There are several quality control tools, but they are focused on issues related to whole genome sequencing. However, in targeted sequencing, quality control should assess experiment, gene and genomic region performances based on achieved coverages.

Here, we propose TarSeqQC R package for quality control in targeted sequencing experiments. The tool is freely available at Bioconductor repository. TarSeqQC was used to analyze two datasets; low-performance primer pools and features were detected, enhancing the quality of experiment results. Read count profiles were also explored, showing TarSeqQC's effectiveness as an exploration tool. Our proposal may be a valuable bioinformatic tool for routinely targeted sequencing experiments in both research and medical genetics.

Keywords: Targeted Sequencing, Experiment Performance, Quality Control, R Package, Cancer Panel, Medical Genetics

INTRODUCTION

Next Generation Sequencing (NGS) is playing a major role in the growth of translational medicine because it provides quantification of genomic variations with high sensitivity (Becker et al., 2013). Targeted Sequencing (TS) is an NGS application for simultaneous exploration of specific genomic regions of a small group of genes (a panel). Such regions are called 'features' and, in general, they are devised to search for known or suspected genetic variations related to a particular disease (Metzker, 2010; Meldrum et al., 2011). Although there are several commercial TS panels available, researchers may design their own (custom) panels to capture other regions (Hadd et al., 2013). Hence, TS is emerging as a cost-effective and versatile methodology to detect actionable genomic variants with clinical significance (Metzker, 2010; Chang and Li, 2013). Particularly, TS panels are playing a key role in cancer, a heterogeneous disease driven by heritable or somatic mutations (Schweiger et al., 2011; Nikiforova et al., 2013). Therefore, accurate detection of genomic variants through TS becomes essential for personalized medicine approaches. In addition, current clinical genomic laboratories and large sequencing projects are involving several patients (Rizzo and Buck, 2012). Thus, it is crucial to have a complete experiment quality control picture instead of single sample portraits.

Every TS experiment starts from a DNA sample collected from blood, tumor or tissue cells, in which features are selected and amplified, generally by polymerase chain reaction (PCR) using specific primers organized in one or several pools (Satya and Di Carlo, 2014). The raw sequencing data is first mapped to a reference genome and then translated into well-mapped reads, from which a list of variants can be derived in order to detect single nucleotide polymorphisms (SNPs) and other genomic alterations (Nielsen et al., 2011). However, a previous evaluation of both quantity and quality of sequencing data is critical because it will limit the accuracy and comprehensiveness of those results. Sequencing data is usually quantitatively assessed by its coverage, which can be

affected by several factors influencing NGS reactions and producing non-uniformly represented genomic regions (Rizzo and Buck, 2012). Qualitatively, uneven coverage can also affect the variant analysis. For example, a deeply sequenced sample with non-uniform coverage could have a portion of the genome under or unsequenced, where the identification of relevant SNPs will not be possible, thereby causing a false negative and a subsequent incorrect clinical report (Rizzo and Buck, 2012). Thus, after read alignment and prior to downstream analysis, it is necessary to perform a coverage-based quality control, which should detect low-performance features, samples, PCR pools or even experiments in advance, to avoid waste of time and loss of money or biological material.

While there are many tools available to check sequence data quality (Martínez-Alcántara et al., 2009; Morgan et al., 2009), they are based on sequencing error rates, per-base Phred scores, and fraction of reads that align to the reference genome. Notably, they only address issues related to sequencing run as a whole. However, in TS experiments, quality control should assess if all the features were sequenced, what the achieved coverages were, what features were consistently underperforming and if some problems arose in the global settings or in each specific pool (Merino et al., 2015). The TEQC R package (Hummel et al., 2011) checks some of these issues. However, the performance measures generated by TEQC are in general at sample level, probably masking effects occurring at pool or feature levels. Neither pool performance nor complementary data can be analyzed. Finally, although TEQC generates a table of feature coverage values, the package does not offer analytical or graphical methods for in-depth analysis. These limitations are also present in tools provided by NGS manufacturers (e.g. Ion Proton™), with the added issue of their restricted access.

Here, we propose the TarSeqQC R package for quality control in TS experiments after read alignment and before variant calling analysis. TarSeqQC allows users to inquire specific and critical aspects, such as overall and specific feature coverage, primer pool performance and coverage consistency across samples. The package also incorporates exploration capabilities providing rapid access and visualization of single features at the nucleotide level.

MATERIAL AND METHODS

TarSeqQC description

The developed R package was built upon the premise that a TS experiment is based on a bed file (https://genome.ucsc.edu/FAQ/FAQformat.html), which defines the panel features, sample alignment BAM files and the reference sequence FASTA file (Figure 1**A**). The package defines two new R classes, `TargetExperiment` and `TargetExperimentList`, and their specific methods. For a detailed description of input file format and TarSeqQC methods, see the package vignette at Bioconductor.

The information and results from each sample, obtained from the mandatory files, are stored in a `TargetExperiment` object. If several samples are involved, all their `TargetExperiment` objects are used to build a `TargetExperimentList` which summarizes experiment results (Figure 1**B**). The `TargetExperiment` constructor computes feature coverage and median read counts, allowing users to choose any of those measures for the analysis. Quality control can be carried out through user-friendly numerical and graphical methods (Figure 1**C**), incorporating coverage/median count intervals and thresholds to further assist the analysis (Figure 1**D**). Here, feature coverage was used as

a performance measure. Hereafter, feature/gene/pool/sample or even experiment that fails to achieve a minimum coverage value will be referred to 'low performance' or 'underperforming'.

The TarSeqQC package allows us to explore the following aspects (Figure 1**C**):

- ***Multiple-samples experiment performance.*** Several methods were implemented that integrate feature coverage results of all the experiment samples. Thus, consistency performance of PCR pools, samples and the experiment as a whole are evaluated, contributing to wet lab operation and future planning.

- ***Single-sample experiment performance****.* It is evaluated using graphical and numerical methods focused on the percentage of reads mapping in targeted regions, feature coverage distribution, coverage comparison among features and over each PCR pool, etc.

- ***Feature performance.*** It is addressed by simultaneously inspecting all feature coverages at an early stage, to identify low-performance or unsequenced features. The involved methods can incorporate user-predefined thresholds or intervals, e.g., *[0,1): unsequenced features; [1,100): low; [100,500): good* and *[500,Inf): excellent coverage,* with the latter interval meaning coverage higher than 500 (*Inf*, infinite number).

- ***Genomic regions exploration****.* Visual inspection of specific regions or even genomic variants is provided through read count profile plots, where the total mapped reads per genomic position and detected nucleotide variations are displayed.

TarSeqQC is available at Bioconductor (http://bioconductor.org/packages/TarSeqQC/). System requirements depend on the panel size and sequencing depth, but since it was developed for multicore machines, a desktop computer is usually sufficient. The run times of the most time-consuming TarSeqQC methods for a typical TS sample using a desktop machine are listed in Supp. Table S1. To demonstrate the tool capabilities, quality control was carried out in two TS datasets after read alignment and before downstream analysis using package version 1.4.1. The R code used to analyze both experiments is available in Supp. File S1 and S2.

Datasets

Colorectal Cancer

This dataset, hereafter CRC, was created to characterize the colorectal cancer molecular profile of an adenomatous polyposis patient, following a protocol approved by Institutional

Review Boards of Udaondo Hospital and Instituto Leloir (available upon request). A written informed consent was obtained from the patient at the moment of enrollment in the study. The experiment involved three DNA samples: Tumor biopsy, Normal (peripheral blood) and cell free DNA (cfDNA). Tumor and Normal samples will be compared to identify the genomic variants contributing to the disease and those that do not. Since circulating cfDNA has a great potential as biomarker for clinical management of cancer, the variants found in cfDNA sample will be compared with Tumor and Normal sample results in order to evaluate both presence of tumor and possible metastasis. In particular, the variant analysis should determine, if present, genomic alterations in *APC* (MIM# 611731), a tumor suppressor gene with reported alterations in adenomatous polyposis colorectal cancer (Segditsas and Tomlinson, 2006; Nieuwenhuis and Vasen, 2007). Thus, the performance of

this gene and its amplicons will be explored in order to ensure analytical utility in reporting possible variants.

Sequencing was performed using the Ion AmpliSeq™ Comprehensive Cancer Panel kit (http://www.thermofisher.com/) and an Ion Proton™ machine. The panel involves 15991 amplicons from 409 genes related to cancer, distributed in four PCR pools with 3996, 4007, 3991 and 3997 amplicons for pools 1 to 4, respectively. Normal and Tumor samples were sequenced together at 1000X average coverage and the cfDNA sample was sequenced at 10000X average coverage. Sequencing reads were aligned using TMAP4, included in the Torrent Suite, with its default configuration (v. 5.0, http://www.thermofisher.com/). The alignment BAM files were processed by Samtools (v. 1.2, Li et al. 2009b) in order to filter unmapped reads and sort the alignments. The data was also processed by TEQC (v. 3.6.0) for comparison purposes. The R script used is available in Supp. File S3.

Breast Cancer

This dataset, hereafter BC, is part of the Ultra-Deep Targeted Sequencing of a set of Cancer Genes project (SRP019940), intended to interpret the molecular profile of 38 breast cancer patients (Harismendy et al., 2013; Yost et al. 2013). This project used a custom panel involving 1736 amplicons of 47 genes that are relevant to cancer patient care (e.g. *TP53* (MIM# 191170), *BRCA1* (MIM# 113705), *PIK3CA* (MIM# 171834)). Sequencing was performed on matched germline and tumor tissues using Illumina MiSeq sequencer under a 151-nt long-paired end reads protocol (Yost et al., 2013). For TarSeqQC demonstration purpose, six patients were randomly selected to form the BC dataset (Supp. Table S2). The raw reads were downloaded from the Short Reads Archive at the NCBI (SRA067610, SRA067611). Then, reads were aligned as in Yost et al. (2013), using the BWA software (Li and Durbin, 2009a). The unmapped reads were filtered and the aligned reads were sorted using Samtools. In their original work, Yost et al. (2013) considered a minimum coverage value of 100 to perform the variant analysis; therefore, this value will be used here as a quality control threshold.

RESULTS

Colorectal Cancer dataset

Experiment performance

The summary statistics of the sample coverages at amplicon and primer pool level were obtained using the `summary` method (Supp. File S1). At amplicon level, the sample's mean coverages were 3585 and 2417 for Tumor and Normal samples respectively and 4219 for the cfDNA sample. Those values suggest good overall performances for Tumor and Normal samples; for the cfDNA sample, however, the achieved value seems to be low compared to the planned sequencing depth of 10000X, suggesting a possible problem. In addition, unread amplicons (coverage=0) were found in all samples. Amplicon coverage distributions across samples did not show differences in median or variance (Supp. Figure S1). The same results were obtained using the TEQC R package (Supp. Figure S2 and Supp. Table S3). However, exploration at pool level showed high variability for the `plotPoolPerformance` method (Supp. Figure S3). Primer pool 1 exhibited the highest average amplicon coverage (5201), with the other three pools exhibiting lower values (2500 – 3000).

Table 1 shows the absolute, relative and cumulative frequencies of amplicons falling into the predefined coverage intervals, obtained using the `summaryIntervals` method. As it can be observed, the three samples had more than 64% of its amplicons with coverage > 1000. In the Tumor sample, more than 90% of its amplicons achieved coverage values exceeding that threshold. This sample had only eight unread amplicons (first coverage interval), whereas the Normal and cfDNA samples had 829 and 130, respectively. In the Tumor and Normal samples, over 31% of their amplicons had coverage < 1000 (first three coverage intervals). Indeed, in the Normal sample, almost 13% of its amplicons had coverage < 100, showing a poor amplicon performance. These results suggest a relationship between those low-performance amplicons and the high variability previously detected in pools 2 and 3. This evidence could not be found by TEQC because this tool does not consider pool information

The suspected low performance of Normal and cfDNA samples was verified using the `plot` method, which generates a coverage tile plot from an $n_x p$ matrix where amplicons (1, …, n) are represented in the x-axis, ordered according to their genomic locations, and samples (1, …, p) in y-axis (Figure 2**A**). In this plot, the color of the $cell_{ij}$ corresponds to the coverage interval into which the coverage of the *i-th* amplicon of the *j-th* sample falls. Note that high coverage values (green bars) are more prevalent in Tumor sample. On the contrary, low coverage values (red bars) are more prevalent in the Normal and cfDNA samples, suggesting that both had low performances. Nevertheless, when the pool information was specified, the resulting tile plot clearly indicated that the previously observed poor sample performances were due to low performances in some of their PCR pools (Figure 2**B**). In fact, more than 90 % of pool 2 amplicons in the Normal sample achieved low coverage values (<1000), resulting in a high prevalence of red shades bars. About the 75% of pool 3 amplicons and 40% of pool 4 amplicons also had low coverage values for the cfDNA sample.

The described graphical behavior was quantitatively complemented using the `plotAttrPerform` method. Figure 2**C** illustrates the cumulative relative frequency for all the samples. In the Tumor sample (Figure 2**CI**), PCR pools presented the same behavior, accumulating more than 70% of its amplicons in good coverage intervals (>1000). Clearly, this scenario tends to the ideal case in which all pools have their amplicons in the so called “excellent” coverage interval (≥10000). The Normal sample showed a clear PCR pool bias towards a low performance for pool 2 (Figure 2**CII**) where 93% of its amplicons achieved coverage < 1000, with 45% (1824 amplicons) of those being even lower than 100. For the cfDNA sample, the situation was even worse (Figure 2**CIII**), with pools 3 and 4 exhibiting lower performances than pools 1 and 2. In particular, pool 3 was the one with lowest performance, with 75% of its amplicons with coverage < 1000, whereas pool 4 had 40 %.

Low-performance amplicons

The previous analysis revealed amplicons with low coverages. For instance, 3037 amplicons from 396 genes were found to have a coverage < 100 in at least one sample (`getLowCtsFeatures` method, Supp. Table S4). These amplicons were mainly from pool 2 (1828) and pool 3 (1007). In particular, 1772 low-performance amplicons from pool 2 had low coverage only in Normal sample and 901 amplicons from pool 3 only in cfDNA. These cases may be explained, for example, as a sample -dependent PCR reaction inhibition. Furthermore, the 91 amplicons with low coverage that appeared in all samples may represent a poor performance of the designed primers. Finally, two consistently unread amplicons, achieving zero coverage in all samples, were identified (Supp. Table

S5). One of them is the 240390110 from *TAL1* (MIM# 187040), which has 12 amplicons in the panel. The other one, 234444604, is part of the *MAP2K4* (MIM# 601335), which has other 20 amplicons.

APC gene performance

The tumor suppressor *APC* gene was explored in detail, revealing that its 94 amplicons achieved mean coverages > 1900 in the three samples (Tumor, 3201; Normal, 1953; cfDNA, 3081). Coverage bar plots are displayed in Figure 3, where amplicons are distributed along the x-axis; bar height represents the achieved coverage and color symbolizes PCR pools. The observed behavior for pool 2 in the Normal sample (Figure 3**B**, green bars) and for pool 3 and 4 in cfDNA sample (Figure 3**C**, cyan and violet bars) was also observed in the *APC* gene. For instance, the 224537542 *APC* amplicon in Tumor and Normal samples (red circle in Figure 3**A** and 3**B**) had coverage > 4500 but less than 360 in the cfDNA sample (red circle in Figure 3**C**). The same occurred with the last amplicon (224540558, blue circles in Figure 3), which achieved values up to 900 in Tumor and cfDNA samples but close to zero in the Normal sample. In particular, these amplicons did not overlap with any other in the panel (Supp. Table S6). Interestingly, cfDNA sample had the highest coverages for pools 1 and 2 (salmon and green bars) but very low for pools 3 and 4, showing a high variability among pools.

Breast Cancer dataset

Experiment performance

The samples of the BC dataset showed mean coverage values from 998 to 2394 (Table 2) with only two samples having at least one unread amplicon, the Tumor sample from AA0943 patient and the Germline sample from UCI9135402 patient. The averaged interquartile range was 784 (±190) and the average coverage range was 5970 (±1385).

The coverage intervals defined for this dataset were: *[0,1), unsequenced; [1,100), low; [100,500), good; [500,1000), very good* and *[1000,Inf), excellent coverage*. The use of the `summaryIntervals` method revealed that in all samples, less than 1% of their amplicons had a coverage < 100, showing a very high amplicon performance (Supp. Table S7). In addition, in almost all samples, more than 68% of their amplicons had a coverage > 1000, highlighting the poor performance of the Germline UCI9135402 sample case, in which 50% of its amplicons achieved coverage values between 500 and 1000.

The above quality control results indicate that all the samples have a good overall performance. This result was confirmed using the `plot` method, revealing a high prevalence (>69%) of excellent coverages (green bars) in all the samples and good and very good values (moss green bars) in the Germline sample of UCI9135402 patient (Supp. Figure S4).

Detection and exploration of low-performance amplicons

The two amplicons with the lowest performance were found using the criteria: “zero read counts in at least one sample” and are listed in Table 3. One of them belongs to *JAK3* (MIM# 600173) gene and achieved coverages < 13 in all samples, showing a consistently poor performance pattern. The other amplicon belongs to *TP53* gene and had a poor performance only in both Tumor and Germline samples from the UCI9135402 patient. In this experiment, consistently unread amplicons were not detected.

Two read profiles plots of these low-performance amplicons were explored using the `plotFeature` method (Figure 4 and Supp. Figure S5). In the first case, the JAK3.12.1.JAK3.1 amplicon in the UCI9135402 Tumor sample (Figure 4**A**) achieved a mean coverage of one (almost no read detected on it), as listed in Table 3, and a mean coverage of 12 in the Germline sample of the AA0943 patient (Figure 4**B**). Thus, the experiment results inappropriate to detect genomic variations in the region covered by that amplicon.

DISCUSSION

In TS experiments, the quality control analysis should ensure that all the involved samples have an average coverage value concordant with the experimental design, and fundamentally that the sequencing process be appropriate for genomic variant downstream analysis. Here, we propose the use of TarSeqQC R package as a light-weight and simple tool to pursue quality assurance in TS experiments. Through its application, we found that the CRC experiment apparently did not reveal differences between samples at amplicon level. The same conclusion was drawn using the TEQC package. However, when pool information was incorporated into TarSeqQC methods, the highest mean amplicon coverage was detected for pool 1 and a high variability was found in pools 2 and 3. These results indicated at a first glance, the occurrence of technical problems during library preparation, which produced the observed pool differences. The incorporation of coverage intervals revealed that those differences existed precisely between Normal and cfDNA samples. Indeed, their low performances would probably influence the detection of germline variants. Nevertheless, by incorporating pool information into TarSeqQC `plot` method we observed that poor performances in Normal and cfDNA samples were due to low performance in some of their PCR pools rather than in the whole samples. In particular, only pool 2 in the Normal sample showed a very low yield compared to the others, and pools 3 and 4 in cfDNA sample were low performance. The TEQC tool was not able to detect this low-performance pools because it does not use pool information in the analysis. Accordingly, TarSeqQC resulted in being a powerful tool to indicate that variant identification in those low-performance pools should not be expected, at least with the same confidence as other well-read areas. Moreover, overall coverage information at first glance can be also useful for checking the performance of different panel designs in samples from different sources. Probably the most useful scenario would be checking a panel designed for intact DNA in FFPE or liquid biopsy samples, since panels for intact DNA are more desirable because of their higher horizontal coverage but usually do not perform as well in vertical coverage when samples have degraded and shorter DNA fragments.

TarSeqQC also enables the identification of amplicons that have coverage of zero or lower than a threshold. This is a very important issue in the context of quality control to avoid erroneous conclusions about variants (presence or absence) in these genomic locations. Namely, if a true SNP located in unread or underperforming amplicon is not reported after variant calling, researchers will assume that it does not exist. However, this conclusion will be incorrect because the SNP exists and, in fact, the experiment failed to find it. When TEQC was used, the only way to detect those features was by inspection of a large table summarizing amplicon coverage values (Supp. Table S8). However, using TarSeqQC this task was user-friendly and was complemented by several diagnostic plots.

In the CRC dataset, TarSeqQC identified 91 low-performance amplicons (coverage < 100), some of them belonging to genes with reported alterations in the colorectal cancer such as *ATM* (MIM# 607585), *ARID1A* (MIM# 603024) and *PIK3CA* genes (Cancer Genome Atlas Network, 2012; Mouradov et al., 2014). In particular, two unread amplicons were reported, 240390110 (*TAL1*) and 234444604 (*MAP2K4*). These results suggest that genomic variants in those regions may not be detected because of a faulty or inherently impossible primer design for this region. In fact, common features observed *in silico* for the genomic regions covered by low coverage amplicons include high GC content and repetitive homopolymeric regions, which may contribute to suboptimal PCR amplification.

Since genomic alterations, principally point mutations, in the *APC* gene characterize the colorectal cancer, the quality control in that gene through TarSeqQC was performed, revealing that some of its amplicons had low coverage in some samples. As a consequence, we suggest that the proposed comparative analysis of genomic variation between Tumor, Normal and cfDNA samples may not be appropriate for the genomic regions covered by those amplicons. For instance, we found that amplicons 223304037 and 224540558, both from pool 2 and without overlap with other amplicons, did not have enough coverage in the Normal sample. In addition, in those regions there are several reported SNPs related to colorectal cancer such as rs2304793 and rs1804197. A similar situation was observed for amplicons 222847417 (pool 3) and 222825282 (pool 4), detected as low-performance in the cfDNA samples. For example, the rs41115 and rs587781816 reported SNPs are in locations not covered by other amplicons. Thus, the quality control over the *APC* gene indicates that this experiment is not useful to analyze those two regions. In order to cover these regions properly, a higher coverage should be attained; alternatively, a new panel design (e.g. including more pools) or a different method of analysis (i.e. Sanger sequencing) would be necessary for this part of the gene.

When TS is performed using custom panels, as in the BC dataset, quality control becomes important to evaluate if the designed primers performed as expected. Accordingly, the proposed tool was able to reveal not only low-performance experiments but also low- performance panels or regions in these panels. The identification of low-performance features and pools may contribute to the design or improvement of the corresponding PCR primers before a new panel use. In the BC dataset, we found that the average coverage achieved for all the samples using the custom TS panel was high, indicating a good panel design and sequencing. TarSeqQC allowed the identification of two unread amplicons in at least one sample. In particular, the JAK3.12.1.JAK3.1 amplicon had a consistently poor performance. Thus, we suggest reconsidering the design of the PCR primer used for its selection. Finally, results revealed the utility of TarSeqQC as a graphical and easy way to quickly explore a genomic region at nucleotide level, without the need to use heavyweight whole genome viewer tools.

The main limitation of TarSeqQC is that the computational resources and execution time depend on the TS panel size and the sequencing depth. In addition, the package can build the whole pileup matrix (at nucleotide level); however, this task also requires additional computational resources. Nevertheless, if they are available, TarSeqQC can be used without limitations. Furthermore, here we illustrate TarSeqQC application only over TS experiments; however, the tool can also be used in any experiments where there are feature-gene relationships. For instance, a bed file specifying several exons could be used to run TarSeqQC without any inconvenient.

In conclusion, our proposal may be a valuable bioinformatic tool for routine TS experiments in both research and medical genetics, allowing easy, simple and fast quality control over a wide range of TS experiments and laboratory regimes.

FUNDING

This work was supported by Argentine grants from the Universidad Católica de Córdoba (BOD/2016 to EAF), Ministerio de Ciencia, Tecnología e Innovación Productiva (FONCYT PPL06/2011 to EAF, FONCYT PPL04/2011 to ALL and OLP, and FONARSEC PBIT 015/13 to ALL), Secretaría de Ciencia y Tecnología-Universidad Nacional de Córdoba (30720150101719CB to EAF) and the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

CONFLICT OF INTEREST

None declare.

REFERENCES

Becker K, Vollbrecht C, Koitzsch U, Koenig K, Fassunke J, Huss S, Nuernberg P, Heukamp LC, Buettner R, Odenthal M, Altmueller J, Merkelbach-Bruse S. 2013. Deep ion sequencing of amplicon adapter ligated libraries: a novel tool in molecular diagnostics of formalin fixed and paraffin embedded tissues. *J Clin Pathol* **66**:803-806.

Cancer Genome Atlas Network. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**: 330-337.

Chang F, Li MM. 2013. Clinical application of amplicon-based next-generation sequencing in cancer. *Cancer Genet* **206**: 413-419.

Hadd AG, Houghton J, Choudhary A, Sah S, Chen L, Marko AC, Sanford T, Buddavarapu K, Krosting J, Garmire L, Wylie D, ShindeR et al. 2013. Targeted, high-depth, next-generation sequencing of cancer genes in formalin-fixed, paraffin-embedded and fine-needle aspiration tumor specimens. *The Journal of Molecular Diagnostics* **15**: 234-247.

Harismendy O, Schwab RB, Alakus H, Yost SE, Matsui H, Hasteh F, Wallace AM, Park HL, Madlensky L, Parker B, Carpenter PM, Jepsen K, et al. 2013. Evaluation of ultra-deep targeted sequencing for personalized breast cancer care. Breast Cancer Research 15: R115

Hummel M, Bonnin S, Lowy E, Roma G. 2011. TEQC: an R package for quality control in target capture experiments. *Bioinformatics* **27**: 1316-1317.

Li H, Durbin R. 2009a. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. and 1000 Genome Project Data Processing Subgroup. 2009b. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Martínez-Alcántara A, Ballesteros E, Feng C, Rojas M, Koshinsky H, Fofanov VY, Havlak P, Fofanov Y. 2009. PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics* **25**: 2438-2439.

Meldrum C, Doyle MA, Tothill RW. 2011. Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin Biochem Rev* **32**: 177-95.

Metzker ML. 2010. Sequencing technologies-the next generation. *Nat Rev Genet* **11**: 31-46.

Merino GA, Fresno C, Koile D, Yankilevich P, Sendoya JM, Oliver J, Llera SA, Fernández EA. 2015. An Exploration Tool for Quality Analysis in Targeted Sequencing Experiments. *IFMBE Proceedings* **49**: 659-662.

Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R. 2009. ShortRead: A bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**: 2607–2608.

Mouradov D, Sloggett C, Jorissen RN, Love CG, Li S, Burgess AW, Arango D, Strausberg RL, Buchanan D, Wormald S, O'Connor L, Wilding JL, et al. 2014. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Res* **74**: 3238-3247.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **126**: 443-451.

Nieuwenhuis MH, Vasen HFA. 2007. Correlations between mutation site in APC and phenotype of familial adenomatous polyposis (FAP): a review of the literature. *Crit Rev Oncol Hematol* **61**: 153-161.

Nikiforova MN, Wald AI, Roy S, Durso MB, Nikiforov YE. 2013. Targeted next-generation sequencing panel (ThyroSeq) for detection of mutations in thyroid cancer. *J Clin Endocrinol Metab* **98**: E1852-E1860.

Rizzo JM, Buck MJ. 2012. Key Principles and Clinical Applications of "Next-Generation" DNA Sequencing. *Cancer Prev Res* **5**: 887-900

Satya RV, Di Carlo J. 2014. Edge effects in calling variants from targeted amplicon sequencing. *BMC Genomics* **15**: 1073.

Schweiger MR, Kerick M, Timmermann B, Isau M. 2011. The power of NGS technologies to delineate the genome organization in cancer: from mutations to structural variations and epigenetic alterations. *Cancer Metastasis Rev* **30**: 199–210

Segditsas S, Tomlinson I. 2006. Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene* **25**: 7531-7537.

Yost SE, Alakus H, Matsui H, Schwab RB, Jepsen K, Frazer KA, Harismendy O. 2013. Mutascope: sensitive detection of somatic mutations from deep amplicon sequencing. *Bioinformatics* **29**: 1908-1909.

Figure 1: TarSeqQC schema. **A)** Sample alignment BAM, panel BED and reference sequence FASTA files are specified. **B)** `TargetExperiment` objects are built, one for each sample. If the experiment has several samples, all the `TargetExperiment` objects are summarized into a single `TargetExperimentList` object. **C)** `TargetExperiment` and `TargetExperimentList` methods are used to perform quality control at different levels: experiment, sample and feature, or even to explore specific features or genomic regions. **D)** An example of the graphics generated by TarSeqQC methods.

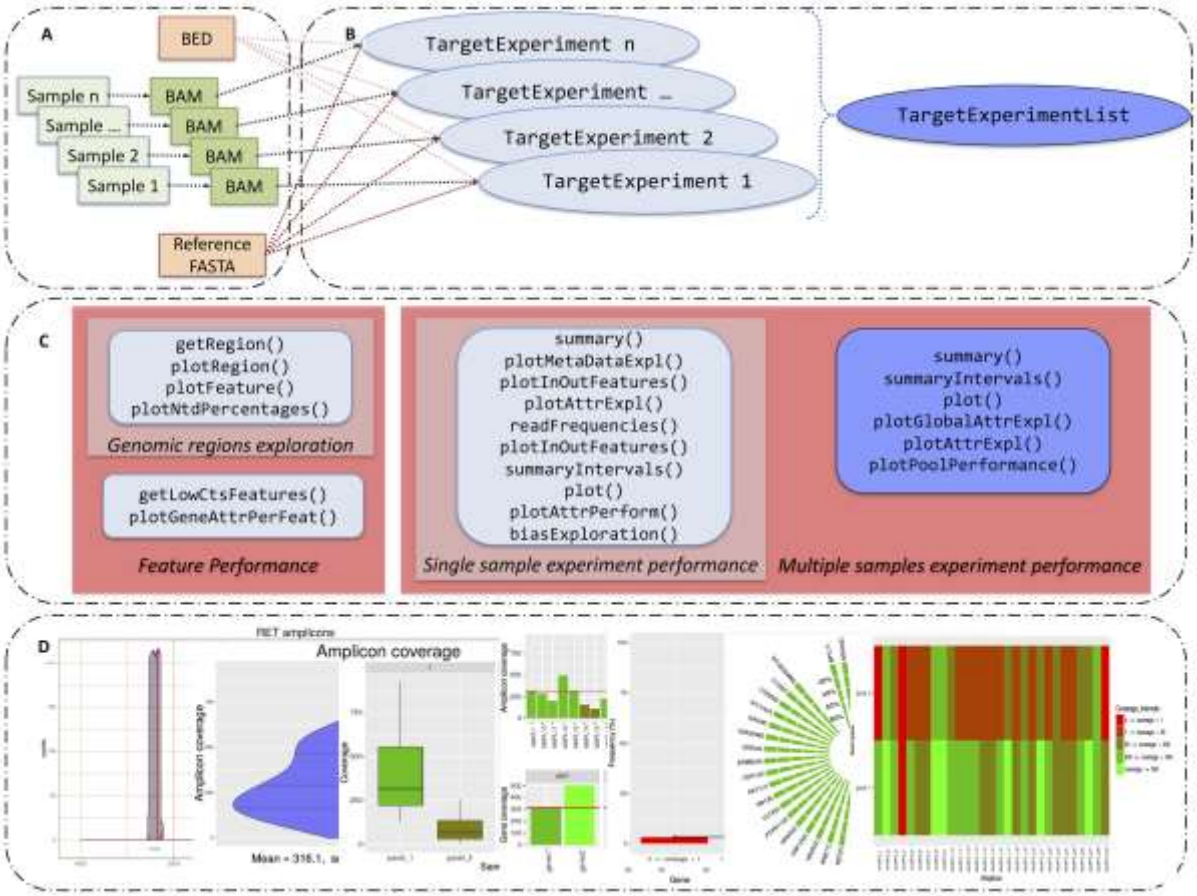

Figure 2: Colorectal cancer dataset coverage tile plot. Amplicons are presented in columns and samples in rows; cells are colored according to the amplicon coverage intervals. **A)** Amplicons ordered by their genomic position; **B)** Amplicons grouped by PCR pools. **C)** Relative cumulative frequency of amplicons falling into coverage intervals for **I)** Tumor, **II)** Normal and **III)** cfDNA samples. Each color represents one PCR pool and the ideal situation is observed in **I** where pool curves show the same behavior, with a high percentage of amplicons falling in good coverage intervals (>1000). In **II** pool 2 has a lower performance, accumulating many amplicons in low coverage intervals (<1000). A similar situation is observed with pools 3 and 4 in cfDNA sample (**III**).

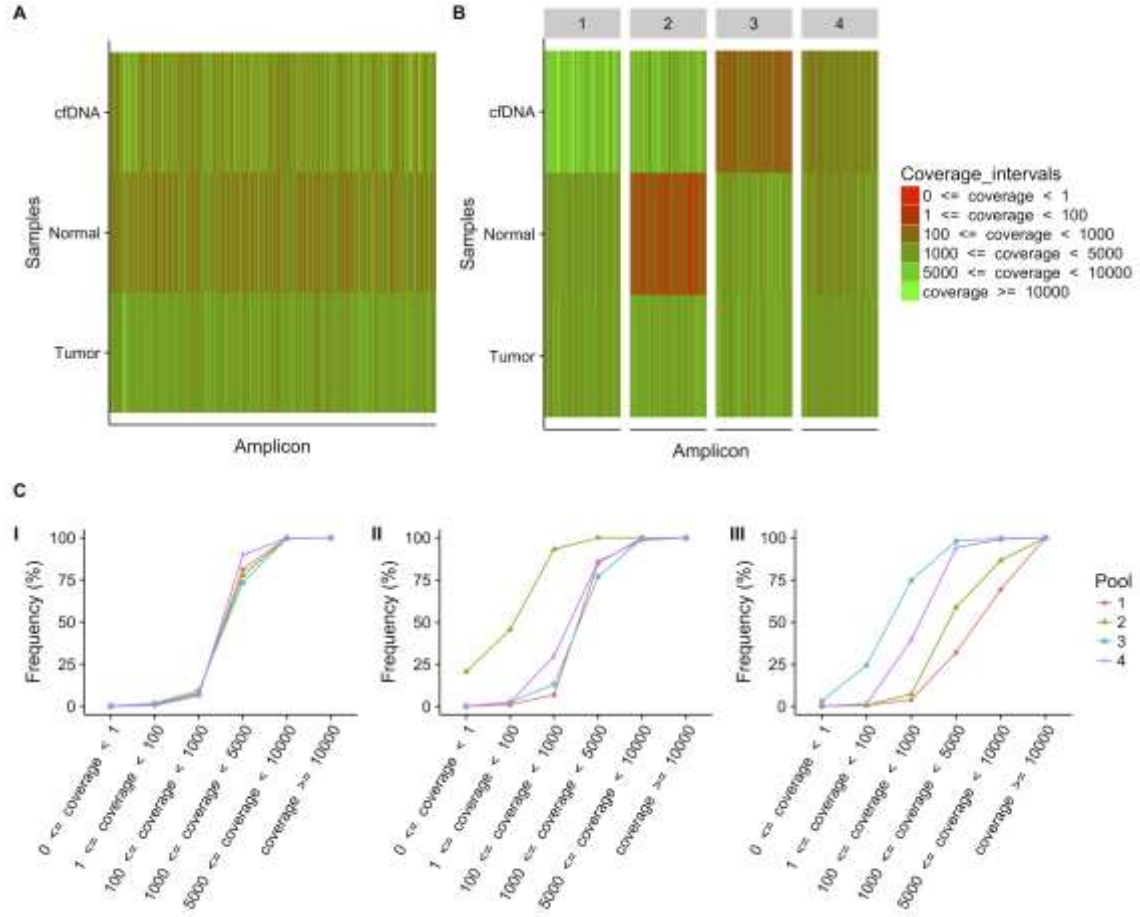

Figure 3: Amplicon coverage for *APC* gene in the Colorectal cancer dataset. **A)** Tumor, **B)** Normal and **C)** cfDNA samples. Amplicons are along x-axis; bar heights represent achieved coverage and bar colors indicate the corresponding PCR pool. The figure was obtained using the `plotGeneAttrPerFeat` method with its default values for `overlap` and `level` parameters.

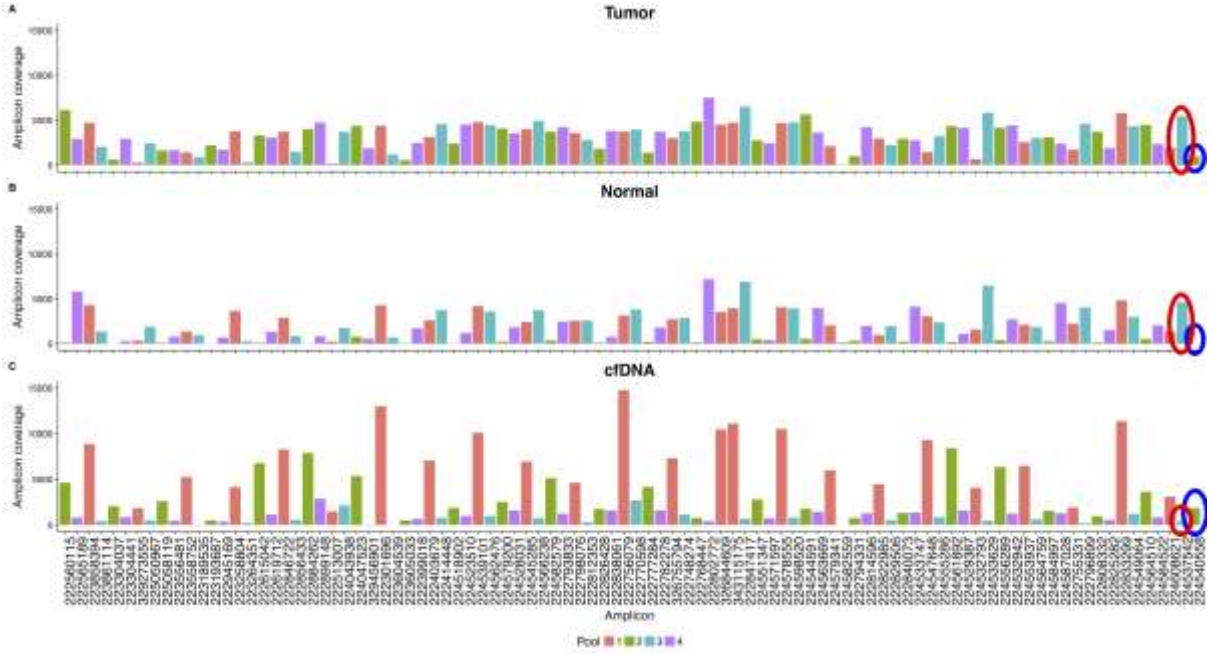

Figure 4: Read counts profile of JAK3.12.1.JAK3.1 amplicon. **A)** Tumor sample of UCI9135402 patient and **B)** Germline sample of AA0943 patient. The x-axis represents genomic position and y-axis, read counts. The violet curve stands for read counts matching with the reference sequence; nucleotide variants are reported as bars colored according to the detailed scale and the gray shadow represents total counts. The horizontal dark cyan segment represents the inspected feature.

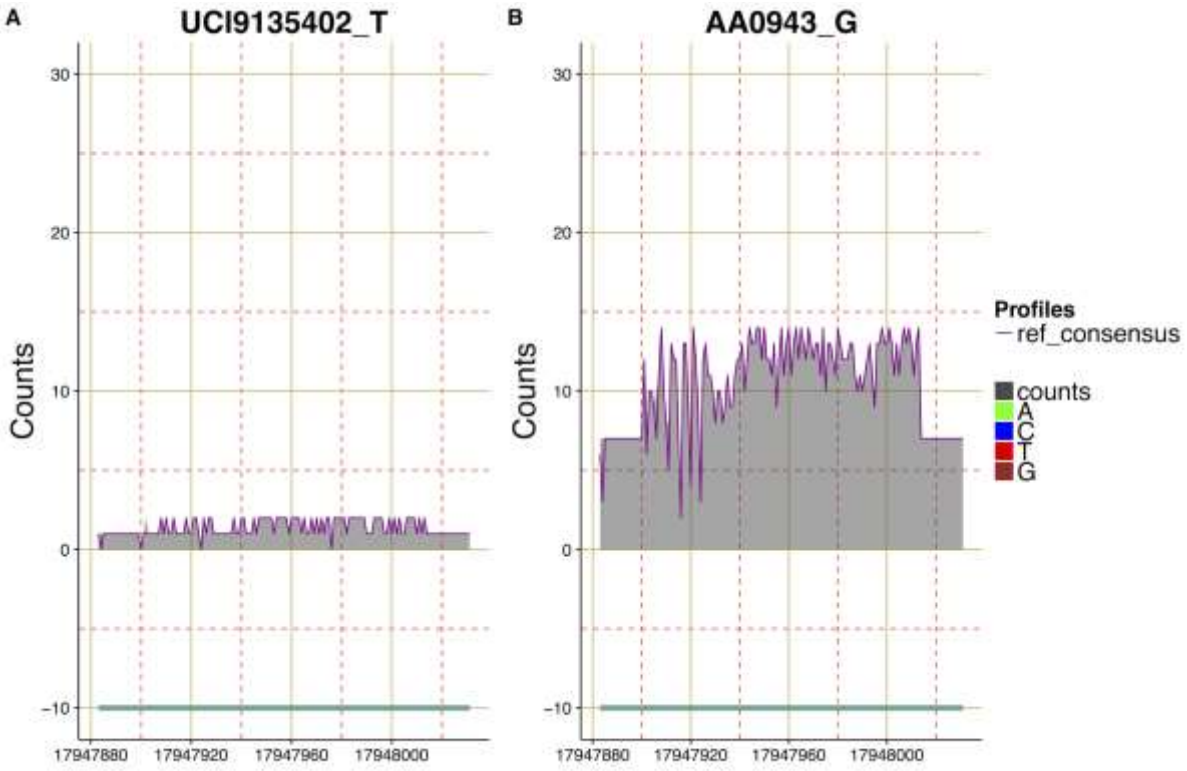

Table 1: Absolute and relative amplicons frequencies falling into the defined coverage intervals for Tumor, Normal and cfDNA samples.

Coverage interval	**Sample**					
	Tumor		*Normal*		*cfDNA*	
	Abs (cum)	Rel (cum)	Abs (cum)	Rel (cum)	Abs (cum)	Rel (cum)
[0, 1)	8 (8)	0.1 (0.1)	829 (829)	5.2 (5.2)	130 (130)	0.8 (0.8)
[1, 100)	168 (176)	1.1 (1.2)	1232 (2061)	7.7 (12.9)	943 (1073)	5.9 (6.7)
[100, 1000)	1063 (1239)	6.6 (7.8)	3661 (5722)	22.9 (35.8)	3942 (5015)	24.7 (31.4)
[1000, 5000)	11637 (12876)	72.8 (80.6)	8187 (13909)	51.2 (87)	6287 (11302)	39.3 (70.7)
[5000, 10000)	3079 (15955)	19.3 (99.9)	1991 (15900)	12.5 (99.5)	2877 (14179)	18 (88.7)
[10000, Inf)	36 (15991)	0.2 (100)	91 (15991)	0.5 (100)	1812 (15991)	11.3 (100)

Abs, absolute; Rel, relative; cum, cumulative; [a, b) indicates a coverage interval defined as a≤coverage<b; Inf, infinite value.

Table 2: Descriptive statistics for amplicon coverage in the Breast Cancer dataset.

			Statistics					
	Patient	Type	*Min*	*1st Qu*	*Median*	*Mean*	*3rd Qu*	*Max*
Samples	*AA1025*	T	1	1192	1535	1612	1956	5335
		G	5	942	1154	1120	1461	4895
	AA0926	T	3	1083	1423	1519	1873	5168
		G	8	1301	1646	1704	2079	4532
	AA0930	T	10	1749	2204	2288	2796	6377
		G	7	1169	1457	1533	1851	6171
	AA0943	T	0	1785	2320	2394	2923	6010
		G	12	1703	2161	2229	2686	7434
	AA0948	T	3	1364	1680	1781	2128	9180
		G	7	1225	1513	1590	1900	7385
	UCI9135402	T	1	1193	1574	1639	2005	5056
		G	0	748	937	998	1209	4155

T, Tumor; G, Germline; Qu, quartile.

Table 3: Coverage of the low-performance amplicons found in the Breast Cancer dataset.

		Samples											
		AA1025		*AA0926*		*AA0930*		*AA0943*		*AA0948*		*UCI9135402*	
		T	G	T	G	T	G	T	G	T	G	T	G
Amplicon	*TP53.15.1.TP53.1*	125	120	455	642	312	326	271	393	108	141	1	0
	JAK3.12.1.JAK3.1	1	5	3	8	10	7	0	12	3	7	1	0

T, Tumor sample; G, Germline sample