An Exploration Tool for Quality Analysis in Targeted Sequencing Experiments

G. Merino[1], C. Fresno[1], D. Koile[2], P. Yankilevich[2], J. Sendoya[3], J. Oliver[3], A. Llera[3], and E. Fernández[1]

[1] Catholic University of Córdoba-CONICET, Córdoba, Argentina
[2] Instituto de Investigación en Biomedicina de Buenos Aires (IBioBA) -CONICET- Partner Institute of the Max Planck Society, Buenos Aires, Argentina
[3] Laboratory of Molecular and Cellular Therapy, Fundación Instituto Leloir, Buenos Aires, Argentina
jmsendoya@leloir.org.ar, {efernandez,cfresno,gmerino}@bdmg.com.ar,
{javiom,allera1966,yankilevich}@gmail.com,
dkoile@ibioba-mpsp-conicet.gov.ar

Abstract— **Amplicon Exome Sequencing allows focusing on exonic regions of a small group of genes. The overall aim is to inquire sequenced regions about the occurrence of mutations, single nucleotide polymorphisms, insertions and deletions, trough Variant Calling analysis. The first steps include sequencing, followed by alignment. Prior to further analysis, it is crucial to evaluate how well the sequencing run was achieved, as well as the quality of the carried experiment. At present, there are several open access tools to perform these tasks. But, most of them were designed for whole genome data. Hence, they are computationally expensive to just analyze a small group of genes. In addition they only offer limited visualization capabilities. Here, we proposed a light-weight amplicon sequencing exploration tool for fast visualization and experiment quality control. The tool was developed under R language and can be executed in parallel in order to provide fast and accurate quality control results within a few minutes. The article presents the tool implementations and capabilities. Finally, we show its application on real amplicon sequencing data.**

Keywords— **Amplicon Sequencing, Quality control, Visualization tool**

I. Introduction

High throughput sequencing (HTS) technologies are commonly used in biology. They are usually applied to explore genomes, transcriptomes and epigenomes. Among the different HTS applications, Targeted Sequencing (TS) allows the exploration of specific genomic regions. In particular, Amplicon Exome Sequencing (AES) allows focusing on exonic regions of a small group of genes [1]. The main goal of AES is to inquire sequenced regions, trough Variant Calling analysis, about the occurrence of known and/or novel mutations, single nucleotide polymorphisms (SNPs), insertions and deletions (indels).

In AES experiments, exonic regions of a DNA sample are copied and amplified by PCR. If an exon of the target region is too large for the specific technology, several primers should be used to read it. Thus, different PCR pools are required to achieve good amplification [2]. Then, all those fragments are sequenced of an HTS machine. As results, millions of short sequence reads are obtained and then aligned against a reference genome. After that, Variant Calling analysis must be done. However, prior to further analysis, it is crucial to evaluate how well the run was achieved, as well as the quality of the carried experiment. In this sense, it is important to know how well the amplicons were sequenced, which amplicon and gene coverages were achieved or if some problems arise in the global setting or by specific pools, among others. All the points should be evaluated in advance prior to perform any biological interpretation.

At present, some of these exploration tasks are performed using open access tools [3]. Those allow visualization and some level of read profiles quantification. Despite the fact that they are very powerful tools, most of them are computational intensive. Hence a supercomputer is needed to perform the analysis. In addition the process is time consuming, whereas only a small fraction of the genome is handled. In this context, AES tools should be light-weight and avoid the need of strong computational hardware.

On the other hand, in AES it is required to evaluate statistics such as the achieved average coverage and its distribution over sequenced amplicons. If PCR pools were used, is also necessary an evaluation and comparison of pool results.

For these reasons, statistical information, as well as visualization capabilities should be available and easily accessed in a compact form. Current genome browsers lack of this feature and do not provide a detailed friendly user exploration capability for this kind of experiments.

Here we present an exploration tool for fast visualization and quality control of AES experiments. This tool is implemented in R language [4] using basic data structures. In addition, our development can be also used to analyze data from others HTS applications like TS, RNA-seq, and DNA-seq. The tool is freely available on request.

A. Braidot and A. Hadad (eds.), *VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014,* IFMBE Proceedings 49, DOI: 10.1007/978-3-319-13117-7_168

II. Materials and Methods

A. Generalities

Visualization must be properly combined with statistical results. Keeping this in mind, our tool was designed and implemented based on statistical R platform [4].

The tool is feed with the aligned reads stored in a *bam* file [5] and the information regarding the designed amplicon regions (ID, start, end, gene, etc.) in a *bed* file. From those files, the pipeline builds the *pileup* and *coverage* files. This task is carried out by means of Samtools [5] executed from R. Thus, all the required information for each genomic position is ready and available for further analysis.

The *Pileup* file contains all the sequenced base pairs (bp) from the genomic regions with additional information, such as chromosome, location, reference bp, amount and type (consensus, variant, or indels) of basis read at this position and their quality. Consequently, the file has a complex structure with a very large size (≈ Giga Bytes). Therefore, parallelization is required for the analysis. The proposed tool can run on a multicore computer, setting the number of available cores. On the contrary, the *Coverage* file is smaller and only contains the total read counts for each position of all amplicons. Hence it can be handled by a single core.

B. Implemented methods

Starting from the *pileup, coverage* files and the regions of interest stored in the *bed* file, the proposed pipeline implements the following stages:

a) Counting feature reads

In the context of HTS experiments, there are several features of interest. They depend on the particular application and process stage that you are working in. Scientists working in AES, are interested in analyzing variations in each bp position of the sequenced regions. However, in early quality control process, the main features could be genes or amplicons as a whole, since they provide a global view of the experiment performance. In this context, we provide:

```
> trb<-loadCoverageBedFile(cov_file)
> feature_total_reads<-featureTotalReads(trb,
+   feature="amplicon", mc.cores=18)
```

These functions allow reading the *coverage* file and obtaining total read counts by feature. If PCR pools were used, this information is also contained into the data structure.

b) Computing feature coverage

A good performance sequencing run indicator is the feature coverage, defined as:

$$Cov = (N*L)/G \qquad (1)$$

where N is the number of aligned reads, L the aligned read length and G the reference specific feature length. In order to obtain the metric the user can call:

```
> featureCoverage(reads.by.bp,feature="amplicon")
```

This function can be also run on a multicore machine.

c) Exploring feature coverage.

In AES is expected that all the genomic regions of interest will be sequenced with a similar coverage. This is key information to evaluate technical defects that could arise from PCR amplification, kit deficiency as well as amplicon efficiency. In this pipeline,

```
> featureCovExploration(cov_by_feature, feature,
+   join, pool)
```

allows exploring feature coverage distribution by boxplot and density plot. In addition, the analysis can also be performed for each PCR pool, in order to detect unexpected coverage effects.

d) Computing summary statistics.

At this point, hundreds of genes and thousands of amplicons have been characterized by their coverage and read counts profiles. Summary statistics are very powerful metrics for quality evaluation. The tool provides, for count reads and coverage at gene/amplicon or pool level, statistics such as maximum, minimum, median, quantiles, etc.

e) Generating read profiles of a particular region.

A global view of the results can be achieved with the above functions. However, the user could be interested in the exploration of a specific genomic region: a gene, an amplicon or a bp position. Thus, the *pileup* file can be summarized in order to obtain read counts at a bp level calling:

```
> getReadProfiles (pileup_file, gene_info )
```

The result is a profile matrix containing the reference sequence, matched read counts, nucleotide variant counts (identified by nucleotide), insertions and deletions counts for each position of the genomic region of interest.

f) Exploring read profiles of a genomic region.

Visual inspection of any genomic region profiles of interest, obtained in the previous section, can be explored using:

```
> plotRegions(counts, start_reg, end_reg)
```

The user can evaluate, for instance, if a particular amplicon was amplified or visualize the sequencing depth achieved.

g) Exploring bp read counts.

Detection of SNP and mutation events requires considerable amount of read counts over each genomic position. In order to explore the distributional characteristics of read counts on a genomic region, here is provided:

```
> plotReadDens(counts, start_reg, end_reg)
```

By mean of this function, the user could have a quick ac-

cess to distributional characteristics of bps in a programmatic way.

h) Exploring a particular gene.

When exploring an AES experiment in a gene by gene basis, the user can apply what is presented in *e)*, *f)* and *g)*. But, coverage uniformity check is also crucial for quality assessment and detection of underrepresented regions. By using:

```
> plotAmpliconCov(amplicon _results, gene,…)
```

it is possible to get a quick view, in a bar graph, showing the achieved coverage efficiency per amplicon.

i) Exploring results of a particular amplicon.

The same approach as in previous section can be applied in an amplicon level. However, when analyzing SNPs and mutation events, it is important to know how well they were represented (in percentages) in a specific position into the amplicon. Thus, a bar plot (each bp position per bar) is provided. Depending on the function parameters, the count read percentage of consensus/variants reads by bp against the total counts of this position, amplicon coverage or median reads can be displayed by means of:

```
> plotPercentagesAmpli (counts, gene, amplicon)
```

III. Results

In order to test the proposed tool, an AES dataset was processed (unpublished results). Tumor SKOV3 cell line was sequenced using Ion AmpliSeq Comprehensive Cancer Panel protocol (Life Technologies®). This panel contains 409 genes involved in cancer disease [2]. DNA sample was sequenced in an Ion PGM® machine under standard protocol with 4 PCR pools.

The resulting *bam* file and the *bed* file, provided by the manufacturer, where used to feed our tool. The developed functions where applied. Read counts, coverage and statistics were computed and explored for all features (bp positions, amplicons and genes).

Fig.1 shows gene and amplicon coverage exploration plots. In the left panel *join* option was set to TRUE. Thus, both boxplot and density plot are shown together. It can be seen that almost all genes were sequenced at least with 150X coverage. In addition, density function of gene coverage shows symmetry and is concentrated near the median value. This suggests that the achieved coverage agrees with both expected value and manufacturer's specifications [2].

Amplicon coverage exploration is shown on the right panel of Fig. 1. In this case, *join = FALSE* and *pool = TRUE*, enabling visualization of amplicon coverage distribution achieved for each PCR pool. All pools show similar distribution coverage behaviors as expected. This plot provides a way to detect pool, sequencing or other technical effects in an early analysis stage.

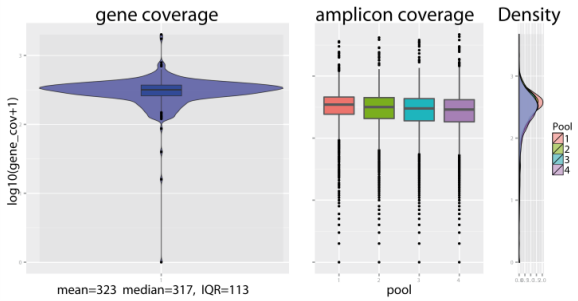

Fig. 1 gene and amplicon coverage exploration plots

On the other hand, different statistical results were obtained and summarized into table files. As an example, Table 1 shows some metrics for two genes. In it, information like chromosome, location, identifier (id) as well as distributional characteristics like the median and interquartile range (IQR), mean and coverage variance for read counts are listed.

Table 1 Gene level summary metrics.

chr	start	End	gene	Med. reads	IQR	Av. cov	sd
1	179076830	179198390	ABL2	361	148.75	357	157
1	243662991	244006498	AKT3	230	175	236	133

Med: Median, Av. cov: Average coverage, sd: standard deviation

Similar information was obtained at amplicon level as listed in Table 2. Pool identifier is also available for amplicons.

Table 2 amplicon level summary metrics.

chr	start	end	Amplicon ID	gene	pool	Med. reads	cov.
1	179076830	179076961	224233178	ABL2	1	279	294
1	179076959	179077043	224190541	ABL2	3	418	449

For genomic region exploration we chose TP53 gene, which is commonly mutated in ovarian cancer patients [6]. First, total read counts by bp within the gene were calculated. Then, consensus and variants read counts profiles were obtained, as depicted in Fig.2. In it is possible to observe some level of amplicon overlapping (read counts peaks).
Key information provided by the above figure is the presence of A and C variants near the 7.577.500 genomic position. If desired, it can be amplified using the same function call, but now defining as the interest region, the variants region.

Exploration of amplicon coverage homogeneity of TP53 gene was done. In Fig. 3 coverage of 19 sequenced amplicons was plotted, highlighting amplicon pool number. It is notable that not all amplicon coverage and pools perform equal.

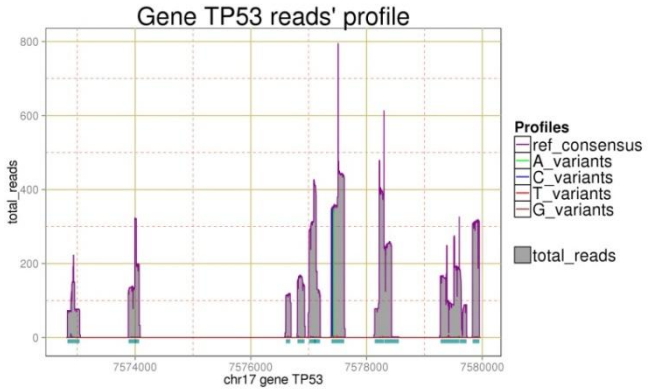

Fig. 2 Read counts profile of TP53 gene

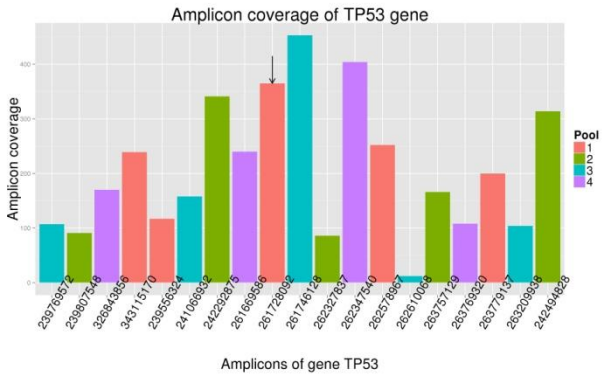

Fig. 3 Bar-plot of coverage achieved for each TP53 amplicon.

Next, we chose previous amplicon showing variants to explore. This corresponds to id 261728092, indicated in Fig. 3 by an arrow. As is observed, the chosen amplicon have a good coverage around 360X. In Fig. 4 read counts profiles over it and percentage of reads in each bp position are shown. Left panel shows similar information than Fig. 2.

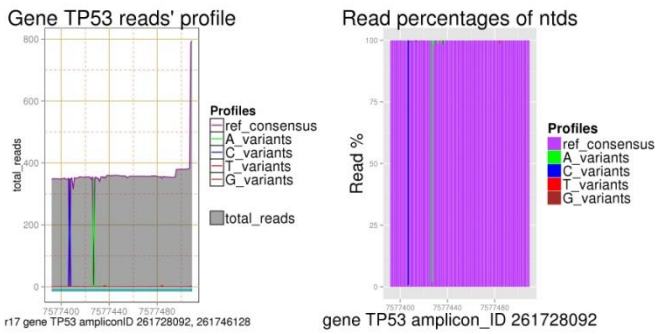

Fig. 4 exploration of amplicon 26178092 of TP53 gene.

Read counts percentages are shown on the right panel of Fig.4. In this case, bar height represents percentage of read counts against total read counts of this bp position. Then, it is possible to note that variants C and A were supported by a 95 percent of reads.

IV. Discussion

In an AES experiment performance evaluation is crucial. This implies to evaluate how well the amplicon regions were sequenced, the level of read counts achieved to perform the downstream analysis, among others. To do this, a versatile, simple, yet powerful tool was presented.

Here we demonstrate that our tool allows checking that most of the genes were sequenced and also good gene and amplicon coverage were achieved in SKOV3 experiment. By means of the proposed tool, we were able to easily and visually evaluate that amplicon coverage distributions in all PCR pools were consistent as expected.

In addition, our tool allows exploration of a particular gene. TP53 was characterized by their read counts profiles, as shown in Fig. 2. In broad terms we observed that all amplicons (light-blue segments in Fig.2) were sequenced. Amplicon coverage was encountered at least around 100X almost all amplicons (see Fig.3). Finally, by amplicon region exploration, two nearby regions exhibiting variants at 95% of reads were identified.

V. Conclusion

Tool based on an exploration pipeline for fast visualization and quality control in AES experiments was designed and tested here. We demonstrate it utility in a real scenario of an AES experiment, where run sequence evaluation, and quality control was performed.

VI. References

1. Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nature biotechnology, 26(10), 1135-1145.
2. Ion Torrent: https://tools.lifetechnologies.com/content/sfs/brochures/Ion_CompCancerPanel_Flyer.pdf
3. Lee, H. C., Lai, K., Lorenc, M. T. et al. (2012). Bioinformatics tools and databases for analysis of next-generation sequence data. Briefings in functional genomics, 11(1): 12-24.
4. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
5. Li, H., Handsaker, B., Wysoker, et al. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16): 2078-2079.
6. Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. Nature, 474(7353): 609-615.