A probabilistic approach for characterizing the marking system of multiplex sequencing in ABI SOLiD platform

F. Lobato[1], P. Machado[1], A. Gonçalves[2], Â. Ribeiro-dos-Santos[2], D. Alencar[2], S. Darret[2], Á. Santana[1]
[1]Technological Institute, Federal University of Pará, Belém, Pará, Brazil
[2]Institute of Biological Sciences, Federal University of Pará, Belém, Pará, Brazil

Abstract— *High-Throughput Sequencers such as Illumina and ABI SOLiD, generate large quantities of data, typically above 10 Gigabytes of text files. These platforms enable multiplex sequencing, that is, the sequencing of multiple samples in a single run, through a marking system. This requires a computational process for separation the data generated, which contains the mixture of all samples in a single output. It is necessary that the quality of the marking system is evaluated to ensure the reliability of this separation. This work proposes measures to characterize the marking system obtained from SOLiD sequencing. In fact, measures presented are proven to be sufficient to describe the sequencing and hence guide the process of filtering the data and the analysis of the sequencing protocol.*

Keywords: Statistical Analysis, Multiplex Sequencing, SOLiD, Barcode System.

1. Introduction

The volume of data generated by new DNA sequencers increased substantially in recent years. Platforms such as Illumina and ABI SOLiD, High-Throughput Sequencers, can generate millions of small reads sequences from different samples in a single multiplex run.

The SOLiD platform has its own peculiarities when compared to other sequencers, particularly, the data representation, which is coded in "two base color encoding", also called colorspace. This represents the transition between two nucleotides of a characteristic color obtained by the detection of fluorochromes: FAM; Cy3; TXR; or Cy5 [1]. What, in turn, implies the need to add a step for converting the data, in order to obtain the DNA sequence of nucleotide bases.

Additionally, the SOLiD sequencing supports up to 256 multiplex samples by means of, among other items, the marking system called barcode [2]. These have a characteristic central to the process of discernment between the samples, the orthogonality, meaning that a barcode of the standard library has no correlation with each other. However, even with all the security surrounding the library of markers, there were failures of common error, known as erroneous color calls, and in the quality aspects associated with the transitions of nucleotide bases [3].

These failures should be identified and, if possible, mitigated, given the importance of accuracy in the recovery of sequences per marker. However, each run held in the SOLiD platform has unique characteristics for the marking system, which ratifies the need to study methods to assess the quality of sequencing protocols.

It is important to consider the impact of a high degree of reliability for the sequencing data due to the fact that failures in barcode systems can cause a shuffle in the sequences of interest. And it implies in a waste of computer processing in genome analysis and eventual errors in results.

In this context, the lack of literature, studying measures to characterize the sequencing as the marking system, motivated this work, in which a statistical analysis is developed in order to identify summary measures to characterize the SOLiD sequencing as the marking system.

Through computational tests, it was defined four measures: median, mode, variance and sequences to barcodes ratio. The results obtained allowed demonstrating the differences in the marking system for each run analyzed. In fact, the previously mentioned measures are sufficient to describe the sequencing and hence, guide the process of filtering data and the analysis of the sequencing protocol.

This paper is organized as follows: section two presents the related work to the analysis made, section three describes the materials and methods used in the development of this work, which is presented in section four. The results obtained are discussed in section five and finally, section six presents the conclusions.

2. Correlated Works

Most studies that involve the filtering of errors or quality assessment of data generated by the SOLiD platform are based on heuristics. Taking for example the work of [4], sequences that show some transition with quality below a predetermined threshold are retained by the filter. In this same study, the default values adopted for filtering independent errors is a Quality Value (QV) <= 10, while the errors of polymorphism is QV >= 25.

Other works like [5] treat the errors of substitutions, insertions and deletions. However, this treatment does not take into account the quality value, because they filter the data in advance using a system based on heuristics.

It should be noted that such heuristics are useful for the analysis of large sequences. Heuristic-based algorithms usually have lower complexity and require less processing

time compared with analytical algorithms. On the other hand, analytical algorithms improve reliability by generating accurate results based on analysis, something required for barcodes.

The differential matter of this study is the adoption of a statistical analysis in order to assess the quality of barcodes, that allow the characterization of sequencing protocols to guide the development of a custom filter to the marking system used in the SOLiD platform.

3. Materials and Methods

The biological material used in the studies was derived from cancer patients, extracting two samples each, with the written consent for study approved by the Committee of Ethics in from the Federal University of Pará, protocol number 14052004 / HUJBB.

The sequencing process is preceded by some essential procedures: sample collection after the DNA is fragmented, as the sequencer can only read fragments from 35 to 50 base; the linkage to known sequence fragments with adapters, named P1 and P2, as shown in Figure 1.

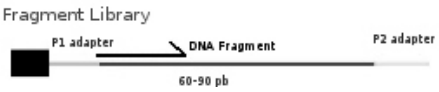

Fig. 1: Schematic drawing of sample preparation.

The adapter has a P1 sequence complementary to the metal beads; and P2 has complemented by a polystyrene coated bead, a material which floats in water; this way, the fragments that do not affect P1 and P2, will be discarded. The templates relating to the selected beads have their 3 'end modified, so they can join covalently to the blade. In the end, they are deposited on the blades and taken for sequencing, as shown in Figure 2.

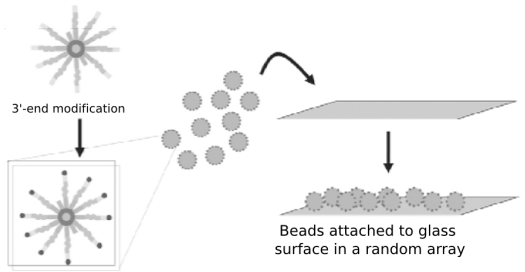

Fig. 2: Modification of the P2 adapter to allow the grip to glass surface.

The markers in each sample are inserted into the adapter P2. The SOLiD Small RNA Expression Kit (Ambion Inc., U.S.), was used for the preparation of fragments. All miRNAs were attached to the library with a specific extension of primers, in this case, the barcode system.

The data generated from three runs consists by the barcodes, samples sequences and quality values files. Together they amounted over 175 Gigabytes (Gb), which about 24 Gb correspond to information from barcodes. These data were analyzed in order to obtain sufficient statistics from the following summary measures: mean, mode and variance, in addition to viewing the probability distribution to facilitate evaluation of multiplex sequencing.

4. Statistical Analysis of Multiplex Sequencing

For this analysis, it is necessary to know the probability of the marking sequence, however, the Applied BioSystems does not provide the mapping function between the value of quality and the associated probability.

In [6] we found a function that adapted for the range of quality values generated by the SOLiD platform and approximate the quality-probability mapping, as follows:

$$P(Q) = 1 - 10^{-(Q+1)/10} \quad (1)$$

The value $P(Q)$ represents the probabilistic degree of confidence of a given transition, as evaluated by their quality value, represented in Equation 1 to (Q +1). This aspect was used for normalization, for Q is comprised in the range from -1 to 35, so the value of -1 would indicate a negative outlook.

To calculate the confidence level of a given sequence (θ), multiply the probabilities of all existing transitions. The result represents the probability that the sequence obtained is, in fact, present in the sample.

$$P(\theta) = \prod P(Q) \quad (2)$$

To optimize the calculation of summary measures and plotting the probability distribution, the results obtained by applying Equation 2 are stored in a data structure, the map [7]. This structure is composed by two fields, one for storing the key and another for the value; in this problem, the likelihood is the access key that points to the number of occurrences, so the probability distribution is easily manipulated. Other relevant information is the relationship between the total amount of marking sequences obtained and those which corresponded to one of the ten barcodes presented in the standard library used in the experiments.

5. Results

The statistics contained in Table 1 have revealed that C1, with an average of 81.23% has a higher confidence than C2 and C3 in the quality of markers and, consequently, the recovery of sequences from samples. The variance of 5.5% was the lowest compared to other runs, indicating a low dispersion of data in relation to the expected value.

It is observed that 30.88% was the most frequent likelihood in the runs C1 and C3. Moreover, C2 had higher

Table 1: Information about Sequencing Analyzed.

Data	First Run(C1)	Second Run(C2)	Third Run(C3)
Number of sequences	142,453,565	19,523,621	27,634,981
Mean	0.8123	0.4715	0.5917
Mode	0.3088	0.3726	0.3088
Variance	0.0553	0.0687	0.0806
Sequence to Barcode ratio	73,42%	1,45%	44,66%

value, of 37.26%, in the value of mode. The proportion the Sequence to Barcode ratio presented in C1 is 73.42%. C3 showed a drop to 44.66%; and in C2, the ratio is extremely low, 1.45%, indicating prior sequencing problems.

However, between C2 and C3, the latter showed the best performance, with an average of 59.17% and 8% of variance. The graphical analysis of these two runs, obtained by comparing Figures 4 and 5, evidences this difference, especially regarding the density of sequences with low confidence, presented in Figure 4 ; even though the mode of this run is higher, it does not reflects the general scenario.

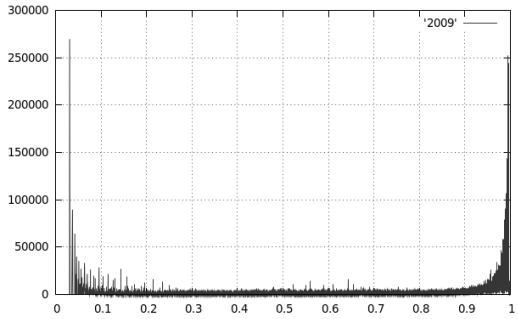

Fig. 3: Probability distribution of first run.

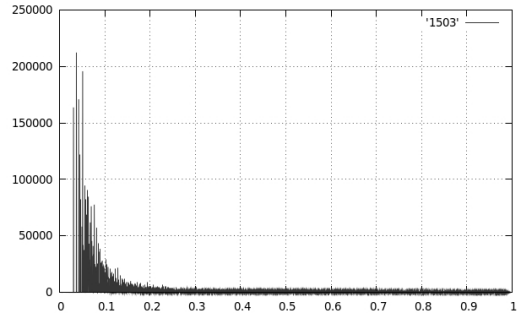

Fig. 4: Probability distribution of second run.

Such statements were pertinent to the sequencing protocol analysis and verification of possible disposal of the data generated in the case of low confidence. For example, the values for runs C1 and C3 are statistically significant, detecting at least 45% of the sequences marked, thus proving trustworthy for the genomic analysis.

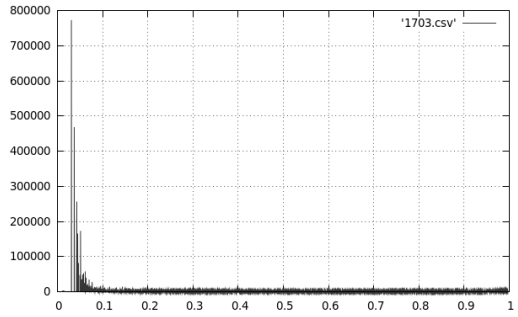

Fig. 5: Probability distribution of third run.

However, the data qualities of C2 are extremely low, with only 2% of recognized barcodes that were actually used in the sequencing; possibly indicating a shuffling of the samples. A scenario that illustrates this problem is the sequencing of two patients, one healthy and another sick; if the marking system has characteristics similar to C2, the sequences of interest can be exchanged between the patients, which causes errors in the results.

Regarding the poor quality of data from C2, among the possible causes of this failure, it can be highlighted the fluctuation of electric power on the sequencing unit .This particular failure could affected the capture of beads, as these are captured through an electromagnetic field, which is extremely sensitive to power quality.

6. Conclusions

The lack of literature on summary measures able to characterize the sequencing with regards to the marking system is one of the aspects that motivated this work. Also, we stress the importance of a high degree of reliability of these data; in particular, because the marking system failures can cause the shuffle of the sequences of interest, which implies on a waste of computational processing for genome analysis and errors in the results.

Seeking to fill these gaps, we developed a statistical analysis that had the following summary measures: median, mode, variance and sequences to barcodes ratio. This allows, among other analysis: the evaluation of protocols used in the preparation of the libraries for labeling in multiplex sequencing; assessment of possible discard of the generated data; and the initial guiding in the process of developing a custom filter for barcodes.

References

[1] H. Breu,"A theoretical understanding of 2 base color codes and its application to annotation, error detection, and error correction," *White Paper* SOLiD System, 2010.

[2] A. Biosystem, "Solid system barcoding," *application note SOLiD*, 2008.

[3] A. Valouvev and et al., "A high-resolution n nucleosome position map of c. elegans reveals a lack of universal sequence-dictated position," *Genome Res.*, vol. 18, pp. 1051–1063, 2008.

[4] A. Sassom and T. P. Michael, “Filtering error from SOLiD output,” *Bioinformatics*, vol. 26, pp. 849–850, 2010.
[5] L. Salmela, “Correction of sequencing errors in a mixed set of reads,” *Bioinformatics*, 2010.
[6] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, “Base-calling of automated sequencer traces using phred. ii. error probabilities,” *Genome Res*, vol. 8, pp. 175–185, 1998.
[7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *“Introduction to Algorithms.”* McGraw-Hill, 2002.