Real-Time Data Quality Feedback for the 5500 Series SOLiD™ Sequencers

Kathleen Perry, Sylvia Chang, Rachel Fish, Laura Lua, Lee Jones, Maria Mariano, Sarah Ngola, Min-Yi Shen, Arjun Vadapalli, Licen Xu, Subbu Yerramelli & Janet Ziegle,
Life Technologies, Foster City, CA 94404

ABSTRACT

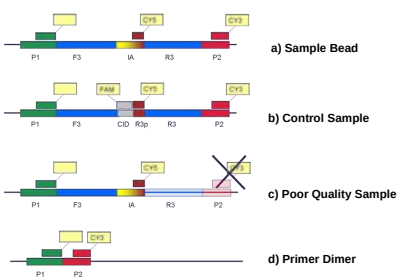
A major problem with next generation sequencing systems is the lack of feedback regarding sample and run quality. Some systems use a dedicated control lane for the run, which reduces the overall sample throughput. With the introduction of the 5500-series SOLiD sequencers, a small percentage of the total sample population is occupied by sequencing control beads, which are included as an internal control to monitor the progress of the run. This minimizes the overall impact on run yield. These beads are comprised of a set of synthetic sequences that will support all run sample types including fragment, paired-end, mate pair, barcoded fragment, whole transcriptome and smRNA. The control template sequences contain a unique identifier that perfectly distinguishes these beads from sample beads. The 5500 software provides an initial bead assessment where the sample bead quality is determined. As the run progresses, the sequencing control provides real time comparison of expected versus measured signal and run quality. Data has been collected with a variety of sample types and a broad range of sample bead quality to demonstrate the dynamic range and utility of the sequencing control beads as a reliable internal standard. Input from existing SOLiD system customers was solicited in designing novel ways to display real-time run progress and data quality. Together, the 5500 internal control and the software user-interface provide a powerful tool for monitoring and determining run quality and performance.

INTRODUCTION

A set of 1024 sequences was designed such that for each cycle of ligation there is a template matching each probe. These sequences were designed using the mate-pair construct with a control-identification sequence (CID) included to allow these beads to be distinguished from the customer sample in a mixture.

All library constructs will now contain an internal adapter sequence (IA). In the absence of this sequence, template quantity on the bead has been measured by the amount of P2 on the bead. This number, however, also includes any beads that have primer-dimer present but no template. (Figure 1)

Figure 1. Design of SOLiD system sequencing control sequences

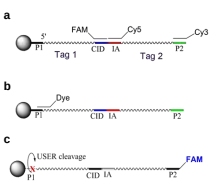


The sequence construct of the **(a)** SOLiD system sample beads containing IA and P2 tags, **(b)** SOLiD system sequencing control beads, containing CID, IA and P2 tags, **(c)** partially extended sample beads, have a good IA but a weak P2 tag, and **(d)** primer dimer, containing only P2 tags. The inclusion of the control sequence can help distinguish between populations of (a) and the undesired (b) and (c).

MATERIALS AND METHODS

The sequencing control beads were prepared by placing each of the 1024 sequences into a separate well and then performing PCR to amplify the material onto the bead. The beads were then assayed using three methods (Figure 2). We have a calibration curve that allows us to convert fluorescence units to number of copies per bead.

Figure 2. Methods used to assay sequencing control beads



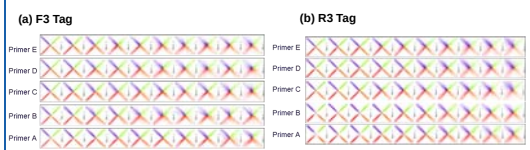
a - Hybridization assay to measure the fluorescence of probes against the CID, IA and P2 simultaneously.

b - Ligation assay to measure the fluorescence of the ligated probe as an estimate for the amount of Tag 1.

c - CE assay to measure fragment size by cleaving off FAM end-labeled material from the bead enzymatically through a deoxyU residue incorporated into P1. Not high-throughput and used in method development only.

RESULTS

Figure 3. Sequencing Test of the Sequencing Control Beads



The satay plots of the mate-pair run of SOLiD 5500's control beads. The build quality of F3 and R3 tags are highly consistent.

Table 2. Matching Statistics of sequencing control beads

MM	F3	R3
0	42.5%	50.3%
1	19.0%	20.1%
2	11.8%	11.7%
3	7.8%	7.1%
4	4.9%	3.9%
5	3.0%	2.1%
6	1.8%	1.1%
NM	9.2%	3.7%
Matching	90.8%	96.3%

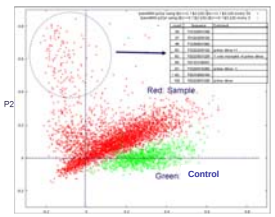
As shown in the satay plots of Figure 3, there are comparable sequencing qualities for the two tags F3 and R3 (see Figure 1). The matching has been broken down into the number of tolerated mismatches (MM) and a population of unmappable beads (NM – no match).

Table 3. Matching Statistics of control spiked into the sample beads

DH10B, Long Frag, PCR cycles	Control Beads	Sample Beads
20 cycles	96.3%	3.71%
30 cycles	89.1%	66.04%
40 cycles	95.8%	76.88%
50 cycles	93.1%	75.23%
60 cycles	93.6%	73.52%
70 cycles	94.8%	77.28%
80 cycles	94.2%	77.65%
90 cycles	93.7%	76.48%

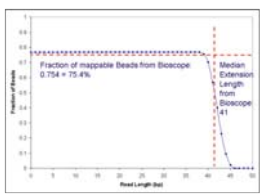
The sequencing accuracy of the control beads is independent of the quality of the sample we spiked into. In this case, the control beads are spiked into eight different samples of various qualities. The matching percentage of the control beads are maintained at ~93% while the sample beads range from 3% to 77% as a function of the number of PCR cycles

Figure 4. Primer Dimer Identification with 5500 SOLiD System control



In Figure 1, we suggested the existence of the primer dimer can be identified by the P2 and IA signals. If P2 >> IA, there is a good chance that it is a primer dimer. In this P2-IA scatter plot, the biggest fraction of the beads with P2>>IA match perfectly with the P2 sequence. In contrast, control beads do not display primer dimers.

Figure 5. The Read Length Predictor

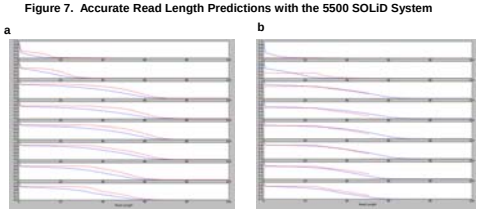


The read length distribution function is the visual representation of the quality of user sample determined by control bead models. At 25 base pairs the fraction is 0.76, meaning ~76% of the beads in this sample can be correctly read up to 25 base. About 50% of the beads can be correctly read up to 42 base pairs based on the control model. The mapping data from bioscope is 41.

Figure 6. User Interface Design

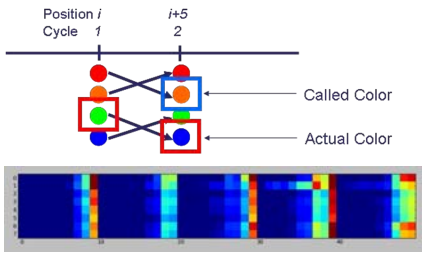
The design of the 5500 SOLiD system user interface is tightly integrated with the sequencing control. Displayed here is the prototype of the 5500 user interface. The gray and pink curves are the predicted read length distribution functions of the user sample. The distribution function is explained in Figure 5.

Figure 7. Accurate Read Length Predictions with the 5500 SOLiD System



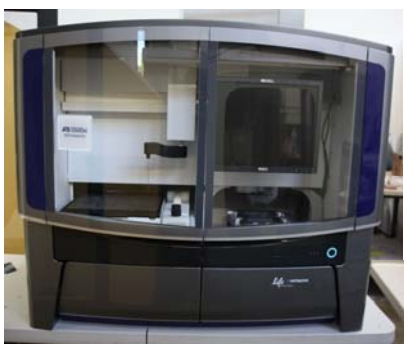
The measured P2-IA signals from sample and control beads can be used as the input features of the predictive modeling for the expected read length of customer's samples. **(a)** the upper (red) and lower (blue) limits of the read length distribution functions. **(b)** comparing the predicted read length (blue) and the actual read length (red dots) determined by Bioscope.

Figure 8. In-run error analysis with the 5500 SOLiD System control sequences



The unique design of the 5500 SOLiD System control sequences enables on-the-fly error analysis. The color transitions for any two consecutive cycles are mathematically correlated. On the example shown on the top panel, we know the "green" must be followed by "blue" by design. If "orange" is called, we know it must be introduced by a sequencing error. The bottom panel exhibits the cycle-by-cycle color call error rate heat map generated by this design, where the horizontal axis is the sequencing cycle, and the vertical axis is the eight samples with control beads spiked in. The color call heat map scale displays increasing error from blue to red, with the upper limit at 30%.

Figure 9. The 5500 SOLiD System



CONCLUSIONS

The first version of the sequencing control beads has been successfully built and tested as an internal reference and for use in predicting sample read length. In-run error analysis is an additional feature that is available with the use of sequencing controls on the 5500.

TRADEMARKS/LICENSING