SOLiD™ System Accuracy

Introduction

The throughput and scalability of next-generation sequencing platforms hold great promise for enabling applications such as large-scale resequencing, digital gene expression, and hypothesis-free chromatin immunoprecipitation (ChIP) and methylation studies. High accuracy, however, is critical for robust detection of genomic variation and will likely be a key differentiator between platforms.

The SOLiD™ System is the only next-generation platform to demonstrate overall accuracy greater than 99.94%. The performance of the SOLiD System is based on three fundamental principles—high fidelity ligase enzymology, primer reset functionality, and 2 base encoding technology. The combination of these attributes produces a highly robust system with the accuracy necessary to support high throughput variation detection.

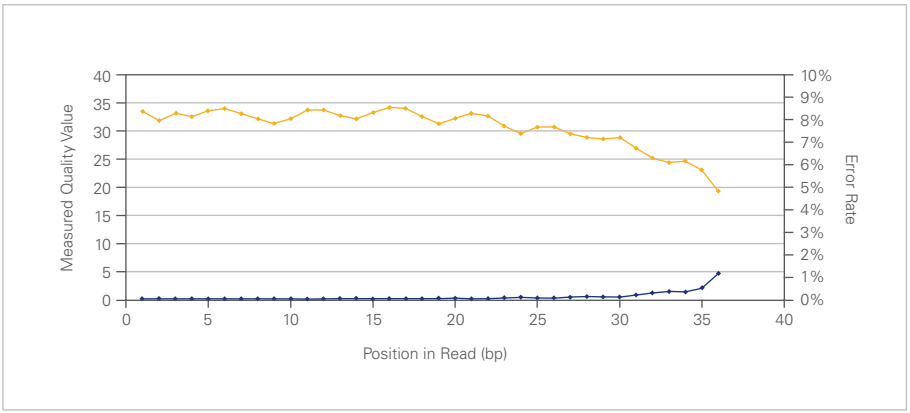

Figure 1. High quality data achieved across entire length of read. Error rates calculated per base position in sequence read after processing with 2 base encoding rules.

Accuracy and Performance of the SOLiD System

Streptococcus suis (genome size = 2 megabases) is a well characterized organism suitable to measure the performance of sequencing platforms. A fragment library was created from *S. suis* and sequencing reads were generated using the SOLiD System. The data was compared to previously generated Sanger reference sequence and error rates calculated for each position in each read after filtering out non-productive beads (Figure 1). This data demonstrates a <0.5% error rate at each individual base out to 25 base pairs. When averaged across the read length, an overall accuracy rate of 99.94% was achieved with 2 base encoding. This unrivalled accuracy, combined with mate-paired analysis, enables detection of sequence variation, including SNPs (single nucleotide polymorphisms), CNVs (copy number variations), inversions, and insertions.

Sequencing by Ligation

The SOLiD System enables massively parallel sequencing of clonally amplified DNA fragments linked to magnetic beads. The sequencing methodology is based on sequential ligation with dye-labeled oligonucleotides. In this method, the DNA sequence is generated by measuring the serial ligation of an oligonucleotide to the DNA by ligase (Figure 2).

The SOLiD System is the only platform to utilize ligation-based sequencing. Ligation-based chemistry demonstrates dramatically reduced error rates compared to polymerase-based sequencing by synthesis approaches (Table 1). The ligation reaction is based on probe recognition, not sequential addition, and is therefore less prone to accumulation of errors. The nature of the chemistry virtually eliminates the possibility of spurious insertions or deletions; an error mode that has challenged many other sequencing systems, because the probes used in the SOLiD System interrogate two bases per reaction. The improvement that this has on accuracy are described on the next page. Two base encoding is only possible using a ligation-based assay. Furthermore, the cleavage activated ligation step and phosphatase treatment of unligated probes prevents dephasing.

TABLE 1. Comparison of ligation-based vs. polymerase-based sequencing by synthesis.

	Ligase	Polymerase
Forward and reverse sequencing from single-stranded template	Yes	No
Cumulative sequential/ stepwise errors	Low constant/independent	High/cumulative
Homopolymeric sequences	Easy	Difficult

Primer Reset Functionality

Primer reset is a function that is inherent in the system chemistry and contributes to a reduction in noise. After seven cycles of ligation, the original primer is stripped from the template and a new primer is hybridized to begin interrogating at the n-1 position (Figure 2). Use of a reset phase allows for reduction in systemic noise and allows for longer read lengths.

Two Base Encoding

Two base encoding is a unique and powerful approach designed to clearly discriminate measurement errors versus true polymorphisms. Two base encoding utilizes four dyes to encode for 16 possible two base combinations. Each probe interrogates two bases, and through multiple rounds of ligation-based sequencing, each base is interrogated twice by two different dye-labeled probes.

Dual interrogation dramatically reduces sequencing errors. In the simplest case of an individual SNP, a true polymorphism will require a change at two adjacent positions in the sequence (Figure 3). Changes at a single position are identified as random errors and can be removed by the software in data analysis. Furthermore, two thirds of the two color changes may also be filtered as they would require multiple changes in the surrounding bases. In the case of more complex variations such as multiple SNPs or INDELs (insertions and deletions), more complicated algorithms are employed. The increased accuracy provided by 2 base encoding is extremely powerful when working with complex genomic DNA templates or regions where average coverage is lower than 15-fold.

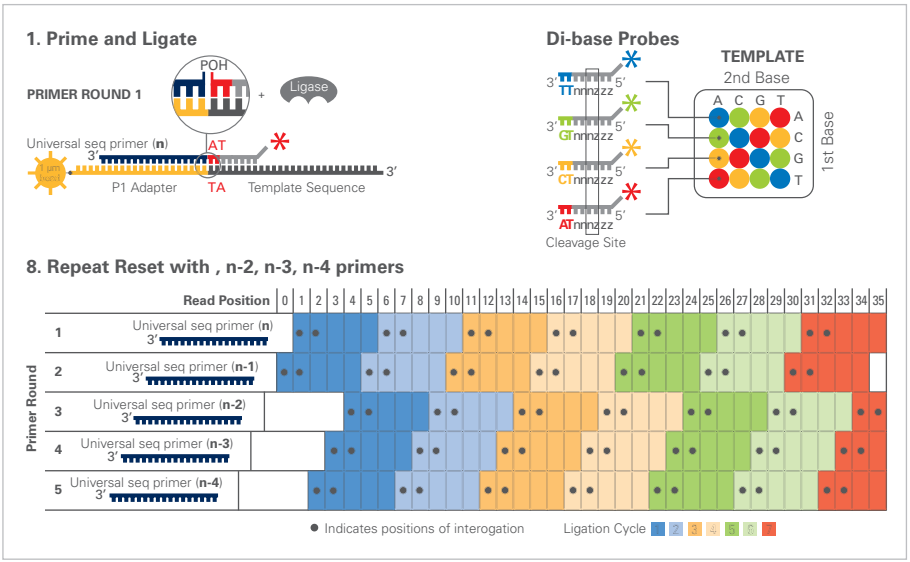

Figure 2. Dual interrogation of a single base through sequential rounds of "frameshift" ligation-based sequencing. Note the 4th and 9th bases are interrogated in both rounds one and two of cyclical ligation (blue rectangles).

Conclusion

The SOLiD System provides a robust platform to support exciting new applications such as large-scale resequencing, digital gene expression, hypothesis-free ChIP and methylation studies. The combination of high fidelity ligase enzymology, primer reset, and 2 base encoding functionality all contribute to the low error rate and reduced systemic noise associated with the system. With overall accuracy greater than 99.94%, the SOLiD System is the only next-generation platform with the performance necessary to support high throughput variation detection, somatic mutation analysis or pooled sample experiments.

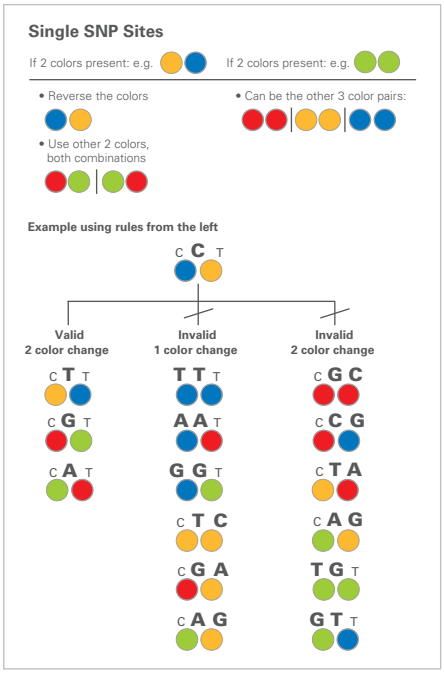

Figure 3. Principles of 2 base encoding. Random and systemic errors are easily identified and filtered out leading to higher accuracy for variation detection.

Printed in the USA. 04/2008 Publication 139AP04-04

Applied Biosystems

Headquarters
850 Lincoln Centre Drive | Foster City, CA 94404 USA
Phone 650.638.5800 | Toll Free 800.345.5224
www.appliedbiosystems.com

International Sales
For our office locations please call the division headquarters or refer to our Web site at
www.appliedbiosystems.com/about/offices.cfm