Principles of Di-Base Sequencing and the Advantages of Color Space Analysis in the SOLiD™ System

Introduction

The SOLiD™ System is the only next generation sequencing system to employ ligation based chemistry with di-base labeled probes. This unique approach enables a method which provides significant advantages in terms of system accuracy and downstream data analysis.

- Unique built-in error checking capability distinguishes between measurement errors and true polymorphisms
- Ability to detect more complicated genomic variation such as adjacent SNPs, insertions, deletions and structural rearrangements
- Final data reported as standard base calls relative to provided reference sequence

The process of 2 base encoding and the benefits of performing analysis in the di-base alphabet (a.k.a. "color space") are described below.

Principles of Ligation Based Chemistry and 2 Base Encoding

The SOLiD™ System enables massively parallel sequencing of clonally amplified DNA fragments linked to beads. This unique sequencing methodology is based on sequential ligation of dye labeled oligonucleotide probes whereby each probe queries two base positions at a time. The system uses four fluorescent dyes to encode for the sixteen possible two base combinations. Multiple ligation cycles

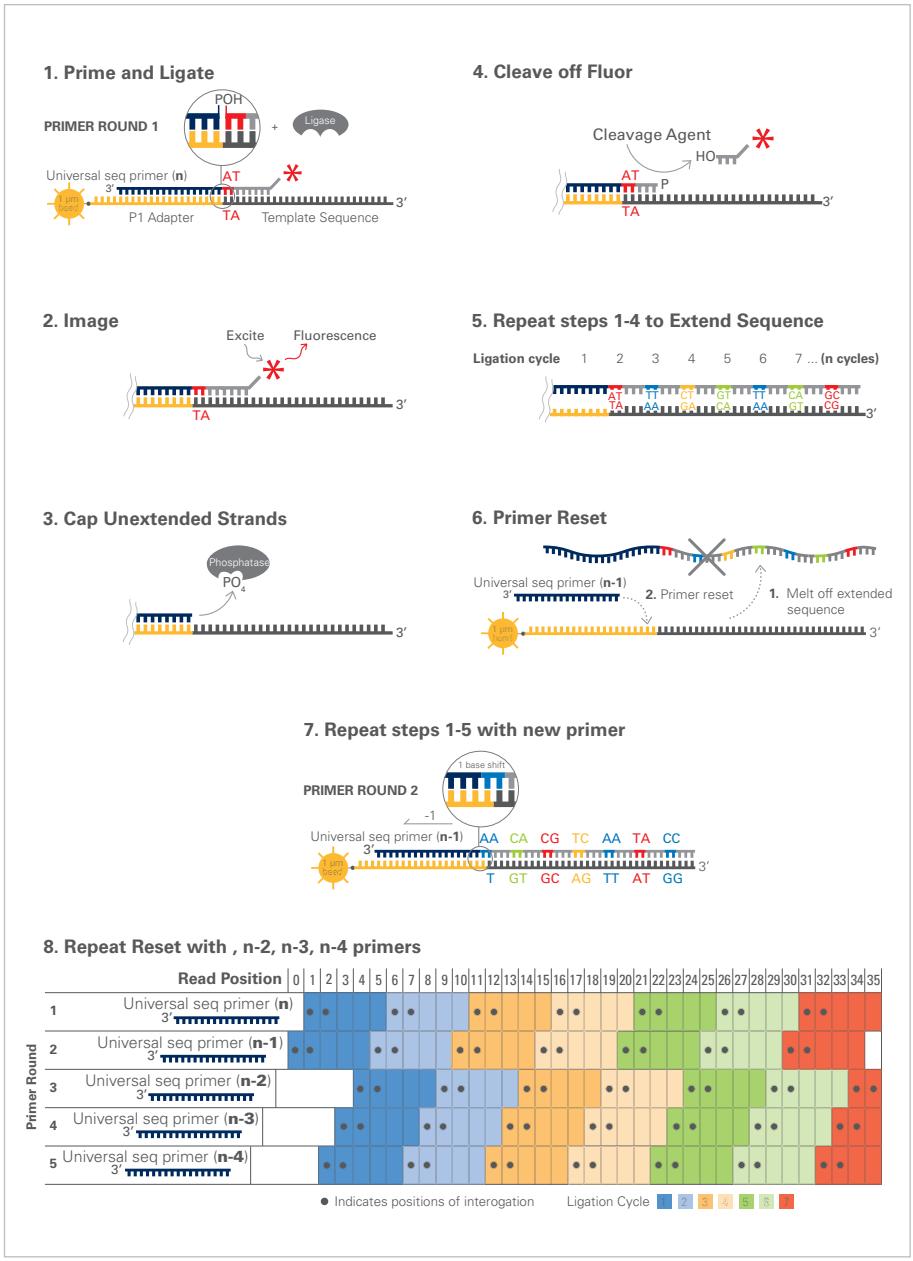

Figure 1: SOLiD™ System – Sequencing by ligation using di-base labeled probes

of probe hybridization, ligation, imaging and analysis are performed to extend the strand from a primer hybridized to a ligated adaptor proximal to the immobilized bead (P1 adaptor). The resulting product is then removed and the process repeated for 5 more extension rounds with primers hybridized to positions n-1, n-2 etc., in the P1 adaptor (Figure 1 on previous page).

There are several fundamental properties unique to ligation based sequencing which contribute to the high accuracy inherent to the SOLiD system. The advantages of these properties and their contribution to data quality are described below.

1. Two bases are interrogated in each ligation reaction providing increased specificity.
2. The primer is periodically reset for 5 independent rounds of extension improving the signal to noise ratio of the system.
3. Each base is interrogated twice in independent primer rounds providing increased confidence in each call.
4. Four dyes are used to encode for sixteen possible two base combinations. The design of the encoding matrix enables built in error checking capability.

Fundamentals of SOLiD Color Space

"Color space" is not a new concept. Data from Sanger sequencing is also encoded in color space by the four dyes used in the sequencing chemistry and displayed as peaks in an electropherogram. One difference between Sanger and SOLiD color space is that previously, in Sanger sequencing, each color represented only a single nucleotide and was automatically translated to A,C,G or T. With the SOLiD System each color now represents 4 potential two base combinations and the conversion into nucleotide base space is usually done after the sequence is aligned to a reference genome transcribed in color space. Alternatively, translation can occur following generation of a consensus

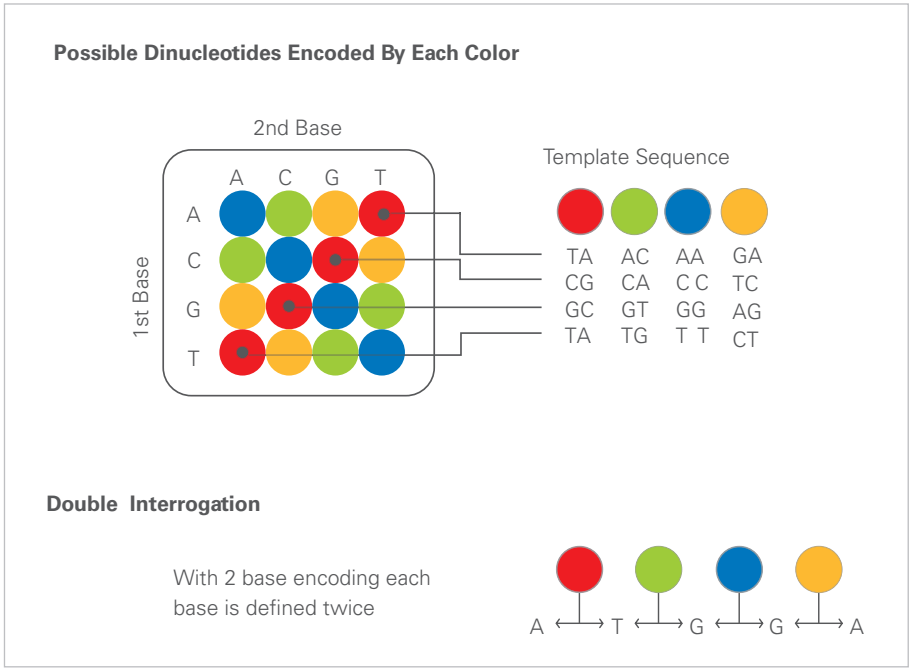

Figure 2. SOLiD Color Space Code — Four dyes encode for sixteen potential two base combinations

TABLE 1. COLOR SPACE TRANSITIONS

	Blue	Green	Yellow	Red
Blue	BB	BG	BY	BR
Green	GB	GG	GY	GR
Yellow	YB	YG	YY	YR
Red	RB	RG	RY	RR

All 2 color possibilities are represented for the four possible dyes. Four sets of allowable changes are classified by color. Changes that fall outside a particular set can be classified as possible errors for further analysis.

sequence. Since this system uses 4 fluorescent dyes, there are 16 possible two color combinations (Figure 2).

Please note that the SOLiD color space schema was specifically designed to have the following properties which enable the unique error checking capability of the SOLiD™ System.

- For each di-base the reverse (e.g., CA and AC) is always in the same color
- For each di-base the complement (e.g., CA and GT) is always in the same color
- For each di-base the reversed complement (e.g., CA and TG) is always in the same color

Advantages of 2 base encoding and color space — higher accuracy for SNP detection

Two base encoding provides higher system accuracy and built-in error checking capability which enables discrimination between measurement errors and true polymorphisms. Since each base is interrogated twice in independent reactions, the information about each base is included in two adjacent pieces of color space data. There are 16 possible two color combinations that may encode for a transition at any given single base position (Table 1). A transition at that position will result in a characteristic change that is limited to a subset of all the possible positions. This restriction allows for anomalies to be easily detected and discarded as errors.

For example, in the simplest case, where you have a single base change, the following rules apply. For any given reference, there are only 3 valid two color changes. The other 12 possibilities — 6 single color and 6 two color changes — are invalid as they would require multiple changes within that genomic region (Figure 3).

The statistical power of this method is tremendous for two reasons;

(1) 4/5s of the possible changes can be immediately filtered out

(2) The probability of having two adjacent totally independent errors is extremely low

The SOLiD System also has the capability to detect more complicated genomic variation such as adjacent SNPs, insertions, deletions, structural rearrangements, etc. (Figure 5 on the following page). For these more complicated scenarios, more complex algorithms are employed.

NOTE: Color space rules are automatically applied within the SOLiD Analysis Tools during secondary analysis. Users are not required to apply these filters manually.

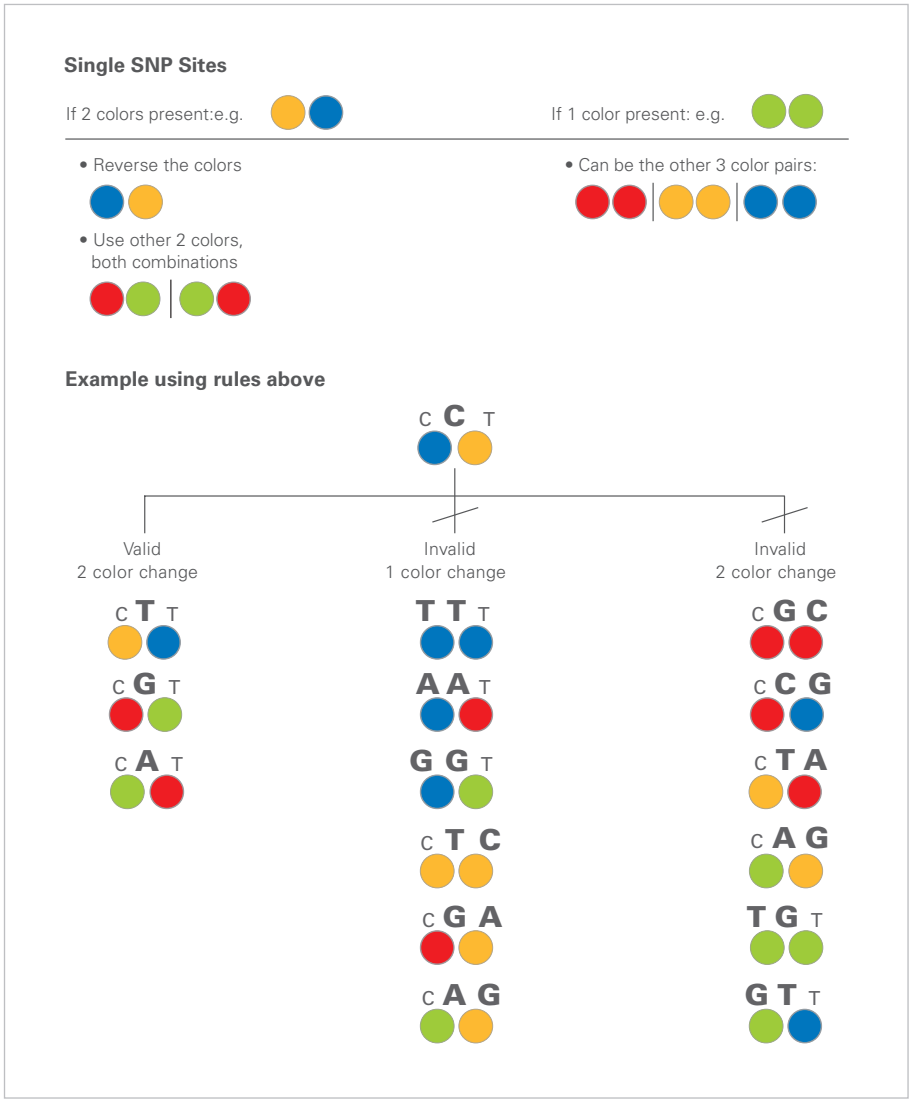

Figure 3. Color Space Rules for Single SNP sites.

Conversion of data from color space to base space

In principle, it is possible to convert data from color space to the corresponding nucleotides or "base space" with knowledge of the identity of any base in the read (Figure 4). In practice however, data from the SOLiD System is automatically aligned to a reference sequence, translated to color space and then analyzed within the SOLiD Data Analysis Tools. It is only after analysis and application of the 2 base encoding rules that the software translates the data back into base space and reports it in traditional file formats for downstream analysis.

NOTE: Conversion from color space is done within the system and final data is reported as standard base calls. Users are not required to convert data manually.

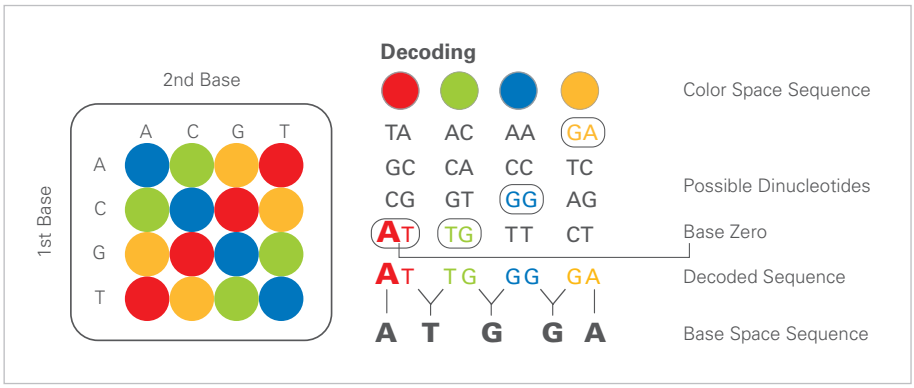

Figure 4. Decoding color space data. In principle, it is possible to convert data from color space to the corresponding nucleotides or "base space" with knowledge of the identity of any base in the read. For example, if we know A is at position 0, and the color positions 0/1 is red, then T is the only possible base that could be at position 1. Furthermore if base 1 is T and the second color is green then position 2 must be a G, and so on.

Primary and secondary data analysis in color space

The SOLiD Data Analysis Tools provide a suite of applications for primary and secondary analysis. Reference sequence may be uploaded and internally converted to a color space reference to facilitate analysis of SOLiD data. Reads from the SOLiD Analyzer are then aligned to the reference and 2 base encoding rules applied. Analysis of the data in color space enables the detection of measurement errors by the application of 2 base encoding rules.

Conclusions

The SOLiD™ System is the only next generation sequencing system to employ ligation based chemistry with di-base probes. This unique approach provides significant advantages in terms of system accuracy and downstream data analysis. Each base is interrogated in two independent primer rounds providing increased confidence in each call. The periodic primer reset contributes to improved signal to noise ratios and enables the potential for longer read lengths. Employment of this 2 base encoding strategy provides higher sequencing accuracy and inherent error checking capability. Analysis of data in color space enables the detection of measurement errors and lowers the false positive rate for variation detection.

Figure 5. Examples of polymorphisms in color space.

Printed in the USA, 04/2008 Publication 139AP10-01

Applied Biosystems

Headquarters
850 Lincoln Centre Drive | Foster City, CA 94404 USA
Phone 650.638.5800 | Toll Free 800.327.3002
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