SOLiD™ Sequencing and 2-Base Encoding

Poster Number 2624

Heather E. Peckham[1], Stephen F. McLaughlin[1], Jingwei N. Ni[2], Michael D. Rhodes[2], Joel A. Malek[1], Kevin J. McKernan[1] and Alan P. Blanchard[1]
1. Applied Biosystems, 500 Cummings Center, Beverly, MA 01915
2. Applied Biosystems, 850 Lincoln Centre Dr, Foster City, CA 94404

ABSTRACT

The next generation of DNA sequencing platforms produces sequencing reads with increased depth of coverage but reduced read length and lower per-base accuracy than data from Sanger-based DNA sequencing. New approaches are needed to overcome these issues and provide accurate mutation discovery and consensus sequences. 2-Base encoding is uniquely enabled by the ligation-based sequencing protocol used in the SOLiD™ system (a massively parallel sequencing technology based on ligation of oligonucleotides). Sequencing is carried out *via* sequential rounds of ligation with high fidelity and high read quality. In this system there are 16 dinucleotide combinations with 4 fluorescent dyes, each dye corresponding to a probe pool of 4 dinucleotides per pool. Using this dinucleotide, 4-dye encoding scheme in conjunction with a sequencing assay that samples every base, each base is effectively probed in two different reactions. The double interrogation of each base causes a SNP to result in a two-color change while a measurement error results in a single color change. In addition, only one-third of all possible two-color combinations are considered valid and result in a base change. 2-Base encoding rules (a single mismatch is a measurement error, only one-third of adjacent mismatches are valid) significantly reduce the raw error rate (30 bp reads have a 45x reduction in raw measurement errors) and this benefit increases 3/2 as the read length is increased. The reduction in raw error rate enabled by 2-base encoding translates into more accurate alignment of short reads, polymorphism discovery and consensus calling.

What is 2-Base Encoding?

The SOLiD Sequencing System uses probes with dual base encoding.

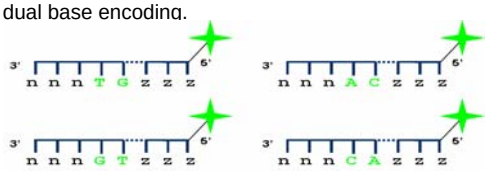

Figure 1. Each probe consists of 8 bases. As shown, the first 3 bases are degenerate (n), and the last 3 are universal (z), with the 4th and 5th bases as the two bases being interrogated. Thus, a single color observation only limits the potential dinucleotide to being four out of the 16 possible dinucleotides. As seen above, a green signal represents a AC, CA, TG or GT.

Double Interrogation

Using this dinucleotide, 4-dye encoding scheme in conjunction with a sequencing assay that samples every base, each base is effectively probed in two different reactions.

Figure 2 demonstrates the principle of double interrogation. Each color measurement represents four possible dinucleotide combinations. For example, the first measured blue represents 'AA' and the third blue represents 'CC'.

Color Space

In order to use 2-base encoding the concept of color space must be used. Instead of using a nucleotide-based reference sequence, a color space reference sequence is used. As color space and base space both consist of four elements (four colors represented as 0, 1, 2, or 3 and A, C, G or T, respectively) existing algorithms can be used for alignment and consensus calling of color space. As will be demonstrated, the properties of 2-base encoding allow significantly enhanced results if 2-base encoding is taken into account and expanded algorithms used.

Decoding

To decode a sequence the decoding matrix in figure 3 is used:

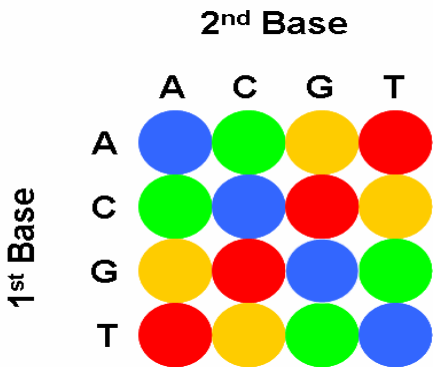

Figure 3. The decoding matrix allows a sequence of dinucleotides to be converted to a base sequence, as long as one of two bases is known. The design of encoding probes has been carefully made, as can be seen by the reversed transition (e.g., A -> T and T -> A is the same color as is the complement A -> G and T -> C).

Single Base Insertions/Deletions

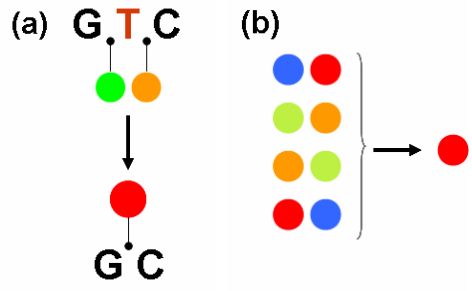

Figure 4. (a) If a deletion occurs in the sequence GTC the result has to be GC. The number of observed transitions will decrease from 2 to1. The single transition must be a G to C thus giving a signature to the event. **(b)** The reverse is true if a single base insertion occurs with the result that only 4 of the potential adjacent transitions can occur for any individual starting transition.

Single Nucleotide Polymorphisms (SNPs)

In many resequencing projects one of the most important objectives is to measure Single Nucleotide Polymorphisms (SNPs) that may be responsible for differences in phenotype. In 2-base encoding most measure errors can be distinguished from potential SNPs as demonstrated below in figure 5:

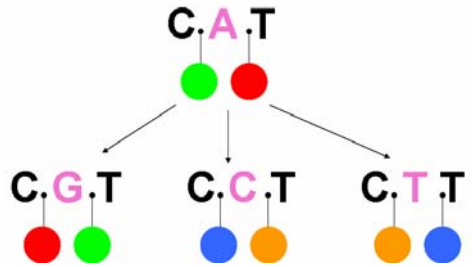

Figure 5. If a SNP occurs in the sequence 'C-A-T' there are only 3 possible results: CGT, CCT and CTT. This means that only 3 dibase combinations are allowed and any other dibase combinations are illegal. Since any base is defined by two nucleotides (e.g., C-A and A-T), then two adjacent changes must be observed for any SNP. Thus, measurement errors are represented by single changes. As there are only 3 alternative bases that can occur when a SNP is observed (i.e., an A can go to C, G or T), there are only three allowed dibase combinations for any starting adjacent transition. The other six possible adjacent combinations are therefore by definition invalid. Thus, when two adjacent measurement errors are seen, only 1/3 of them could be mistaken for a real SNP, prior to applying any consensus rules. Since the two surrounding combinations contain information about the incorrect combination it is possible to have support for the hypothesis that the reference sequence is unchanged even if a single changed combination is seen and discarded.

SOLiD System™ Accuracy

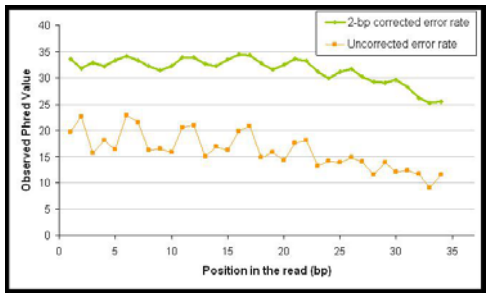

Figure 6: SOLiD™ System's error rate per base position in sequence read.

Conclusion

The ability to use 2-base encoding to recognize and eliminate measurement errors from subsequent analysis has been demonstrated. In numerous experiments, a minimum error reduction of 20-fold has been seen. Only sequencing by ligation offers the ability to use 2-base encoding. Thus, SOLiD sequencing systems offer the best solution to many applications.