An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System

Brian D. Ondov1,2, Charles Cochran3, Mark Landers4, Gavin D. Meredith4, Miroslav Dudas4 and Nicholas H. Bergman1,2,*

1National Biodefense Analysis and Countermeasures Center, 110 Thomas Johnson Drive, Frederick, MD 21702, 2School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332-0230, 3Life Technologies, 850 Lincoln Centre Drive, Foster City, CA 94404 and 4Invitrogen, a division of Life Technologies Corporation, Genetic Systems Business Unit, 5791 Van Allen Way, Carlsbad, CA 92008, USA

ABSTRACT

Summary: Bisulfite sequencing allows cytosine methylation, an important epigenetic marker, to be detected via nucleotide substitutions. Since the Applied Biosystems SOLiD System uses a unique di-base encoding that increases confidence in the detection of nucleotide substitutions, it is potentially advantageous platform for this application. However, the di-base encoding also makes reads with many nucleotide substitutions difficult to align to a reference sequence with existing tools, preventing the platform’s potential utility for bisulfite sequencing from being realized. Here, we present SOCS-B, a reference-based, un-gapped alignment algorithm for the SOLiD System that is tolerant of both bisulfite-induced nucleotide substitutions and a parametric number of sequencing errors, facilitating bisulfite sequencing on this platform. An implementation of the algorithm has been integrated with the previously reported SOCS alignment tool, and was used to align CpG methylation-enriched Arabidopsis thaliana bisulfite sequence data, exhibiting a 2-fold increase in sensitivity compared to existing methods for aligning SOLiD bisulfite data. Availability: Executables, source code, and sample data are available at http://solidsoftwaretools.com/gl/project/socs/

Contact: bergmann@nbacc.net

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on March 19, 2010; revised on May 28, 2010; accepted on May 28, 2010

Cytosine methylation is a major epigenetic marker in eukaryotes, performing functions such as transcriptional regulation and transposon silencing. It is now possible to create genome-wide maps of this type of DNA modification at single-nucleotide resolution using a technique termed bisulfite sequencing, or BS-Seq. The method utilizes high-throughput sequencing technologies in conjunction with selective nucleotide substitutions. These substitutions are induced by bisulfite, which converts cytosine residues to uracil residues, but occurs at a much slower rate for 5-methylcytosine (the most common type of methylated cytosine). After an appropriate amount of bisulfite treatment and subsequent PCR amplification, 5-methylcytosine residues will be represented by cytosine (or guanine on the complementary strand) and cytosine residues by thymine (or adenine on the complementary strand). By sequencing the converted DNA and aligning with a reference sequence, methylation can be inferred from these substitutions (Frommer et al., 1992).

While any sequencing method can be employed for BS-Seq, the Applied Biosystems SOLiD System is attractive for this application because it is designed around the reliable detection of nucleotide substitutions. This reliability arises not from the absence of sequencing errors, but from the ability to discern most of them from true substitutions. The system achieves this by querying overlapping dinucleotides rather than single nucleotides, such that each nucleotide of each read is ultimately queried by two independent ligation events. The caveat is that each dinucleotide is reported as a color that could represent any of the four dinucleotides, and alignment must be performed using these colors (in ‘color-space’). In order for sequencing errors to be distinguished. In color-space, a nucleotide substitution appears as a specific pattern of two adjacent color-space mismatches. This increased divergence is not a major problem for applications in which reads will typically only contain one nucleotide substitution, such as single nucleotide polymorphism (SNP) detection. However, in BS-Seq, bisulfite-induced nucleotide substitutions (BINS) are ubiquitous, causing most reads to contain too many color-space mismatches relative to the reference sequence to be aligned using the standard color-space alignment tools.

There are ways to avoid this issue and align a portion of the reads with standard tools. One is to create reference sequences that represent the original sequence and complete bisulfite conversion of both the Watson and Crick strands of the sequence, as done in previous bisulfite experiments (Lister et al., 2008). Reads can then be aligned using existing SOLiD alignment tools. The problem with this approach is that reads containing both methylated and unmethylated cytosines could contain numerous color-space mismatches against either reference sequence. Since areas such as CpG islands are dense with potential methylation sites, many reads that are of particular interest would not be aligned within realistic error tolerances. The other possible method is to convert SOLiD reads from their di-nucleotide encoding into nucleotide strings. The reads can then be aligned with an existing tool that is tolerant of BINS, such as the one developed by Cokus et al. (2008). The issue here is that the conversion process assumes the absence of sequencing errors,
As a control, the reads were first aligned using the alignment tool provided by Applied Biosystems (mapreads) against reference sequences representing the fully bisulfite converted Watson and Crick strands and the unconverted Watson strands of Arabidopsis thaliana and phage lambda. This approach is analogous to that employed by previous BS-Seq studies (Lister et al., 2008). Alignment was then performed against only the unconverted genomes using SOCS-B, which exhibited a 2-fold increase in total sensitivity for reads with three or fewer errors (Table 1). Using reads that uniquely aligned to the lambda genome, the bisulfite conversion rate was estimated to be 99%, indicating that the increase in sensitivity (and thus the abundance of heterogeneously converted reads) was due to complex methylation patterns, rather than incomplete conversion. Furthermore, since the most biologically relevant methylation sites in the genome occur in dense clusters, it seems possible that the additional reads aligned by SOCS-B might be of more biological significance than those aligned with mapreads. Because of the algorithm’s inherent lack of bias, SOCS-B also showed increased specificity when aligning simulated reads (Supplementary Table S2).

SOCS-B alignment took 30 h using an Apple Mac Pro (dual 2.93 GHz Quad-Core Intel Xeon with hyper-threading, 32 GB RAM). Since color-space errors are fairly abundant in SOLiD reads, more reads can be aligned by allowing more mismatches, either at the expense of run time (by increasing the sensitivity) or of specificity (by increasing the tolerance). For larger reference genomes or datasets, or for higher error sensitivity, SOCS has features that facilitate distributed processing. Executable versions, source code, sample datasets, and usage instructions are available at http://solidsoftwaretools.com/gf/project/socs/.

**ACKNOWLEDGEMENTS**

We thank Ryan Lister and Joe Ecker of the Salk Research Institute (La Jolla, CA, USA) for providing A. thaliana genomic DNA, Hank Tu of Life Technologies (Foster City, CA, USA) for assisting with distributed computing, and George Marnellos (Invitrogen, Carlsbad, CA) for helpful discussions.

**Funding:** DHHS contract N266200400059C/N01-AI-40059.

**Conflict of Interest:** none declared.

**REFERENCES**

Cokus,S.J. et al. (2008) Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452, 215–219.

Fromme,M. et al. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl Acad. Sci. USA.*, 89, 1827–1831.

Karp,R.M. and Rabin,M.O. (1987) Efficient randomized pattern-matching algorithms. *IBM J. Res. Dev.*, 31, 249–260.

Lister,R. et al. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133, 523–536.

Ondov,B.D. et al. (2008) Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics*, 24, 2776–2777.

| Errors permitted | Mapreads (reads aligned) | SOCS-B (reads aligned) | SOCS-B increase factor |
|------------------|-------------------------|-----------------------|------------------------|
| 0                | 1 150 578               | 8 701 800             | 7.56                   |
| 1                | 3 283 347               | 13 856 042            | 4.22                   |
| 2                | 6 091 811               | 18 764 850            | 2.80                   |
| 3                | 11 139 673              | 22 656 148            | 2.03                   |

Table 1. Sensitivity of SOCS-B in aligning SOLiD bisulfite sequence data

|| DOGS-Seq (reads aligned) | DOGS-B (reads aligned) | DOGS-B increase factor |
|--------------------------|------------------------|------------------------|
| 0                        | 1 150 578              | 8 701 800              | 7.56                   |
| 1                        | 3 283 347              | 13 856 042             | 4.22                   |
| 2                        | 6 091 811              | 18 764 850             | 2.80                   |
| 3                        | 11 139 673             | 22 656 148             | 2.03                   |

Table 2. Trends in alignment rate and the alignment quality with DOGS-B and DOGS-Seq.