BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 23 2008, pages 2776–2777
doi:10.1093/bioinformatics/btn512

Sequence analysis

Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications

Brian D. Ondov[1], Anjana Varadarajan[1], Karla D. Passalacqua[1] and Nicholas H. Bergman[1,2,*]

[1]School of Biology, Georgia Institute of Technology, 310 Ferst Dr., Atlanta, GA 30332-0230 and [2]Electro-Optical Systems Laboratory, Georgia Tech Research Institute, 925 Dalney St., Atlanta, GA 30332-0810, USA

Received on May 2, 2008; revised on September 26, 2008; accepted on October 1, 2008

Advance Access publication October 7, 2008

Associate Editor: John Quackenbush

ABSTRACT

Summary: Here, we report the development of SOCS (short oligonucleotide color space), a program designed for efficient and flexible mapping of Applied Biosystems SOLiD sequence data onto a reference genome. SOCS performs its mapping within the context of 'color space', and it maximizes usable data by allowing a user-specified number of mismatches. Sequence census functions facilitate a variety of functional genomics applications, including transcriptome mapping and profiling, as well as ChIP-Seq.

Availability: Executables, source code, and sample data are available at http://socs.biology.gatech.edu/

Contact: nickbergman@gatech.edu

Supplementary information: Supplementary data are available at *Bioinformatics* Online.

Recent advances in DNA sequencing technology have made it possible to collect sequence data on a much larger scale than in previous years, and several sequencing platforms are now capable of generating >1 Gb of sequence data in a single run. Although *de novo* genome sequencing with these systems remains a challenge because of difficulties in assembling short reads, their extremely high throughput makes next-generation sequencing methods an increasingly attractive option for a variety of functional genomics applications, including transcriptome profiling, global identification of protein–DNA interactions and single nucleotide polymorphism (SNP) discovery. Several recent studies have demonstrated the feasibility and advantages of a sequencing-based approach to these applications (Johnson *et al*., 2007; Nagalakshmi *et al*., 2008; Torres *et al*., 2008; Wilhelm *et al*., 2008). Although there are computational challenges in dealing with the massive volumes of data produced by these systems (chiefly in mapping individual sequence reads to a reference genome), there has been significant progress made in these areas as well (Li *et al*., 2008; Smith *et al*., 2008), and overall it appears that high-throughput sequencing will be an increasingly powerful option for functional genomics.

One of the newest next-generation sequencing platforms is the Applied Biosystems SOLiD system. This platform generates significantly more sequence data than previously described systems—6 or more Gb per run, in 25–35 nt reads—and uses a unique ligation-mediated sequencing strategy that is less prone to some of the problems that have been associated with high-throughput sequencing-by-synthesis strategies, such as inaccurate recording of homopolymer sequences (Shendure *et al*., 2005, see Applied Biosystems website for a complete description of the platform). In addition, the SOLiD system uses a two-base encoding scheme in which each data point represents two adjacent bases, and each base is interrogated twice, which helps in discriminating between sequencing errors and true polymorphisms. Collectively, these attributes make the SOLiD sequencing system particularly well suited to a variety of functional genomics applications.

In contrast to other sequencing systems, SOLiD data are not collected directly as DNA sequences, but instead are recorded in 'color space', in which the individual values (colors) within a read provide information about (but not a definite identification of) two adjacent bases. Without a decoding step, in which color data are converted to sequence data, they cannot be mapped to a reference genome using conventional alignment tools. Direct conversion of color data to sequence data, however, has a significant drawback—reads that contain sequencing errors cannot be converted accurately (in translating a color space string, all bases after a sequencing error will be translated incorrectly). Given this, there is a clear incentive to map sequence reads to a reference genome within color space, and there have been several software tools developed recently to perform this task [e.g. MAQ (http://maq.sourceforge.net/), Shrimp (http://compbio.cs.toronto .edu/shrimp/), Mosaik (http://bioinformatics.bc.edu/marthlab /Mosaik), as well as ABI's SOLiD Alignment Browser].

One of the challenges facing these alignment tools is that the ABI SOLiD system, like other ultra high-throughput short read sequencing systems, has an error rate that is significantly higher than traditional Sanger sequencing, and sequence reads containing one or more mismatches relative to the reference genome are very common in SOLiD datasets. These reads are much more difficult to map than reads that match the reference exactly, and because of this, existing tools generally only map reads that have $\leq$3 mismatches relative to the reference genome. This allows for rapid runtimes, but also leaves a sizable fraction (>50% in some cases) of each dataset unused. Since much of these remaining data can be unambiguously

*To whom correspondence should be addressed.

Table 1. Performance of SOCS in mapping SOLiD sequence data

Mismatch tolerance	Time required	Number of additional reads mapped (percentage)
0	10.3 min	4 004 404 (14.3%)
1	11.9 min	4 664 183 (16.7%, 31.0% total)
2	15.7 min	3 583 141 (12.8%, 43.8% total)
3	35.4 min	2 706 247 (9.7%, 53.5% total)
4	3.5 h	2 054 061 (7.4%, 60.9% total)
5	22.1 h	1 594 608 (5.7%, 66.6% total)

SOCS was tested using a sample dataset containing 27 942 602 35-bp reads generated by the SOLiD sequencing system. The reads were drawn from an experiment in which an mRNA sample isolated from *B. anthracis* was sequenced, and they were mapped to the *B. anthracis* Ames Ancestor genome sequence. SOCS was run on an Apple Mac Pro (2×3.0 GHz Dual-core Xeon, 4 GB of RAM). Times shown are the totals required for both mapping and scoring functions at the specified mismatch tolerance, and they reflect a single-threaded execution. Multithreading improved overall runtimes considerably, particularly at mismatch tolerances ≥3.

mapped despite having ≥4 mismatches (Table 1), and are therefore useful for sequence census methods, we sought to develop a tool that would allow mapping of SOLiD sequence data in a more flexible, mismatch-tolerant context that would maximize the number of usable sequences within a given dataset.

Here, we describe SOCS (short oligonucleotide color space), a program for efficient mapping of SOLiD sequence data to a reference genome within color space. SOCS is built on an iterative variation of the Rabin–Karp string algorithm (Karp and Rabin, 1987), which uses hashing to accelerate the process of matching sequence reads to the reference genome (see Supplementary Material for a more extensive description of the algorithm). Our hash function enumerates a subset of the sequence being hashed using 2 bits per color (the size of the subset is constrained by memory limitations on the hash table). The overall algorithm is similar to that used by software tools developed for analysis of Illumina-Solexa data (Li *et al*., 2008; Smith *et al*., 2008); briefly, to match all sequence reads with n mismatches relative to the reference genome, $n+1$ partial hashes are used, which ensures that at least one partial hash will match a partial hash from the reference string. The mismatch tolerance is specified by the user, with higher tolerances resulting in more usable data and longer run times (as the tolerance increases, the fragments used for each partial hash get smaller, and thus their hashes are less unique). To help offset this time increase, SOCS maps at lower tolerances first, reducing the data to be mapped at higher tolerances.

During the mapping process, if a read maps to two or more non-identical genomic substrings within the maximum tolerance, quality scores and mismatch counts are used in determining the optimal match (see Supplementary Material). If the genomic substrings are identical, all matching locations are recorded and flagged as ambiguous. Once optimal matches are determined, coverage maps of each reference chromosome are calculated. For each read mapped, the coverage scores of the nucleotides covered by that read are increased by 1. Essentially, each coverage score represents the number of times a given nucleotide in the reference genome is represented within the pool of sequence reads (with each strand considered independently). Scores for reads flagged as ambiguous are recorded in a separate file—in this way, unambiguously mapped data can be kept separate from data for which uncertainty exists. Finally, to aid in SNP discovery, SOCS finds all color space differences that indicate isolated mismatches between the sequenced nucleotides and the reference genome. The position and base transition of the indicated mismatches are recorded in an additional set of score files.

We tested SOCS using a SOLiD dataset obtained in sequencing an mRNA sample isolated from *Bacillus anthracis*. Our test dataset contained 27 942 602 reads, and we mapped them to the *B. anthracis* Ames Ancestor genome obtained from GenBank. The times required for each iteration of the algorithm are shown in Table 1, along with the number of reads successfully mapped at each step. The times required at a mismatch tolerance of ≤3 are comparable to those reported for other recently developed tools (Li *et al*., 2008), and it should be noted that although setting the tolerance above three results in a significantly increased run time, the amount of usable sequence data increases dramatically as well. A mismatch tolerance of five, for instance, yields 24.5% more usable data than a tolerance of three, and a mismatch tolerance of eight yields 65.8% more data (data not shown). This is a significant advantage for applications such as transcriptome profiling, where sequencing errors or polymorphisms are irrelevant as long as each read can be unambiguously mapped to the genome.

SOCS is written in C++, and runs well on Mac OS and Linux/Unix systems. The program supports multithreading, and is able to use multiple processors efficiently (mapping at a tolerance of five mismatches runs ∼3.6× faster with four threads than with a single thread). Further, for efficient mapping of SOLiD data to large reference genomes (since runtime will scale in a roughly linear way with both read number and reference genome size), SOCS can be implemented on a cluster—we have mapped a 32 million read data set to the complete human genome (Build 36.3) at a tolerance of four mismatches in ∼17 h on an eight node (64 core) cluster. Executable versions, source code, sample datasets, usage instructions, and scripts that facilitate implementation of SOCS on a cluster are available at http://socs.biology.gatech.edu/.

ACKNOWLEDGEMENTS

We thank Martin Storm for assistance in collecting SOLiD sequence data, Terry Turner and the Georgia Tech OIT group for assistance in implementing and testing SOCS on the PACE cluster and the Bergman laboratory for helpful discussions.

Funding: DHHS contract (N266200400059C/N01-AI-40059); New Opportunities award from the Southeast RCE for Biodefense and Emerging Infectious Diseases.

Conflict of Interest: none declared.

REFERENCES

Johnson,D.S. *et al*. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

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

Li,R. *et al*. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Nagalakshmi,U. *et al*. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Shendure,J. *et al*. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.

Smith,A.D. *et al*. (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, **9**, 128.

Torres,T.T. *et al*. (2008) Gene expression profiling by massively parallel sequencing. *Genome Res.*, **18**, 172–177.

Wilhelm,B.T. *et al*. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.