BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 20 2008, pages 2395–2396
doi:10.1093/bioinformatics/btn429

*Sequence analysis*

# SeqMap: mapping massive amount of oligonucleotides to the genome

Hui Jiang[1] and Wing Hung Wong[2,*]

[1]Institute for Computational and Mathematical Engineering and [2]Department of Statistics, Stanford University, Stanford, California 94305, USA

Received on July 21, 2008; revised and accepted on August 8, 2008
Advance Access publication August 12, 2008
Associate Editor: Limsoon Wong

**ABSTRACT**

**Summary:** SeqMap is a tool for mapping large amount of short sequences to the genome. It is designed for finding all the places in a reference genome where each sequence may come from. This task is essential to the analysis of data from ultra high-throughput sequencing machines. With a carefully designed index-filtering algorithm and an efficient implementation, SeqMap can map tens of millions of short sequences to a genome of several billions of nucleotides. Multiple substitutions and insertions/deletions of the nucleotide bases in the sequences can be tolerated and therefore detected. SeqMap supports FASTA input format and various output formats, and provides command line options for tuning almost every aspect of the mapping process. A typical mapping can be done in a few hours on a desktop PC. Parallel use of SeqMap on a cluster is also very straightforward.

**Contact:** whwong@stanford.edu

## 1 INTRODUCTION

With the rapid development of sequencing technology, there are now several next generation sequencing platforms available. Enormous amount of short sequences can be generated by these sequencing machines in a short time. For example, the Illumina-Solexa system can generate over 50 million sequences of length 30–50 nt in a single run taking <3 days. Using a different technology, the ABI-SOLiD system can generate data at a similar rate. The Roche-454 system generates fewer, but longer sequences. To date, the sequencing technology is still in the developing phase with a very fast pace of increase in throughput.

Due to the large amount of data generated by the above systems, efficient algorithms for mapping short oligonucleotides to a reference genome are in great demand. Popular alignment programs that are designed to align smaller amount of longer sequences, such as BLAST (Altschul *et al.*, 1997) and its successor BLAT (Kent, 2002), are not the best options to accomplish this task. Therefore, several short sequence mapping programs have been developed recently for this particular objective. They use effective techniques such as hashing and short key indexing, but differ in algorithm, implementation details and capability. ELAND (Cox, unpublished software), developed by Solexa, can map reads of length 20–32 nt to the genome, allowing up to two substitutions in the mapping. SOAP (Li *et al.*, 2008) can also handle up to two substitutions, or a gap of 1–3 nt without any other substitution. It can also handle longer reads or pair-end reads. RMAP (Smith *et al.*, 2008) is another program for ungapped mapping, which takes read qualities into account.

Compared to these existing programs, SeqMap offers more flexibility in the mapping. It allows up to five mixed substitutions and inserted/deleted nucleotides in the mapping, which is considered sufficient for most mapping applications. FASTA input format and various output formats (e.g. the ELAND format) are supported by SeqMap for the convenience of users. It also provides many command line options for tuning almost every aspect of the mapping process. For instance, SeqMap allows sequences to contain *N*'s, and to have unequal lengths. Such flexibility is beneficial for the analysis since both sequencing errors and SNPs may cause substitutions and inserted/deleted nucleotides in the reads. It is especially useful for the analysis of cancer genomes where substitutions and insertions/deletions happen more often. For reads that are longer than 30 nt, the 3 bp mismatch mapping gives mostly true signal rather than noise. We map 11M RNA-Seq reads of length 30 nt from (Mortazavi *et al*., 2008) to mouse chr19 using 2 bp and 3 bp mismatch mapping, respectively. The 3 bp mismatch mapping gives 18.5% more uniquely mapped reads. This is achieved without large drop in specificity—60.3% of the uniquely mapped reads in 3 bp mismatch mapping are mapped to RefSeq genes versus 63.6% in 2 bp mismatch mapping. Moreover, there are other applications in which this flexibility is helpful, such as mapping of exon tiling array probes in (Xing *et al.*, 2008) for probe level cross-hybridization analysis.

A website (http://biogibbs.stanford.edu/~jiangh/SeqMap/) has been setup for maintaining the SeqMap program, its source code and documentations.

## 2 METHODS

Many of the short sequence mapping programs, including ELAND, SOAP and RMAP, are based on the pigeonhole principle. This principle was also used in detecting near-duplicated web pages (Manku *et al*., 2007). The idea, also used by SeqMap, is to split each read into several parts. By requiring some of the parts instead of all of them to be perfectly matched in the mapping, the noncandidates can be filtered out very quickly. For example, for a mapping up to two substitutions, we can split each read into four parts. Since the substitutions can only occur in at most two of the four parts, at least two of the four parts will be perfectly matched to the target. Therefore, we

[*]To whom correspondence should be addressed.

can use the sequence combined from the two perfectly matched parts as the key to index all the candidates. A hash table is an effective and efficient data structure to implement this task. By enumerating all combinations of two parts chosen from all of the four parts, we need only six scans to find all the candidates. After that, a second stage of filtering is done between the sequence to be mapped and all the candidates from the first stage to determine all the targets in the genome. Insertions and deletions can also be incorporated into this algorithm in a similar fashion if carefully implemented. The only change will be that more rounds of scans are needed in the first stage of filtering. To allow one substitution and one insertion/deletion, each read will still be split into four parts. However, we need to scan not only all of the combinations of the two of the four parts, but also the combinations of the two parts with one of them shifted one nucleotide to its left or to its right.

SeqMap is written in ANSI C++. It uses bit operations to accelerate the mapping. Each nucleotide is encoded into 2 bits in the memory. In common with ELAND or RMAP, SeqMap indexes and hashes the reads before scanning the reference genome. This is different from some other alignment programs such as BLAT and SOAP, which indexes and hashes the reference genome instead. Hashing the genome usually needs large memory (e.g. SOAP needs 14 GB memory when mapping to the human genome) and therefore prohibits the program from running on most desktop PCs. As a comparison, SeqMap runs smoothly and quickly on a PC with 2 GB memory and a single 32-bit CPU when working with the human genome. It can also be compiled and run on any other platform, including 64-bit workstations with >16 GB memory, where it can take advantage of the large memory and the high performance CPU. Furthermore, by simply splitting the reads or the genome or both into several parts, SeqMap can be used in parallel on large scale data sets to speed up the mapping process.

## 3 RESULTS

To evaluate the mapping efficiency and effectiveness of SeqMap, we take more than 11 million Solexa reads from RNA-Seq experiments (Mortazavi *et al.*, 2008), and use several short sequence mapping programs including ELAND, SOAP, RMAP and SeqMap to map these reads to mouse chrX. ELAND, SOAP and RMAP are known to be among the best available short sequence mapping programs. Other alignment programs such as BLAST and BLAT are not evaluated because they have been shown to be much slower than the programs designed for mapping short reads (Li *et al.*, 2008). We take the first 25 nt of the reads and do a mapping with up to two substitutions, since this is supported by all these programs. As we can see from Table 1, ELAND is the fastest, and SeqMap is faster than SOAP and RMAP. Given the fact that ELAND is optimized for <32 nt reads and up to two substitutions mapping only, it is reasonable that SeqMap is slower than ELAND since SeqMap can map longer reads, with more substitutions and even several additional insertions and deletions. In terms of the mapping accuracy, all four programs give similar results. SeqMap and ELAND gives the most number of mapped reads. The memory usage of the programs is also given in Table 1, where we can see that SeqMap and RMAP use more memory than ELAND and SOAP. We need to point out that the comparison with SOAP is very data and parameter dependent, since SOAP indexes the reference genome rather than the reads as in the other three programs.

To show that SeqMap can deal with more substitutions and also insertion/deletions, we randomly generate a DNA sequence of a length of 1 Mb, add to it 100 Kb random substitutions, *N*'s and insertion/deletions, and then randomly sample 100k short reads of 30 nt from it. Finally, we map these short reads back to the original DNA sequence using all four programs. The results we get are shown in Table 2. We can see that SeqMap is able to detect a much larger number of matches.

**Table 1.** Benchmark results of SeqMap, ELAND, SOAP and RMAP

| Software | Running time | Memory used | Mapped reads |
|---|---|---|---|
| SeqMap | 2213 s | 3.0 GB | 455 384 |
| ELAND | 345 s | 721 MB | 455 384 |
| SOAP | 5464 s | 979 MB | 452 005 |
| RMAP | 14 h | 3.1 GB | 321 651 |

11 530 816 Solexa reads (25 nt) are mapped to mouse chrX (166 650 296 bp) using SeqMap, ELAND, SOAP and RMAP, respectively. The running time, memory usage, and number of mapped reads for each program are reported. For each program, up to two substitutions are allowed and no gap is allowed. The experiments are done on a machine with 3 GHz Intel Xeon CPU and 32 GB memory, running 64-bit Linux.

**Table 2.** Mapping 100 000 randomly perturbed reads with SeqMap, ELAND, SOAP and RMAP

| Software | Running time (s) | Memory used (MB) | Mapped reads |
|---|---|---|---|
| SeqMap | 82 | 923 | 78 211 |
| ELAND | 3 | 261 | 27 561 |
| SOAP | 2 | 142 | 38 256 |
| RMAP | 4 | 232 | 31 891 |

## ACKNOWLEDGEMENTS

We thank Karen Kapur, Wenxiu Ma and Yi Xing for helpful advice and Ying Xu for the (Manku *et al.*, 2007) reference.

*Funding*: National Institute of Health (R01HG003903 and U54GM62119 to W.H.W.)

*Conflict of Interest*: none declared.

## REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

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

Manku,G.S. *et al*. (2007) Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web.* ACM, New York, NY, USA, pp. 141–150.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods*, **5**, 621–628.

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

Xing,Y. *et al.* (2008) MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. *RNA*, **14**, 1470–1479.