BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 15 2009, pages 1966–1967
doi:10.1093/bioinformatics/btp336

*Sequence analysis*

# SOAP2: an improved ultrafast tool for short read alignment

Ruiqiang Li[1,2,†], Chang Yu[1,†], Yingrui Li[1], Tak-Wah Lam[3], Siu-Ming Yiu[3], Karsten Kristiansen[2] and Jun Wang[1,2,*]

[1]Beijing Genomics Institute at Shenzhen, Shenzhen, 518083, China, [2]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, DK-5230, Denmark and [3]Department of Computer Science, University of Hong Kong, Hong Kong, China

Received on January 23, 2009; revised on April 27, 2009; accepted on May 24, 2009

Advance Access publication June 3, 2009

Associate Editor: Joaquin Dopazo

## ABSTRACT

**Summary:** SOAP2 is a significantly improved version of the short oligonucleotide alignment program that both reduces computer memory usage and increases alignment speed at an unprecedented rate. We used a Burrows Wheeler Transformation (BWT) compression index to substitute the seed strategy for indexing the reference sequence in the main memory. We tested it on the whole human genome and found that this new algorithm reduced memory usage from 14.7 to 5.4 GB and improved alignment speed by 20–30 times. SOAP2 is compatible with both single- and paired-end reads. Additionally, this tool now supports multiple text and compressed file formats. A consensus builder has also been developed for consensus assembly and SNP detection from alignment of short reads on a reference genome.

**Availability:** http://soap.genomics.org.cn

**Contact:** soap@genomics.org.cn

Next-generation DNA sequencing technologies, including Illumina/Solexa and AB/SOLiD, have been dominant tools for genomic data collection. Various applications have been developed using these technologies to promote biological research, such as detecting genetic variation through whole genome or target region resequencing, refining gene annotation by whole transcriptome sequencing, profiling mRNA and miRNA expression and studying DNA methylation. One of the common key data analysis steps of these applications is to align huge amounts of short reads onto a reference genome. New efficient programs have been developed to meet the challenges for such alignment. Among them, SOAP (Short Oligonucleotide Alignment Program; Li *et al*., 2008) has been used widely for these types of analyses due to its fast speed and richness of features.

With further improvement on sequencing throughput and the launch of big research projects, much faster short-read alignment methods are required to handle the data analysis of such large-scale sequence production. For example, the 1000 Genomes project that aims to create the most detailed and medically useful human genetic variation map, will generate about 15 Tb of sequence using next-generation sequencing technologies. With even the fastest programs currently available, one would need ~1000 CPU months to align these short reads onto the human reference genome. Additionally, new methods are now needed to support longer reads as the existing methods were primarily designed for very short reads with typical lengths shorter than 50 bp. With improvements in sequencing chemistry and data processing algorithms, the Illumina Genome Analyzer can now generate up to 75–100 bp high-quality reads, and longer reads are expected in the near future.

Here, we have developed an improved version of SOAP, called SOAP2. The new program uses the Burrows Wheeler Transformation (BWT) compressed index instead of the seed algorithm that was used in the previous version for indexing the reference sequence in the main memory. Use of BWT substantially improved alignment speed; additionally, it significantly reduced memory usage.

Big eukaryotic genomes always consist of a large number of repetitive sequences (e.g. 45% of the human genome). Suffix trees and suffix arrays are considered the most appropriate methods for indexing DNA sequence, through which only one alignment is needed to for repetitive sequence with multiple identical copies in the genome. The complexity in space and time of such index construction has limited such algorithm usage in only small genomes. But the recent development of compressed indexing has reduced the space complexity from $O(n)$ bytes to $O(n)$ bits. Among these is the BWT (Burrow and Wheeler, 1994), a reversible data compression algorithm, which was found to be the most efficient. The space complexity of BWT is $n/4$ bytes, and only 1 GB memory in RAM is required for indexing the whole human genome. This algorithm has been used for efficient whole-genome comparison and indexing for Smith–Waterman local alignment to the human genome (Lam *et al*., 2008).

Using this in our alignment method, we determined an exact match, by constructing a hash table to accelerate searching for the location of a read in the BWT reference index. For example, if we use a 13mer on the hash, then the reference index would be partitioned into $2^{26}$ blocks, and very few search interactions are sufficient to identify the exact location inside the block. For inexact (both mismatch and indel) alignment, we applied a ‘split-read strategy’. To allow one mismatch, a read was split into two fragments. The mismatch can exist in, at most, one of the two fragments at the same

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

**Table 1.** Benchmark of short read alignment tools

| Software | Reads aligned (%) | Time (paired, s) | Time (single, s) | Memory usage (GB) |
|---|---|---|---|---|
| SOAP2 | 93.6 | 828 | 478 | 5.4 |
| SOAP | 93.8 | 19 234 | 14 328 | 14.7 |
| MAQ | 93.2 | 22 506 | 19 847 | 1.2 |
| Bowtie | 91.7 | – | 405 | 2.3 |

We used a query dataset of one million read pairs generated by the Illumina Genome Analyzer on a human DNA sample to test the software performance. The read length was 44 bp. The paired-end insert size was about 200 bp. The human reference genome was NCBI build 36.1. At most four mismatches were allowed in SOAP2, at most two mismatches were allowed in 5′-end 35 bp with low-quality trimming in SOAP, and mapping quality cutoff '-e 80' was used for MAQ and Bowtie. Evaluation was performed on a computer server with CPU Intel Xeon E5335 (2.0 GHz) and 16 GB RAM installed.

time. Likewise, we split a read into three fragments to search for hits that allow two mismatches. This enumeration algorithm was used to identify mutation sites on the reads.

In paired-end alignment mode, we first independently aligned the two reads belonging to a pair then searched for the pair of hits with the correct orientation relationship and proper distance. Similar to SOAP, we preferentially select the best hit of each read or read pair, which have the lowest number of mismatches or small gaps. In general practice, a user can also choose the option to report all hits that satisfy their selected preset similarity rate. For most analyses, to guarantee alignment accuracy, we recommend allowing at most two mismatches or one continuous gap in the high-quality part of a read. For the low-quality regions of a read (the 3′-end, which can contain a higher rate of sequencing errors), we provided an option that allows more mismatches within this defined 3′-end region. Since sequencing read length is getting longer and longer, the SOAP2 program is now compatible with read lengths up to 1024 bp.

We evaluated the performance of this software on a dataset containing one million read pairs generated from a human Asian individual (Wang *et al*., 2008). Although SOAP2 was designed for improved Illumina GA sequencing with read length over 50 bp, we chose a 44 bp read length in this evaluation, which is compatible with the tools SOAP (Li,R. *et al*., 2008), MAQ (Li,H. *et al*., 2008) and the recently developed BWT-based alignment tool Bowtie (Langmead *et al*., 2009). SOAP2 takes 7200 s to build a BWT index for the human reference genome, which is 12 times slower than building the seed index that was implemented in SOAP. Thus, we prebuild the index on a hard disk, and then load it into RAM directly when starting a new alignment job for that genome. The memory usage was reduced from 14.7 GB in SOAP to 5.4 GB in SOAP2. SOAP2 was more than 20 times faster than SOAP and MAQ with similar amount of reads aligned (Table 1). SOAP2 and Bowtie have comparable speed on aligning single-end reads, while Bowtie cannot always find the best alignment hits and cannot align paired-end reads (Langmead *et al*., 2009). It should be made aware that the alignment sensitivity was determined by sequencing quality and the parameter setting of each alignment tool, so the percent of reads aligned would vary in different datasets.

SOAP2 supports multiple input and output file formats. The reference sequence can be loaded as either a text or a gzipped FASTA format, and the query reads can be in either FASTA or FASTQ format. The output formats include a SOAP tab-delimited text table, a gzipped text table, a Sequence Alignment/Map (SAM) format and its binary equivalent (BAM) that is recommended by the 1000 Genomes Consortium, and a Consed format that fits with the assembly viewer.

As SOAP2 is specifically designed for ultrafast alignment of short reads onto a reference sequence for large-scale resequencing projects, we have developed a companion assembler for consensus assembly of the sequenced individual based on the alignment of the reads on the reference sequence. The assembler has been included in the SOAP software package and is also freely available from the website. With this, we can detect SNPs by comparing the assembled sequence to the reference genome. The assembler uses Bayes' theorem to infer the genotype of each base pair from the aligned reads and sequencing quality scores. The estimated SNP rate between the sequenced individual and the reference genome is used as prior probability, the raw sequencing qualities were recalibrated according to the alignment, the reads generated from potential duplicate clones were removed, and finally the genotype was called from the posterior probabilities with a Phred-like score transformed from the probability to indicate its accuracy. The tool has been used in analyzing the Asian genome data and showed over 99.9% accuracy (Wang *et al*., 2008).

## ACKNOWLEDGEMENTS

We are indebted to the many SOAP users who tested the program, reported bugs and proposed improvements to make it more powerful and user friendly. Laurie Goodman edited the manuscript.

*Funding*: National Natural Science Foundation of China (30725008); Shenzhen local government of China; Grants from the Danish Natural Science Research Council (272-05-0344 and 272-07-0196).

*Conflict of Interest*: none declared.

## REFERENCES

Burrow,M. and Wheeler,D.J. (1994) A block-sorting lossless data compression algorithm. *Technical Report 124,* Digital Equipment Corporation, CA.

Lam,T.W. *et al*. (2008) Compressed indexing and local alignment of DNA. *Bioinformatics*, **24**, 791–797.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,H. *et al*. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

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

Wang,J. *et al*. (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.