BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009

Advance Access publication May 18, 2009

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The enormous amount of short reads generated by the new DNA sequencing technologies call for the development of fast and accurate read alignment programs. A first generation of hash table-based methods has been developed, including MAQ, which is accurate, feature rich and fast enough to align short reads from a single individual. However, MAQ does not support gapped alignment for single-end reads, which makes it unsuitable for alignment of longer reads where indels may occur frequently. The speed of MAQ is also a concern when the alignment is scaled up to the resequencing of hundreds of individuals.

Results: We implemented Burrows-Wheeler Alignment tool (BWA), a new read alignment package that is based on backward search with Burrows–Wheeler Transform (BWT), to efficiently align short sequencing reads against a large reference sequence such as the human genome, allowing mismatches and gaps. BWA supports both base space reads, e.g. from Illumina sequencing machines, and color space reads from AB SOLiD machines. Evaluations on both simulated and real data suggest that BWA is ∼10–20× faster than MAQ, while achieving similar accuracy. In addition, BWA outputs alignment in the new standard SAM (Sequence Alignment/Map) format. Variant calling and other downstream analyses after the alignment can be achieved with the open source SAMtools software package.

Availability: http://maq.sourceforge.net

Contact: rd@sanger.ac.uk

1 INTRODUCTION

The Illumina/Solexa sequencing technology typically produces 50–200 million 32–100 bp reads on a single run of the machine. Mapping this large volume of short reads to a genome as large as human poses a great challenge to the existing sequence alignment programs. To meet the requirement of efficient and accurate short read mapping, many new alignment programs have been developed. Some of these, such as Eland (Cox, 2007, unpublished material), RMAP (Smith *et al.*, 2008), MAQ (Li *et al.*, 2008a), ZOOM (Lin *et al.*, 2008), SeqMap (Jiang and Wong, 2008), CloudBurst (Schatz, 2009) and SHRiMP (http://compbio.cs.toronto.edu/shrimp), work by hashing the read sequences and scan through the reference sequence. Programs in this category usually have flexible memory footprint, but may have the overhead of scanning the whole genome when few reads are aligned. The second category of software, including SOAPv1 (Li *et al.*, 2008b), PASS (Campagna *et al.*, 2009), MOM (Eaves and Gao, 2009), ProbeMatch (Jung Kim *et al.*, 2009), NovoAlign (http://www.novocraft.com), ReSEQ (http://code.google.com/p/re-seq), Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik) and BFAST (http://genome.ucla.edu/bfast), hash the genome. These programs can be easily parallelized with multi-threading, but they usually require large memory to build an index for the human genome. In addition, the iterative strategy frequently introduced by these software may make their speed sensitive to the sequencing error rate. The third category includes slider (Malhis *et al.*, 2009) which does alignment by merge-sorting the reference subsequences and read sequences.

Recently, the theory on string matching using Burrows–Wheeler Transform (BWT) (Burrows and Wheeler, 1994) has drawn the attention of several groups, which has led to the development of SOAPv2 (http://soap.genomics.org.cn/), Bowtie (Langmead *et al.*, 2009) and BWA, our new aligner described in this article. Essentially, using backward search (Ferragina and Manzini, 2000; Lippert, 2005) with BWT, we are able to effectively mimic the top-down traversal on the prefix trie of the genome with relatively small memory footprint (Lam *et al.*, 2008) and to count the number of exact hits of a string of length m in $O(m)$ time independent of the size of the genome. For inexact search, BWA samples from the implicit prefix trie the distinct substrings that are less than k edit distance away from the query read. Because exact repeats are collapsed on one path on the prefix trie, we do not need to align the reads against each copy of the repeat. This is the main reason why BWT-based algorithms are efficient.

In this article, we will give a sufficient introduction to the background of BWT and backward search for exact matching, and present the algorithm for inexact matching which is implemented in BWA. We evaluate the performance of BWA on simulated data by comparing the BWA alignment with the true alignment from the simulation, as well as on real paired-end data by checking the fraction of reads mapped in consistent pairs and by counting misaligned reads mapped against a hybrid genome.

2 METHODS

2.1 Prefix trie and string matching

The prefix trie for string X is a tree where each edge is labeled with a symbol and the string concatenation of the edge symbols on the path from a leaf to

*To whom correspondence should be addressed.