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

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 Aligner 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 Some of these, such as Eland (Cox, 2007, unpublished material), RMAP (Smith et al., 2008), MAQ (Li et al., 2008a), ZOOM (Lin et al., 2008), SeqMap (Jiang and Wong, 2008), CloudBurst (Schatz, 2009) and SHRIMP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), NovaAlign (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 mismapped and 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...
2.2 Burrows–Wheeler transform

Let \( \Sigma \) be an alphabet. Symbol \( S \) is not present in \( \Sigma \) and is lexicographically smaller than all the symbols in \( \Sigma \). A string \( X = a_0 \ldots a_{n-1} \) is always ended with symbol \( S \) (i.e. \( a_{n-1} = S \)) and this symbol only appears at the end. Let \( X(i) = a_i, i = 0, 1, \ldots n-1 \), be the \( i \)-th symbol of \( X \). \( X[i..j] = a_i \ldots a_{j-1} \) is a substring of \( X \). Suffix array \( S \) of \( X \) is a permutation of the integers \( 0 \ldots n-1 \) such that \( S(i) \) is the start position of the \( i \)-th smallest suffix.

The BWT of \( X \) is defined as \( \mathbf{B}(W) \) when \( S(i) = 0 \) or \( B[i..j] = X[S(i)-1] \) otherwise. We also define the length of string \( X \) as \( |X| \) and therefore \( |X| = |B| = n \).

Fig. 2 gives an example on how to construct BWT and suffix array. The algorithm shown in Figure 2 is quadratic in time and space. However, this is not necessary. In practice, we usually construct the suffix array first and then generate BWT. Most algorithms for constructing suffix array require at least \( n \log n \) bits of working space, which amounts to 12 GB for human genome. Recently, Hon et al. (2007) gave a new algorithm that uses \( n \) bits of working space and only requires \( < 1 \) GB memory at peak time for constructing the BWT of human genome. This algorithm is implemented in BWT-SW (Lam et al., 2008). We adapted its source code to make it work with BWA.

2.3 Suffix array interval and sequence alignment

If string \( W \) is a substring of \( X \), the position of each occurrence of \( W \) in \( X \) will occur in an interval in the suffix array. This is because all the suffixes that have \( W \) as prefix are sorted together. Based on this observation, we define:

\[
\begin{align*}
\mathbf{B}(W) &= \min \{ k : W \text{ is the prefix of } X[k] \} \\
\mathbf{R}(W) &= \max \{ k : W \text{ is the prefix of } X[k] \}
\end{align*}
\]

In particular, if \( W \) is an empty string, \( \mathbf{B}(W) = 1 \) and \( \mathbf{R}(W) = n - 1 \). The interval \( [\mathbf{B}(W), \mathbf{R}(W)] \) is called the SA interval of \( W \) and the set of positions of all occurrences of \( W \) in \( X \) is \( \{ S[k] : \mathbf{B}(W) \leq k \leq \mathbf{R}(W) \} \). For example in Figure 2, the SA interval of string ‘go’ is \([1, 2]\). The suffix array values in this interval are 3 and 0 which give the positions of all the occurrences of ‘go’.

Knowing the intervals in suffix array we can get the positions. Therefore, sequence alignment is equivalent to searching for the SA intervals of substrings of \( X \) that match the query. For the exact matching problem, we can find only such one interval; for the inexact matching problem, there may be many.

2.4 Exact matching: backward search

Let \( C(a) \) be the number of symbols in \( X[0..n-2] \) that are lexicographically smaller than \( a \) in \( \Sigma \) and \( O(a, i) \) the number of occurrences of \( a \) in \( B[0..i] \)

Ferragina and Manzini (2000) proved that if \( W \) is a substring of \( X \):

\[
\begin{align*}
\mathbf{B}(aW) &= C(a) + O(a, \mathbf{B}(W) - 1) + 1 \\
\mathbf{R}(aW) &= C(a) + O(a, \mathbf{R}(W))
\end{align*}
\]

and that \( \mathbf{B}(aW) \leq \mathbf{B}(a) \) if and only if \( aW \) is a substring of \( X \). This result makes it possible to test whether \( W \) is a substring of \( X \) and to count the occurrences of \( W \) in \( O(|W|) \) time by iteratively calculating \( \mathbf{B} \) and \( \mathbf{R} \) from the end of \( W \). This procedure is called backward search.

It is important to note that Equations (3) and (4) actually realize the top-down traversal on the suffix trie of \( X \) given that we can calculate the SA interval of a child node in constant time if we know the interval of its parent. In this sense, backward search is equivalent to exact string matching on the prefix trie, but without explicitly putting the trie in the memory.
2.5 Inexact matching: bounded traversal/backtracking

Figure 3 gives a recursive algorithm to search for the SA intervals of substrings that match the query string W with no more than \( \ell \) differences (mismatches or gaps). Essentially, this algorithm uses backward search to sample distinct substrings from the genome. This process is bounded by the \( D(\ell) \) array where \( D(\ell) \) is the lower bound of the number of differences in string \( W[0,j] \).

The \( \text{CALCULATE}(D) \) procedure in Figure 3 effectively mimics a depth-first search (DFS) on the prefix trie, while BWA implements a breadth-first search (BFS) using this heap-like data structure. Third, we adopt an iterative strategy: if the top interval is repetitive, we do not search for suboptimal intervals by default; if the top interval is unique and has \( z \) differences, we only search for hits with up to \( z \) differences. This iterative strategy accelerates BWA while retaining the ability to generate mapping quality. However, this also makes BWA's speed sensitive to the mismatch rate between the reads and the reference because finding hits with more differences is usually slower.

Fourth, we allow to set a limit on the maximum allowed differences in the first few tens of base pairs on a read, which we call the seed sequence. Given 70 bp simulated reads, alignment with maximum two differences in the 32 by seed is 2.5× faster than without seeding. The alignment error rate, which is the fraction of wrong alignments out of confident mappings in simulation (see also Section 3.2), only increases from 0.08% to 0.11%. Seeding is less effective for shorter reads.

2.6 Reducing memory

The algorithm described above needs to load the occurrence array \( O \) and the suffix array \( S \) in the memory. Holding the full \( O \) and \( S \) arrays requires huge memory. Fortunately, we can reduce the memory by only storing a small fraction of the \( O \) and \( S \) arrays, and calculating the rest on the fly. BWT-SW (Lam et al., 2008) and Bowtie (Langmead et al., 2009) use a similar strategy which was first introduced by Ferragina and Manzini (2000).

Given a genome of size \( n \), the occurrence array \( O(\cdot) \) requires \( 4n(\log n) \) bits as each integer takes \( \log n \) bits and there are \( 4n \) of them in the array. In practice, we store in memory \( O(\cdot,k) \) for \( k = 1 \) that is a factor of 128 and calculate the rest of elements using the BWT string \( B \). When we use two bits to represent a nucleotide, \( B \) takes 2n bits. The memory for backward search is
thus 2n+⌈log_2 n⌉/32 bits. As we also need to store the BWT of the reverse genome to calculate the bound, the memory required for calculating intervals is doubled, or about 2.3 GB for a 3 GB genome.

Enumerating the position of each occurrence requires the suffix array S. If we put the entire S in memory, it would use s⌈log_2 n⌉ bits. However, it is also possible to reconstruct the entire S when knowing part of it. In fact, S and inverse compressed suffix array (inverse CSA) \( \Phi^{-1} \) (Grossi and Vitter, 2000) satisfy:

\[
S(k) = S(\Phi^{-1}(j)) + j
\]  

(5)

where \( \Phi^{-1}(j) \) denotes repeatedly applying the transform \( \Phi^{-1} \) for \( j \) times. The inverse CSA \( \Phi^{-1} \) can be calculated with the occurrence array \( \Psi \):

\[
\Psi^{-1}(i) = C[i] + i \times B[i]
\]  

(6)

In BWA, we only store in memory \( \Phi^{-1} \) with \( k \leq \frac{2\%}{\text{uniform base error rate}} \) may contain differences more than \( \Psi(i) \) = 0 after \( \Psi = \Phi^{-1} \) of the alignment being incorrect. The algorithm is similar to MAQ's except for genome <4 Gb. This includes the memory for the BWT string, partial genome to calculate the bound, the memory required for calculating intervals \( S \) and inverse compressed suffix array (inverse CSA) \( \Phi^{-1} \) (Grossi and Vitter, 2000).

For each alignment, BWA calculates a mapping quality score, which is the Phred-scaled probability of a mutation, \( q_{i} \) and function \( g_{b} = g_{b}^{l} \) gives the color corresponding to the two adjacent nucleotides \( b \) and \( k \). Essentially, we pay a penalty \( q_{i} \) if \( b \neq k \) and a penalty \( q_{i} \) if \( b = k \) and \( k \neq 1 \). This optimization can be done by dynamic programming because the best decoding beyond position \( i \) only depends on the choice of \( b_{i} \) and \( \hat{b}_{i} \). Let the best decoding score up to \( i \). The iteration equations are:

\[
f_{i}(b_{i}) = q_{i} \cdot \min_{b_{i}} \left\{ f_{i-1}(b_{i-1}) + g_{b} \left[ 1 - \delta_{b_{i},a_{i}} \right] \right\}
\]

(7)

where \( q_{i} \) is the Phred-scaled probability of a mutation, \( q_{i} \) is the Phred quality of color \( \hat{b}_{i} \) and \( H \) is the Hamming distance. BWA approximates base qualities as follows. Let \( q_{i} \equiv q_{i} \cdot e^{-\frac{q_{i}^{2}}{2}} \) and \( g_{b} = g_{b}^{l} \) gives the color corresponding to the two adjacent nucleotides \( b \) and \( k \). Essentially, we pay a penalty \( q_{i} \). Let the best decoding score up to \( i \). The iteration equations are:

\[
f_{i}(b_{i}) = q_{i} \cdot \min_{b_{i}} \left\{ f_{i-1}(b_{i-1}) + g_{b} \left[ 1 - \delta_{b_{i},a_{i}} \right] \right\}
\]

(7)

where \( q_{i} \) is the Phred-scaled probability of a mutation, \( q_{i} \) is the Phred quality of color \( \hat{b}_{i} \) and \( H \) is the Hamming distance. BWA approximates base qualities as follows. Let \( q_{i} = \hat{q}_{i} \). The i-th base quality \( \hat{q}_{i} \) is calculated as:

\[
\hat{q}_{i} = \begin{cases} 
q_{i} + q_{i} & \text{if } c_{i-1} = c_{i} = c_{i} \neq \hat{c}_{i} \\
q_{i} & \text{if } c_{i-1} = \hat{c}_{i} = c_{i} \neq \hat{c}_{i} \\
q_{i} - q_{i} & \text{if } c_{i-1} = \hat{c}_{i} = c_{i} \neq \hat{c}_{i} \\
0 & \text{otherwise}
\end{cases}
\]

(8)

BWA uses the sequence \( b_{i} \) with quality \( \hat{q}_{i} \) as the final result for SODI mapping.

3 RESULTS

3.1 Implementation

We implemented BWA to do short read alignment based on the BWT of the reference genome. It performs gapped alignment for single-end reads, supports paired-end mapping, generates mapping quality and gives multiple hits if required. The default output alignment format is SAM (Sequence Alignment/Map format). Users can use SAMtools (http://samtools.sourceforge.net) to extract alignments in a region, merge/sort alignments, get single nucleotide polymorphism (SNP) and indel calls and visualize the alignment. BWA is distributed under the GNU General Public License (GPL).

Documentation and source code are freely available at the MAQ web site: http://maq.sourceforge.net.
3.2 Evaluated programs

To evaluate the performance of BWA, we tested additional three alignment programs: MAQ (Li et al., 2008a), SOAPv2 (http://soap.genomics.org.cn) and Bowtie (Langmead et al., 2009). MAQ indexes reads with a hash table and scans through the genome. It is the software package we developed previously for large-scale read mapping. SOAPv2 and Bowtie are the other two BWT-based short read aligners that we are aware of. The latest SOAP-2.1.7 (Li et al., unpublished data) uses 2way-BWT (Lam et al., unpublished data) for alignment. It tolerates more mismatches beyond the 35 bp seed sequence and supports gapped alignment limited to one gap open. Bowtie (version 0.9.2) deploys a similar algorithm to BWA. Nonetheless, it does not reduce the search space by bounding the search with $D(k)$, but by cleverly doing the alignment for both original and reverse read sequences to bypass unnecessary searches towards the root of the prefix trie. By default, Bowtie performs a DFS on the prefix trie and stops when the first qualified hit is found. Thus, it may miss the best inexact hit even if its seeding strategy is disabled. It is possible to make Bowtie perform a BFS by applying ‘-best’ at the command line, but this makes Bowtie slower. Bowtie does not support gapped alignment at the moment.

All the four programs, including BWA, randomly place a repetitive read across the multiple equally best positions. As we are mainly interested in confident mappings in practice, we need to rule out repetitive hits. SOAPv2 gives the number of equally best hits of a read. Only unique mappings are retained. We also ask SOAPv2 to limit the possible gap size to at most 3 bp. We run Bowtie with the command-line option ‘-best -k 2’, which renders Bowtie to output the top two hits of a read. We discard a read alignment if the second best hit contains the same number of mismatches as the best hit. MAQ and BWA generate mapping qualities. We use mapping quality threshold 1 for MAQ and 10 for BWA to determine confident mappings. We use different thresholds because we know that MAQ’s mapping quality is underestimated, while BWA’s is overestimated.

3.3 Evaluation on simulated data

We simulated reads from the human genome using the wgsim program that is included in the SAMtools package and ran the four programs to map the reads back to the human genome. As we know the exact coordinate of each read, we are able to calculate the alignment error rate.

Table 1 shows that BWA and MAQ achieve similar alignment accuracy. BWA is more accurate than Bowtie and SOAPv2 in terms of both the fraction of confidently mapped reads and the error rate of confident mappings. Note that SOAP-2.1.7 is optimized for reads longer than 35 bp. For the 32 bp reads, SOAP-2.0.1 outperforms the latest version.

On speed, SOAPv2 is the fastest and actually it would be 30–80% faster for paired-end mapping if gapped alignment was disabled. Bowtie with the default option (data not shown) is several times faster than the current setting ‘-best -k 2’ on single-end mapping. However, the speed is gained at a great cost of accuracy. For example, with the default option, Bowtie can map the two million single-end 32 bp reads in 151 s, but 6.4% of confident mappings are wrong. This high alignment error rate may complicate the detection of structural variations and potentially affect SNP accuracy. Between BWA and MAQ, BWA is 6–18× faster, depending on the read length. MAQ’s speed is not affected by read length because internally it treats all reads as 128 bp. It is possible to accelerate BWA by not checking suboptimal hits similar to what Bowtie and SOAPv2 are doing. However, calculating mapping quality would be impossible in this case and we believe generating proper mapping quality is useful to various downstream analyses such as the detection of structural variations.

On memory, SOAPv2 uses 5.4 GB. Both Bowtie and BWA uses 2.3 GB for single-end mapping and about 3 GB for paired-end, larger than MAQ’s memory footprint 1 GB. However, the memory usage of all the three BWT-based aligners is independent of the number of reads to be aligned, while MAQ’s is linear in it. In addition, all BWT-based aligners support multi-threading, which reduces the memory per CPU core on a multi-core computer. On modern computer servers, memory is not a practical concern with the BWT-based aligners.

3.4 Evaluation on real data

To assess the performance on real data, we downloaded about 12.2 million pairs of 51 bp reads from European Read Archive (ACERR000589). These reads were produced by Illumina for NA12750, a male included in the 1000 Genomes Project (http://www.1000genomes.org). Reads were mapped to the human genome NCBI build 36. Table 2 shows that almost all confident mappings from MAQ and BWA exist in consistent pairs although MAQ gives fewer confident alignments. A slower mode of BWA (no seeding; searching for suboptimal hits even if the top hit is a repeat) did even better. In that mode, BWA confidently mapped 89.2% of all reads in 6.3 hours with 99.2% of confident mappings in consistent pairs.

In this experiment, SOAPv2 would be twice as fast with both percent confident mapping (Conf) and percent paired (Paired)
Table 2. Evaluation on real data

<table>
<thead>
<tr>
<th>Program</th>
<th>Time (h)</th>
<th>Conf (%)</th>
<th>Paired (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowtie</td>
<td>5.2</td>
<td>84.4</td>
<td>96.3</td>
</tr>
<tr>
<td>BWA</td>
<td>4.0</td>
<td>88.9</td>
<td>98.8</td>
</tr>
<tr>
<td>MAQ</td>
<td>94.9</td>
<td>86.1</td>
<td>98.7</td>
</tr>
<tr>
<td>SOAP2</td>
<td>3.4</td>
<td>88.3</td>
<td>97.5</td>
</tr>
</tbody>
</table>

The 12.2 million read pairs were mapped to the human genome. CPU time in hours on a single core of a 2.3 GHz Xeon E5420 processor (Time), percent confidently mapped reads (Conf) and percent confident mappings with the mates mapped in the correct orientation and within 300 bp (Paired), are shown in the table.

dropping by 1% if gapped alignment was disabled. In contrast, BWA is 1.4 times as fast when it performs ungapped alignment only. But even with BWT-based gapped alignment disabled, BWA is still able to recover many short indels with Smith–Waterman alignment given paired-end reads.

We also obtained the chicken genome sequence (version 2.1) and aligned these 12.2 million read pairs against a human–chicken hybrid reference sequence. The percent confident mappings is almost unchanged in comparison to the human-only alignment. As for the number of reads mapped to the chicken genome, Bowtie mapped 2640, BWA 2942, MAQ 3005 and SOAPv2 mapped 4531 reads to the wrong genome. If we consider that the chicken sequences take up one-quarter of the human–chicken hybrid reference, the alignment error rate for BWA is about 0.06% (=2942 × 4/12.2M/0.889).

Note that such an estimate of the alignment error rate may be underestimated because wrongly aligned human reads tend to be related to repetitive sequences in human and to be mapped back to the human sequences. The estimate may also be overestimated due to the presence of highly conservative sequences and the incomplete assembly of human or misassembly of the chicken genome.

If we want fewer errors, the mapping quality generated by BWA and MAQ allows us to choose alignments of higher accuracy. If we increased the mapping quality threshold in determining a confident hit to 25 for BWA, 86.4% of reads could be aligned confidently with 1927 reads mapped to the chicken genome, outperforming Bowtie in terms of both percent confident mappings and the number of reads mapped to the wrong genome.

4 DISCUSSION

For short read alignment against the human reference genome, BWA is an order of magnitude faster than MAQ while achieving similar alignment accuracy. It supports gapped alignment for single-end reads, which is increasingly important when reads get longer and tend to contain indels. BWA outputs alignment in the SAM format to take advantage of the downstream analyses implemented in SAMtools. BWA plus SAMtools provides most of functionality of the MAQ package with additional features.

In comparison to speed, memory and the number of mapped reads, alignment accuracy is much harder to evaluate on real data as we do not know the ground truth. In this article, we used three criteria for evaluating the accuracy of an aligner. The first criterion, which can only be evaluated with simulated data, is the combination of the number of confident mappings and the alignment error rate out of the confident mappings. Note that the number of confident mappings alone may not be a good criterion: we can map more at the cost of accuracy. The second criterion, which is the combination of the number of aligned reads and the number of reads mapped in consistent pairs, works on real data on the assumption that the mating information from the experiment is correct and that structural variations are rare. Although this criterion is related to the way an aligner defines pair ‘consistency’, in our experience it is highly informative if the pairing parameters are set correctly. The third criterion is the fraction of reads mapped to the wrong reference sequence if we intentionally add reference sequences from a diverged species.

Although in theory BWA works with arbitrarily long reads, its performance is degraded on long reads especially when the sequencing error rate is high. Furthermore, BWA always requires the full read to be aligned, from the first base to the last one (i.e. global with respect to reads), but longer reads are more likely to be interrupted by structural variations or misalignments in the reference genome, which will fail BWA. For long reads, a possibly better solution would be to divide the read into multiple short fragments, align the fragments separately with the algorithm described above and then join the partial alignments to get the full alignment of the read.

ACKNOWLEDGEMENTS

We are grateful to Mark DePristo and Jared Maguire from the Broad Institute for their suggestions on standardizing the criteria for evaluating alignment programs, and to the three anonymous reviewers whose comments helped us to improve the manuscript. We also thank the members of the Durbin research group for the comments on the initial draft.

Funding: Wellcome Trust/077192/Z/05/Z.

Conflict of Interest: none declared.

REFERENCES


