Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

Heng Li
Broad Institute of Harvard and MIT

BACKGROUND

Most short-read mappers initially developed for ~36bp reads perform end-to-end alignment (i.e. every read base to be aligned) and require a hammering or edit distance threshold. However, end-to-end alignment rejects reads bridging a break point caused by structural variations, and an edit distance threshold forbids long INDELs. Both scenarios occur more often with increasing read lengths, which make many short-read mappers less preferred for longer reads.

Although several long-read mappers have been developed recently, they all have certain limitations: BWA-SW (Li and Durbin, 2010) is slow for 100-200bp reads without achieving high accuracy, while Bowtie2 (Langmead and Salzberg, 2012) and Cushaw2 (Lu and Schmidt, 2012) are slow for reads over 500bp and does not well support split alignment. A fast, accurate and feature rich aligner accepting sequences with a wide range of lengths is still lacking.

SUMMARY

BWA-MEM (Li, 2013) is a new alignment algorithm for aligning sequence reads or long query sequences against a large reference genome such as human. It automatically chooses between local and end-to-end alignments, supports paired-end reads and performs split alignment. The algorithm is robust to sequencing errors and applicable to a wide range of sequence lengths from 70bp to a few megabases. For short-read mapping, BWA-MEM shows better performance than several state-of-art read aligners to date. For long reads, it is several times as fast as Bowtie2 and Cushaw2.

METHODS

FMD-index

FMD-index of a DNA sequence is the FM-index of the concatenation of the sequence and its reverse complement. It is essentially equivalent to the bi-directional BWT (Lam et al. 2009) used by SOAP2 and Bowtie2, but is more advantageous:

1. Forward and backward strands aligned simultaneously, which is faster than aligning the two strands separately.

Supermaximal Exact Match

Maximal exact match (MEM): an exact match that cannot be extended further in either direction. Supermaximal exact match (SMEM): a MEM that is not contained in any other MEMs on the query coordinate (Li, 2012). At any query position, the longest exact match covering the position is 1-maximal.

There are usually much fewer SMEMs than MEMs. With forward-backward search, SMEMs can be found very quickly:

Reference: AC CGTAGC
Read: AC CGTAGC

Round 1: AC
Round 2: CC TAGC

AC CGTAGC (2 hits)

x-maximal intervals

Given query P and reference T, let O([j,k]) be the occurrences of substring P[j..k] in T. Query interval [j,k] is called x-maximal if O([j,k]) and there does not exist [l,j] that contains [j,k] with O([j,k])<O([l,j]). In particular, query subsequences in SMEMs are 1-maximal.

The SMEM algorithm can be modified to find all x-maximal intervals with SMEM being the special case.

RESULTS AND DISCUSSIONS

Simulated 101bp SE/PE

One million pairs of 101bp reads simulated with 1.5% uniform substitution sequencing error rate and 0.2% small insertion/deletion (indel) variants.

Whole-genome alignment between E. coli strains

Two E. coli strains (NC_000913 and NC_001852) are aligned with both BWA-MEM and MUMmer (Kurtz et al. 2004). MUMmer identified 105,505 substitution differences, while BWA-MEM identified 104,321 of which 102,241 overlap. Most differences unique to one aligner lie in short regions of high divergence. BWA-MEM is 5 times as slow, but is scaled well to large genomes.

SNP calling for 35X NA12878

35X 101bp HiSeq paired-end reads were aligned the human genome with BWA, Bowtie2 and BWA-MEM. SNPs on chr20 were called with SAMtools. The table below shows the number of Q10 SNPs calls on chr20 and their transition/transversion (transv) ratio:

<table>
<thead>
<tr>
<th>Mapper</th>
<th>noBAQ, noFLT</th>
<th>BAQ, no FLT</th>
<th>BAQ, BAQ, filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>BWA</td>
<td>83,019/2.04</td>
<td>77,635/2.17</td>
<td>75,910/2.22</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>87,135/2.02</td>
<td>80,234/2.14</td>
<td>78,556/2.22</td>
</tr>
<tr>
<td>BWA-MEM</td>
<td>79,999/2.17</td>
<td>76,693/2.22</td>
<td>76,939/2.23</td>
</tr>
</tbody>
</table>

- Aggressive BWA pairing (right fig). Red reads have exact matches elsewhere. False SNPs like these should have similar transv ratio to true SNPs.
- With clipping penalty. BWA-MEM usually gives cleaner alignment around indels.
- Some problems in mapping can be fixed by BAQ and post filtering, but others not.

REFERENCES