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 hamming 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 higher accuracy, while Bowtie2 (Langmead and Salzberg, 2012) and Cushaw2 (Liu 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:

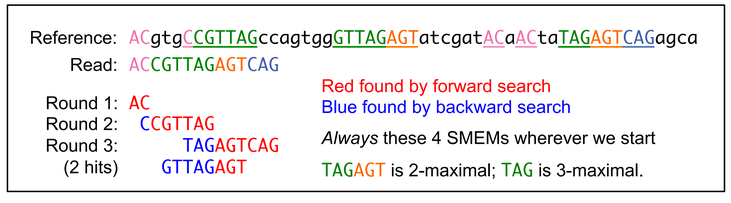
1. Forward and backward strands aligned simultaneously, which is faster than aligning the two strands separately.
2. More straightforward forward-backward search.
3. More advanced seeding algorithm.

Supermaximal Exact Match

Maximal exact match (MEM): an exact match that cannot be extended further in either direction

Super-maximal 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 must be a SMEM.

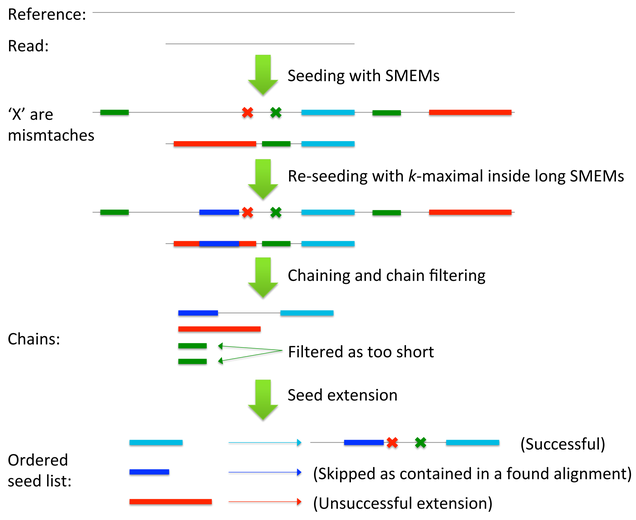
There are usually much fewer SMEMs than MEMs. With forward-backward search, SMEMs can be found very quickly:



k-maximal intervals

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

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



Seed-and-extend

BWA-MEM follows the seed-and-extend paradigm. It advances Bowtie2 and Cushaw2 for long query alignment in that (the following were first implemented in BWA-SW):

- Chaining and chain filtering. After seeding, BWA-MEM chains colinear seeds that are close to each other and filters out very short chains that are significantly overlap with long chains.
- Extension with banded dynamic programming (DP). Bowtie2 and Cushaw2 perform SSE2 Smith-Waterman (SW; Farrar, 2007), while GEM runs the bit-parallelism algorithm for k-different hits.

Seed extension: Z-dropoff

Banded DP may align through the bad region if the flanking good alignments have higher score.

Solution: stop if score drops significantly from the best (X-dropoff). Z-dropoff is similar but not penalizing long gaps in one of the query/reference.

Seed extension: clipping penalty

```
CGATG--GCTAGCATAGCTAGAGTTC
   |||  ||||||||||||||||
   ATGATGCTAGCATAGCTAGACAC
```

Under the BWA-MEM scoring, the best local hit is the green region with both ends clipped off. True variants may be clipped.

BWA-MEM gives a bonus to an extension reaching the end. It may prefer to reach the left end.

Paired-end mapping

BWA-MEM also uses SW to rescue hits missed during single-end (SE) alignment, but different from BWA and BWA-SW which only perform SW for the mate of a unique read, BWA-MEM also apply SW to repetitive hits.

BWA-MEM jointly considers alignment scores, insert size and the possibility of chimera in pairing. For the *i*-th hit for the first read and *j*-th of the second read, let S_i be the SW score of hit *i* and S_j the score of *j*, and d_{ij} be their distance if they are in the right orientation or infinity otherwise. BWA-MEM scores the hit pair (*i*,*j*) as $S_{ij}=S_i+S_j$ - min{-*a*log$_4$*P*(d_{ij}),*U*}, where *P*(*d*) is the probability of observing an insert size larger than *d*, *a* is the matching score and *U* is a threshold balancing the competition between pairing and alignment scores. If both a proper pair and a chimeric pair exist, the proper pair may still be rejected if the sum of the SW scores of the two ends in the chimeric pair is much better.

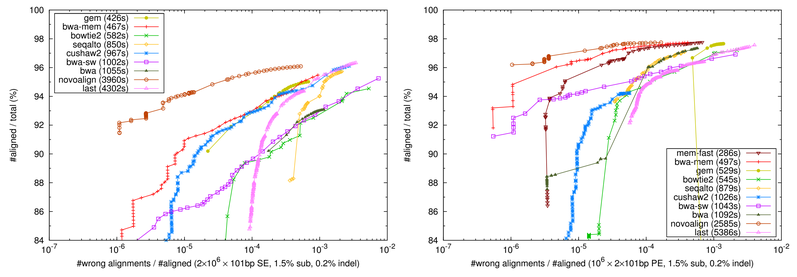
Possible acceleration for long query sequences

Banded DP is the bottleneck for long sequences. It is possible to make BWA-MEM faster by using SSE2 based DP and by restricting DP to regions not covered by a long exact match.

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.



650bp SE

250,000 reads with average length 650bp were aligned to the human genome. BWA-MEM takes 244 seconds, GEM takes 393s, Bowtie2 1454s and Cushaw2 takes 1634s.

Whole-genome alignment between *E. coli* strains

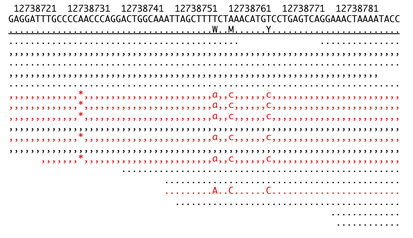
Two *E. coli* strains (NC_000913 and NC_008253) are aligned with both BWA-MEM and MUMmer (Kurtz *et al.*, 2004). MUMmer identified 105,505 substitution differences, while BWA-MEM identified 104,321 of which 102,241 overlap. Most differences unique to one aligner lie in short regions of high divergence. 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 SNP calls on chr20 and their transition:transversion ratio (ts/tv):

Mapper	noBAQ, noFLT	noBAQ, FLT	BAQ, noFLT	BAQ, filtered
Bowtie2	83,019/2.04	77,459/2.17	75,910/2.22	74,484/2.25
BWA	87,133/2.02	80,824/2.14	78,556/2.22	76,673/2.24
BWA-MEM	79,999/2.17	76,693/2.22	76,939/2.23	75,400/2.25

- Aggressive BWA pairing (right fig). Red reads have exact matches elsewhere. False SNPs like these should have similar ts/tv 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

- Farrar, M. (2007). *Bioinformatics,* 23:156–61.
- Kurtz, S. et al. (2004). *Genome Biol,* 5:R12.
- Lam,T.W. *et al.* (2009) In *BIBM,* Washington, DC, USA. pp. 31–36.
- Langmead, B. and Salzberg, S. L. (2012). *Nat Methods,* 9:357–9.
- Li, H. (2012). *Bioinformatics,* 28:1838–44.
- Li, H. (2013). *Submitted*.
- Li, H. and Durbin, R. (2009). *Bioinformatics,* 25:1754–60.
- Li, H. and Durbin, R. (2010). *Bioinformatics,* 26:589–95.
- Liu, Y. and Schmidt, B. (2012). *Bioinformatics,* 28:i318–i324.
- Marco-Sola, S. et al. (2012). *Nat Methods,* 9:1185–8.