

BWA.aln Documentation

Description:	A fast light-weight tool that aligns sequences up to ~200 bp to a sequence database.
Author:	Heng Li, Broad Institute
BWA Version:	0.5.9
Contact:	Marc-Danie Nazaire, gp-help@broadinstitute.org

Summary

Burrows-Wheeler Aligner (BWA.aln) is a fast, light-weight tool that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome. It works for query sequences shorter than 200bp, and does gapped alignment. BWA.aln is usually faster and more accurate on queries with low error rates.

This document is adapted from the BWA documentation for release 0.5.9. For more information about BWA.aln, see the [BWA project site](#). BWA.aln was developed at the Wellcome Trust Sanger Institute and the Broad Institute.

Speed

Speed of alignment is largely determined by the error rate of the query sequences, faster with near-perfect hits and slower for higher error rates. Pairing is slower for shorter reads, mostly because shorter reads have more spurious hits.

In experimental runs, BWA was able to map 2 million 32bp reads to:

- a bacterial genome in several minutes
- the human X chromosome in 8-15 minutes
- the human genome in 15-25 minutes

References

BWA manual page: <http://bio-bwa.sourceforge.net/bwa.shtml>.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*. 2009;25:1754-1760. [PMID: 19451168] (<http://www.ncbi.nlm.nih.gov/pubmed/19451168>)

Parameters

Name	Description
prebuilt.BWA.index	<p>An indexed genome. A number of pre-built indexes are available:</p> <ul style="list-style-type: none"> • <i>A. thaliana</i>, TAIR8 • <i>B. taurus</i>, UMD Freeze 3.0 • <i>E. coli</i> • <i>C. elegans</i>, WormBase, WS200 • <i>H. sapiens</i>, UCSC hg19 • <i>H. sapiens</i>, UCSC hg18 • <i>M. musculus</i>, UCSC mm9 • <i>M. musculus</i>, UCSC mm8 • <i>M. musculus</i>, NCBI v.37 • <i>S. cerevisiae</i> <p>If this list does not include the genome the user requires, an indexed genome can be generated using BWA.indexer. Either a prebuilt or a custom BWA index must be specified.</p>
custom.BWA.index	<p>A ZIP archive containing BWA index files. Either a prebuilt or a custom BWA index must be specified.</p>
reads.pair.1 (required)	<p>Unpaired reads file or first mate for paired reads. This can be a file in FASTA, FASTQ, or BAM format. Note: the FASTA or FASTQ can be gzipped.</p>
reads.pair.2	<p>Second mate for paired reads. This can be a file in FASTA, FASTQ, or BAM format. Note: the FASTA or FASTQ can be gzipped.</p>
bam.mapping	<p>Specifies how to map BAM input. Options include single, first, second, or paired.</p>
max.edit.distance	<p>The maximum edit distance. This specifies a threshold of the maximum number of deletions, insertions, and substitutions needed to transform the reference sequence into the read sequence.</p>

GenePattern

max.num.gap	Maximum number of gap opens. This specifies a threshold of the maximum number of gaps that can be initiated to match a given read to the reference.
max.gap.extension	Maximum number of gap extensions. This specifies a threshold of the maximum number of bases by which gaps in a read can be extended.
max.deletion.length	Disallow a long deletion within this many bp of the 3' end.
max.indel.length	Disallow an indel within this many bp of the ends.
seed.length	The set of bases determined by this option in the high-quality (left) end of the read is the seed.
max.seed.edit.distance	Maximum edit distance in the seed; that is, the maximum number of changes required to transform the reference sequence of the seed into the read sequence of the seed.
mismatch.penalty	Specifies the mismatch penalty.
gap.open.penalty	Gap open penalty. The gap open penalty is the score taken away for the initiation of the gap in sequence. To make the match more significant you can try to make the gap penalty larger.
gap.extension.penalty	Gap extension penalty. The gap extension penalty is added to the standard gap penalty for each base or residue in the gap. To reduce long gaps, increase the extension gap penalty. A few long gaps are expected, rather than many short gaps, so the gap extension penalty should be lower than the gap penalty. (The exception to this rule is where one or both sequences are single reads with possible sequencing errors, in which case many single base gaps are expected. To cope with this, try setting the gap open penalty very low and using the gap extension penalty to control gap scoring.)

GenePattern

max.best.hits	Proceed with suboptimal alignments if there are no more than this many equally best hits. This option only affects paired-end mapping. Increasing this threshold helps to improve the pairing accuracy at the cost of speed, especially for short reads (~32 bp).
reverse.query (required)	Specifies whether to reverse the query sequence but not complement it. This is required for alignment in the color space. Default: no
iterative.search (required)	Specifies whether to disable iterative search. Enabling this will slow the alignment process. Default: no
trim.reads	Specifies a quality threshold for read trimming. The trimming algorithm in BWA scans from the right of the read, accumulating a penalty sum (or "area") for each position that is lower quality than this threshold and reducing this area for each position above that threshold. The read is trimmed to the position where the penalty area is greatest.
Illumina.1.3. format (required)	The input is in the Illumina 1.3+ read format. Default: no
barcode.length	Length of the barcode starting from the 5' end.
max.insert.size	Specifies the maximum insert size for a read pair to be considered to be mapped properly.
max. occurrences	Specifies the maximum occurrences of a read for pairing.
max. alignments	Specifies the maximum number of alignments to output in the XA tag for reads paired properly.
max.dc. alignments	Specifies the maximum number of alignments to output in the XA tag for discordant read pairs (excluding singletons).
num threads (required)	Number of threads. Default: 3

GenePattern

output.prefix (required)	Prefix to use for the output file name.
-----------------------------	---

Output Files

1. SAM file

The aligned sequences are output in SAM format. For more details on this alignment file, see the SAM format specification at <http://samtools.sourceforge.net/SAM-1.3.pdf>.

Platform Dependencies

Module type:	RNA-seq
CPU type:	any
OS:	Macintosh, Linux
Language:	C++, Perl