

GenePattern

BWA.bwasw Documentation

Description:	A fast light-weight tool that aligns long sequences (>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

BWA.bwasw is a fast, light-weight tool that aligns long reads (>200 bp). It performs heuristic Smith-Waterman-like alignment to find high-scoring local hits, and thus chimera. BWA.bwasw can also be used to align ~100bp reads, but it is slower than the BWA.aln module.

This document is adapted from the BWA documentation for release 0.5.9. For more information about BWA.bwasw, see the [BWA project site](#). BWA.bwasw 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. On 350-1000bp reads, BWA.bwasw is several to tens of times faster than other existing programs.

References

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

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26:589-595. [PMID: 20080505]
(<http://bioinformatics.oxfordjournals.org/content/26/5/589.long>)

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>
read.file (required)	<p>Read file for single-end reads. This can be a file in FASTA or FASTQ format.</p>
match.score	<p>The score of the match.</p>
mismatch.penalty	<p>Specifies the mismatch penalty.</p>
gap.open.penalty	<p>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.</p>

GenePattern

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.)
band.width	Band width in the banded alignment.
min.score.threshold	Minimum score threshold divided by the match score.
threshold.coefficient	Coefficient for threshold adjustment according to query length.
z.best.heuristics	Z-best heuristics. Specifying a higher number increases accuracy at the cost of speed.
max.sa.interval.sizes	Maximum SA interval size for initiating a seed. Specifying a higher number increases accuracy at the cost of speed.
min.num.seeds	The minimum number of seeds contained in the best alignment from the forward-forward alignment process that allows the algorithm to skip performing the reverse alignment. BWA.bwasw tends to be faster and more accurate if the alignment is supported by more seeds.
num threads (required)	Number of threads. Default: 4
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>.

GenePattern

Platform Dependencies

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