GenePattern

TopHat Documentation

Description:	A fast splice junction mapper for RNA-seq reads.
Author:	Cole Trapnell et al, University of Maryland Center for Bioinformatics and Computational Biology
TopHat Version	release 1.3.3
Contact:	Marc-Danie Nazaire, gp-help@broadinstitute.org

Summary

TopHat is a fast splice junction mapper for RNA-seq reads. It aligns RNA-seq reads to mammalian-sized genomes and then analyzes the mapping results to identify splice junctions between exons. The software is optimized for reads 75bp or longer.

TopHat 1.3.3 does not allow short (fewer than a few nucleotides) insertions and deletions in the alignments it reports. Also, mixing paired- and single-end reads together is not supported.

TopHat requires Python version 2.4 or higher.

TopHat was created at the University of Maryland Center for Bioinformatics and Computational Biology. This document is adapted from the TopHat documentation for release 1.3.3. For more information, see the TopHat documentation.

IMPORTANT NOTES:

The first time you run a job with a given prebuilt index, the job may fail. Please re-run your job. It should work on the second run. If you encounter a recurring problem with your jobs failing in TopHat, contact gp-help@broadinstitute.org.

TopHat is memory intensive and takes several hours to run. Test runs with 20GB of RAM available on the GenePattern Public server took 3 hours or more. On servers with less available memory, it was not unusual for a test run of TopHat to take upwards of 12 hours.

References

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009;25:1105-11. (http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btp120)

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10:R25. (http://genomebiology.com/2009/10/3/R25)

Links

TopHat: http://tophat.cbcb.umd.edu/

TopHat documentation: http://tophat.cbcb.umd.edu/manual.html

Example data http://tophat.cbcb.umd.edu/downloads/test_data.tar.gz

Parameters

Name	Description
prebuilt.bowtie.index	An indexed genome. A number of pre-built indexes are available: • *A. thaliana*, TAIR8 • *B. taurus*, UMD Freeze 3.0 • *E. coli* • *C. elegans,* WormBase, WS200 • *H. sapiens*, UCSC hg19 • *H. sapiens*, UCSC hg18 • *M. musculus*, UCSC mm9 • *M. musculus*, UCSC mm8 • *M. musculus*, NCBI 37 • *S. cerevisiae* If this list does not include the genome the user requires, an indexed genome can be generated using Bowtie.indexer. **Either** a prebuilt or a custom Bowtie index **must** be specified.
custom.bowtie.index	A ZIP archive containing Bowtie index files. **Either** a prebuilt or a custom Bowtie index **must** be specified.
reads.pair.1 (required)	Unpaired reads file or first mate for paired reads. This can be a file in FASTA or FASTQ format (bz2 and gz compressed files are supported), a ZIP archive containing FASTA or FASTQ files, or a directory that is accessible to the GenePattern server containing FASTA or FASTQ files. For more information on the FASTA format, see the NIH description here: http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml. For more information on the FASTQ format, see the specification here: http://nar.oxfordjournals.org/content/early/2009/12/16/nar.gkp1137.full.
reads.pair.2 (optional)	Second mate for paired reads. This can be a file in FASTA or FASTQ format, a ZIP archive containing FASTA or FASTQ files.
mate.inner.dist (optional)	The expected mean inner distance between mate pairs. For example, for paired-end runs with fragments selected at 300 bp, where each end is 50 bp, you should set this to be 200. Default: 50

mate.std.dev (optional)	The standard deviation for the distribution on inner distances between mate pairs. This does not have to be specified for paired end reads.
library.type (optional)	Library type for strand specific reads. Options include: • Standard Illumina • dUTP, NSR, NNSR • Ligation, Standard SOLiD
GTF.file (optional)	A GTF or GFF file containing a list of gene model annotations. (for more information on GTF format, see the specification: http://mblab.wustl.edu/GTF22.html, for more information on GFF format, see the specification: http://www.sequenceontology.org/gff3.shtml) The exon records in this file will be used to: • build a set of known splice junctions for each gene • attempt to align reads to these junctions even if they would not normally be covered by the initial mapping
raw.junctions.file (optional)	A file containing raw junctions. Junctions are specified one per line, in a tab-delimited format.
find.novel.junctions (optional)	Default: *yes*. If you select *no*, then the module will only look for junctions indicated in the GTF file supplied in the *GTF file* parameter. (This parameter is ignored when no GTF 2.2 file is specified.) Default: yes
min.anchor.length (optional)	The anchor length. TopHat will report junctions spanned by reads with at least this many bases on each side of the junction. Note that individual spliced alignments may span a junction with fewer than this many bases on one side. However, every junction involved in spliced alignments is supported by at least one read with this many bases on each side. This value must be at least 3. Default: 8
splice.mismatches (optional)	The maximum number of mismatches that may appear in the "anchor" region of a spliced alignment. Default: 0
min.intron.length (optional)	The minimum intron length. TopHat will ignore donor/acceptor pairs closer than this many bases apart. Default: 70
max.intron.length (optional)	The maximum intron length. When searching for junctions *ab initio*, TopHat will ignore donor/acceptor pairs farther than this many bases apart, except when such a pair is supported by a split segment alignment of a long read. Default: 500000

indel.search (optional)	Whether to allow indel search. Default: yes
max.insertion.length (optional)	The maximum insertion length. Default: 3.
max.deletion.length (optional)	The maximum deletion length. Default: 3.
use.solexa.scale (optional)	Use the Solexa scale for quality values in FASTQ files. For more information on quality values, see http://maq.sourceforge.net/fastq.shtml. Default: no
use.solexa.1.3.scale (optional)	As of the Illumina GA pipeline version 1.3, quality scores are encoded in Phred-scaled base-64. Use this option for FASTQ files from 1.3 or later. Default: no
quality.value.files.1 (optional)	A ZIP file containing separate quality value files for single end reads or the first pair of paired end reads. Colorspace read files (CSFASTA) come with separate quality value files.
quality.value.files.2 (optional)	A ZIP file containing separate quality value files for the second pair of paired end reads. Colorspace read files (CSFASTA) come with separate quality value files.
integer.quals (optional)	Quality values are space-delimited integer values; this becomes the default when you select *Yes* for *colorspace reads*. Default: no
colorspace.reads (optional)	Uses colorspace reads. Note that this option uses a colorspace Bowtie index. Colorspace is the characteristic output format of Applied Biosystems' SOLiD system. In a colorspace read, each character is a color rather than a nucleotide, where a color encodes a class of dinucleotides. For instance, the color blue encodes any of the dinucleotides: AA, CC, GG, TT. Colorspace has the advantage of (often) being able to distinguish sequencing errors from SNPs once the read has been aligned. Prebuilt colorspace indexes are available on the Bowtie website: http://bowtie-bio.sourceforge.net/index.shtml Default: no

min.isoform.fraction (optional)	Filters out junctions supported by too few alignments. Suppose a junction spanning two exons, is supported by S reads. Let the average depth of coverage of exon A be D, and assume that it is higher than B. If S/D is less than the minimum isoform fraction, the junction is not reported. A value of zero disables the filter. Default: 0.15
max.multihits (optional)	Allows up to the specified alignments to the reference for a given read, and suppresses all alignments for reads with more than this many alignments. Default: 40
initial.read.mismatch (optional)	Specifies the number of mismatches allowed in the initial read mapping in each read alignment. Default: 2
closure.search (optional)	Enables or disables the mate pair closure-based search for junctions. Closure-based search should only be used when the expected inner distance between mates is small (≤50 bp). Default: no
coverage.search (optional)	Enables or disables the coverage-based search for junctions. Use when coverage search is disabled by default (such as for reads ≥75 bp), for maximum sensitivity. Default: no
microexon.search (optional)	Attempts to find alignments incident to microexons. Works only for reads ≥50 bp. Default: no
butterfly.search (optional)	Enables or disables a slower, but potentially more sensitive, algorithm to find junctions in addition to its standard search. Consider using this if you expect that your experiment produced a lot of reads from pre-mRNA, that fall within the introns of your transcripts. Default: *no*.
num.threads (optional)	Specify how many parallel search threads to launch. All threads find alignments in parallel, increasing alignment throughput by approximately a multiple of the number of threads (though in practice, speedup is somewhat worse than linear). Default: 1 NOTE: There is a known issue that specifying more than 1 thread will cause the results to be non-deterministic.
output.prefix (required)	The prefix to use for the output file

Output Files

1. <output.prefix>.bam

A list of read alignments in BAM format. BAM is the binary equivalent of SAM, a compact short read alignment format. For more information on the SAM/BAM formats, see the specification at: http://samtools.sourceforge.net.

2. <output.prefix>.junctions.bed

   A BED file of junctions reported by TopHat (for more information on the BED format, see: http://genome.ucsc.edu/FAQ/FAQformat.html). Each junction consists of two connected BED blocks, where each block is as long as the maximal overhang of any read spanning the junction. The score is the number of alignments spanning the junction.

3. <output.prefix>.insertions.bed
   UCSC BED tracks of insertions reported by TopHat.
   insertions.bed - chromLeft refers to the last genomic base before the insertion.
4. <output.prefix>.deletions.bed.

   UCSC BED tracks of deletions reported by TopHat.
   deletions.bed - chromLeft refers to the first genomic base of the deletion.

5. left_kept_reads.info and right_kept_reads.info

   Contain statistics about the reads that are parsed from the FASTA/FASTQ input files. *min_read* and *max_read* are the minimum and maximum lengths of the read sequences. *reads_in* is the total number of reads that were found. *reads_out* is the total number of reads that were kept after filtering out according to parameter settings.

Platform Dependencies

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

GenePattern Module Version Notes

Version	Description
v.5	TopHat module v.5 contains updates to TopHat versions 1.3.1, 1.3.2, and 1.3.3, using SAMtools version 0.1.17. See the TopHat documentation (http://tophat.cbcb.umd.edu/) for more information about these versions. Improvements, modifications, and bug fixes include: • Fixed the sorting of read files so that paired read files are correctly matched • Changed the default value of the indel search parameter to *yes* • Added an output prefix parameter so that the output files can be renamed