TopHat

A spliced read mapper for RNA-Seq

McKUSICK-NATHANS Institute of Genetic Medicine

CBCB

Manual

What is TopHat?

TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splice junctions. It is built on the ultrafast short read mapping program Bowtie. TopHat runs on **Linux** and **OS X**.

What types of reads can I use TopHat with?

TopHat was designed to work with reads produced by the Illumina Genome Analyzer, although users have been successful in using TopHat with reads from other technologies. In TopHat 1.1.0, we began supporting Applied Biosystems' Colorspace format. The software is optimized for reads 75bp or longer.

Mixing paired- and single- end reads together is **not** supported.

How does TopHat find junctions?

TopHat can find splice junctions without a reference annotation. By first mapping RNA-Seq reads to the genome, TopHat identifies potential exons, since many RNA-Seq reads will contiguously align to the genome. Using this initial mapping information, TopHat builds a database of possible splice junctions and then maps the reads against these junctions to confirm them.

Short read sequencing machines can currently produce reads 100bp or longer but many exons are shorter than this so they would be missed in the initial mapping. TopHat solves this problem mainly by splitting all input reads into smaller segments which are then mapped independently. The segment alignments are put back together in a final step of the program to produce the end-to-end read alignments.

TopHat generates its database of possible splice junctions from two sources of evidence. The first and strongest source of evidence for a splice junction is when two segments from the

Site Map

Home
Getting started
Manual
Index and annotation downloads
FAQ
Protocol

News and updates

New releases and related tools will be announced through the **mailing list**

Getting Help

Questions about TopHat should be sent to **tophat.cufflinks@gmail.com**. Please do not email technical questions to TopHat contributors directly.

Releases

version 2.0.6 (BETA) 11/02/2012
- Source code
- Linux x86_64 binary
- Mac OS X x86_64 binary

Related Tools

Cufflinks: Isoform assembly and quantitation for RNA-Seq
Bowtie: Ultrafast short read alignment
TopHat-Fusion: An algorithm for Discovery of Novel Fusion Transcripts
CummeRbund: Visualization of RNA-Seq differential analysis

Pre-built indexes

H. sapiens, UCSC hg18	**2.7 GB**
colorspace: full	
H. sapiens, UCSC hg19	**2.7 GB**
colorspace: full	
M. musculus, UCSC mm9	**2.4 GB**

same read (for reads of at least 45bp) are mapped at a certain distance on the same genomic sequence or when an internal segment fails to map - again suggesting that such reads are spanning multiple exons. With this approach, "GT-AG", "GC-AG" and "AT-AC" introns will be found ab initio. The second source is pairings of "coverage islands", which are distinct regions of piled up reads in the initial mapping. Neighboring islands are often spliced together in the transcriptome, so TopHat looks for ways to join these with an intron. We only suggest users use this second option (--coverage-search) for short reads (< 45bp) and with a small number of reads (<= 10 million). This latter option will only report alignments across "GT-AG" introns

Prerequisites

To use TopHat, you will need the following programs in your `PATH`:

- `bowtie2 and bowtie2-align (or bowtie)`
- `bowtie2-inspect (or bowtie-inspect)`
- `bowtie2-build (or bowtie-build)`
- `samtools`

Because TopHat outputs and handles alignments in BAM format, you will need to download and install the SAM tools. You may want to take a look at the Getting started guide for more detailed installation instructions, including installation of SAM tools and Boost.

You will also need Python **version 2.6 or higher**.

Obtaining and installing TopHat

You can download the latest source release and precompiled binaries for Linux and Mac OSX here. See the Getting started guide for detailed instructions about installing TopHat from the binary package or building TopHat and its dependencies from source.

To install TopHat from source package, unpack the tarball and change directory to the package directory as follows:

```
tar zxvf tophat-2.0.0.tar.gz
cd tophat-2.0.0/
```

Configure the package, specifying the install path and the library dependencies as needed (see the Getting started guide for details):

```
./configure --prefix=<install_prefix> --with-
boost=<boost_install_prefix> --with-bam=<samtools_install_prefix>
```

Finally, build and install TopHat:

```
make
```

colorspace: full

All indexes are for assemblies, not contigs. Unplaced or unlocalized sequences and alternate haplotype assemblies are excluded.

Some unzip programs cannot handle archives >2 GB. If you have problems downloading or unzipping a >2 GB index, try downloading in two parts.

Check .zip file integrity with MD5s.

Pre-built indexes are compatible with Bowtie versions 0.9.8 and later. For older indexes, please contact us.

Publications

Trapnell C, Pachter L, Salzberg SL. **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* doi:10.1093/bioinformatics/btp120

Langmead B, Trapnell C, Pop M, Salzberg SL. **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biology* **10**:R25.

Kim D and Salzberg SL. **TopHat-Fusion: an algorithm for discovery of novel fusion transcripts**. *Genome Biology* 2011, 12:R72

Contributors

Cole Trapnell
Daehwan Kim
Geo Pertea
Harold Pimentel
Ryan Kelley
Lior Pachter
Steven Salzberg

Links

University of Maryland
University of California, Berkeley
UMD CBCB
UMD Computer Science
Berkeley Mathematics
UMIACS

```
make install
```

As detailed in the Getting started guide, if you want to install TopHat 2 without overwriting a previous version of TopHat already installed on your system you should specify a new, separate <install_prefix> for the ./configure command above, and after the 'make install' step just copy the tophat2 script from <install_prefix>/bin to a directory that is in your shell's PATH, so you can invoke this new version of TopHat with the command 'tophat2'.

Below you will find a detailed list of command-line options you can use to control TopHat. Beginning users should take a look at the Getting started guide for a tutorial on installing and running TopHat and its prerequisites.

Please Note TopHat has a number of parameters and options, and their default values are tuned for processing **mammalian** RNA-Seq reads. If you would like to use TopHat for another class of organism, we recommend setting some of the parameters with more strict, conservative values than their defaults. Usually, setting the maximum intron size to 4 or 5 Kb is sufficient to discover most junctions while keeping the number of false positives low.

Using `TopHat`

The following is a detailed description of the options used to control the `tophat` script:

```
Usage: tophat [options]* <index_base> <reads1_1[,...,readsN_1]>
[reads1_2,...readsN_2]
```

When running TopHat with paired ends, it is **critical** that the *_1 files an the *_2 files appear in separate comma separated lists, and that the order of the files in the two lists is the same.

NOTE: TopHat can align reads that are up to 1024 bp, and it handles paired-end reads, but we **do not recommend** mixing different types of reads in the same TopHat run. For example, mixing 100bp single end reads and 2x27bp paired ends into the same TopHat run may give bad results. If you'd like to combine results from data sets with different types of RNA-Seq reads, you can follow a protocol like this:

- run TopHat on the first set of reads, with the appropriate parameters for this data set
- use **bed_to_juncs** to convert the **junctions.bed** file obtained in this first run to a junction file usable by Tophat's

-j option

- run Tophat on the 2nd set of reads using the **-j** option to supply the junctions file produced by bed_to_juncs in the previous step

Arguments:

`<ebwt_base>`	The basename of the index to be searched. The basename is the name of any of the five index files up to but not including the first period. Bowtie first looks in the current directory for the index files, then looks in the `indexes` subdirectory under the directory where the currently-running `bowtie` executable is located, then looks in the directory specified in the `BOWTIE_INDEXES` (or `BOWTIE2_INDEXES`) environment variable.
`<reads1_1[,...,readsN_1]>`	A comma-separated list of files containing reads in FASTQ or FASTA format. When running TopHat with paired-end reads, this should be the *_1 ("left") set of files.
`<[reads1_2,...readsN_2]>`	A comma-separated list of files containing reads in FASTA or FASTA format. Only used when running TopHat with paired end reads, and contains the *_2 ("right") set of files. The *_2 files **MUST** appear in the same order as the *_1 files.

Options:

`-h/--help`	Prints the help message and exits
`-v/--version`	Prints the TopHat version number and exits
`-N/--read-mismatches`	Final read alignments having

	more than these many mismatches are discarded. The default is 2.
`--read-gap-length`	Final read alignments having more than these many total length of gaps are discarded. The default is 2.
`--read-edit-dist`	Final read alignments having more than these many edit distance are discarded. The default is 2.
`--read-realign-edit-dist`	Some of the reads spanning multiple exons may be mapped incorrectly as a contiguous alignment to the genome even though the correct alignment should be a spliced one - this can happen in the presence of processed pseudogenes that are rarely (if at all) transcribed or expressed. This option can direct TopHat to re-align reads for which the edit distance of an alignment obtained in a previous mapping step is above or equal to this option value. If you set this option to `0`, TopHat will map every read in all the mapping steps (transcriptome if you provided gene annotations, genome, and finally splice variants detected by TopHat), reporting the best possible alignment found in any of these mapping steps. This may greatly increase the mapping accuracy at the expense of an increase in running time. The default value for this option is set such that TopHat will not try to realign reads already mapped in earlier steps.
`--bowtie1`	Uses Bowtie1 instead of Bowtie2. If you use colorspace

	reads, you need to use this option as Bowtie2 does not support colorspace reads.
`-o/--output-dir <string>`	Sets the name of the directory in which TopHat will write all of its output. The default is "./tophat_out".
`-r/--mate-inner-dist <int>`	This is the expected (mean) inner distance between mate pairs. For, example, for paired end runs with fragments selected at 300bp, where each end is 50bp, you should set -r to be 200. The default is 50bp.
`--mate-std-dev <int>`	The standard deviation for the distribution on inner distances between mate pairs. The default is 20bp.
`-a/--min-anchor-length <int>`	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 must be at least 3 and the default is 8.
`-m/--splice-mismatches <int>`	The maximum number of mismatches that may appear in the "anchor" region of a spliced alignment. The default is 0.
`-i/--min-intron-length <int>`	The minimum intron length. TopHat will ignore donor/acceptor pairs closer than this many bases apart. The default is 70.
`-I/--max-intron-length <int>`	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. The default is 500000.
`--max-insertion-length <int>`	The maximum insertion length. The default is 3.
`--max-deletion-length <int>`	The maximum deletion length. The default is 3.
`--solexa-quals`	Use the Solexa scale for quality values in FASTQ files.
`--solexa1.3-quals`	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 pipeline 1.3 or later.
`-Q/--quals`	Separate quality value files - colorspace read files (CSFASTA) come with separate qual files.
`--integer-quals`	Quality values are space-delimited integer values, this becomes default when you specify -C/--color.
`-C/--color`	Colorspace reads, note that it uses a colorspace bowtie index and requires Bowtie 0.12.6 or higher. Common usage: tophat --color --quals [other options]* <colorspace_index_base> <reads1_1[,...,readsN_1]> [reads1_2,...readsN_2] <quals1_1[,...,qualsN_1]> [quals1_2,...qualsN_2]
`-p/--num-threads <int>`	Use this many threads to align reads. The default is 1.
`-g/--max-multihits <int>`	Instructs TopHat to allow up to this many alignments to the

	reference for a given read, and choose the alignments based on their alignment scores if there are more than this number. The default is 20 for read mapping. Unless you use --report-secondary-alignments, TopHat will report the alignments with the best alignment score. If there are more alignments with the same score than this number, TopHat will randomly report only this many alignments. In case of using --report-secondary-alignments, TopHat will try to report alignments up to this option value, and TopHat may randomly output some of the alignments with the same score to meet this number.
`--report-secondary-alignments`	By default TopHat reports best or primary alignments based on alignment scores (AS). Use this option if you want to output additional or secondary alignments (up to 20 alignments will be reported this way, this limit can be changed by using the -g/--max-multihits option above).
`--no-discordant`	For paired reads, report only concordant mappings.
`--no-mixed`	For paired reads, only report read alignments if both reads in a pair can be mapped (by default, if TopHat cannot find a concordant or discordant alignment for both reads in a pair, it will find and report alignments for each read separately; this option disables that behavior).
`--no-coverage-search`	Disables the coverage based

	search for junctions.
`--coverage-search`	Enables the coverage based search for junctions. Use when coverage search is disabled by default (such as for reads 75bp or longer), for maximum sensitivity.
`--microexon-search`	With this option, the pipeline will attempt to find alignments incident to micro-exons. Works only for reads 50bp or longer.
`--library-type`	TopHat will treat the reads as strand specific. Every read alignment will have an `XS` attribute tag. Consider supplying library type options below to select the correct RNA-seq protocol.

Library Type	Examples	Description
fr-unstranded	`Standard Illumina`	Reads from the left-most end of the fragment (in transcript coordinates) map to the transcript strand, and the right-most end maps to the opposite strand.
fr-firststrand	`dUTP, NSR, NNSR`	Same as above except we enforce the rule that the right-most end of the fragment (in transcript coordinates) is the first sequenced (or only sequenced for single-end reads). Equivalently, it is assumed that only the strand

		generated during first strand synthesis is sequenced.
fr-secondstrand	Ligation, Standard SOLiD	Same as above except we enforce the rule that the left-most end of the fragment (in transcript coordinates) is the first sequenced (or only sequenced for single-end reads). Equivalently, it is assumed that only the strand generated during second strand synthesis is sequenced.

Advanced Options:

--bowtie-n	TopHat uses "-v" in Bowtie for initial read mapping (the default), but with this option, "-n" is used instead. Read segments are always mapped using "-v" option.
--segment-mismatches	Read segments are mapped independently, allowing up to this many mismatches in each segment alignment. The default is 2.
--segment-length	Each read is cut up into segments, each at least this long. These segments are mapped independently. The default is 25.
--min-segment-intron	The minimum intron length that may be found during split-segment search. The default is 50.
--max-segment-intron	The maximum intron length that may be found during split-segment search. The default is 500000.
--min-coverage-intron	The minimum intron length that may be

found during coverage search. The default is 50.

`--max-coverage-intron` The maximum intron length that may be found during coverage search. The default is 20000.

`--keep-tmp` Causes TopHat to preserve its intermediate files produced during the run (mostly useful for debugging). The default is to delete these temporary files.

`--keep-fasta-order` In order to sort alignments in the same order in the genome fasta file, the option can be used. But this option will make the output SAM/BAM file incompatible with those from the previous versions of TopHat (1.4.1 or lower).

`--no-sort-bam` Output BAM is not coordinate-sorted.

`--no-convert-bam` Do not convert to bam format. Output is <output_dir>/accepted_hit.sam. Implies --no-sort-bam.

`-R/--resume <string>` In case a TopHat run was terminated prematurely (process failure due to external factors, e.g. running out of memory because of other processes running on the same machine, or the disk getting full), users can attempt to resume the interrupted TopHat run by just providing this option with the output directory for that run. TopHat sets several checkpoints after every lengthy operations in the pipeline and when this option is provided, it will attempt to resume the pipeline from the last successful checkpoint. This special usage of TopHat only requires this option, e.g. the command line could simply be:

`tophat -R tophat_out` *(or your TopHat output directory if you used the -o/--output-dir option)*

Note that none of the original options used for the original TopHat run should be provided, TopHat will find all the original options (and the checkpoint info) in the logs/run.log file found in the specified directory.

`-z/--zpacker`	Manually specify the program used for compression of temporary files; default is `gzip`; use `-z0` to disable compression altogether. Any program that is option-compatible with gzip can be used (e.g. bzip2, pigz, pbzip2).

Bowtie 2 specific options:

Bowtie 2 provides many options so that users can have more flexibility as to how reads are mapped. TopHat 2 allows users to pass many of these options to Bowtie 2 by preceding the Bowtie 2 option name with the **`--b2-`** prefix. Please refer to the Bowtie2 website for detailed information.

Preset options in --end-to-end mode (local alignment is not used in TopHat2):

Tophat 2 option:	*Corresponding Bowtie 2 option:*
`--b2-very-fast`	--very-fast
`--b2-fast`	--fast
`--b2-sensitive`	--sensitive
`--b2-very-sensitive`	--very-sensitive

Alignment options:

`--b2-N`	The default is 0.
`--b2-L`	The default is 20.
`--b2-i`	The default is S,1,1.25.
`--b2-n-ceil`	The default is L,0,0.15.
`--b2-gbar`	The default is 4.

Scoring options:

`--b2-mp`	The default is 6,2.
`--b2-np`	The default is 1.
`--b2-rdg`	The default is 5,3.
`--b2-rfg`	The default is 5,3.
`--b2-score-min`	The default is L,-0.6,-0.6.

Effort options:

`--b2-D`	The default is 15.

`--b2-R`	The default is 2.

Fusion mapping options:

Reads can be aligned to potential fusion transcripts if the --fusion-search option is specified. The fusion alignments are reported in SAM format using custom fields XF and XP (see the output format) and some additional information about fusions will be reported (see fusions.out). Once mapping is done, you can run tophat-fusion-post to filter out fusion transcripts (see the TopHat-Fusion website for more details).

`--fusion-search`	Turn on fusion mapping
`--fusion-anchor-length`	A "supporting" read must map to both sides of a fusion by at least these many bases. The default is 20.
`--fusion-min-dist`	For intra-chromosomal fusions, TopHat-Fusion tries to find fusions separated by at least this distance. The default is 10000000.
`--fusion-read-mismatches`	Reads support fusions if they map across fusion with at most these many mismatches. The default is 2.
`--fusion-multireads`	Reads that map to more than these many places will be ignored. It may be possible that a fusion is supported by reads (or pairs) that map to multiple places. The default is 2.
`--fusion-multipairs`	Pairs that map to more than these many places will be ignored. The default is 2.
`--fusion-ignore-chromosomes`	Ignore some chromosomes such as chrM when detecting fusion break points. Please check the correct names for chromosomes, that is, mitochondrial DNA is represented as chrM or M depending on the annotation you use.

Supplying your own transcript annotation data:

The options below allow you validate your own list of known transcripts or junctions with your RNA-Seq data. Note that the chromosome names in the files provided with the options below **must match** the names in the Bowtie index. These names are **case-senstitive**.

`-j/--raw-juncs <.juncs file>`

Supply TopHat with a list of raw junctions. Junctions are specified one per line, in a tab-delimited format. Records look like:

```
<chrom> <left> <right> <+/->
```

`left` and `right` are zero-based coordinates, and specify the last character of the left sequenced to be spliced to the first character of the right sequence, inclusive. That is, the last and the first positions of the flanking exons. Users can convert `junctions.bed` (one of the TopHat outputs) to this format using **bed_to_juncs < junctions.bed > new_list.juncs** where bed_to_juncs can be found under the same folder as tophat

`--no-novel-juncs`

Only look for reads across junctions indicated in the supplied GFF or junctions file. (ignored without -G/-j)

`-G/--GTF <GTF/GFF3 file>`

Supply TopHat with a set of gene model annotations and/or known transcripts, as a GTF 2.2 or GFF3 formatted file. If this option is provided, TopHat will first extract the transcript sequences and use Bowtie to align reads to this virtual transcriptome first. Only the reads that do not fully map to the transcriptome will then be mapped on the genome. The reads that did map on the transcriptome will be converted to genomic mappings (spliced as needed) and merged with the novel mappings and junctions in the final tophat output.

Please note that the values in the first column of the provided GTF/GFF file (column which

indicates the chromosome or contig on which the feature is located), **must match** the name of the reference sequence in the Bowtie index you are using with TopHat. You can get a list of the sequence names in a Bowtie index by typing:

```
bowtie-inspect --names
your_index
```

So before using a known annotation file with this option please make sure that the 1st column in the annotation file uses the **exact same chromosome/contig names (case sensitive)** as shown by the `bowtie-inspect` command above.

`--transcriptome-index <dir/prefix>`

When providing TopHat with a known transcript file (-G/--GTF option above), a transcriptome sequence file is built and a Bowtie index has to be created for it in order to align the reads to the known transcripts. Creating this Bowtie index can be time consuming and in many cases the same transcriptome data is being used for aligning multiple samples with TopHat. A transcriptome index and the associated data files (the original GFF file) can be thus reused for multiple TopHat runs with this option, so these files are only created for the first run with a given set of transcripts. If multiple TopHat runs are planned with the same transcriptome data, TopHat should be first run with the -G option and with the --transcriptome-index option pointing to a directory and a name prefix which will indicate

where the transcriptome data files will be stored. Then subsequent TopHat runs using the same --transcriptome-index option value will directly use the transcriptome data created in the first run (no -G option needed for subsequent runs).
For example the first TopHat run could look like this:

```
tophat -o out_sample1 -G
known_genes.gtf \
--transcriptome-
index=transcriptome_data/known \
hg19 sample1_1.fq.z
```

In this example the first run will create the transcriptome_data directory if it doesn't exist, and files known.fa, known.gff and known.*ebwt (Bowtie index files) will be generated in that directory. Then for subsequent runs with the same genome and known transcripts but different reads (e.g. sample2_2.fq.z etc.), TopHat will no longer spend time building the transcriptome index because it can directly use the previously built transcriptome index, so the -G option can be discarded for subsequent runs (however using it again will not force TopHat to build the transcriptome index files again if they are already present)

```
tophat -o out_sample2 \
--transcriptome-
index=transcriptome_data/known \
hg19 sample2_1.fq.z
```

(The following options in this section are only used when the transcriptome search was activated with `-G/--GTF` and/or `--transcriptome-index`)

`-T/--transcriptome-only` — Only align the reads to the transcriptome and report only those mappings as genomic mappings.

`-x/--transcriptome-max-hits`	Maximum number of mappings allowed for a read, when aligned to the transcriptome (any reads found with more then this number of mappings will be discarded).
`-M/--prefilter-multihits`	When mapping reads on the transcriptome, some repetitive or low complexity reads that would be discarded in the context of the genome may appear to align to the transcript sequences and thus may end up reported as mapped to those genes only. This option directs TopHat to first align the reads to the whole genome in order to determine and exclude such multi-mapped reads (according to the value of the `-g/--max-multihits` option).

Supplying your own insertions/deletions:

The options below allow you validate your own indels with your RNA-Seq data. Note that the chromosome names in the files provided with the options below **must match** the names in the Bowtie index. These names are **case-senstitive**.

`--insertions/--deletions <.juncs file>`	Supply TopHat with a list of insertions or deletions with respect to the reference. Indels are specified one per line, in a tab-delimited format, identical to that of junctions. Records look like: `<chrom> <left> <right> <+/->` `left` and `right` are zero-based coordinates, and specify the last character of the left sequenced to be spliced to the first character of the right sequence, inclusive.
`--no-novel-indels`	Only look for reads across indels in the supplied indel

file, or disable indel detection when no file has been provided.

TopHat Output

The `tophat` script produces a number of files in the directory in which it was invoked. Most of these files are internal, intermediate files that are generated for use within the pipeline. The output files you will likely want to look at are:

1. `accepted_hits.bam`. A list of read alignments in SAM format. SAM is a compact short read alignment format that is increasingly being adopted. The formal specification is here.
2. `junctions.bed`. A UCSC BED track of junctions reported by TopHat. 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. `insertions.bed and deletions.bed`. UCSC BED tracks of insertions and deletions reported by TopHat.
   Insertions.bed - chromLeft refers to the last genomic base before the insertion.
   Deletions.bed - chromLeft refers to the first genomic base of the deletion.

This research was supported in part by NIH grants R01-LM06845 and R01-GM083873 and NSF grant CCF-0347992.

Administrator: Cole Trapnell. Design by David Herreman