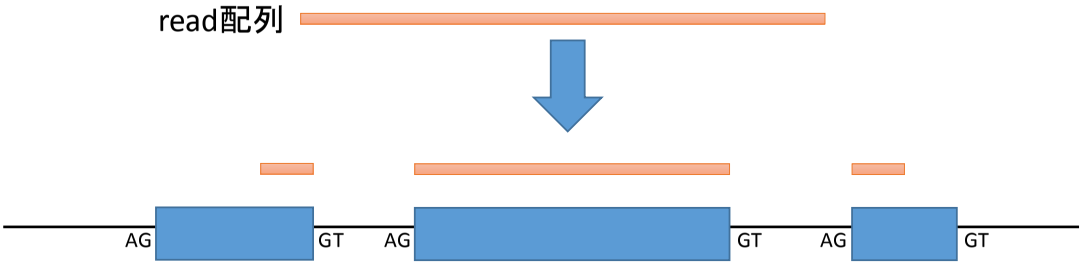
基生研ゲノムインフォマティックストレーニングコース　2015秋
2015.09.09-2015.09.11

# RNA-Seqパイプライン
# ゲノムベースの解析法

基礎生物学研究所・生物機能解析センター
山口勝司

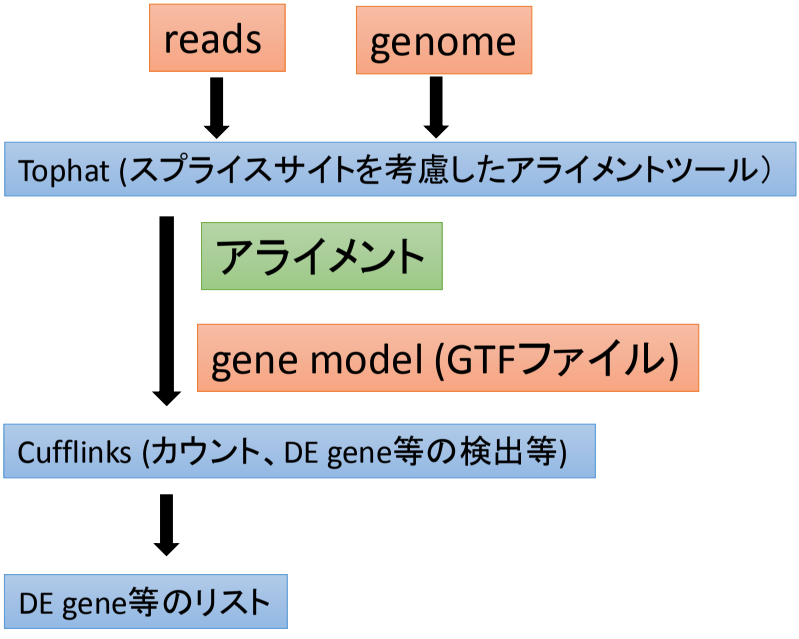
## genomeをレファレンスとする場合

レファレンスがゲノム配列の場合
イントロン配列のスプライシングを考慮した
アライメントを行う必要がある。
TopHatを用いる
他　Blat, SpliceMap, MapSplice, GSMAP, QPALMA

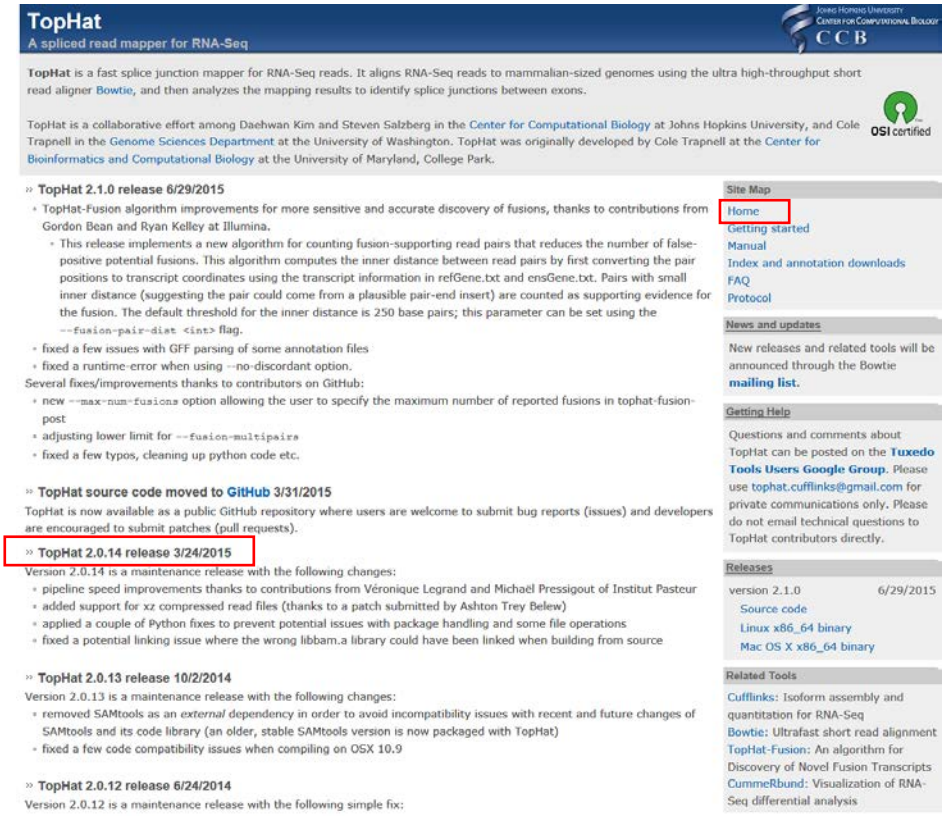


# 実際こんな感じにアラインされる


# 本トレーニングコースでの流れ



TopHat2になりalignerとして
Bowtie2に対応
indelを考慮したアライメントが
可能になった　2012.4



TopHat
A spliced read mapper for RNA-Seq

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort among Daehwan Kim and Steven Salzberg in the Center for Computational Biology at Johns Hopkins University, and Cole Trapnell in the Genome Sciences Department at the University of Washington. TopHat was originally developed by Cole Trapnell at the Center for Bioinformatics and Computational Biology at the University of Maryland, College Park.

OSI certified

» TopHat 2.1.0 release 6/29/2015

- TopHat-Fusion algorithm improvements for more sensitive and accurate discovery of fusions, thanks to contributions from Gordon Bean and Ryan Kelley at Illumina.
  - This release implements a new algorithm for counting fusion-supporting read pairs that reduces the number of false-positive potential fusions. This algorithm computes the inner distance between read pairs by first converting the pair positions to transcript coordinates using the transcript information in refGene.txt and ensGene.txt. Pairs with small inner distance (suggesting the pair could come from a plausible pair-end insert) are counted as supporting evidence for the fusion. The default threshold for the inner distance is 250 base pairs; this parameter can be set using the `--fusion-pair-dist <int>` flag.
- fixed a few issues with GFF parsing of some annotation files
- fixed a runtime-error when using --no-discordant option.

Several fixes/improvements thanks to contributors on GitHub:

- new `--max-num-fusions` option allowing the user to specify the maximum number of reported fusions in tophat-fusion-post
- adjusting lower limit for `--fusion-multipairs`
- fixed a few typos, cleaning up python code etc.

» TopHat source code moved to GitHub 3/31/2015

TopHat is now available as a public GitHub repository where users are welcome to submit bug reports (issues) and developers are encouraged to submit patches (pull requests).

» TopHat 2.0.14 release 3/24/2015

Version 2.0.14 is a maintenance release with the following changes:

- pipeline speed improvements thanks to contributions from Véronique Legrand and Michaël Pressigout of Institut Pasteur
- added support for xz compressed read files (thanks to a patch submitted by Ashton Trey Belew)
- applied a couple of Python fixes to prevent potential issues with package handling and some file operations
- fixed a potential linking issue where the wrong libbam.a library could have been linked when building from source

» TopHat 2.0.13 release 10/2/2014

Version 2.0.13 is a maintenance release with the following changes:

- removed SAMtools as an *external* dependency in order to avoid incompatibility issues with recent and future changes of SAMtools and its code library (an older, stable SAMtools version is now packaged with TopHat)
- fixed a few code compatibility issues when compiling on OSX 10.9

» TopHat 2.0.12 release 6/24/2014

Version 2.0.12 is a maintenance release with the following simple fix:

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 Bowtie **mailing list.**

Getting Help
Questions and comments about TopHat can be posted on the **Tuxedo Tools Users Google Group**. Please use tophat.cufflinks@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

Releases
version 2.1.0 6/29/2015
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

Kim *et al. Genome Biology* 2013, **14**:R36
http://genomebiology.com/2013/14/4/R36

Genome Biology

METHOD Open Access

# TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

Daehwan Kim[1,2,3*], Geo Pertea[3], Cole Trapnell[5,6], Harold Pimentel[7], Ryan Kelley[8] and Steven L Salzberg[3,4]

**Abstract**

TopHat is a popular spliced aligner for RNA-sequence (RNA-seq) experiments. In this paper, we describe TopHat2, which incorporates many significant enhancements to TopHat. TopHat2 can align reads of various lengths produced by the latest sequencing technologies, while allowing for variable-length indels with respect to the reference genome. In addition to *de novo* spliced alignment, TopHat2 can align reads across fusion breaks, which can occur after genomic translocations. TopHat2 combines the ability to identify novel splice sites with direct mapping to known transcripts, producing sensitive and accurate alignments, even for highly repetitive genomes or in the presence of pseudogenes. TopHat2 is available at http://ccb.jhu.edu/software/tophat.

# TopHat

A spliced read mapper for RNA-Seq

Johns Hopkins University Center for Computational Biology CCB

**TopHat** is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort among Daehwan Kim and Steven Salzberg in the Center for Computational Biology at Johns Hopkins University, and Cole Trapnell in the Genome Sciences Department at the University of Washington. TopHat was originally developed by Cole Trapnell at the Center for Bioinformatics and Computational Biology at the University of Maryland, College Park.

OSI certified

» **TopHat 2.1.0 release 6/29/2015**

- TopHat-Fusion algorithm improvements for more sensitive and accurate discovery of fusions, thanks to contributions from Gordon Bean and Ryan Kelley at Illumina.
  - This release implements a new algorithm for counting fusion-supporting read pairs that reduces the number of false-positive potential fusions. This algorithm computes the inner distance between read pairs by first converting the pair positions to transcript coordinates using the transcript information in refGene.txt and ensGene.txt. Pairs with small inner distance (suggesting the pair could come from a plausible pair-end insert) are counted as supporting evidence for the fusion. The default threshold for the inner distance is 250 base pairs; this parameter can be set using the `--fusion-pair-dist <int>` flag.
- fixed a few issues with GFF parsing of some annotation files
- fixed a runtime-error when using --no-discordant option.

Several fixes/improvements thanks to contributors on GitHub:

- new `--max-num-fusions` option allowing the user to specify the maximum number of reported fusions in tophat-fusion-post
- adjusting lower limit for `--fusion-multipairs`
- fixed a few typos, cleaning up python code etc.

» **TopHat source code moved to GitHub 3/31/2015**

TopHat is now available as a public GitHub repository where users are welcome to submit bug reports (issues) and developers are encouraged to submit patches (pull requests).

» **TopHat 2.0.14 release 3/24/2015**

Version 2.0.14 is a maintenance release with the following changes:

- pipeline speed improvements thanks to contributions from Véronique Legrand and Michaël Pressigout of Institut Pasteur
- added support for xz compressed read files (thanks to a patch submitted by Ashton Trey Belew)
- applied a couple of Python fixes to prevent potential issues with package handling and some file operations
- fixed a potential linking issue where the wrong libbam.a library could have been linked when building from source

» **TopHat 2.0.13 release 10/2/2014**

Version 2.0.13 is a maintenance release with the following changes:

- removed SAMtools as an *external* dependency in order to avoid incompatibility issues with recent and future changes of SAMtools and its code library (an older, stable SAMtools version is now packaged with TopHat)
- fixed a few code compatibility issues when compiling on OSX 10.9

» **TopHat 2.0.12 release 6/24/2014**

Version 2.0.12 is a maintenance release with the following simple fix:

**Site Map**

- Home
- Getting started
- Manual
- Index and annotation downloads
- FAQ
- Protocol

**News and updates**

Getting startedで、とりあえず使ってみる

**Tools Users Google Group**. Please use tophat.cufflinks@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

**Releases**

version 2.1.0 6/29/2015

- 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

## Getting started

- Install quick-start
- Test the installation
- Preparing your reference
- Preparing your reads
- Running TopHat
- Examining your results

» **Install quick-start**

**Download and extract the latest Bowtie 2 (or Bowtie) releases.**

Note that you can use either Bowtie 2 (the default) or Bowtie (--bowtie1) and you will need the following Bowtie 2 (or Bowtie) programs in your PATH:

- `bowtie2 (or bowtie)`
- `bowtie2-build (or bowtie-build)`
- `bowtie2-inspect (or bowtie-inspect)`

**Installing a pre-compiled binary release**

In order to make it easy to install TopHat we provide a few binary packages to save users from the occasionally frustrating process of building TopHat themselves, which requires a certain development environment and the Boost libraries installed. To use the binary packages, simply download the appropriate one for your platform, unpack it, and make sure the `tophat` binaries are in a directory in your PATH environment variable (or create a symbolic link to the included tophat2 script somewhere in your PATH, see below)

**Note:** if you want to be able to install and run this new version without overwriting a previous TopHat version already installed on your system, make sure you unpack the new version into a different directory from the old version, then instead of copying the new programs in a directory in your PATH just create a symbolic link from the `tophat2` wrapper script in this new directory to a directory in your shell's PATH. For example, assuming the ~/bin directory is in your PATH and you unpack tophat-2.0.0.Linux_x86_64.tar.gz under your home directory:

```
cd
tar xvfz tophat-2.0.0.Linux_x86_64.tar.gz
cd ~/bin
ln -s ~/tophat-2.0.0.Linux_x86_64/tophat2 .
```

Now you can start the new version of TopHat with the `tophat2` command, while the previous version, if present, can still be launched with the regular "tophat" command (assuming this is how you used it before).

**Building TopHat from source**

In order to build TopHat2 you must have the following installed on your system:

- the Boost C++ libraries (we recommend version 1.47 or higher so you can use it for building Cufflinks as well)

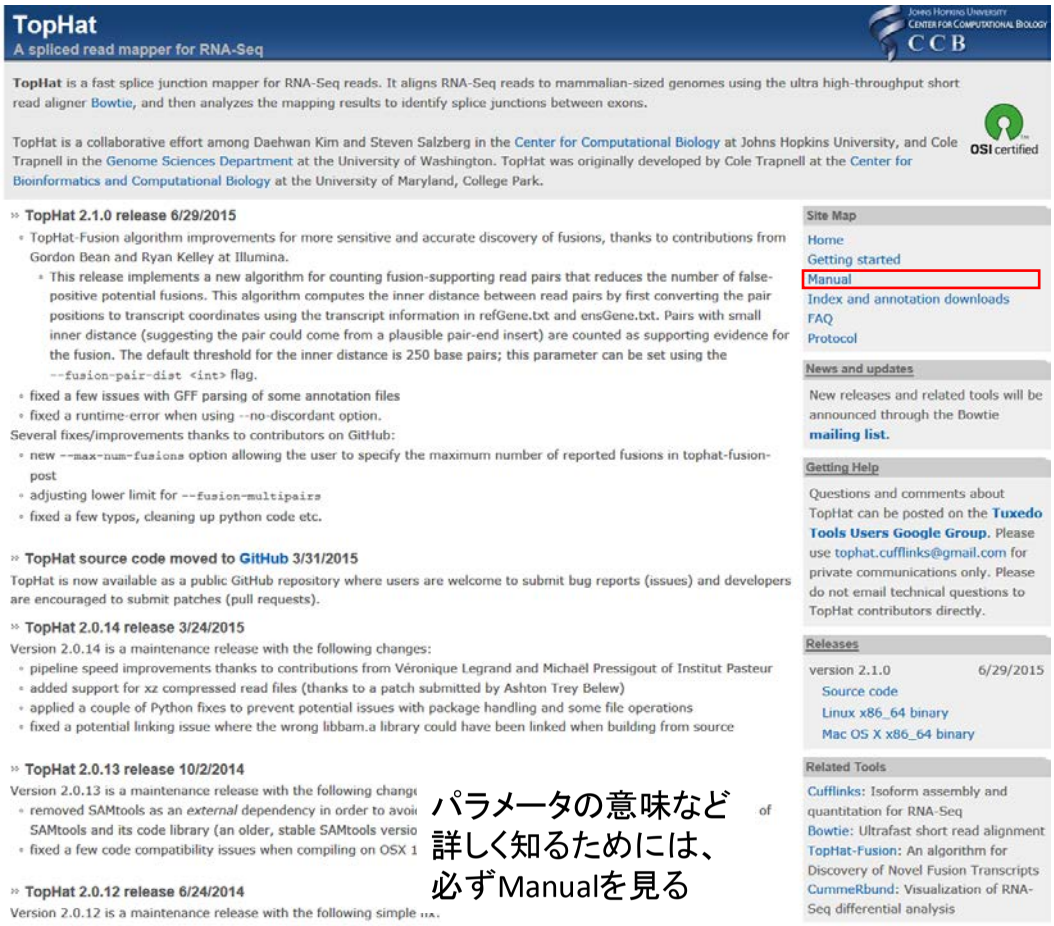
インストールの方法・
必要ツールなどの記載・
テストデータ等での極く簡単な
解析手順に関する記載がある

必要ツール
・bowtie2
・samtools

TopHat2はあらかじめコンパイルしたバイナリーファイルが配布されているので、自分でmakeする必要はない。
自分でソースからmakeする場合は
・SAMtools lib
・Boost C++ library
が必要

testデータが用意されている

```
tar zxvf test_data.tar.gz
cd test_data
tophat -r 20 test_ref reads_1.fq reads_2.fq
```



# TopHat

A spliced read mapper for RNA-Seq

Johns Hopkins University Center for Computational Biology CCB

**TopHat** is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort among Daehwan Kim and Steven Salzberg in the Center for Computational Biology at Johns Hopkins University, and Cole Trapnell in the Genome Sciences Department at the University of Washington. TopHat was originally developed by Cole Trapnell at the Center for Bioinformatics and Computational Biology at the University of Maryland, College Park.

OSI certified

» **TopHat 2.1.0 release 6/29/2015**

- TopHat-Fusion algorithm improvements for more sensitive and accurate discovery of fusions, thanks to contributions from Gordon Bean and Ryan Kelley at Illumina.
  - This release implements a new algorithm for counting fusion-supporting read pairs that reduces the number of false-positive potential fusions. This algorithm computes the inner distance between read pairs by first converting the pair positions to transcript coordinates using the transcript information in refGene.txt and ensGene.txt. Pairs with small inner distance (suggesting the pair could come from a plausible pair-end insert) are counted as supporting evidence for the fusion. The default threshold for the inner distance is 250 base pairs; this parameter can be set using the `--fusion-pair-dist <int>` flag.
- fixed a few issues with GFF parsing of some annotation files
- fixed a runtime-error when using --no-discordant option.

Several fixes/improvements thanks to contributors on GitHub:

- new `--max-num-fusions` option allowing the user to specify the maximum number of reported fusions in tophat-fusion-post
- adjusting lower limit for `--fusion-multipairs`
- fixed a few typos, cleaning up python code etc.

» **TopHat source code moved to GitHub 3/31/2015**

TopHat is now available as a public GitHub repository where users are welcome to submit bug reports (issues) and developers are encouraged to submit patches (pull requests).

» **TopHat 2.0.14 release 3/24/2015**

Version 2.0.14 is a maintenance release with the following changes:

- pipeline speed improvements thanks to contributions from Véronique Legrand and Michaël Pressigout of Institut Pasteur
- added support for xz compressed read files (thanks to a patch submitted by Ashton Trey Belew)
- applied a couple of Python fixes to prevent potential issues with package handling and some file operations
- fixed a potential linking issue where the wrong libbam.a library could have been linked when building from source

» **TopHat 2.0.13 release 10/2/2014**

Version 2.0.13 is a maintenance release with the following change

- removed SAMtools as an *external* dependency in order to avoi
  SAMtools and its code library (an older, stable SAMtools versio
- fixed a few code compatibility issues when compiling on OSX 1

of

» **TopHat 2.0.12 release 6/24/2014**

Version 2.0.12 is a maintenance release with the following simple

パラメータの意味など
詳しく知るためには、
必ずManualを見る

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 Bowtie **mailing list.**

Getting Help
Questions and comments about TopHat can be posted on the **Tuxedo Tools Users Google Group**. Please use tophat.cufflinks@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

Releases
version 2.1.0 6/29/2015
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

・75base以上のreadに最適化
・リファレンス annotationなしでも
スプライスジャンクションを見つける

## Manual

- What is TopHat?
- Prerequisites
- Using TopHat

### » 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.

### » 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 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

Illumina has provided the RNA-Seq user community with a set of genome sequence indexes (including Bowtie indexes) as well as GTF transcript annotation files. These files can be used with TopHat and Cufflinks to quickly perform expression analysis and gene discovery. The annotation files are augmented with the `tss_id` and `p_id` GTF attributes that Cufflinks needs to perform differential splicing, CDS output, and promoter user analysis. We recommend that you download your Bowtie indexes and annotation files from this page. More information about Illumina's iGenomes project can be found here.

| Organism | Data source | Version | Size | Last Modified |
|---|---|---|---|---|
| Homo sapiens | Ensembl | GRCh37 | 17297 MB | May 14 17:23 |
| | NCBI | build36.3 | 15814 MB | May 14 19:36 |
| | | build37.1 | 15850 MB | May 14 19:04 |
| | | build37.2 | 21450 MB | May 14 17:54 |
| | UCSC | hg18 | 17349 MB | May 14 15:31 |
| | | hg19 | 21058 MB | May 14 15:36 |
| Mus musculus | Ensembl | NCBIM37 | 14428 MB | May 14 22:13 |
| | NCBI | build37.1 | 15260 MB | May 15 17:53 |
| | | build37.2 | 15725 MB | May 14 22:52 |
| | UCSC | mm9 | 14537 MB | May 14 21:12 |
| | | mm10 | 14193 MB | Jun 14 11:29 |
| Rattus norvegicus | Ensembl | RGSC3.4 | 13725 MB | May 15 22:33 |
| | NCBI | RGSC_v3.4 | 14234 MB | May 15 23:58 |
| | UCSC | rn4 | 13710 MB | May 15 22:32 |
| Bos taurus | Ensembl | Btau_4.0 | 13315 MB | May 11 14:18 |
| | | UMD3.1 | 14042 MB | May 11 12:41 |
| | NCBI | Btau_4.2 | 13357 MB | May 11 14:11 |
| | | Btau_4.6.1 | 13448 MB | May 11 16:09 |
| | | UMD_3.1 | 13990 MB | May 11 16:08 |

**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 Bowtie **mailing list**.

**Getting Help**

Questions and comments about TopHat can be posted on the **Tuxedo Tools Users Google Group**. Please use tophat.cufflinks@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

**Releases**

version 2.0.12 6/24/2014
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

メジャーな生物種ではindexファイルやannotationファイル等が配布されているので有効活用できる

## Frequently Asked Questions

- How to control the alignment of reads in terms of number of mismatches, gap length etc. ?
- How can I maximize the accuracy of spliced mapping in TopHat?
- I don't know the mate inner distance (-r/--mate-inner-dist option) for my paired reads, what value should I use?
- I am not sure which library type to use (fr-firststrand or fr-secondstrand), what should I do?
- What should I do if I see a message like "Too many open files"?

FAQも参考に

### How to control the alignment of reads in terms of number of mismatches, gap length etc. ?

You can use three options: `--read-mismatches`, `--read-gap-length` and `--read-edit-dist`. For instance, if you want read alignments with at most 2 base mismatches and no gaps then you can specify:

```
--read-mismatches 2 --read-gap-length 0 --read-edit-dist 2
```

Or if you want read alignments with total length of indels (alignment gaps) of at most 3bp and at most 2 base mismatches you can use these options:

```
--read-mismatches 2 --read-gap-length 3 --read-edit-dist 3
```

### How can I maximize the accuracy of spliced mapping in TopHat?

Based on real RNA-seq samples we found out that in the genome mapping step of TopHat a high portion of reads spanning several exons can incorrectly be aligned to processed pseudogenes that are rarely (if any) transcribed or expressed, instead of the genes where they originate from. You can use either of the options below to improve the accuracy of spliced mapping in TopHat:

- If a good gene annotation is available (as the case with the human genome), use it with the -G option
- For poorly annotated genomes you might want to consider using the "`--read-realign-edit-dist 0`" option
  With the realignment option users can choose to remap some (or all) of the mapped reads with mapping edit distance equal to or above user-specified "remapping" edit distance (see `--read-realign-edit-dist` option). Setting "`--read-realign-edit-dist 0`" will map every read against transcriptome, genome, and splice variants (or splice junctions) that are detected by TopHat, no matter whether it is mapped or not in any mapping step. With this remapping strategy, this "pseudogene" problem can be effectively handled. If you use a genome that has processed pseudogenes and you cannot provide good gene annotation to TopHat, you may want to consider using this option for accurate mapping results.

NATURE PROTOCOLS | PROTOCOL

# Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

Affiliations | Contributions | Corresponding author

*Nature Protocols* **7**, 562–578 (2012) | doi:10.1038/nprot.2012.016
Published online 01 March 2012



## Abstract

Abstract • Accession codes • References • Author information

Recent advances in high-throughput cDNA sequencing (RNA-seq) can reveal new genes and splice variants and quantify expression genome-wide in a single assay. The volume and complexity of data from RNA-seq experiments necessitate scalable, fast and mathematically principled analysis software. TopHat and Cufflinks are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. Together, they allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. This protocol describes in detail how to use TopHat and Cufflinks to perform such analyses. It also covers several accessory tools and utilities that aid in managing data, including CummeRbund, a tool for visualizing RNA-seq analysis results. Although the procedure assumes basic informatics skills, these tools assume little to no background with RNA-seq analysis and are meant for novices and experts alike. The protocol begins with raw sequencing reads and produces a transcriptome assembly, lists of differentially expressed and regulated genes and transcripts, and publication-quality visualizations of analysis results. The protocol's execution time depends on the volume of transcriptome sequencing data and available computing resources but takes less than 1 d of computer time for typical experiments and ~1 h of hands-on time.

protocol論文も出ている

ただし今となっては少し古い

Freeではない

# tophat基本コマンド

**TopHat** is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

```
> tophat -G gene.gtf -o out_dir genome read_1.fastq read_2.fastq
```

-G/--GTF <GTF/GFF3 file>
まずgtfに基づき、トランスクリプトにmapさせ、ゲノム位置として戻す。
mapしないリードはゲノムから探す

# tophatの出力

prep_reads.info
align_summary.txt
deletions.bed
insertions.bed
junctions.bed
accepted_hits.bam
unmapped.bam

sam/bam フォーマットのファイル
accepted_hits.bamファイルがこの後必要

実習1
tophatを用いて2D_1のfastqファイルをgenome_chr4にmapさせよ、
GTPファイルとしてgenes_chr4.gtfを用いる

例）
```
> tophat -p 4 -G genes_chr4.gtf -o 2D_1 genome_chr4 2D_1_R1.fastq 2D_1_R2.fastq
```

出力を確認しよう。
例えば、align_summary.txtを見ればどの程度mapしたか分かる。
これでRNA-Seqのリード配列がゲノム配列にアラインできた。

# cufflinksを用いてアラインされたreadを数える

定義した方法でのカウントが可能
gene単位
トランスクリプト単位
エキソン単位

- cufflinks
-BEDTools
-HTseq
が利用できる

今回はCufflinksを利用
そもそもTopHat → Cufflinksの解析系は同じ開発元、非常に良く使われている。

ローカスアノテーション情報を記載したgtfファイルを用意しておけば、
それに基づいて、genes単位、isoforms単位での解析を進めてくれる。

簡易的に、特定ローカスの解析などを進めたい場合や、
gtfファイルがない場合などは、BEDToolsも有用
gtfファイルを自分で作製するのは結構大変だが、bedファイルは比較的容易

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

http://cole-trapnell-lab.github.io/cufflinks/

# Cufflinks

*Transcriptome assembly and differential expression analysis for RNA-Seq.*

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.

Cufflinks was originally developed as part of a collaborative effort between the Laboratory for Mathematical and Computational Biology, led by Lior Pachter at UC Berkeley, Steven Salzberg's computational genomics group at the Institute of Genetic Medicine at Johns Hopkins University, and Barbara Wold's lab at Caltech. The project is now maintained by Cole Trapnell's lab at the University of Washington.

Cufflinks is provided under the OSI-approved Boost License

## News

*To get the latest updates on the Cufflinks project and the rest of the "Tuxedo tools", please subscribe to our **mailing list***

| | |
|---|---|
| Cufflinks has moved to GitHub | DECEMBER 10, 2014 |
| Cufflinks 2.2.1 released | MAY 05, 2014 |
| Cufflinks 2.2.0 released | MARCH 25, 2014 |
| Cufflinks 2.1.1 released | APRIL 11, 2013 |

Protocol

# Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold & Lior Pachter

Affiliations | Contributions | Corresponding author

*Nature Biotechnology* **28**, 511–515 (2010) | doi:10.1038/nbt.1621
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

PDF | Citation | Reprints | Rights & permissions | Article metrics

**High-throughput mRNA sequencing (RNA-Seq) promises simultaneous transcript discovery and abundance estimation[1, 2, 3]. However, this would require algorithms that are not restricted by prior gene annotations and that account for alternative transcription and splicing. Here we introduce such algorithms in an open-source software program called Cufflinks. To test Cufflinks, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series. We detected 13,692 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species. Over the time series, 330 genes showed complete switches in the dominant transcription start site (TSS) or splice isoform, and we observed more subtle shifts in 1,304 other genes. These results suggest that Cufflinks can illuminate the substantial regulatory flexibility and complexity in even this well-studied model of muscle development and that it can improve transcriptome-based genome annotation.**

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

# Cufflinks

*Transcriptome assembly and differential expression analysis for RNA-Seq.*

Cufflinks is available for Linux and Mac OS X. You can find the full list of releases below.

The Cufflinks source code for each point release is available below as well. If you want to grab the current code, check out the Cufflinks GitHub repository.

Star 23 Fork 12

## Cufflinks Releases

| Version | Date | | | |
|---|---|---|---|---|
| 2.2.1 | May 05, 2014 | Linux | Mac OS X | Source |
| 2.2.0 | March 25, 2014 | Linux | Mac OS X | Source |
| 2.1.1 | April 11, 2013 | Linux | Mac OS X | Source |
| 2.1.0 | April 10, 2013 | Linux | Mac OS X | Source |

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

# Cufflinks

*Transcriptome assembly and differential expression analysis for RNA-Seq.*



自分でソースからmakeする場合は
・Samtools
・Boost C++ library
が必要

```
cufflinks ./test_data.sam
```

これでツールが動くことを確認

## Install quick-start

### Installing a pre-compiled binary release

In order to make it easy to install Cufflinks, we provide a few binary packages to save users from occasionally frustrating process of building Cufflinks, which requires that you install the Boost libraries. To use the binary packages, simply download the appropriate one for your machine, untar it, and make sure the cufflinks,cuffdiff and cuffcompare binaries are in a directory in your PATH environment variable.

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

# Cufflinks

*Transcriptome assembly and differential expression analysis for RNA-Seq.*

## Bowtie: ultrafast short read alignment

Bowtie is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

Bowtie is provided under the OSI-approved Artistic License 2.0.

## TopHat: alignment of short RNA-Seq reads

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is provided under the OSI-approved Artistic License 2.0.

## CummeRbund: visualization of RNA-Seq differential analysis

CummeRbund is an R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output.

CummeRbund is provided under the OSI-approved Artistic License 2.0.

## Monocle: Differential expression for single-cell RNA-Seq and qPCR.

Monocle is a toolkit for analyzing single-cell gene expression experiments. Monocle was designed for RNA-Seq, but can also work with single cell qPCR. It performs differential expression analysis, and can find genes that differ between cell types or between cell states. When used to study an ongoing biological process such as cell differentiation, Monocle learns that process and places cells in order according to their progress through it. Monocle finds genes that are dynamically regulated during that process.

Monocle is provided under the OSI-approved Artistic License (version 2.0)

Cufflinksの関連ツール
Bowtie, TopHatは説明済み

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

# Cufflinks

*Transcriptome assembly and differential expression analysis for RNA-Seq.*

## The Cufflinks RNA-Seq workflow

The Cufflinks suite of tools can be used to perform a number of different types of analyses for RNA-Seq experiments. The Cufflinks suite includes a number of different programs that work together to perform these analyses. The complete workflow, performing all the types of analyses Cufflinks can execute, is summarized in the graph below. The left side illustrates the "classic" RNA-Seq workflow, which includes read mapping with TopHat, assembly with Cufflinks, and visualization and exploration of results with CummeRbund. A newer, more advanced worfklow was introduce with Cufflinks version 2.2.0, and is shown on the right. Both are still supported. You can read about the classic workflow in detail in our protocol paper.


## Cufflinks

Cufflinks is both the name of a suite of tools and a program within that suite. Cufflinks the program assembles transcriptomes from RNA-Seq data and quantifies their expression.

## Cuffcompare

After assembling a transcriptome from one or more samples, you'll probably want to compare your assembly to known transcripts. Even if there is no "reference" transcriptome for the organism you're studying, you may want to compare the transcriptomes assembled from different RNA-Seq libraries. Cuffcompare helps you perform these comparisons and assess the quality of your assembly.

## Cuffmerge

When you have multiple RNA-Seq libraries and you've assembled transcriptomes from each of them, we recommend that you merge these assemblies into a master transcriptome. This step is required for a differential expression analysis of the new transcripts you've assembled. Cuffmerge performs this merge step.

## Cuffquant

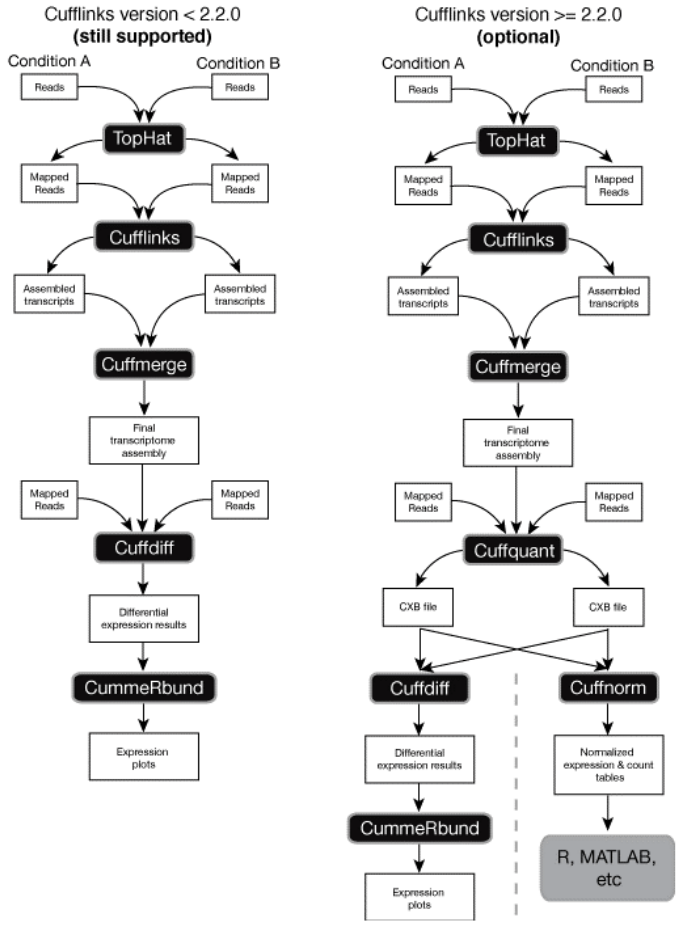
Quantifying gene and transcript expression in RNA-Seq samples can be computationally expensive. Cuffquant allows you to compute the gene and transcript expression profiles and save these profiles to files that you can analyze later with Cuffdiff or Cuffnorm. This can help you distribute your computational load over a cluster and is recommended for analyses involving more than a handful of libraries.

## Cuffdiff

Comparing expression levels of genes and transcripts in RNA-Seq experiments is a hard problem. Cuffdiff is a highly accurate tool for performing these comparisons, and can tell you not only which genes are up- or down-regulated between two or more conditions, but also which genes are differentially spliced or are undergoing other types of isoform-level regulation.

## Cuffnorm

Sometimes, all you want to do is normalize the expression levels from a set of RNA-Seq libraries so that they're all on the same scale, facilitating downstream analyses such as clustering. Expression levels reported by Cufflinks in FPKM units are usually comparable between samples, but in certain situations, applying an extra level of normalization can remove sources of bias in the data. Cuffnorm normalizes a set of samples to be on as similar scales as possible, which can improve the results you obtain with other downstream tools.




cufflink

cufflinks
cuffmerge
cuffcompare
cuffquant
cuffnorm
cuffdiff
の6つのプログラムから構成

cuffquant, cuffnormは
ver2.2.0(20140325)
から実装

MacOSX版のバイナリーはver2.2.0以降は
バグがありsegmentation errorでまともに
動かないようです。

今回の実習ではver2.1.1を使用し、
cuffquant, cuffnormは簡単な説明のみ
に留めます。


**Cufflinks**

*Transcriptome assembly and differential expression analysis for RNA-Seq.*

Cufflinks is an ongoing research project as well as a suite of tools. Here are the papers that describe the science behind the programs. If you use Cufflinks, **please cite these papers** in your work!

**Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**

Cole Trapnell, Brian Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Jeltje van Baren, Steven Salzberg, Barbara Wold, Lior Pachter.

***Nature Biotechnology***, 2010

High-throughput mRNA sequencing (RNA-Seq) promises simultaneous transcript discovery and abundance estimation. However, this would require algorithms that are not restricted by prior gene annotations and that account for alternative transcription and splicing. Here we introduce such algorithms in an open-source software program called Cufflinks. To test Cufflinks, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series. We detected 13,692 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species. Over the time series, 330 genes showed complete switches in the dominant transcription start site (TSS) or splice isoform, and we observed more subtle shifts in 1,304 other genes. These results suggest that Cufflinks can illuminate the substantial regulatory flexibility and complexity in even this well-studied model of muscle development and that it can improve transcriptome-based genome annotation.

doi:10.1038/nbt.1621

***Note:*** *This is the original Cufflinks paper. Please cite this paper if you use Cufflinks in your work.*

**Improving RNA-Seq expression estimates by correcting for fragment bias**

Adam Roberts, Cole Trapnell, Julie Donaghey, John L. Rinn, Lior Pachter.

***Genome Biology***, 2011

どうやって動いているか

まず動いて使えそうな感じに
なったら詳細を把握していく

# cufflinks基本コマンド

## Cufflinksコマンド

cufflinks -o out_directory -G hoge.gtf tophat_directory/accepted_hits.bam

cufflinksを実行してパラメータを確認しよう。

考慮すべきパラメーター例

| | |
|---|---|
| -o | 出力の指定、TopHatの出力と同じ場所にしておくのが分かりやすいだろう |
| -p | CPUスレッド数の指定(デフォルトは1)、結構時間がかかるので使える数を指定すると良いだろう |
| -G | GTFファイルに記載されたアノテーションのみについて計算 |
| -g | GTFファイルに記載されたアノテーションをガイドにしてアセンブルする |
| -M | 無視したいトランスクリプト（rRNAなど）を指定 |

# cufflinks出力

出力

skipped.gtf
transcripts.gtf
genes.fpkm_tracking
isoforms.fpkm_trancking

実習2
先のtophatの結果を用いてcufflinksにかけてみよう

例）
```
> cufflinks -p 4 -o 2D_1 -G genes_chr4.gtf accepted_hits.bam
```

出力を確認しよう。
geneごと、isoformごとにFPKM値が計算されているのが分かる。
-gを用いてcufflinksにかけると新規の発現領域が存在するのが分かる

# cuffcompareコマンド

Cufflinks includes a program that you can use to help analyze the transfrags you assemble. The program cuffcompare helps you:
Compare your assembled transcripts to a reference annotation
Track Cufflinks transcripts across multiple experiments (e.g. across a time course)
From the command line, run cuffcompare as follows:

cuffcompare [options]* <cuff1.gtf> [cuff2.gtf] ... [cuffN.gtf]

今回はすでにあるgtfファイルの情報を用いるので、意識的に使う必要はない。

# cuffmergeコマンドと出力

個々のサンプルのアセンブルモデルを統合する。

```
Usage:
    cuffmerge [Options] <assembly_GTF_list.txt>

Options:
    -h/--help                          Prints the help message and exits
    -o                 <output_dir>    Directory where merged assembly will be written  [ default: ./merged_asm  ]
    -g/--ref-gtf                       An optional "reference" annotation GTF.
    -s/--ref-sequence  <seq_dir>/<seq_fasta> Genomic DNA sequences for the reference.
    --min-isoform-fraction <0-1.0>     Discard isoforms with abundance below this       [ default:           0.05 ]
    -p/--num-threads   <int>           Use this many threads to merge assemblies.       [ default:             1  ]
    --keep-tmp
```

統合ファイルリストを事前に作製する必要がある(例 assemblies,txt)

cuffmerge -s $REFSEQ -g $GTF assemblies.txt

例 assemblies,txt

~/arabi_2D_2/transcripts.gtf
~/arabi_2D_3/transcripts.gtf
~/arabi_2D2L_2/transcripts.gtf
~/arabi_2D2L_3/transcripts.gtf

出力
merged.gtf

Cufflinks includes a script called cuffmerge that you can use to merge together several Cufflinks assemblies. It handles also handles running Cuffcompare for you, and automatically filters a number of transfrags that are probably artfifacts. If you have a reference GTF file available, you can provide it to the script in order to gracefully merge novel isoforms and known isoforms and maximize overall assembly quality. The main purpose of this script is to make it easier to make an assembly GTF file suitable for use with Cuffdiff.

# cuffdiffコマンド

DE gene等を統計計算で取り出す
コマンド入力して使用法を確認してみよう

```
Usage:   cuffdiff [options] <transcripts.gtf> <sample1_hits.sam> <sample2_hits.sam> [... sampleN_hits.sam]
Supply replicate SAMs as comma separated lists for each condition:
sample1_rep1.sam,sample1_rep2.sam,...sample1_repM.sam
General Options:
  -o/--output-dir             write all output files to this directory             [ default:     ./ ]
  -L/--labels                 comma-separated list of condition labels
  --FDR                       False discovery rate used in testing                 [ default:   0.05 ]
```

```
cuffdiff -o out_file merged.gtf bam1,bam2,bam3 bam4,bam5,bam6
```

Version 2.2.0以降は先のcuffquantで得られたcxbファイルをbamファイルの代わりに用いる。
cuffdiffにかかる時間やメモリー使用量が軽減される。

# cuffdiffの出力

bias_params.info
run.info
read_groups.info
var_model.info
cds.read_group_tracking
cds.fpkm_tracking
cds.count_tracking
genes.read_group_tracking
genes.fpkm_tracking
genes.count_tracking
isoforms.read_group_tracking
isoforms.count_tracking
isoforms.fpkm_tracking
tss_groups.read_group_tracking
tss_groups.fpkm_tracking
tss_groups.count_tracking

gene_exp.diff
cds_exp.diff
cds.diff
isoform_exp.diff
promoters.diff
splicing.diff
tss_group_exp.diff

diffの付いたファイルがそれぞれの
違いの情報を記載したファイル

# .diffファイルの内容

| Column number | Column name | Example | Description |
|---|---|---|---|
| 1 | Tested id | XLOC_000001 | A unique identifier describing the transcipt, gene, primary transcript, or CDS being tested |
| 2 | gene | Lypla1 | The gene_name(s) or gene_id(s) being tested |
| 3 | locus | chr1:4797771-4835363 | Genomic coordinates for easy browsing to the genes or transcripts being tested. |
| 4 | sample 1 | Liver | Label (or number if no labels provided) of the first sample being tested |
| 5 | sample 2 | Brain | Label (or number if no labels provided) of the second sample being tested |
| 6 | Test status | NOTEST | Can be one of OK (test successful), NOTEST (not enough alignments for testing), LOWDATA (too complex or shallowly sequenced), HIDATA (too many fragments in locus), or FAIL, when an ill-conditioned covariance matrix or other numerical exception prevents testing. |
| 7 | $FPKM_x$ | 8.01089 | FPKM of the gene in sample *x* |
| 8 | $FPKM_y$ | 8.551545 | FPKM of the gene in sample *y* |
| 9 | log2 ($FPKM_y$/$FPKM_x$) | 0.06531 | The (base 2) log of the fold change *y*/*x* |
| 10 | test stat | 0.860902 | The value of the test statistic used to compute significance of the observed change in FPKM |
| 11 | p value | 0.389292 | The **uncorrected** *p*-value of the test statistic |
| 12 | q value | 0.985216 | The **FDR-adjusted** *p*-value of the test statistic |
| 13 | significant | no | Can be either "yes" or "no", depending on whether *p* is greater then the FDR **after** Benjamini-Hochberg correction for multiple-testing |

# cuffquantコマンドと出力(ver2.2.0以降)

bamの内容からgene/transcriptレベルで定量化し、バイナリー出力する

cuffquant -o out_directory hoge.gtf accepted_hits.bam

cuffquantを実行してパラメータを確認しよう。

考慮すべきパラメーター例
-o　　出力ディレクトリーの指定
-p　　CPUスレッド数の指定(デフォルトは1)、結構時間がかかるので使える数を指定
-M　　無視したいトランスクリプト（rRNAなど）を指定
他にもestimationに関わる　-b　-u パラメータがある。

出力
abundances.cxb

```
> cuffquant -p 4 -o 2D_1 genes_chr4.gtf accepted_hits.bam
```

新たにcxbファイルが作製されていることが分かる。
出力ファイルはこの１つだけ

cuffdiffの前にcuffquantを行い、cxbファイルを作製することで
cuffdiffを速くできる。

# cuffnormコマンドと出力(ver2.2.0以降)

## Cuffnormコマンド

Cuffnorm, which simply computes
a normalized table of expression values for genes and transcripts.

```
> cuffnorm -o out_file genes_chr4.gtf bam1,bam2,bam3 bam4,bam5,bam6
```

cuffnorm [options]* <transcripts.gtf>
<sample1_replicate1.sam[,...,sample1_replicateM.sam]>
<sample2_replicate1.sam[,...,sample2_replicateM.sam]>...
[sampleN.sam_replicate1.sam[,...,sample2_replicateM.sam]]

sam/bamかcxbファイルどちらも入力可能。ただし混在は不可

# cuffnormの出力(ver2.2.0以降)

cds.attr_table
cds.count_table
cds.fpkm_table
cuffnorm.tree
genes.attr_table
genes.count_table
genes.fpkm_table
isoforms.attr_table
isoforms.count_table
isoforms.fpkm_table
run.info
samples.table
tss_groups.attr_table
tss_groups.count_table
tss_groups.fpkm_table

たくさんのサンプルで発現プロットやクラスター図を書きたい場合便利。

# tophat -> cufflinksの解析系を使用する際の注意

It does not perform differential expression analysis. To assess the significance of changes in expression for genes and transcripts between conditions, use Cuffdiff. Cuffnorm's output files are useful when you have many samples and you simply want to cluster them or plot expression levels of genes important in your study.

Cuffnorm will report both FPKM values and **normalized**, estimates for the number of fragments that originate from each gene, transcript, TSS group, and CDS group. Note that because these counts are already normalized to account for differences in library size, they should not be used with downstream differential expression tools that require **raw** counts as input.

tophat -> cufflinksは一連の解析系

cufflinksの出力はすでにノーマライズされたものなので、rawデータを要求するedgeRなどの別のツールのinputには利用できない。

nature protocols

Home | Current issue | Archive | Authors & referees | About the journal

NATURE PROTOCOLS | PROTOCOL

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber & Mark D Robinson

Affiliations | Contributions | Corresponding authors

Nature Protocols 8, 1765–1786 (2013) | doi:10.1038/nprot.2013.099
Published online 22 August 2013

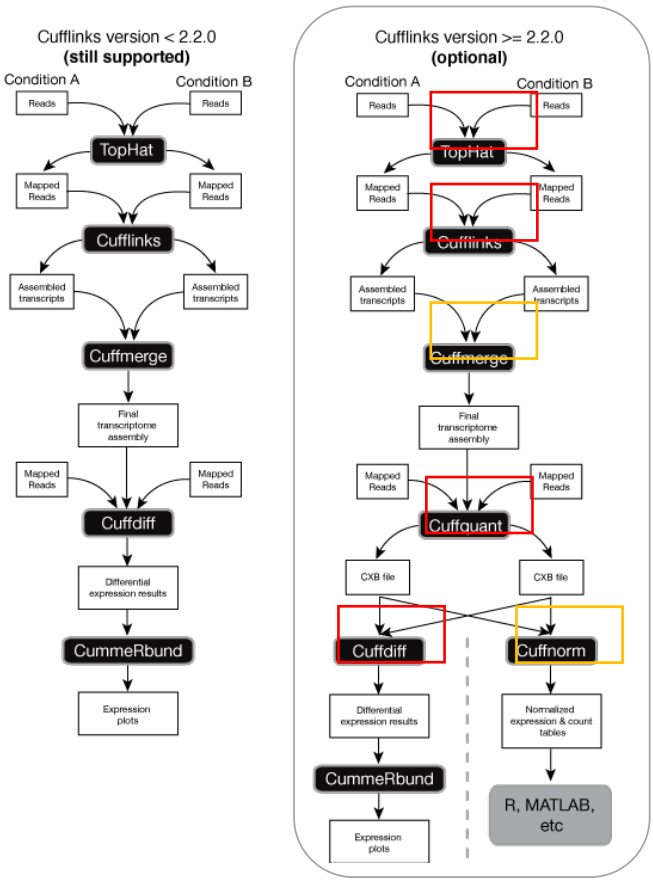
Full text | PDF | Citation | Reprints | Rights & permissions | Article metrics

Abstract

Abstract • Accession codes • References • Author information • Supplementary information

RNA sequencing (RNA-seq) has been rapidly adopted for the profiling of transcriptomes in many areas of biology, including studies into gene regulation, development and disease. Of particular interest is the discovery of differentially expressed genes across different conditions (e.g., tissues, perturbations) while optionally adjusting for other systematic factors that affect the data-collection process. There are a number of subtle yet crucial aspects of these analyses, such as read counting, appropriate treatment of biological variability, quality control checks and appropriate setup of statistical modeling. Several variations have been presented in the literature, and there is a need for guidance on current best practices. This protocol presents a state-of-the-art computational and statistical RNA-seq differential expression analysis workflow largely based on the free open-source R language and Bioconductor software and, in particular, on two widely used tools, DESeq and edgeR. Hands-on time for typical small experiments (e.g., 4–10 samples) can be <1 h, with computation time <1 d using a standard desktop PC.

# versionによる違いまとめ



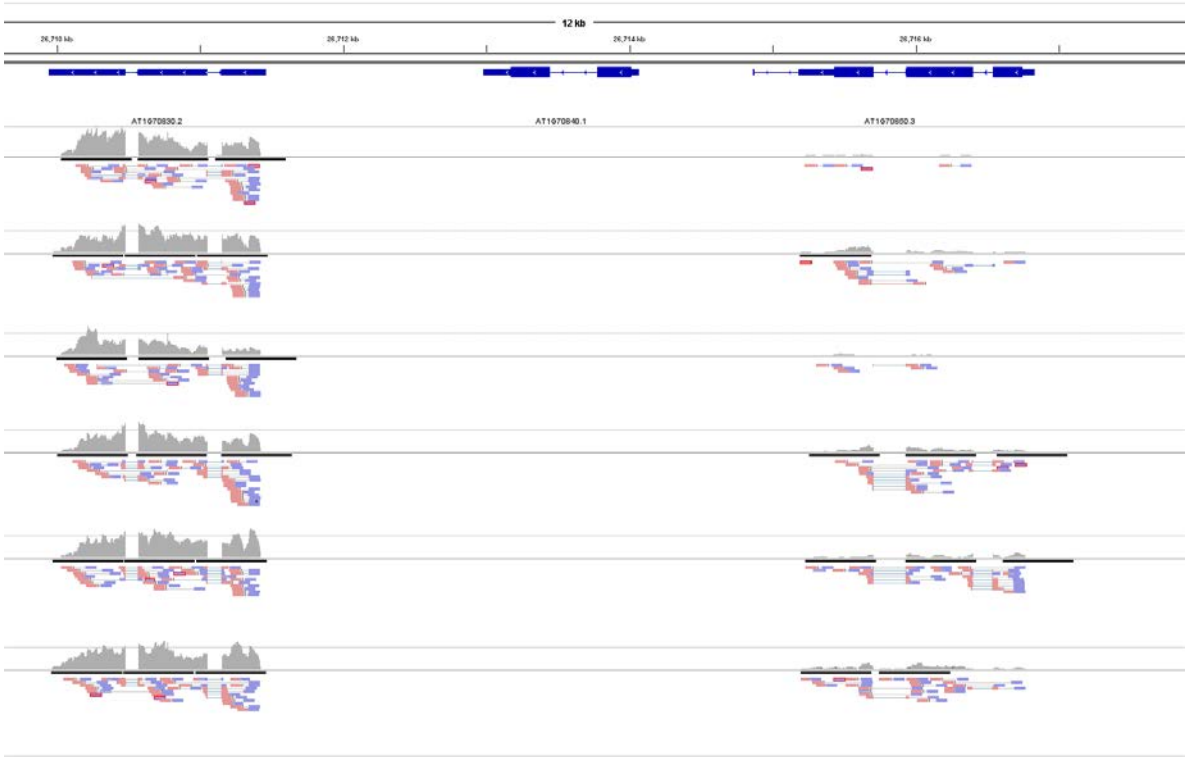
# tophat, cufflinksの実習

1. TopHatを用いて、paired-endのtest data
2D_2_R1.fastq,2D_2_R2.fastq
をリファレンスgenome_chr4にマップさせよ
オプション -Gの有無に
よる違いを確認しよう。

2.Cufflinksを用いて、
2D_2のカウントをしよう。
-Gと-gの違いを確認しよう。

# 結果をIGVで可視化してみよう

TAIR10の配列を呼び出し、TopHatで得られたBAMファイルを読み込む



# Excelを使って結果を確認してみよう

gene_exp.diffファイルを読み込んでみる
tab区切りテキストファイルなのでそのまま読み込める
Excelのsort機能を使ってq値でsortしてみる

q値でsort

| test_id | gene_id | gene | locus | sample_1 | sample_2 | status | value_1 | value_2 | log2(fold_chan | test_stat | p_value | q_value | significant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| XLOC_000047 | XLOC_000047 | KEA1 | 1:284609-291094 | q1 | q2 | OK | 12.8356 | 47.6879 | 1.89347 | 4.44122 | 5.00E-05 | 0.000325 | yes |
| XLOC_000091 | XLOC_000091 | BXL2 | 1:564204-567769 | q1 | q2 | OK | 112.839 | 21.5634 | -2.38762 | -6.02938 | 5.00E-05 | 0.000325 | yes |
| XLOC_000148 | XLOC_000148 | PSB27 | 1:898875-899655 | q1 | q2 | OK | 194.744 | 691.64 | 1.82844 | 7.10401 | 5.00E-05 | 0.000325 | yes |
| XLOC_000310 | XLOC_000310 | PSBP-1 | 1:2047824-2049418 | q1 | q2 | OK | 588.195 | 3147.84 | 2.42 | 7.92975 | 5.00E-05 | 0.000325 | yes |
| XLOC_000404 | XLOC_000404 | NPQ1 | 1:2706923-2709531 | q1 | q2 | OK | 21.2494 | 78.5734 | 1.88662 | 3.26377 | 5.00E-05 | 0.000325 | yes |
| XLOC_000419 | XLOC_000419 | CSD1 | 1:2827060-2838469 | q1 | q2 | OK | 503.523 | 181.545 | -1.47173 | -5.38312 | 5.00E-05 | 0.000325 | yes |
| XLOC_000450 | XLOC_000450 | CSP41B | 1:3015327-3018234 | q1 | q2 | OK | 113.687 | 650.406 | 2.51627 | 8.83387 | 5.00E-05 | 0.000325 | yes |
| XLOC_000487 | XLOC_000487 | LRR XI-23 | 1:3252239-3255693 | q1 | q2 | OK | 26.4081 | 49.6396 | 0.910512 | 2.30664 | 5.00E-05 | 0.000325 | yes |
| XLOC_000598 | XLOC_000598 | ATGLX1 | 1:3995168-3997907 | q1 | q2 | OK | 60.1583 | 162.387 | 1.4326 | 3.26419 | 5.00E-05 | 0.000325 | yes |
| XLOC_000600 | XLOC_000600 | AT1G11860 | 1:4001112-4003442 | q1 | q2 | OK | 319.6 | 756.582 | 1.24323 | 4.18318 | 5.00E-05 | 0.000325 | yes |
| XLOC_000614 | XLOC_000614 | AT1G12080 | 1:4084161-4085045 | q1 | q2 | OK | 1884.29 | 67.9613 | -4.79316 | -9.20293 | 5.00E-05 | 0.000325 | yes |
| XLOC_000616 | XLOC_000616 | CHL1-1 | 1:4105232-4109545 | q1 | q2 | OK | 107.267 | 57.7917 | -0.892267 | -2.70294 | 5.00E-05 | 0.000325 | yes |
| XLOC_000624 | XLOC_000624 | AT1G12230 | 1:4147961-4151056 | q1 | q2 | OK | 102.049 | 50.9296 | -1.00268 | -2.40566 | 5.00E-05 | 0.000325 | yes |
| XLOC_000680 | XLOC_000680 | CYP71B7 | 1:4467219-4469033 | q1 | q2 | OK | 17.1443 | 84.588 | 2.30272 | 4.53043 | 5.00E-05 | 0.000325 | yes |
| XLOC_000724 | XLOC_000724 | AT1G13930 | 1:4761011-4762666 | q1 | q2 | OK | 94.6747 | 2483.48 | 4.71324 | 10.4968 | 5.00E-05 | 0.000325 | yes |
| XLOC_000749 | XLOC_000749 | AT1G14345 | 1:4899144-4899979 | q1 | q2 | OK | 38.3992 | 157.145 | 2.03295 | 4.49341 | 5.00E-05 | 0.000325 | yes |
| XLOC_000765 | XLOC_000765 | AT1G14670 | 1:5037611-5040528 | q1 | q2 | OK | 84.8105 | 44.439 | -0.932415 | -2.66978 | 5.00E-05 | 0.000325 | yes |
| XLOC_000835 | XLOC_000835 | NDF1 | 1:5489297-5493772 | q1 | q2 | OK | 20.0548 | 104.567 | 2.3824 | 4.27443 | 5.00E-05 | 0.000325 | yes |
| XLOC_000884 | XLOC_000884 | HCF173 | 1:5723087-5727312 | q1 | q2 | OK | 7.34039 | 112.227 | 3.93442 | 5.2414 | 5.00E-05 | 0.000325 | yes |
| XLOC_000916 | XLOC_000916 | FUG1 | 1:5885082-5890470 | q1 | q2 | OK | 48.9638 | 105.457 | 1.10687 | 3.5512 | 5.00E-05 | 0.000325 | yes |
| XLOC_001003 | XLOC_001003 | NDF6 | 1:6460597-6462224 | q1 | q2 | OK | 45.3045 | 185.555 | 2.03412 | 2.97075 | 5.00E-05 | 0.000325 | yes |
| XLOC_001030 | XLOC_001030 | LHCA6 | 1:6612748-6613972 | q1 | q2 | OK | 52.6816 | 153.395 | 1.54188 | 4.09397 | 5.00E-05 | 0.000325 | yes |
| XLOC_001063 | XLOC_001063 | PUP14 | 1:6832346-6833837 | q1 | q2 | OK | 37.731 | 91.5568 | 1.27892 | 3.13218 | 5.00E-05 | 0.000325 | yes |
| XLOC_001076 | XLOC_001076 | ATLFNR2 | 1:6942716-6945018 | q1 | q2 | OK | 87.7487 | 1025.37 | 3.54662 | 10.0816 | 5.00E-05 | 0.000325 | yes |
| XLOC_001099 | XLOC_001099 | AT1G20390 | 1:7065493-7071561 | q1 | q2 | OK | 45.6232 | 15.9769 | -1.51378 | -4.22277 | 5.00E-05 | 0.000325 | yes |
| XLOC_001170 | XLOC_001170 | AT1G21680 | 1:7613004-7615339 | q1 | q2 | OK | 27.146 | 80.96 | 1.57647 | 3.93831 | 5.00E-05 | 0.000325 | yes |

GTFファイルに記載された遺伝子ごとの発現カウントに対して倍率、p値、q値が計算される。

# RをつかってMA plotを書いて見よう

gene_exp.diffファイルを読み込んでみる
tab区切りテキストファイルなのでread.delim関数で読み込む
M, Aをそれぞれ計算する
plot関数を使って描画
colorのパラメータをsignifitureの値で色分けさせてみる。

例)

```
dat <- read.delim("gene_exp.diff")
A<-1/2*(log2(dat$value_1+1)+log2(dat$value_2+1))
M<-log2(dat$value_1+1)-log2(dat$value_2+1)
plot(A,M,col=dat$significant, pch=16, cex=0.4, ylim=c(-8,8))
```

# 簡易スクリプトを使って、結果を成形してみよう

Awkは便利な簡易スクリプト
　１行記述でもできる

例）
　q_valueが0.05以下のもののみリストアップするには？
　　q_valueの記載は13列目だから・・・

　awk '$13<=0.05 {print $0}' gene_exp.diff
　と記述すればOK
　$で列番号を指定できる
　$0は行全体を意味する

その他
　grep, head, sort, cut, uniq等のUnixコマンドも活用しよう

# 実践演習課題

データセット
2D_1, 2D_2, 2D_3と2D2L_1, 2D2L_2, 2D2L_3をTopHat→Cufflinksの系を用いて、
2D（2days dark条件で育てた芽生え)
2D2L（その後2days light条件で育てた芽生え）
でのDE gene等を確認せよ。

GTFファイルとしてgenes_chr4.gtf
fastaファイルとしてgenome_chr4.fa
を利用する。
（アラビドプシスTAIR10の配列だが計算時間を考慮して、
それぞれChr4のみになっている）

RNA-Seqパイプライン -ゲノムベースの解析法-の最終3スライドを参考に、
マッピングデータのIGVでの可視化、
エクセルでの確認、
Rを用いたM-A plotの描画、
簡易スクリプトを用いたデータ抽出をせよ。