

# **Cufflinks.cuffdiff Documentation**

**Description:** Finds significant changes in transcript expression, splicing, and

promoter use.

**Author:** Cole Trapnell et al, University of Maryland Center for

**Bioinformatics and Computational Biology** 

**Cufflinks Version:** Release 1.3.0

**Contact:** qp-help@broadinstitute.org

# Summary

Cufflinks.cuffdiff finds significant changes in transcript expression, splicing, and promoter use. The Cufflinks.cuffdiff module takes a GTF file of transcripts as input, along with two or more SAM or BAM files containing the fragment alignments for two or more samples.

For more information on GTF format, see the specification at <a href="http://mblab.wustl.edu/GTF22.html">http://mblab.wustl.edu/GTF22.html</a>. SAM is a standard short read alignment that allows aligners to attach custom tags to individual alignments. BAM is the binary equivalent of SAM. For more information about the SAM/BAM format, see the specification at <a href="http://samtools.sourceforge.net/SAM1.pdf">http://samtools.sourceforge.net/SAM1.pdf</a>.

Cufflinks.cuffdiff produces a number of output files that contain test results for changes in expression at the level of transcripts, primary transcripts, and genes. It also tracks changes in the relative abundance of transcripts sharing a common transcription start site, and in the relative abundances of the primary transcripts of each gene. Tracking the former shows changes in splicing, and the latter shows changes in relative promoter use within a gene.

Cufflinks.cuffdiff was created at the University of Maryland Center for Bioinformatics and Computational Biology. This document is adapted from the Cufflinks documentation for release 0.9.3. For more information about Cufflinks.cuffdiff, see the Cufflinks Web site.

### **Important Notes:**

There are known issues that prevent Cufflinks.cuffdiff from running on the Mac Mini and possibly other Macintosh hardware.

This module may produce some empty files. This does not mean that the algorithm has failed; it is generally the result when no transcripts with differential expression are detected. In particular, this may occur if there is no differential expression.

## **Preparing to Run Cufflinks.cuffdiff**

In the case where there are two SAM/BAM input files, these can be specified directly as input parameters to the module. However, if there are more than two SAM/BAM files, a list of input SAM/BAM files should be specified in a text file. The text file can be passed to the module via the input.file.list parameter. The files listed **must** be either available on the same file system as the server or accessible via URL. In the text file, each filename should include its full path or URL. Files that are on the same line and are commaseparated are considered to be replicates of a *single* sample; files pertaining to *different* samples should appear on separate lines.



Cufflinks.cuffdiff requires that transcripts in the input GTF be annotated with certain attributes in order to look for changes in primary transcript expression, splicing, coding output, and promoter use. These attributes are:

- tss\_id: The ID of this transcript's inferred start site. Determines which primary transcript this processed transcript is believed to come from.
- p\_id: The ID of the coding sequence this transcript contains. This attribute is
  attached to Cufflinks.cuffcompare output by Cufflinks.cuffcompare only when it is run
  with a reference annotation that include coding sequence (CDS) records. Further,
  differential CDS analysis is only performed when all isoforms of a gene have p\_id
  attributes, because neither Cufflinks nor Cufflinks.cuffcompare attempt to assign an
  open reading frame to transcripts.

#### References

Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. <u>Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation</u>

Nat Biotechnol. 2010;28:511-515. (http://dx.doi.org/10.1038/nbt.1621)

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

### Links

Cufflinks: http://cufflinks.cbcb.umd.edu/

Cufflinks documentation: http://cufflinks.cbcb.umd.edu/manual.html

#### **Parameters**

Name	Description
first.SAM.or. BAM.file	Input file of aligned RNA-seq reads in SAM or BAM format. For more information about the SAM/BAM format, see the specification at <a href="http://samtools.sourceforge.net/SAM1.pdf">http://samtools.sourceforge.net/SAM1.pdf</a> . EITHER these two SAM/BAM files must be specified OR an input files list of SAM/BAM files must be specified.
second.SAM.or. BAM.file	Input file of aligned RNA-seq reads in SAM or BAM format. EITHER these two SAM/BAM files must be specified OR an input file list of SAM/BAM files must be specified.

# GenePattern

input.files.list	If you are specifying more than two SAM or BAM files, list the absolute pathnames of the input SAM/BAM files in a TXT file. Each line in the TXT file corresponds to a different sample. If there are replicate SAM/BAM files for the same sample, list them on the same line in the TXT file, separated by commas. If SAM/BAM files for more than two samples are specified, Cufflinks.cuffdiff tests for differential expression and regulation between all pairs of samples.
GTF.file (required)	A reference annotation transcript GTF or GFF file produced by Cufflinks, Cufflinks.cuffcompare, or other source. For more information on GTF/GFF format, see the specification at <a href="http://mblab.wustl.edu/GTF22.html">http://mblab.wustl.edu/GTF22.html</a> .
output.label (optional)	A TXT file containing a label for each sample, one label per line.
time.series (optional)	Analyze the provided samples as a time series, rather than testing for differences between all pairs of samples. Default: no
upper.quartile. norm (optional)	Tell Cufflinks to normalize by the upper quartile of the number of fragments mapping to individual loci instead of the total number of sequenced fragments. This can improve robustness of differential expression calls for less abundant genes and transcripts. Default: no
total.hits. norm (optional)	Tell Cufflinks to count all fragments, including those not compatible with any reference transcript, towards the number of mapped fragments used in the FPKM denominator. This option can be combined with <code>-N/upper-quartile-norm</code> . It is inactive by default. Default: no
compatible.hits. norm (optional)	Tell cufflinks to count only those fragments compatible with some reference transcript towards the number of mapped fragments used in the FPKM denominator. This option can be used with upper.quartile.norm. Using this mode is generally recommended in Cufflinks.cuffdiff to reduce certain types of bias caused by differential amounts of ribosomal reads which can create the impression of falsely differentially expressed genes. Default: yes

# GenePattern

frag.bias.correct (optional)	Analyze the provided samples as a time series, rather than testing for differences between all pairs of samples. Default: no
multi.read.correct (optional)	Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome. Default: no
min.alignment. count (optional)	The minimum number of alignments in a locus needed to conduct significance testing on changes in that locus observed between samples. If no testing is performed, changes in the locus are deemed not significant, and the locus's observed changes do not contribute to correction for multiple testing. Default: 500 fragment alignments (up to 1000 paired reads)
FDR (optional)	The allowed false discovery rate. Default: 0.05
mask.file (optional)	This file tells Cufflinks.cuffdiff to ignore all reads that could have come from transcripts in this GTF/GFF file. It is recommended that annotated rRNA, mitochondrial transcripts, and other abundant transcripts you want to ignore in your analysis be included in this file.



## **Output Files**

For more information on the formats of the individual output files, see the <u>Cufflinks Web</u> <u>site</u>.

## 1. FPKM tracking files

Cufflinks.cuffdiff calculates the FPKM of each transcript, primary transcript, and gene in each sample. Primary transcript and gene FPKMs are computed by summing the FPKMs of transcripts in each primary transcript group or gene group. There are **four** FPKM tracking files:

- isoforms.fpkm\_tracking: Transcript FPKMs
- genes.fpkm\_tracking: Gene FPKMs. Tracks the summed FPKM of transcripts sharing each gene\_id.
- cds.fpkm\_tracking: Primary transcript FPKMs. Tracks the summed FPKM of transcripts sharing each tss\_id.
- tss\_groups.fpkm\_tracking: Coding sequence FPKMs. Tracks the summed FPKM of transcripts sharing each p id, independent of tss id.

### 2. Differential expression tests

These tab-delimited files list the results of differential expression testing between samples for spliced transcripts, primary transcripts, genes, and coding sequences. For each pair of samples *x* and *y*, **four** files are created:

- tss\_group\_exp.diff: Primary transcript differential FPKM. Tests differences in the summed FPKM of transcripts sharing each tss\_id.
- isoform\_exp.diff: Transcript differential FPKM.
- gene\_exp.diff: Gene differential FPKM. Tests difference sin the summed FPKM of transcripts sharing each gene\_id.
- cds\_exp.diff: Coding sequence differential FPKM. Tests differences in the summed FPKM of transcripts sharing each p\_id independent of tss\_id.
- 3. Differential splicing tests: splicing.diff

This tab-delimited file lists, for each primary transcript, the amount of overloading detected among its isoforms, i.e., how much differential splicing exists between isoforms processed from a single primary transcript. Only primary transcripts from which two or more isoforms are spliced are listed in this file.

### 4. Differential coding output: cds.diff

This tab-delimited file lists, for each gene, the amount of overloading detected among its primary transcripts, i.e., how much differential promoter use exists between samples. Only genes producing two or more distinct primary transcripts (i.e., multi-promoter genes) are listed here.

### 5. Differential promoter use: promoters.diff

This tab-delimited file lists, for each gene, the amount of overloading detected among its coding sequences, i.e., how much differential CDS output exists between samples. Only genes producing two or more distinct CDS (i.e., multi-protein genes) are listed here.



# **Platform Dependencies**

Module type: RNA-seq

**CPU type:** any

**OS:** Macintosh, Linux

Language: C++, Perl

# **GenePattern Module Version Notes**

Version Description

3.0 Version 3.0 of this module contains Cuffdiff v1.3.0. For

more information about Cuffdiff v1.3.0, see the links

listed in the "Links" section above.