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 2.0.2
Contact:	gp-help@broadinstitute.org

Summary

Cufflinks.cuffdiff finds significant changes in transcript expression, splicing, and promoter use. You can use it to find differentially expressed genes and transcripts, as well as genes that are being differentially regulated at the transcriptional and post-transcriptional level.

To identify a gene or transcript as differentially expressed, Cufflinks.cuffdiff tests the observed log fold change in its expression against the null hypothesis of no change (i.e., that the true log fold change = zero). Because measurement error, technical variability, and cross-replicate biological variability might result in an observed log fold change that is nonzero even if the gene/transcript is not differentially expressed, Cufflinks.cuffdiff also assesses the significance of each comparison. For more information on the model, see Trapnell et al (2013) or see the "How It Works" page on the Cufflinks site.

Cufflinks.cuffdiff also groups transcripts into biologically meaningful groups, such as transcripts that share the same transcription start site (TSS), in order to identify genes that are differentially regulated at the transcriptional or post-transcriptional level.

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 2.0.2.

Usage

The Cufflinks.cuffdiff module takes two or more fragment alignment SAM/BAM files (from TopHat or other read aligner), as well as a GTF file containing reference genome annotations (such as merged.gtf from Cufflinks.cuffmerge) as input.

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.

For more information on using RNA-seq modules in GenePattern, see the RNA-seq Analysis page.

Important Notes:

Cufflinks.cuffdiff jobs can be very resource intensive. If your job does not complete within a day, retry it on a server with more available memory, or, if you are running on the GenePattern public server, see this FAQ.

There are known issues that prevent Cufflinks.cuffdiff from running on the Mac Mini. If you have issues that prevent Cufflinks.cuffdiff from running on your machine, contact gp-help@broadinstitute.org.

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 available on the same file system as the server. In the text file, each filename should include its full path. Files that are on the same line and are comma-separated are considered to be replicates of a *single* sample; files pertaining to *different* samples should appear on separate lines.

If you want Cufflinks.cuffdiff to look for changes in primary transcript expression, splicing, coding output, and promoter use, the input GTF transcript file needs to be annotated with certain attributes. These attributes are:

- tss_id: The ID of a transcript's inferred start site; this determines which primary transcript this processed transcript is believed to come from.
- p_id: The ID of the coding sequence this transcript contains.

Important note on Cufflinks.cuffdiff results

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

It may also be the result of using an input GTF transcript file that does not have p_id annotation. This attribute is attached to Cufflinks.cuffmerge output only when it is run with a reference annotation that includes coding sequence (CDS) records. Differential CDS analysis in Cufflinks.cuffdiff is only performed when all isoforms of a gene have p_id attributes. The CDS output files will be empty if there is no p_id attribute in the input GTF.

References

Trapnell C, Hendrickson D,Sauvageau S, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*. 2013;31:46-53.

- http://www.nature.com/nbt/journal/v31/n1/full/nbt.2450.html

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 2012;7;562–578.

- http://www.nature.com/nprot/journal/v7/n3/full/nprot.2012.016.html

Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*. 2011 Sep 1;27(17):2325-9.

- http://bioinformatics.oxfordjournals.org/content/27/17/2325.long

Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010;28:511-515.

- http://dx.doi.org/10.1038/nbt.1621

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25:1105-1111.

- http://bioinformatics.oxfordjournals.org/cgi/content/abstract/btp120

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

- http://genomebiology.com/2009/10/3/R25

Links

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 Input Files section. **EITHER** these SAM/BAM files must be specified **OR** an *input files list* of SAM/BAM files must be specified.
second SAM or BAM file	Second input file of aligned RNA-seq reads in SAM or BAM format. **EITHER** these SAM/BAM files must be specified **OR** an *input files list* of SAM/BAM files must be specified.
input files list	If you are specifying more than two SAM/BAM files, list the absolute pathnames of the input SAM/BAM files in a text file. Each line in the 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 text file, separated by commas. The files listed **must** be available on the same file system as the server. 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 GTF or GFF file containing reference genome annotations. For more information on GTF and GFF formats, see the Input Files section.

sample labels (optional)	A text file containing a label for each sample, one label per line. These labels replace the default "q0, q1, ...qN" labeling for each sample in the tracking output files. While this parameter is optional, using it may make downstream analysis of your samples easier.
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.cuffdiff 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.cuffdiff 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 *upper quartile norm*. Default: no
compatible hits norm (optional)	Tell Cufflinks.cuffdiff to count only those fragments compatible with some reference transcript toward 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 Cuffdiff to reduce certain types of bias caused by differential amounts of ribosomal reads that can create the impression of falsely differentially expressed genes. Default: yes
frag bias correct (optional)	A genome reference multi-FASTA file for the bias detection and correction algorithm. For more information on this algorithm, see "How It Works" on the Cufflinks website.
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.
library type	The library type used to generate reads. The choices are inferred, fr-unstranded, fr-firststrand, fr-secondstrand, ff-unstranded, ff-firststrand, ff-secondstrand, and transfrags. The default is inferred, meaning that no library type information is passed.

Input Files

1. EITHER two SAM/BAM files containing aligned RNA-seq reads
   OR
   a text file containing the absolute pathnames of more than two SAM/BAM files

   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 http://samtools.sourceforge.net/.
2. GTF/GFF file containing reference genome annotations

   A common file used as input here is merged.gtf from Cufflinks.cuffmerge. For more information on GTF format, see the specification at http://mblab.wustl.edu/GTF22.html.
3. Genome reference multi-FASTA file (optional, submitted in the *frag bias correct* parameter)
   This reference genome file instructs Cufflinks.cuffdiff to run the bias detection and correction algorithm. For more information on this algorithm, see "How It Works" on the Cufflinks website.
4. mask.file (optional, submitted in the *mask file* parameter)
   A tab-delimited GTF/GFF file that specifies transcripts to be ignored.

Output Files

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

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. For more information on the FPKM_tracking format, see the file format page.
   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: Coding sequence FPKMs. Tracks the summed FPKM of transcripts sharing the p_id (ID of the coding sequence each transcript), independent of tss_id.
   - tss_groups.fpkm_tracking: Primary transcript FPKMs. Tracks the summed FPKM of transcripts sharing each tss_id (transcription start site [TSS] ID), which is the ID of the transcript's inferred start site, determining which primary transcript this processed transcript is believed to come from).

2. Count tracking files

   Cufflinks.cuffdiff estimates the number of fragments that originated from each transcript, primary transcript, and gene in each sample. Primary transcript and gene counts are computed by summing the counts of transcripts in each primary transcript group or gene group. The results are output in the format described here. There are four count tracking files:

   - isoforms.count_tracking: Transcript counts.
   - genes.count_tracking: Gene counts. Tracks the summed counts of transcripts sharing each gene ID.
   - cds.count_tracking: Coding sequence counts. Tracks the summed counts of transcripts sharing each p_id, independent of tss_id.
   - tss_groups.count_tracking: Primary transcript counts. Tracks the summed counts of transcripts sharing each tss_id.

3. Read group tracking files

   Cuffdiff calculates the expression and fragment count for each transcript, primary transcript, and gene in each replicate. The results are output in per-replicate tracking files in the format described here. There are four read group tracking files:

   - isoforms.read_group_tracking: Transcript read group tracking.
   - genes.read_group_tracking: Gene read group tracking. Tracks the summed expression and counts of transcripts sharing each gene ID in each replicate.
   - cds.read_group_tracking: Coding sequence FPKMs. Tracks the summed expression and counts of transcripts sharing each p_id, independent of tss_id in each replicate.
   - tss_groups.read_group_tracking: Primary transcript FPKMs. Tracks the summed expression and counts of transcripts sharing each tss_id in each replicate.

4. 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:

   - isoform_exp.diff: Transcript differential FPKM.
   - gene_exp.diff: Gene differential FPKM. Tests differences in 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.
   - tss_group_exp.diff: Primary transcript differential FPKM. Tests differences in the summed FPKM of transcripts sharing each tss_id.

5. 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.

6. Differential coding output: cds.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.

7. Differential promoter use: promoters.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.

8. Read group information: read_groups.info

   This tab-delimited file lists, for each replicate, key properties used by Cufflinks.cuffdiff during quantification, such as library normalization factors.

9. Run information: run.info

   This tab-delimited file lists information about a Cufflinks.cuffdiff run to help track what options were provided.

Platform Dependencies

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

GenePattern Module Version Notes

Version	Release Date	Description
4	5/6/13	Updated to Cufflinks 2.0.2.
3	1/13/2012	Updated to Cufflinks 1.3.0.
2	12/23/2011	Updated to Cufflinks 1.2.1.
1	4/11/11	Initial version.