GenePattern

Cufflinks Documentation

Description:	Assembles transcripts and estimates their abundances in RNA-seq samples.
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 assembles transcripts and estimates their abundances in RNA-seq samples. It accepts aligned RNA-seq reads, then assembles the alignments into a parsimonious set of transcripts, reporting as few full-length transcript fragments [transfrags] as are needed to explain the data. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one.

Cufflinks 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

Cufflinks takes a file of alignments in SAM or BAM (the binary equivalent of SAM) format as input. For more details on the SAM/BAM format, see the Input Files section and/or the specification at http://samtools.sourceforge.net/. The RNA-seq read mapper TopHat produces BAM output, and is recommended for use with Cufflinks. However Cufflinks will accept SAM/BAM alignments generated by any read mapper.

Optionally, a reference genome annotation file can be submitted as well. If it is sent to the *GTF* parameter, Cufflinks will use this file to estimate isoform expression and will not assemble novel transcripts; the program will ignore alignments not structurally compatible with any reference transcript. It can also be sent to the *GTF guide* parameter to enable Cufflinks to use the reference annotation based transcript (RABT) assembly algorithm. This guide file is used to generate faux-reads against which the actual reads are tiled so that every reference transcript position is covered by multiple reads, and the information in the faux-reads is merged with the data from the sequenced reads. For more information, see Roberts et al (2011) or the "How It Works" page on the Cufflinks site. The reference genome annotation GTF can be sent to one or both of these parameters.

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

Important Notes:

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

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

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

Parameters

Name	Description
input file (required)	Input file of RNA-seq read alignments in SAM/BAM format. For more information on SAM/BAM format, see the Input Files section. Cufflinks requires that the supplied alignments have custom tags. See Input formats (in the Cufflinks documentation) for more details.
transfrag label (optional)	A label for the transcribed fragments (transfrags) in the output files.

GTF (optional)	Reference annotation file in GTF/GFF format. For more information on GTF/GFF format, see the Input Files section. Cufflinks will use this file to estimate isoform expression. It will not assemble novel transcripts, and the program will ignore alignments not structurally compatible with any reference transcript.
GTF guide (optional)	Annotation file in GTF/GFF format, used to guide reference annotation based transcript (RABT) assembly. Reference transcripts will be tiled with faux-reads to provide additional information in assembly. Output will include all reference transcripts as well as any novel genes and isoforms that are assembled.
mask file (optional)	GTF/GFF file specifying transcripts to be ignored. The Cufflinks team recommends including any annotated rRNA, mitochondrial transcripts, and other abundant transcripts you wish to ignore in this file. Due to variable efficiency of mRNA enrichment methods and rRNA depletion kits, masking these transcripts often improves the overall robustness of transcript abundance estimates.
frag bias correct (optional)	Providing Cufflinks with a FASTA/FA file via this option instructs it to run a bias detection and correction algorithm that can significantly improve accuracy of transcript abundance estimates.
multi read correct (optional)	Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.
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.
min frags per transfrag (optional)	Assembled transfrags supported by fewer than this many aligned RNA-Seq fragments are not reported.

Input Files

1. SAM or BAM file (required)
   File of RNA-seq read alignments in SAM (a tab-delimited format) or BAM (a compressed binary version of SAM) format. SAM is a standard short read alignment that allows aligners to attach custom tags to individual alignments. This file is the output of a read mapping application, such as TopHat, and the alignment section contains information regarding the mapped location of each sequenced RNA-seq read on a reference genome.
   For more information on the SAM format, see the specification:
   http://samtools.sourceforge.net/
2. GTF file (optional)
   A tab-delimited reference annotation file in GTF format. This file is used by Cufflinks to estimate abundances of isoforms. (This file can be the same file submitted in the *GTF guide* parameter.) These reference annotation files can be downloaded for many genomes from sites like UCSC Genome Browser.
   For more information on the GTF format, see the specification:
   http://cufflinks.cbcb.umd.edu/gff.html
3. GTF guide file (optional)
   A tab-delimited reference annotation file in GTF format. This file is used by Cufflinks to guide RABT assembly. (This file can be the same file submitted in the *GTF* parameter.) Reference annotation files can be downloaded from
4. mask.file (optional)
   A tab-delimited GTF file that specifies transcripts to be ignored.
5. frag bias correct (optional)
   Reference multi-FASTA file for bias detection and correction algorithm. For more information on this format, see this description:
   http://www.bioperl.org/wiki/FASTA_multiple_alignment_format

Output Files

1. transcripts.gtf

   This GTF file contains Cufflinks' assembled isoforms. The first 7 columns are standard GTF, and the last column contains attributes, some of which are also standardized ("gene_id" and "transcript_id"). There is one GTF record per row, and each record represents either a transcript or an exon within a transcript.

2. genes.fpkm_tracking

   This is a tab-delimited file containing one row per gene; the columns contain the attributes in the GTF file. This file contains gene-level coordinates and expression values. Note that since the output for Cufflinks is for a single sample, the "q" numbering format (see the file format information) is not used.

3. isoforms.fpkm_tracking

   This is a tab-delimited file containing one row per isoform; the columns contain the attributes in the GTF file. This file contains transcript-level coordinates and expression values. Note that since the output for Cufflinks is for a single sample, the "q" numbering format (see the file format information) is not used.

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 version 1.3.0, using SAMtools version 0.1.17. Improvements, modifications, and bug fixes include: • Fixed the sorting of read files so that paired read files are correctly matched • Changed the default value of the indel search parameter to *yes* • Added an output prefix parameter so that the output files can be renamed
2	12/23/2011	Updated to Cufflinks 1.2.1.
1	4/11/11	Initial version.