Cuffdiff is a program that uses the Cufflinks transcript quantification engine to calculate gene and transcript expression levels in more than one condition and test them for signficant differences. 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.

How does Cuffdiff test for differential expression and regulation?

Cuffdiff takes a GTF file of transcripts as input, along with two or more SAM files containing the fragment alignments for two or more samples. It 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 allows one to see changes in splicing, and the latter lets one see changes in relative promoter use within a gene. 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:

Attribute	Description
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 is attribute is attached to Cuffcompare output by Cuffcompare only when it is run with a reference annotation that include CDS records. Further, differential CDS analysis is **only** performed when all isoforms of a gene have `p_id` attributes, because neither Cufflinks nor Cuffcompare attempt to assign an open reading frame to transcripts.

The above attributes, along with the `gene_id` required by the GTF specification, make each transcript a member of a "gene group", "primary transcript group", and "CDS group". Transcripts with the same `gene_id` are part of the same gene group, and similarly, those with the same `tss_id` and `p_id` are part of the same primary transcript group and CDS group. Cuffdiff tracks changes not only in absolute transcript, primary transcript, CDS, and gene FPKMs, but also in the *relative* expression changes within these groups.

To test whether an observed difference in a gene's expression is significant, Cuffdiff compares the log ratio of gene's expression in two conditions against the log of one. Suppose we write the ratio of expression of a transcript "t" in condition a versus condition b as

$$Y = \frac{FPKM_a}{FPKM_b}$$

The log of the ratio (T) of expression in two conditions can actually be used as a test statistic, because the quantity:

$$T = \frac{E[\log(Y)]}{Var[\log(Y))]}$$

is approximately normally distributed and can be calculated as

$$T = \frac{E[\log(Y)]}{\mathrm{Var}[\log(Y)]} \approx \frac{\log\left(\frac{FPKM_a}{FPKM_b}\right)}{\sqrt{\frac{\mathrm{Var}[FPKM_a]}{FPKM_a^2} + \frac{\mathrm{Var}[FPKM_b]}{FPKM_b^2}}}$$

Note that in order to calculate the test statistic T, we need to know the *variance* of the expression level in each condition. The variance needs to include the variability in the number of fragments generated by the transcript across replicates, and should also incorporate any *uncertainty* in the expression estimate itself. If the transcript *t* is the only isoform of the gene it belongs to (and all its reads map to it uniquely), there's no uncertainty in the expression estimate. However, if the transcript is one of several isoforms, there will

be some uncertainty about each isoform's expression level. Cuffdiff calculates the variance in a transcript's expression level as

$$\mathrm{Var}[FPKM_t] = \left(\frac{10^9}{\tilde{l}(t)\bar{M}}\right)^2 (\mathrm{Var}[X_t])$$

Where

$$\mathrm{Var}[X_t]$$

is the variance in the number of fragments that come from the transcript across replicates in the experiment. If there are no replicates, the variance is measured across the conditions under the assumption that most transcripts are not differentially expressed. However, if your experiment has replicates for each condition, Cuffdiff can better estimate the variance for fragment counts and thus provides more accurate differential expression calls. For single isoform genes, Cuffdiff models the variance in fragment counts across replicates using the negative binomial distribution, similar to the method described by Anders and Huber (2010). When a gene has multiple isoforms, Cuffdiff must incorporate its own uncertainty in how it assigns fragment to transcripts in its estimate of the variance for the counts. Cuffdiff uses the beta negative binomial to model overdispersion and fragment assignment uncertainty simultaneously. The beta negative binomial is a distribution that arises when mixing several negative binomials, so in a sense, Cuffdiff generalizes previous count based methods that use the negative binomial distribution. For each gene, Cuffdiff fits the parameters of the beta negative binomial (called α, β, and r) by solving the following system of equations:

$$\frac{r(1-p)}{p} = A$$

$$\frac{r(1-p)}{p^2} = B$$

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{A^4}{B^4} \cdot \frac{C}{r^2}.$$