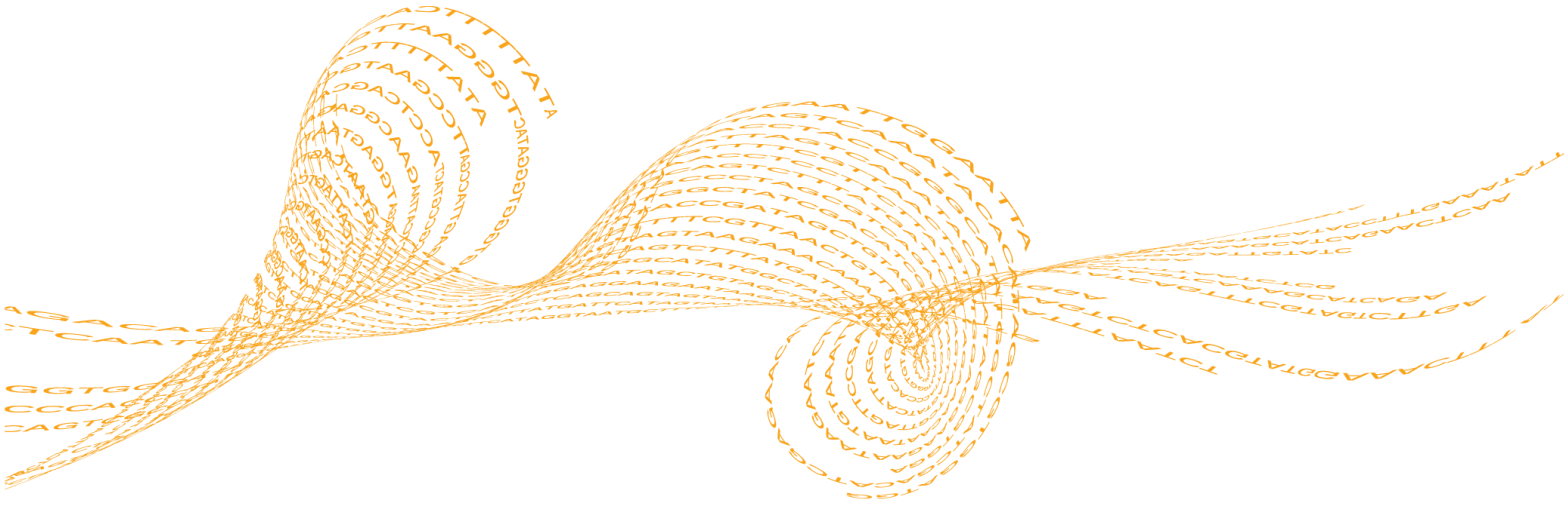
Cufflinks Assembly & DE v2.0

BaseSpace App Guide

For Research Use Only. Not for use in diagnostic procedures.



ILLUMINA PROPRIETARY

Doc # 1000000006108 v00
February 2016

illumina®

Introduction

Using the analysis results from the RNA-Seq Alignment App, the BaseSpace® Cufflinks Assembly & DE v2.0 App can quickly assess novel transcript isoforms and gene expression levels. The app uses Cuffmerge, Cuffquant, Cuffnorm, and Cufflinks tools to perform novel transcript merging. Also, the app uses Cuffmerge, Cuffquant, and Cuffdiff tools to perform differential expression analysis.

Compatible Libraries

See the BaseSpace support page for a list of library types that are compatible with the Cufflinks Assembly & DE App.

Versions

The following components are used in the Cufflinks Assembly & DE App.

Software	Version
Isis (Analysis Software)	2.6.25.12
Bedtools	2.17.0
Cufflinks	2.2.1
BLAST	2.2.26+

Workflow Requirements

- Requires alignment results from the RNA-Seq Alignment App.
- Supports a maximum of 120 alignment results across all the groups.
- Supports a maximum of 200 GB size of alignment results for each group.

Reference Genomes

The following reference genomes are available:

- *Homo sapiens* UCSC hg19 (RefSeq & Gencode gene annotations)
- *Homo sapiens* UCSC hg38 (RefSeq & Gencode gene annotations)
  The human reference genome is PAR-Masked, which means that the Y chromosome sequence has the Pseudo Autosomal Regions (PAR) masked (set to N) to avoid mismapping of reads in the duplicate regions of sex chromosomes.
- *Arabidopsis thaliana* Ensembl TAIR10 (Ensembl gene annotation)
- *Bos taurus* UCSC bosTau6 (RefSeq gene annotation)
- *Caenorhabditis elegans* UCSC ce10 (RefSeq gene annotation)
- *Danio rerio* UCSC danRer7 (RefSeq gene annotation)
- *Drosophila melanogaster* UCSC dm3 (RefSeq gene annotation)
- *Gallus gallus* UCSC galGal4 (RefSeq gene annotation)
- *Mus musculus* UCSC mm9 (RefSeq gene annotation)
- *Mus musculus* UCSC mm10 (RefSeq gene annotation)
- *Oryza sativa japonica* Ensembl IRGSP-1.0 (Ensembl gene annotation)
- *Rattus norvegicus* UCSC rn5 (RefSeq gene annotation)
- *Saccharomyces cerevisiae* Ensembl R64-1-1 (Ensembl gene annotation)

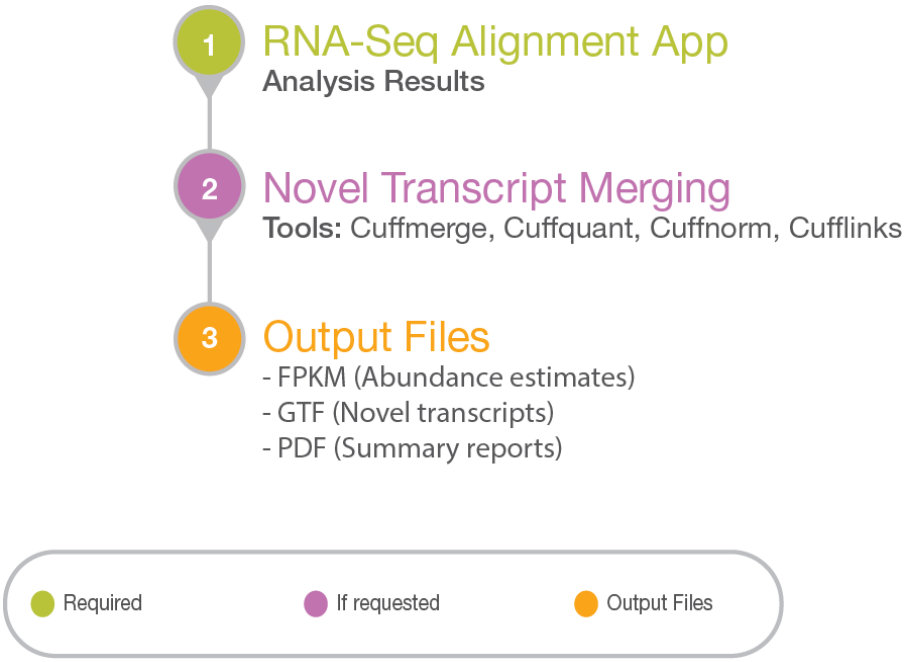
- *Sus scrofa* UCSC susScr3 (RefSeq gene annotation)
- *Zea mays* Ensembl AGPv3 (Ensembl gene annotation)

Workflow

- **Novel Transcript Merging**—When the RNA-Seq Alignment App performs novel transcript assembly, the Cuffmerge tool merges the assemblies within each sample group, and then combines them with the known gene models from annotation. The Cuffcompare tool compares the assembled transcripts to a reference annotation. For example, Cuffcompare identifies whether an assembled transcript overlaps with a known transcript. Lastly, Cuffcompare re-estimates abundances for each known or novel transcript
- **Differential Expression**—The Cuffdiff tool calculates differential expression between 2 sample groups and estimates variance from the individual samples in the 2 groups. For more information, see cole-trapnell-lab.github.io/cufflinks/cuffdiff/
  - For Cuffdiff to calculate differential expression, it requires a consensus set of transcripts that is compared between different sample groups. When you select novel transcript assembly, Cuffmerge merges the consensus set of transcripts from each sample group, and then it merges with the known gene models from annotation. When novel transcript is not performed, the consensus set of transcripts is set to the reference annotation.

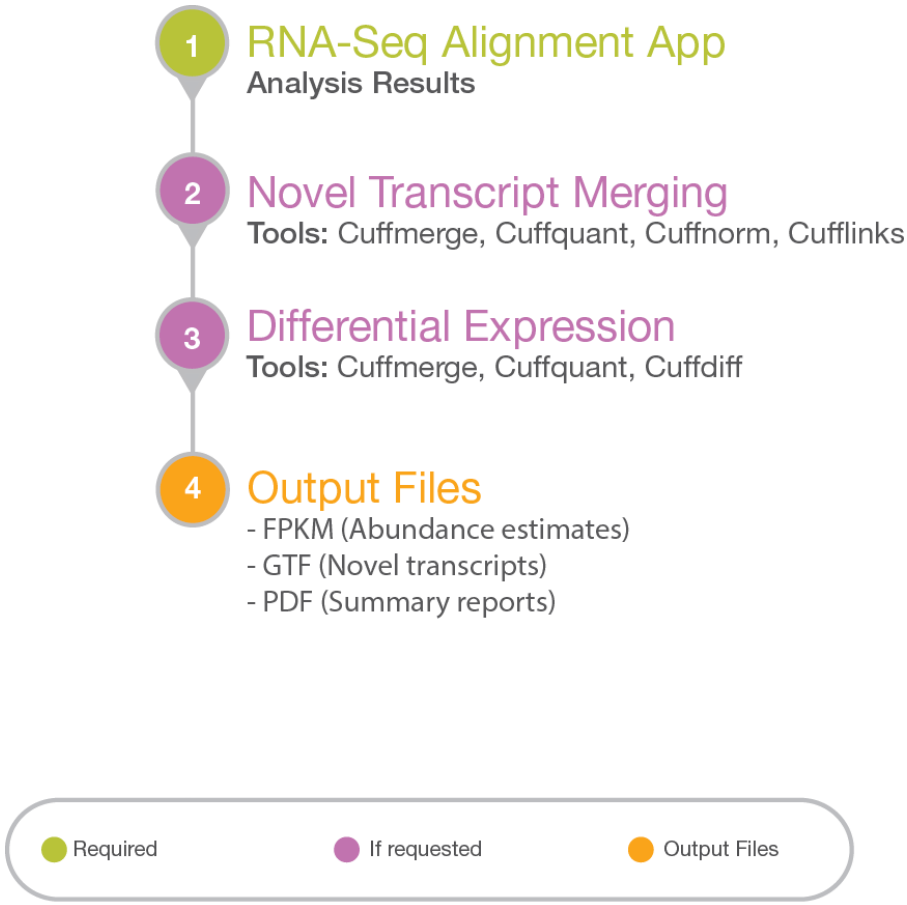
Workflow Diagram for Global Analysis

Figure 1 Cufflinks Global Analysis Workflow



Workflow Diagram for Differential Expression

Figure 2 Cufflinks Differential Expression Workflow



Set Analysis Parameters

1. Navigate to BaseSpace, and then click the **Apps** tab.
2. In Categories, click **RNA-Seq**, and then click **Cufflinks Assembly & DE**.
3. From the drop-down list, select **version 2.0.0**, and then click **Launch** to open the app.
4. In the App Session Name field, enter the analysis name.
   By default, the analysis name includes the app name, followed by the date and time that the analysis session starts.
5. From the Save Results To field, select the project that stores the app results.

NOTE
Step 6 to 10 are filter options that are listed in the RNA-Seq Alignment App. Set all the filters before you can select the app results in the App Result Groups section.

6. From the Reference Genome field, select the reference genome and set of gene annotations.
   Only the RNA-Seq Alignment app results that aligned to this reference genome are available for Cufflinks analysis.
7. From the Strandedness field, select from the following options:
   - First Strand
   - Not Strand Specific
   - Second Strand
8. From the Aligner field, select from the following methods:
   - STAR
   - TopHat (Bowtie)
   - TopHat (Bowtie2)
9. From the Novel Transcript Assembly field, select from the following options:
   - Novel Assembly was not completed in RNA-Seq Alignment
   - Novel Assembly was completed in RNA-Seq Alignment
10. From the Panel field, select from the following options:
    - None
    - TruSight RNA Pan-Cancer
11. In the App Result Groups section, from the Label field, enter the group label name.
    Pairwise differential expression analysis requires at least 2 groups.
12. Click the plus (+) sign to add additional groups.
13. From the POLYA field, select whether the samples are prepared with polyA selection.
    This option is available if the RNA-Seq Alignment App performed novel transcript assembly.
14. From the Select App Results field, select the results created from the RNA-Seq Alignment App for this group.
    Multiple selections are treated as a single group with replicates.
15. [Optional] Select the **Enable Pairwise Differential Expression Analysis** checkbox for pairwise differential expression analysis.
    By default, the app performs differential expression for all the pairs of groups.
16. [Optional] Complete the following steps to specify the groups you want to compare.

a Select the **Select Pairs of Groups for Differential Expression Analysis** checkbox.

b Enter the control group name.

c Enter the comparison group name.

d Click the plus (+) sign to add additional pairs of groups for pairwise differential expression analysis.

The comparison group name and the control group name should be the same as one of the group label names in the App Result Groups section.

17 [Optional] Select the **Set Advanced Options** checkbox to enable the advanced options and then specify the values for the appropriate options.

18 Click **Continue**.
The Cufflinks Assembly & DE App begins analysis.
When analysis is complete, the app updates the status of the session and sends an email to notify you.

Advanced Options

Advanced Options provides the following optional settings to refine analysis results.

Table 1 Cuffdiff Options Table

Option	Description
Hits Normalization	**Compatible**—Cuffdiff counts only those fragments compatible with some reference transcript towards the number of mapped fragments used in the FPKM denominator. We recommend this option to reduce certain types of bias that cause differential amounts of ribosomal reads, which can create the impression of falsely differentially expressed genes. This option is the default. **Total**—Cuffdiff counts all fragments, including the fragments not compatible with any reference transcript, towards the number of mapped fragments used in the FPKM denominator.
False Discovery Rate	Enter the rate. The default is 0.05.
Dispersion Method	**Pooled**—This method builds a model for every replicated condition, and then all the models are averaged into one global model for all conditions in the experiment. This option is the default. **Per-Condition**—This method builds a model for each replicated condition. This option is available when all conditions have replicates.
Minimum Isoform Fraction	For best results, do not change the default value. The default is 1e-5. Cuffdiff rounds the abundance of alternative isoforms quantified at below the specified fraction of the major isoforms down to 0. This process comes after maximum-likelihood estimation (MLE) estimation and before maximum a posteriori (MAP) estimation to improve robustness of confidence interval generation and differential expression analysis.

Table 2 Cuffdiff/Cuffnorm Options Table

Option	Description
Library Normalization Method	Select the method of library sizes normalization (i.e. sequencing depths).

Table 3 Cuffnorm/Cufflinks Options Table

Option	Description
Hits Normalization	**Compatible**—Cuffnorm/Cufflinks counts only those fragments compatible with some reference transcript towards the number of mapped fragments used in the FPKM denominator. This option is the default. **Total**—Cuffnorm/Cufflinks counts all fragments, including the fragments not compatible with any reference transcript, towards the number of mapped fragments used in the FPKM denominator.

Table 4 Cuffquant/Cufflinks Options Table

Option	Description
Fragment Bias Correction	Select to run bias detection and correction algorithm, which can improve accuracy of transcript abundance estimates.
Multi-read Correction	Select to assess reads mapping to multiple locations in the genome more accurately.

Table 5 Cuffquant/Cufflinks/Cuffdiff Options Table

Option	Description
No Effective Length Correction	Select to disable effective length normalization to transcript FPKM.

Analysis Methods

The Cufflinks Assembly & DE App uses Cufflinks to analyze the sequencing data.

Cufflinks

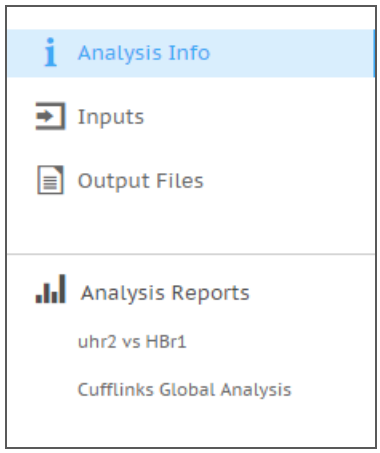
Cufflinks assembles aligned RNA-Seq reads into transcripts, estimates their abundances, and test for differential expression and regulation of transcriptome.

For more information, see cole-trapnell-lab.github.io/cufflinks/.

Analysis Output

To view the results, click the **Projects** tab, then the project name, and then the analysis.

Figure 3 Cufflinks Assembly & DE Output Navigation Bar



Use the left navigation bar to access the following analysis output:

- **Analysis Info**—Information about the analysis session, including log files.
- **Inputs**—Overview of input settings.
- **Output Files**—Output files for the samples.
- **Analysis Reports**
  - **control samples vs comparison samples**—Differential expression analysis for the control vs. comparison samples.
  - **Cufflinks Global Analysis**—Analysis metrics for the aggregate results.

Analysis Info

The Analysis Info page displays the analysis settings and execution details.

Row Heading	Definition
Name	Name of the analysis session.
Application	App that generated this analysis.
Date Started	Date and time the analysis session started.
Date Completed	Date and time the analysis session completed.
Duration	Duration of the analysis.
Session Type	Multi-Node or Single-Node
Status	Status of the analysis session. The status shows either Running or Complete and the number of nodes used.

Log Files

File Name	Description
CompletedJobInfo.xml	Contains information about the completed analysis session.
Logging.zip	Contains all detailed log files for each step of the workflow.
SampleSheet.csv	Sample sheet.
SampleSheetUsed.csv	A copy of the sample sheet.
WorkflowError.txt	Contains error messages created when running the workflow.
WorkflowLog.txt	Contains details about workflow steps, command line calls with parameters, timing, and progress.

Output Files

The Output Files page provides access to the output files for each sample analysis.

- FPKM Files
- DIFF Files
- GTF Files

The app produces GTF files if you selected the Novel Transcript Assembly option in the RNA-Seq Alignment App.

FPKM Files

Fragments Per Kilobase of sequence per Million mapped reads (FPKM) normalizes the number of aligned reads by the size of the sequence feature and the total number of mapped reads.

In each output directory, this app creates the following output files:

- genes.fpkm_tracking—Quantifies the expression of genes specified in the GTF annotation file.
- isoforms.fpkm_tracking—Quantifies the expression of transcripts specified in the GTF annotation file.

DIFF Files

The Cufflinks App creates several DIFF files that describe the differential expression. This tab delimited file lists the results of differential expression testing between samples for spliced transcripts, primary transcripts, genes, and coding sequences.

File Name	Description
isoform_exp.diff	Transcript differential FPKM.
gene_exp.diff	Gene differential FPKM. Tests differences in the summed FPKM of transcripts sharing each `gene_id`.
tss_group_exp.diff	Primary transcript differential FPKM. Tests differences in the summed FPKM of transcripts sharing each `tss_id`.
cds_exp.diff	Coding sequence differential FPKM. Tests differences in the summed FPKM of transcripts sharing each `p_id` independent of `tss_id`.

The DIFF file has the following format.

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 the following: OK (test successful) NOTEST (not enough alignments for testing) LOWDATA (too complex or shallowly sequenced) HIDATA (too many fragments in locus) FAIL, when an ill-conditioned covariance matrix or other numerical exception prevents testing.
7	FPKMx	8.01089	FPKM of the gene in sample x
8	FPKMy	8.551545	FPKM of the gene in sample y
9	log2 (FPKMy/FPKMx)	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 than the FDR after Benjamini-Hochberg correction for multiple-testing

For more information, see the Cufflinks manual at cole-trapnell-lab.github.io/cufflinks/manual/.

GTF Files

The Gene transfer format (GTF) file provides the set of merged transcripts. Each line contains an annotation field ("class_code") that describes the nature of the overlap of this transcript with transcripts from the reference annotation. The table below, taken from the cufflinks manual, (cole-trapnell-lab.github.io/cufflinks/manual/), provides a description of the possible class codes.

=	Match
j	New isoform
e	A single exon transcript overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment
i	A single exon transcript falling entirely with a reference intron
r	Repeat, currently determined by looking at the reference sequence and applied to transcripts where at least 50% of the bases are lower case
p	Possible polymerase run-on fragment

u	Unknown, intergenic transcript
o	Unknown, generic overlap with reference
.	Tracking file only, indicates multiple classifications

Transcripts annotated with the i, j, u, or o class codes represent novel transcripts of potential interest.

For more information, see the Cufflinks manual at cole-trapnell-lab.github.io/cufflinks/manual/.

Analysis Reports

The Cufflinks Assembly & DE App provides an aggregate summary for all the samples and a summary of statistics per sample pair.

Pairewise Differential Expression Analysis

The Cufflinks Assembly & DE App provides the differential expression statistics for the control and comparison sample groups in the Analysis Reports pages. To download the statistics, click **PDF Summary Report**.

Overview

Table 6 Overview Table

Statistic	Definition
Control samples (group name)	Links to RNA-Seq Alignment App results for the control group.
Comparison samples (group name)	Links RNA-Seq Alignment App results for the comparison group.
FPKM tables	Links to gene and transcript FKPM tables.

Assembly

The Assembly table is available if you selected the Novel Transcript Assembly option in the RNA-Seq Alignment App. The app provides analysis for the control group, comparison group, and the merging of the control and comparison groups. Some descriptions are taken from cole-trapnell-lab.github.io/cufflinks/cuffcompare/#transfrag-class-codes.

Table 7 Assembly Table

Statistic	Description
Gene Count	The number of identified genes per sample.
Transcript Count	The number of identified transcripts per sample.
Link to gene models	Link to the gene transfer format (GTF) files results.
Equal (=)	The complete match of intron chain.

Statistic	Description
Potential novel (j)	The potentially novel isoform (fragment); at least one splice junction is shared with a reference transcript.
Unknown, intergenic (u)	The unknown, intergenic transcript.
Overlap with opposite-strand exon (x)	The exonic overlap with reference on the opposite strand.
Other	Other types of transcripts.

Differential Expression

The differential expression table shows only the on-target genes if you selected a panel.

Table 8 Differential Expression Table

Statistic	Definition
Annotation Gene Count	The number of genes in annotation and novel genes.
Assessed Gene Count	The number of genes that are tested successfully.
ΔGene Count	The number of genes with significant differential expression.
Annotation Transcript Count	The number of transcripts in annotation and novel transcripts.
Assessed Transcript Count	The number of transcripts that are tested successfully.
ΔTranscript Count	The number of transcripts with significant differential expression.
Cuffdiff results	Links to selected Cuffdiff results.

Differential Expression Heat Map

The Differential Expression Heat Map shows the level of expression of genes across a number of comparable samples. The heat map shows only the on-target genes if you selected a panel.

The gene dendrogram shows the clustering of genes and the correlation between 2 gene clusters.

The expressed genes are assigned to different colors.

- Red—Genes with low expression.
- Black—Genes with average expression.
- Neon green—Genes with high expression.

To specify a gene, expand the **Jump to Gene** drop-down list, and then select the gene.

To specify a sample, expand the **Jump to Sample** drop-down list, and then select the sample.

To save plot as scalable vector graphics (SVG), click **Save Plot as SVG**.

To export data from plot as TeamViewer Session, click **Export Heatmap Data (TVS)**.

Gene Browser

The Differential Expression Gene Browser shows an interactive scatter plot of the log2 (FPKM) counts of genes for the control samples and comparison samples groups. The table shows only the on-target genes if you selected a panel. You can filter the results for the following metrics:

Table 9 Individual Gene Results Table

Statistic	Definition
Test ID	A unique identifier describing the tested gene.
Gene	The gene symbol.
Locus	The genomic coordinates for browsing easily to tested genes and transcripts.
Status	The status result of the test is set to the following: • OK—The test is successful. • NOTEST—There are not enough alignments for testing. • FAIL—When there is an ill-conditioned covariance matrix or other numerical exception that prevents testing. • HIDATA—There are too many fragments in locus. • LOWDATA—The tested region is either too complex or too shallowly sequenced to support a reliable calculation of abundance.
log2 (control group)	The log FPKM of the control group.
log2 (comparison group)	The log FPKM of the comparison group.
log(Ratio)	The log fold-change ratio from the comparison and control groups.
q Value	Multiple-testing adjusted p-value for differential expression. This value is used for filtering significant differential expression.
Significant	• True—q Value is less than the false discovery rate. • False—q Value is higher than or equal to the false discovery rate. The default false discovery rate is 0.05.

Cufflinks Global Analysis

The Cufflinks Assembly & DE App provides an aggregate summary for all the samples. To download the statistics, click **PDF Summary Report**.

Overview

Table 10 Overview Table

Statistic	Definition
Sample group X (e.g. 1, 2, 3...)	Links to RNA-Seq Alignment App results for group X.
FPKM tables	Links to gene and transcript FKPM tables.

Summary

Table 11 Summary Table

Statistic	Definition
Sample ID	The sample ID.
Read Length	The length of reads.
Number of Reads	The total number of reads passing filter for this sample.
% Total Aligned	The percentage of reads passing filter that aligned to the reference, including abundant reads.
% Abundant	The percentage of reads that align to abundant transcripts, such as mitochondrial and ribosomal sequences.
% Unaligned	The percentage of reads that do not align to the reference.
Median CV Coverage Uniformity	The median coefficient of variation of coverage of the 1000 most highly expressed transcript, as reported by the CollectRnaSeqMetrics utility from Picard tools. Ideal value = 0.
% Stranded	The percentage of reads that are stranded.

Assembly

The Assembly table is available if you selected the Novel Transcript Assembly option in the RNA-Seq Alignment App. The app provides assembly analysis for the aggregate sample groups. Some descriptions are taken from cole-trapnell-lab.github.io/cufflinks/cuffcompare/#transfrag-class-codes.

Table 12 Assembly Table

Statistic	Description
Gene Count	The number of identified genes per sample.
Transcript Count	The number of identified transcripts per sample.
Link to gene models	Link to the gene transfer format (GTF) files results.
Equal (=)	The complete match of intron chain.
Potential novel (j)	The potentially novel isoform (fragment); at least one splice junction is shared with a reference transcript.

Statistic	Description
Unknown, intergenic (u)	The unknown, intergenic transcript.
Overlap with opposite-strand exon (x)	The exonic overlap with reference on the opposite strand.
Other	Other types of transcripts.

Summary Plots

- To save plot as scalable vector graphics (SVG), click **Save Plot as SVG**.
- To export data from plot as comma-separated values (CSV), click **Export Data as CSV**.

Table 13 Summary Plots Table

Plot Name	Description
Sample Correlation	The Sample Correlation matrix shows the similarity of the samples, which are based on the correlation of expression levels. The app performs hierarchical clustering based on the correlation. The correlation levels are expressed in a range between +1 and -1. • 1—Total positive correlation. • 0—No correlation. • -1—Negative correlation. The correlation points represent different colors. • White: -1 • Green: -0.33 • Yellow: 0.33 • Red: 1 The color between two adjacent points are interpolated; greenish yellow represents the 0 point.
PCA	The Principal Component Analysis (PCA) 2-D scatter plot and 3-D plot show samples along the 2 and 3 components that capture the most variance. The app generates a 3-D plot when there are at least 3 principal components. The 3-D plot can be rotated and zoomed in and out. You can save the plot as portable network graphics (PNG).

Revision History

Document	Date	Description of Change
Document # 1000000006108 v00	February 2016	Supports Cufflinks Assembly & DE; parameter settings updates; analysis updates

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 14 Illumina General Contact Information

Website	www.illumina.com
Email	techsupport@illumina.com

Table 15 Illumina Customer Support Telephone Numbers

Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Japan	0800.111.5011
Australia	1.800.775.688	Netherlands	0800.0223859
Austria	0800.296575	New Zealand	0800.451.650
Belgium	0800.81102	Norway	800.16836
China	400.635.9898	Singapore	1.800.579.2745
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850	Switzerland	0800.563118
Germany	0800.180.8994	Taiwan	00806651752
Hong Kong	800960230	United Kingdom	0800.917.0041
Ireland	1.800.812949	Other countries	+44.1799.534000
Italy	800.874909		

Safety data sheets (SDSs)—Available on the Illumina website at support.illumina.com/sds.html.

Product documentation—Available for download in PDF from the Illumina website. Go to support.illumina.com, select a product, then select **Documentation & Literature**.

Illumina
5200 Illumina Way
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com