Using Illumina BaseSpace® Apps to Analyze RNA Sequencing Data

The Illumina TopHat Alignment and Cufflinks Assembly and Differential Expression apps make RNA data analysis accessible to any user, regardless of bioinformatics experience.

Introduction

The Illumina TopHat Alignment and Cufflinks Assembly and Differential Expression apps provide the most widely adopted suite of RNA data analysis tools in a simple click-and-go user interface. These apps make RNA data analysis accessible to any user—regardless of bioinformatics experience—and include the tools required for a range of common transcriptome data analysis needs. TopHat 2 provides high-confidence alignment for abundance measurement, and the detection of splice junctions, gene fusions, and cSNPs. Cuffdiff enables sensitive transcript discovery and differential expression analysis. TopHat Fusion delivers robust, high-confidence detection of gene fusions, while Isaac delivers reliable variant calling. These tools are packaged into a user interface designed to be accessible to an informatics novice. Simple-to-follow prompts guide users through the entire process, starting from selecting the files generated by the sequencer to filtering and analyzing your data.

This technical note provides a step-by-step run-through of a typical analysis workflow, from choosing your sequencing runs through the selection of analysis parameters and analysis of results. It highlights examples of the easily interpretable tables and graphs that the software generates, including descriptions of the differential expression, gene fusion detection, variant calling, and other reports that can be directly reported in a manuscript.

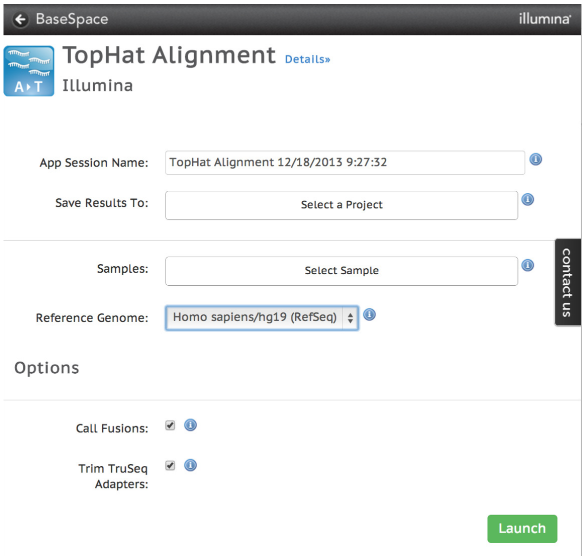
Use the TopHat App to Generate Alignments, Variant Calls, and Fusion Calls

1. Navigate to the demonstration data set.
2. Select **TopHat Alignment** from the app drop-down menu.
3. Select an app session and give it an appropriate name.
4. Click **Select a Project** and create a new project to hold the data from this analysis session.
5. Click **Select Sample** to select samples. Both nonstranded and stranded samples are supported.
6. Select a reference genome to use for alignment from the Illumina sets.
7. [Optional] Select the **Call Fusions** checkbox.

   Alternatively, depending on the study, you can call fusions with the TopHat Fusion app.
8. Click **Launch** to begin the alignment.

   The Session Info page opens, where you can track run progress.

Figure 1: TopHat Alignment Input Parameters



App Session Name: TopHat Alignment 12/18/2013 9:27:32

Save Results To: Select a Project

Samples: Select Sample

Reference Genome: Homo sapiens/hg19 (RefSeq)

Options

Call Fusions: ☑

Trim TruSeq Adapters: ☑

Session Info Page

The Session Info page includes the Inputs tab, which shows the parameters that were used in running this analysis (selected during the analysis configuration). After the analysis is complete, the Completed Application Result tab contains the report that was created.

TopHat Analysis Report

The report has a summary page, and individual detailed reports for each sample.

Summary Page

The report contains a summary page that shows several critical quality metrics across all the analyzed samples (Table 1), and the following graphs that allow you to compare the samples visually.

- **Insert Length Distribution** (paired-end reads only)–Shows distribution of insert lengths based on paired-end sequence information (Figure 2)
- **Alignment Distribution**–Gives distribution of reads across gene region functional annotation categories (Figure 3)
- **Transcript Coverage Distribution**–Provides visual representation of coverage uniformity across transcripts (Figure 4)

Figure 2: TopHat Insert Length Distribution Graph

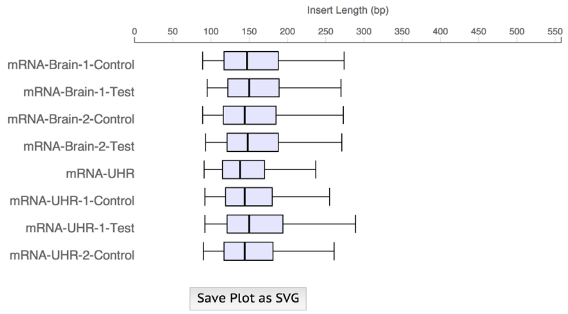


Figure 3: TopHat Alignment Distribution Graph

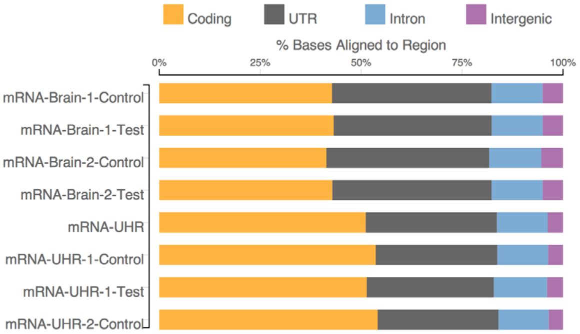
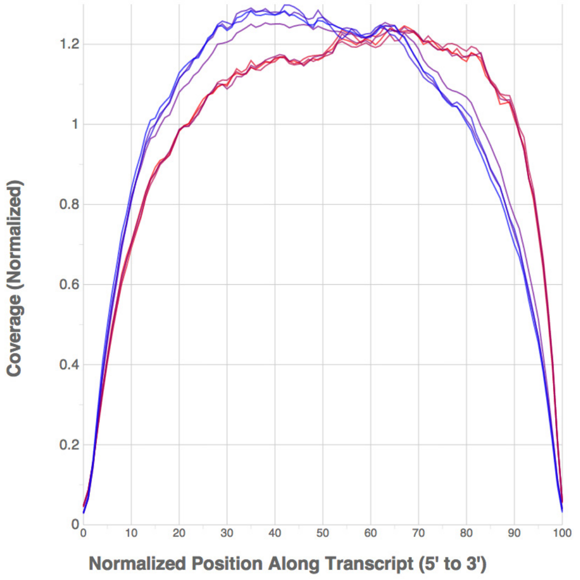


Figure 4: TopHat Transcript Coverage Graph



Individual Detailed Report

The Individual Detailed Report tab shows various metrics specific to individual samples within the run (Table 2). It also includes the following graphs that provide a visual representation of some of the data:

- **Insert Length Distribution Histogram**—For samples with paired reads, shows the full distribution of insert sizes determined from the alignments.
- **Alignment Distribution**—Captures the information present in the "Alignment Information" table, showing the total number of sequences aligned to the respective genomic regions.
- **Transcript Coverage Histogram**—Visualizes the data behind the statistics reported by in the "Coverage Uniformity" table. Specifically, these plots detail a normalized value of coverage along the length of highly expressed genes.

Lastly, the Downloadable Files tab provides links to files generated in the run that may be required for additional downstream analysis, such as for input into secondary analysis software on BaseSpace or into IGV for visualization. The files include:

- Alignments—Include mapping info for aligned reads
- Alignment Coverage—Provide positional fold coverage information
- Reference FPKM Values (Genes)—Provide FPKM-normalized abundance info at gene level
- Reference FPKM Values (Transcripts)—Provide FPKM-normalized abundance info at transcript level
- Genome VCF—Variant call format text file includes variant mapping information

Use the Cufflinks App to Perform Novel Transcript Assembly and Differential Expression

1. Navigate to the project that holds the TopHat analysis results and launch the Cufflinks Assembly and Differential Expression app.
2. Select the same reference genome as used during TopHat alignment and specify whether the samples are nonstranded or stranded.
3. Select the Novel Transcript Assembly checkbox.

   This option causes Cufflinks to detect novel transcripts from the aligned reads. The assembly takes advantage of knowledge about transcripts in the reference assembly.
4. Assign alphanumeric group labels for the Control and Comparison groups.
5. Select TopHat Alignment App Results for the Control and Comparison groups.
6. [Optional] Select the Adjust transcript assembly for samples without polyA selection checkbox, if desired.

   This option is useful for total RNA/riboZero samples, for which potential intronic alignments should be filtered out to reduce the assembly of spurious transcripts.
7. Review the Cufflinks Output Report.

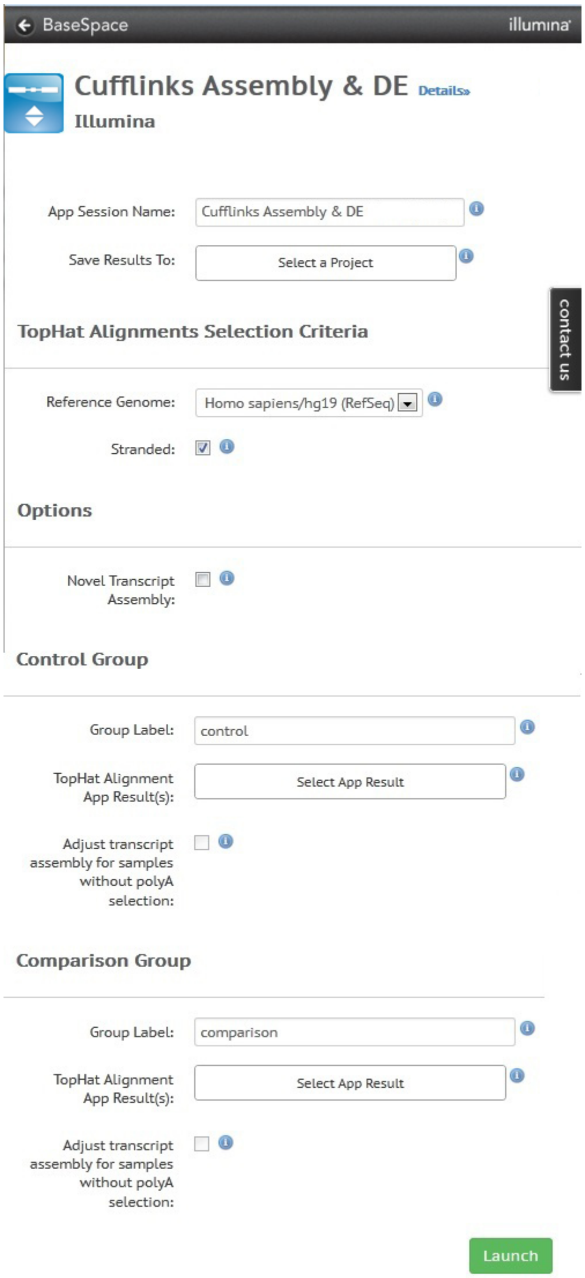
Table 1: Summary Table Data

Row	Description
Number of Reads	Total number of reads generated for a given sample
% of Total Aligned Reads	Portion of reads that aligned to the selected reference genome, including abundant sequences
% of Abundant Reads	Portion of reads coming from abundant RNA sequences such as ribosomal RNA
% of Unaligned Reads	Portion of reads that failed to align to the reference sequence
Median CV	Measure of the uniformity of coverage across transcripts
% Stranded	Portion of reads aligned to correct strand based on comparison with reference genome annotations

Table 2: Individual Detailed Report Table Data

Primary Analysis Information	
Read Length	Length of reads
Number of Reads	Total number of reads generated for a given sample
Bases	Total number of bases for a given sample
Q30 Bases	Total number of bases with a quality score of 30 or higher
Insert Information	
Insert Length Mean	Median length of the sequenced fragment
Insert Length SD	Standard deviation of the sequenced fragment length
Duplicates (% Reads)	Portion of paired reads that have duplicates
Alignment Quality	
Total Aligned Reads (% Reads)	Portion of reads that aligned to the selected reference genome
Abundant Reads (% Reads)	Portion of reads coming from abundant RNA sequences such as ribosomal and mitochondrial RNA
Unaligned Reads (% Reads)	Portion of reads that failed to align to the reference sequence
Reads with Spliced Alignment	Portion of reads overlapping splice sites
Reads Aligned at Multiple Loci	Portion of reads that aligned to the genome, but not uniquely
Coverage Uniformity	
Median CV	Measure of the uniformity of coverage across transcripts
Median 3’	Measure of coverage uniformity at 3’ end of transcript sequence
Median 5’	Measure of coverage uniformity at 5’ end of transcript sequence
Reads Aligned to Correct Strand	Portion of reads aligned to correct strand based on comparison with reference genome annotations
Alignment Information	
Fold coverage of and % of bases represented by each gene region annotation category	
Variant Calls	
Homozygous Reference	Number of reads in which no variants were detected relative to the reference sequence
Heterozygous	Number of reads in which one or more variants were detected
Homozygous Variant	Number of read pairs in which homozygous variants relative to reference sequence were detected
SNV	Total number of single nucleotide variants detected
Indel	Total number of insertion / deletion variants detected
Tn/Tv	Ratio of transition to transversion variants detected

Figure 1: CuffLinks Assembly and Differential Expression Input Parameters



Cufflinks Assembly & DE Details»

Illumina

App Session Name: Cufflinks Assembly & DE

Save Results To: Select a Project

TopHat Alignments Selection Criteria

Reference Genome: Homo sapiens/hg19 (RefSeq)

Stranded: ☑

Options

Novel Transcript Assembly: ☐

Control Group

Group Label: control

TopHat Alignment App Result(s): Select App Result

Adjust transcript assembly for samples without polyA selection: ☐

Comparison Group

Group Label: comparison

TopHat Alignment App Result(s): Select App Result

Adjust transcript assembly for samples without polyA selection: ☐

Launch

Cufflinks Output Report

The Cufflinks Output Report contains the following sections: Overview, Transcript Assembly Table, Differential Expression Table, Sample Correlation Plot, and Differential Expression Gene Browser.

Overview Section

The Overview section details the alignments used in the report, and the grouping and labels. For more detailed information on an individual alignment result, right-click the alignment link to bring up a previous alignment report. Normalized abundance (FPKM) values for reference transcripts are summarized across all samples for both genes and transcripts, in the corresponding links.

Transcript Assembly Table

The Transcript Assembly Table provides more detailed information on transcript and gene level abundance, including novel features identified and cases of differential expression. It provides detailed information on the 3 separate transcript assemblies: one for each sample group, plus a third that is used to conduct the differential expression analysis (Table 3).

Table 3: Transcript Assembly Table Data

Row	Description
Gene Count	Number of genes detected
Transcript Count	Number of transcripts detected
Gene Models	Links to gene models providing more detailed information on transcript assembly
Equal (-)	Number of assembled transcripts that exactly match a known transcript
Potentially novel (j)	Assembled transcripts where at least one splice junction is shared with a reference transcript
Unknown, intergenic (u)	Assembled transcripts with no relation to a known transcript
Overlap with opposite-strand exon (x)	Assembled transcripts overlapping the opposite strand of a known transcript
Other	All other annotation classes for assembled transcripts

Additional information about these annotation categories is available in the **Relation to reference transcripts** section of the Cufflinks online documentation at cufflinks.cbcb.umd.edu/manual.html#class_codes.

Differential Expression Table

The Differential Expression Table provides high-level information about the differential expression results (Table 4).

Table 4: Differential Expression Table Data

Row	Description
Gene Count	Total number of genes in annotation
Gene Count (changed)	Total number of differentially expressed genes detected
Transcript Count	Total number of transcripts in annotation
Transcript Count (changed)	Total number of differentially expressed transcripts detected
Cuffdiff results	Links to the Cuffdiff (differential expression results) file, which can be used as input into downstream secondary analysis programs

Sample Correlation Plot

The Sample Correlation Plot provides an overview of the similarity between different samples involved in the analysis. This information is captured both in a correlation heat plot, and a clustering dendrogram).

Differential Expression Gene Browser

The Differential Expression Gene Browser can be used to filter and plot the differential expression results dynamically (Table 5). The data can be sorted by clicking any column heading and can be saved as either an SVG graphic or CSV table.

Table 5: Differential Expression Gene Browser Filter Options

Filter	Description
Log Ratio Cutoff	Allows the differentially expressed gene table to be filtered based on the observed log ratio of differential expression.
Significance	Filters for results achieving statistical significance at a q-value < 0.05.
Status	Filters based on the reliability of the result. If the result passes all stringency filters, the status is returned as "OK". If the result is flagged due to one of several factors (such as insufficient read depth), a failure mode of NOTEST, LOWDATA, or FAIL is returned.

Additional information is available at http://cole-trapnell-lab.github.io/cufflinks.

Additional Secondary Analysis

As a complement to the BaseSpace Apps for RNA Analysis, the BaseSpace Informatics Suite also includes a growing community of software solutions for visualization, analysis, and sharing. The BaseSpace Apps for RNA Analysis provide output files that may be directly input into a broad range of available secondary analysis solutions. Because Illumina NGS technology is the most established and broadly adopted sequencing solution, users can harness the industry-leading Illumina NGS ecosystem, including the world's largest collection of commercial and open-source data analysis software tools. Further, cross-functional compatibility within the ecosystem enables researchers to integrate results from various applications.

Explore BaseSpace Apps

Use the TopHat Alignment and Cufflinks Assembly and Differential Expression apps to analyze your RNA sequencing data at www.illumina.com/basespacerna.

Learn More

For a comprehensive description of the software features, reports, and output files, review the user guide or online help. You can also contact us at support@illumina.com for more information.

AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCA
AATCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACC
AACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTA
TTAACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGA
AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCA
AAGATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGA
AACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAC

AATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTA
AACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAA
ACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGT
CGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACA
AATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTA
TTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACGAA
TATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTAAC

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

Pub. No. 470-2013-005 Current as of 07 April 2016