RNA-Seq Analysis with TopHat and StringTie

The workflow sample, described below, takes FASTQ files with paired-end RNA-Seq reads and process them as follows:

- Improve reads quality with Trimmomatic
- Provide FastQC quality reports
- Map improved reads to a reference sequence with TopHat
- Assemble transcripts and generate gene abundance output with StringTie
- Produce a common gene abundance report (one for several input samples)

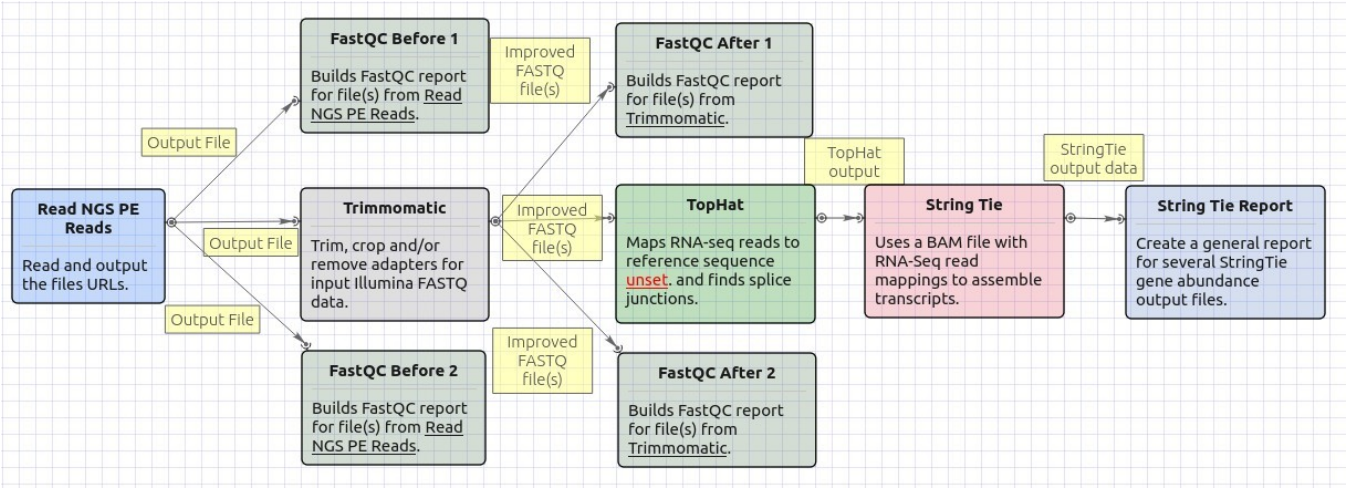
> **How to Use This Sample**
>
> If you haven't used the workflow samples in UGENE before, look at the "How to Use Sample Workflows" section of the documentation.

Workflow Sample Location

The workflow sample "RNA-Seq Analysis with TopHat and StringTie" can be found in the "NGS" section of the Workflow Designer samples.

Workflow Image

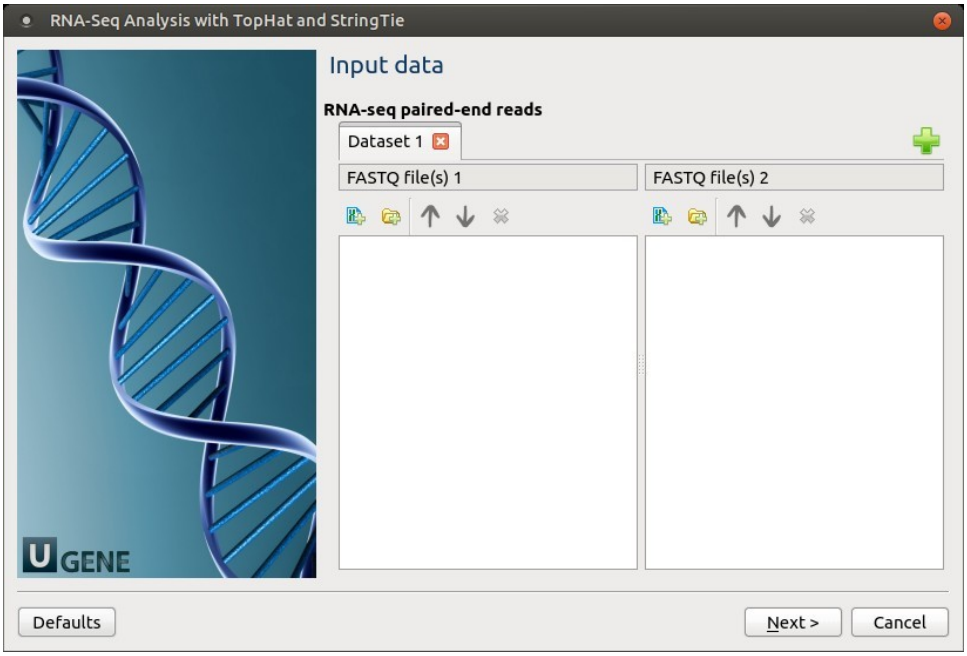
The opened workflow looks as follows:



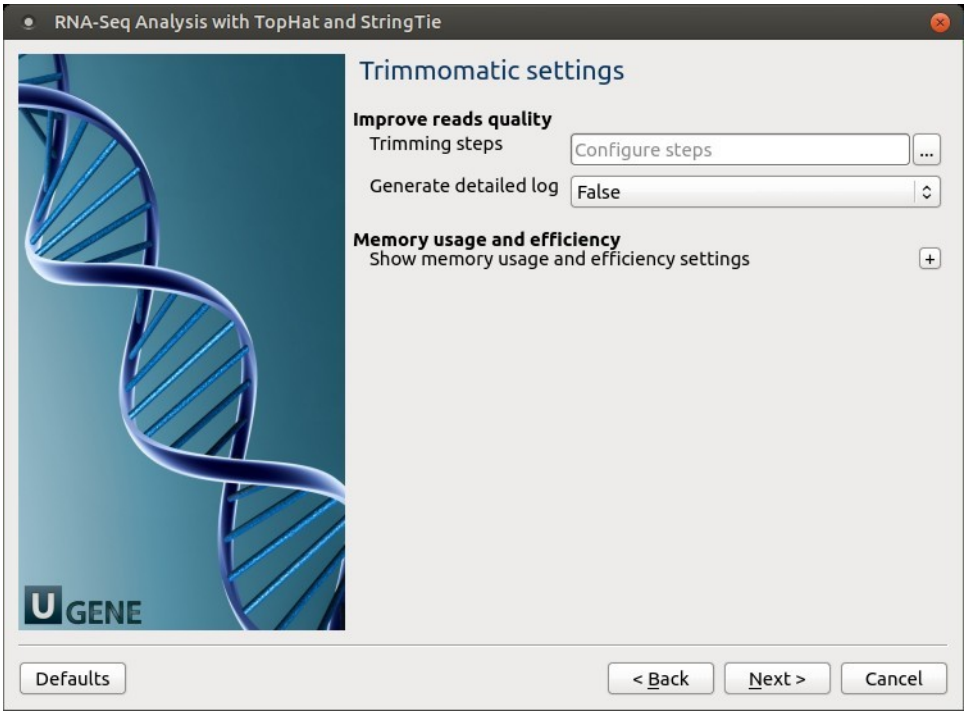
Workflow Wizard

The wizard has 5 pages.

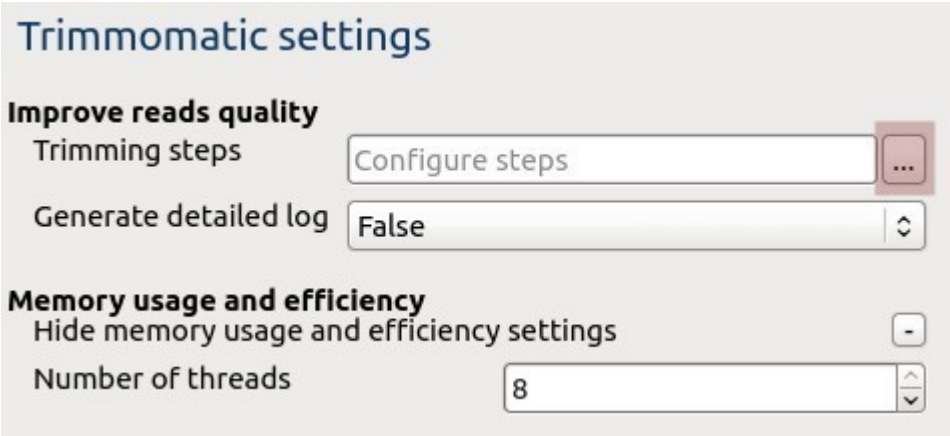
1. Input data: RNA-seq paired-end reads: On this page, files with RNA-seq paired-end reads must be set.



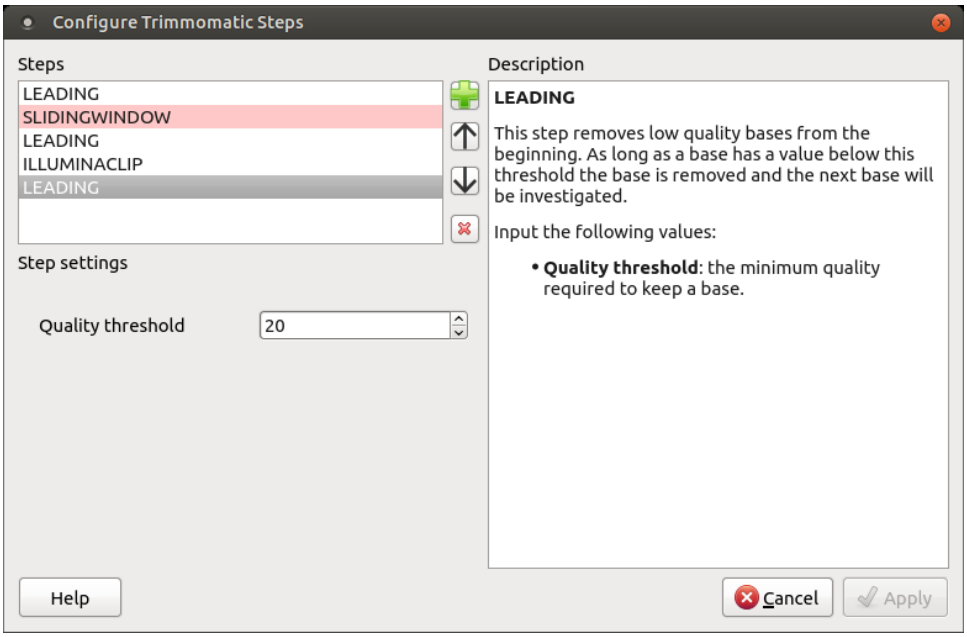
2. Trimmomatic settings: The Trimmomatic parameters can be changed here.



To configure trimming steps use the following button:



The following dialog will appear:



Click the *Add new* step button and select a step. The following options are available:

- ILLUMINACLIP: Cutadapterandotherillumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality.
- TRAILING: Cut bases off the end of a read, if below a threshold quality.
- CROP: Cut the read to a specified length.
- HEADCROP: Cut the specified number of bases from the start of the read.
- MINLEN: Drop the read if it is below a specified length.
- AVGQUAL: Drop the read if the average quality is below the specified level.
- TOPHRED33: Convert quality scores to Phred-33.
- TOPHRED64: Convert quality scores to Phred-64.

Each step has the own parameters:

AVGQUAL

This step drops a read if the average quality is below the specified level.

Input the following values:

- Quality threshold: the minimum average quality required to keep a read.

CROP

This step removes bases regardless of quality from the end of thread, so that the readhas maximally the specified length after this step has been performed. Steps performed after CROP might of course further shorten the read.

Input the following values:

- Length: the number of bases to keep, from the start of the read.

HEADCROP

This step removes the specified number of bases, regardless of quality, from the beginning of the read.

Input the following values:

- Length: the number of bases to remove from the start of the read.

ILLUMINACLIP

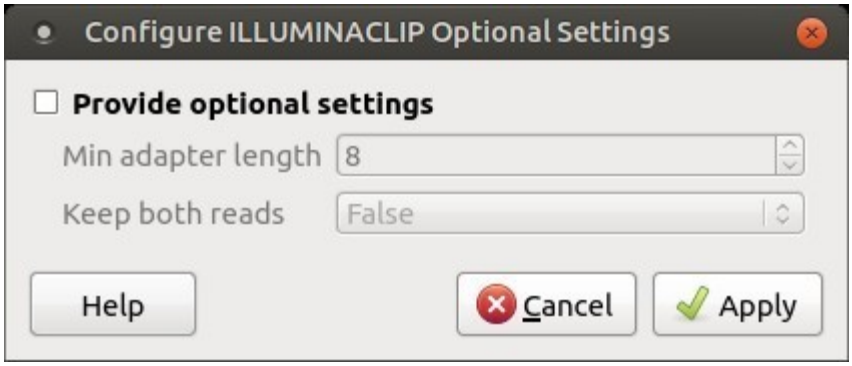
This step is used to find and remove Illumina adapters.

Trimmomatic first compares short sections of an adapter and a read. If they match enough, the entire alignment between the read and adapter is scored. For paired-end reads, the "palindrome" approach is also used to improve the result. See Trimmomatic manual for details.

Input the following values:

- Adapter sequences: a FASTA file with the adapter sequences. Files for TruSeq2 (GAII machines), TruSeq3 (HiSeq and MiSeq machines) and Nextera kits for SE and PE reads are now available by default. The naming of the various sequences within the specified file determines how they are used.
- Seed mismatches: the maximum mismatch count in short sections which will still allow a full match to be performed.
- Simple clip threshold: a threshold for simple alignment mode. Values between 7 and 15 are recommended. A perfect match of a 12 base sequence will score just over 7, while 25 bases are needed to score 15.
- Palindrome clip threshold: a threshold for palindrome alignment mode. For palindromic matches, a longer alignment is possible. Therefore the threshold can be in the range of 30. Even though this threshold is very high (requiring a match of almost 50 bases) Trimmomatic is still able to identify very, very short adapter fragments.

There are also two optional parameters for palindrome mode: Min adapter length and Keep both reads. Use the following dialog. To call the dialog press the *Optional* button.



LEADING

This step removes low-quality bases from the beginning. As long as a base has a value below this threshold the base is removed and the next base will be investigated.

Input the following values:

- Quality threshold: the minimum quality required to keep a base.

MAXINFO

This step performs an adaptive quality trim, balancing the benefits of retaining longer reads against the costs of retaining bases with errors. See Trimmomatic manual for details.

Input the following values:

- Target length: the read length which is likely to allow the location of the read within the target sequence. Extremely short reads, which can be placed into many different locations, provide little value. Typically, the length would be in the order of 40 bases, however, the value also depends on the size and complexity of the target sequence.
- Strictness: the balance between preserving as much read length as possible vs. removal of incorrect bases. A low value of this parameter (0.8) favours read correctness.

MINLEN

This step removes reads that fall below the specified minimum length. If required, it should normally be after all other processing steps. Reads removed by this step will be counted and included in the "dropped reads" count.

Input the following values:

- Length: the minimum length of reads to be kept.

SLIDINGWINDOW

This step performs a sliding window trimming, cutting once the average quality within the window falls below a threshold. By considering multiple bases, a single poor quality base will not cause the removal of high-quality data later in the read.

Input the following values:

- Window size: the number of bases to an average across.
- Quality threshold: the average quality required.

TOPHRED33

This step (re)encodes the quality part of the FASTQ file to base 33.

TOPHRED64

This step (re)encodes the quality part of the FASTQ file to base 64.

TRAILING

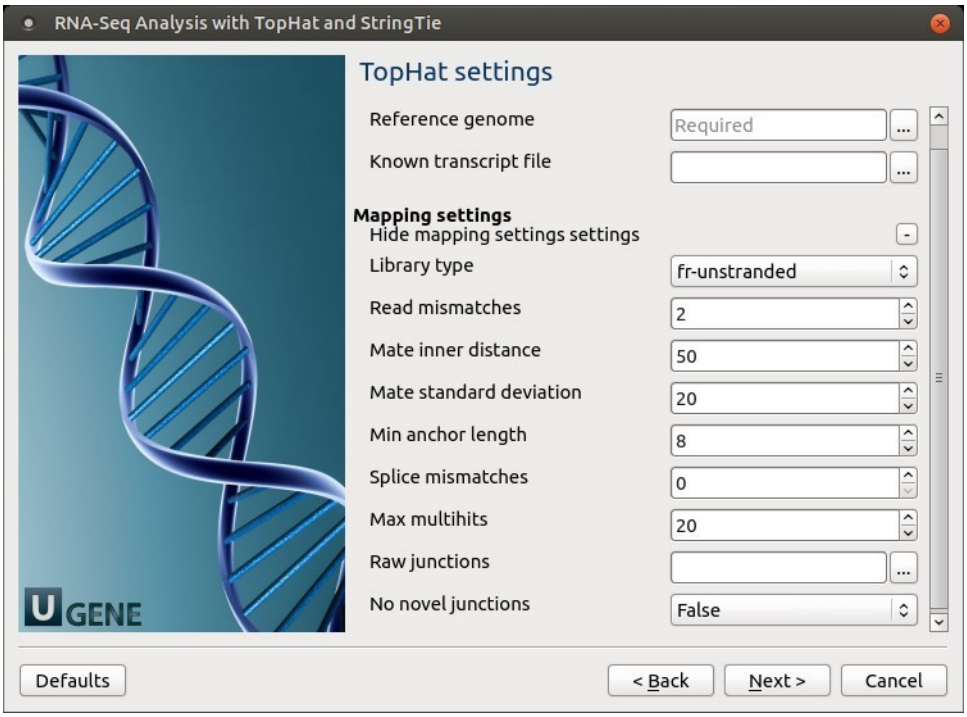
This step removes low-quality bases from the end. As long as a base has a value below this threshold the base is removed and the next base (i. e. the preceding one) will be investigated. This approach can be used removing the special Illumina " low-quality segment" regions (which are marked with a quality score of 2), but SLIDINGWINDOW or MAXINFO are recommended instead.

Input the following values:

- Quality threshold: the minimum quality required to keep a base.

To remove a step use the *Remove selected step* button. The pink highlighting means the required parameter has not been set.

3. TopHat settings: TopHat parameters can be set here.

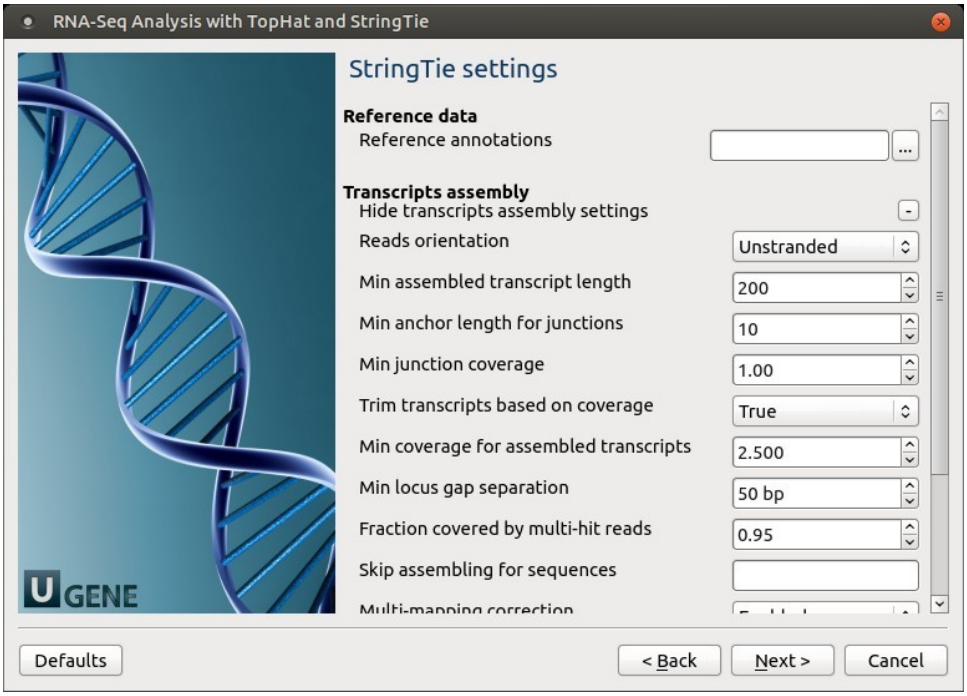


The following parameters are available:

Reference genome	Path to the indexed reference genome.
Known transcript file	A set of gene model annotations and/or known transcripts.
Library type	Specifies RNA-Seq protocol.

Read mismatches	Final read alignments having more than these many mismatches are discarded.
Mate inner distance	The expected (mean) inner distance between mate pairs.
Mate standard deviation	The standard deviation for the distribution on inner distances between mate pairs.
Min anchor length	The anchor length. TopHat will report junctions spanned by reads with at least this many bases on each side of the junction. Note that individual spliced alignments may span a junction with fewer than this many bases on one side. However, every junction involved in spliced alignments is supported by at least one read with this many bases on each side.
Splice mismatches	The maximum number of mismatches that may appear in the anchor region of a spliced alignment.
Max multihits	Instruct TopHat to allow up to this many alignments to the reference for a given read and suppresses all alignments for reads with more than this many alignments.
Raw junctions	The list of raw junctions.
No novel junctions	Only look for reads across junctions indicated in the supplied GFF or junctions file. This parameter is ignored if Raw junctions or Known transcript file is not set.

4. StringTie settings: StringTie parameters can be set here.

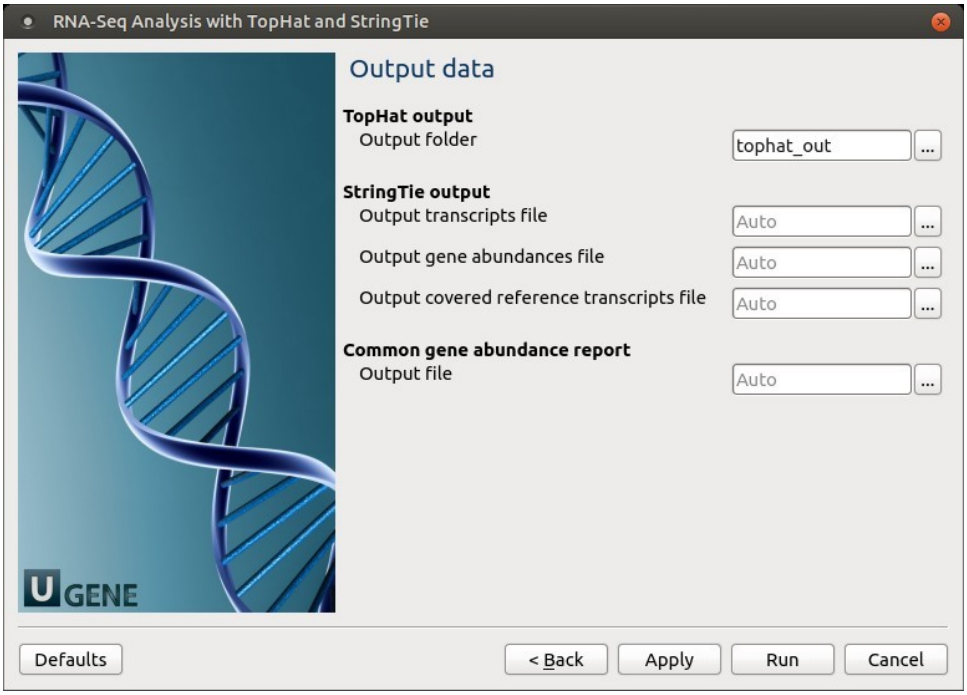


The following parameters are available:

Reference annotations	Use the reference annotation file (in GTF or GFF3 format) to guide the assembly process (-G). The output will include expressed reference transcripts as well as any novel transcripts that are assembled.
Reads orientation	Select the NGS libraries type: unstranded, stranded fr-secondstrand (--fr), or stranded fr-firststand (--rf).
Min	Specify the minimum length for the predicted transcripts (-m).

assembled transcript length	
Min anchor length for junctions	Junctions that don't have spliced reads that align them with at least this amount of bases on both sides is filtered out (-a).
Min junction coverage	There should be at least this many spliced reads that align across a junction (-j). This number can be fractional since some reads align in more than one place. A read that aligns in n places will contribute 1/n to the junction coverage.
Trim transcripts based on coverage	By default StringTie adjusts the predicted transcript's start and/or stop coordinates based on sudden drops in coverage of the assembled transcript. Set this parameter to "False" to disable the trimming at the ends of the assembled transcripts (-t).
Min coverage for assembled transcripts	Specifies the minimum read coverage allowed for the predicted transcripts (-c). A transcript with a lower coverage than this value is not shown in the output. This number can be fractional since some reads align in more than one place. A read that aligns in n places will contribute 1/n to the coverage.
Min locus gap separation	Reads that are mapped closer than this distance are merged together in the same processing bundle (-g).
Fraction covered by multi-hit reads	Specify the maximum fraction of muliple-location-mapped reads that are allowed to be present at a given locus (-M). A read that aligns in n places will contribute 1/n to the coverage.
Skip assembling for sequences	Ignore all read alignments (and thus do not attempt to perform transcript assembly) on the specified reference sequences (-x). The value can be a single reference sequence name (e.g. "chrM") or a comma-delimited list of sequence names (e.g. "chrM,chrX,chrY"). This can speed up StringTie especially in the case of excluding the mitochondrial genome, whose genes may have very high coverage in some cases, even though they may be of no interest for a particular RNA-Seq analysis. The reference sequence names are case sensitive, they must match identically the names of chromosomes/contigs of the target genome against which the RNA-Seq reads were aligned in the first place.
Multi-mapping correction	Enables or disables (-u) multi-mapping correction.
Verbose log	Enable detailed logging, if required (-v). The messages will be written to the UGENE log (enabling of "DETAILS" and "TRACE" logging may be required) and to the dashboard.
Label	Use the specified string as the prefix for the name of the output transcripts (-l).

5. Output Files Page: On this page, output directories can be selected:


RNA-Seq Analysis with TopHat and StringTie
Output data
TopHat output
Output folder
tophat_out
StringTie output
Output transcripts file
Auto
Output gene abundances file
Auto
Output covered reference transcripts file
Auto
Common gene abundance report
Output file
Auto
Defaults
< Back
Apply
Run
Cancel