RSEM Documentation

Includes Name, Synopsis, Arguments, Basic Options, Advanced Options, Description of RSEM, Outputs, and Examples.
Source: http://deweylab.biostat.wisc.edu/rsem/rsem-calculate-expression.html

---

NAME: rsem-calculate-expression

SYNOPSIS:

rsem-calculate-expression [options] upstream_read_file(s) reference_name sample_name

rsem-calculate-expression [options] --paired-end upstream_read_file(s) downstream_read_file(s) reference_name sample_name

rsem-calculate-expression [options] --sam/--bam [--paired-end] input reference_name sample_name

ARGUMENTS:

upstream_read_files(s)
Comma-separated list of files containing single-end reads or upstream reads for paired-end data. By default, these files are assumed to be in FASTQ format. If the --no-qualities option is specified, then FASTA format is expected.

downstream_read_file(s)
Comma-separated list of files containing downstream reads which are paired with the upstream reads. By default, these files are assumed to be in FASTQ format. If the --no-qualities option is specified, then FASTA format is expected.

input
SAM/BAM formatted input file. If "-" is specified for the filename, SAM/BAM input is instead assumed to come from standard input. RSEM requires all alignments of the same read group together. For paired-end reads, RSEM also requires the two mates of any alignment be adjacent. See Description section for how to make input file obey RSEM's requirements.

reference_name
The name of the reference used. The user must have run 'rsem-prepare-reference' with this reference_name before running this program.

sample_name
The name of the sample analyzed. All output files are prefixed by this name (e.g., sample_name.genes.results)

BASIC OPTIONS

--paired-end
Input reads are paired-end reads. (Default: off)

--no-qualities
Input reads do not contain quality scores. (Default: off)

--strand-specific
The RNA-Seq protocol used to generate the reads is strand specific, i.e., all (upstream) reads are derived from the forward strand. This option is equivalent to --forward-prob=1.0. With this option set, if RSEM runs the Bowtie/Bowtie 2 aligner, the '--norc' Bowtie/Bowtie 2 option will be used, which disables alignment to the reverse strand of transcripts. (Default: off)

--bowtie2
Use Bowtie 2 instead of Bowtie to align reads. Since currently RSEM does not handle indel, local and discordant alignments, the Bowtie2 parameters are set in a way to avoid those alignments. In particular, we use options '--sensitive --dpad 0 --gbar 99999999 --mp 1,1 --np 1 --score-min L,0,-0.1' by default. The last parameter of '--score-min', '-0.1', is the negative of maximum mismatch rate. This rate can be set by option '--bowtie2-mismatch-rate'. If reads are paired-end, we additionally use options '--no-mixed' and '--no-discordant'. (Default: off)

--star
Use STAR to align reads. Alignment parameters are from ENCODE3's STAR-RSEM pipeline. To save computational time and memory resources, STAR's Output BAM file is unsorted. It is stored in RSEM's temporary directory with name as 'sample_name.bam'. Each STAR job will have its own private copy of the genome in memory. (Default: off)

--star-path <path>
The path to STAR's executable. (Default: the path to STAR executable is assumed to be in user's PATH environment variable)

--sam
Input file is in SAM format. (Default: off)

--bam
Input file is in BAM format. (Default: off)

-p/--num-threads <int>
Number of threads to use. Both Bowtie/Bowtie2, expression estimation and 'samtools sort' will use this many threads. (Default: 1)

--no-bam-output
Do not output any BAM file. (Default: off)

--output-genome-bam
Generate a BAM file, 'sample_name.genome.bam', with alignments mapped to genomic coordinates and annotated with their posterior probabilities. In addition, RSEM will call samtools (included in RSEM package) to sort and index the bam file. 'sample_name.genome.sorted.bam' and 'sample_name.genome.sorted.bam.bai' will be generated. (Default: off)

--sampling-for-bam
When RSEM generates a BAM file, instead of outputing all alignments a read has with their posterior probabilities, one alignment is sampled according to the posterior probabilities. The sampling procedure includes the alignment to the "noise" transcript, which does not appear in the BAM file. Only the sampled alignment has a weight of 1. All other alignments have weight 0. If the "noise" transcript is sampled, all alignments appeared in the BAM file should have weight 0. (Default: off)

--seed <uint32>
Set the seed for the random number generators used in calculating posterior mean estimates and credibility intervals. The seed must be a non-negative 32 bit interger. (Default: off)

--calc-pme
Run RSEM's collapsed Gibbs sampler to calculate posterior mean estimates. (Default: off)

--calc-ci
Calculate 95% credibility intervals and posterior mean estimates. The credibility level can be changed by setting '--ci-credibility-level'. (Default: off)

-q/--quiet
Suppress the output of logging information. (Default: off)

-h/--help
Show help information.

--version
Show version information.

ADVANCED OPTIONS

--sam-header-info <file>
RSEM reads header information from input by default. If this option is on, header information is read from the specified file. For the format of the file, please see SAM official website. (Default: "")

--seed-length <int>

Seed length used by the read aligner. Providing the correct value is important for RSEM. If RSEM runs Bowtie, it uses this value for Bowtie's seed length parameter. Any read with its or at least one of its mates' (for paired-end reads) length less than this value will be ignored. If the references are not added poly(A) tails, the minimum allowed value is 5, otherwise, the minimum allowed value is 25. Note that this script will only check if the value >= 5 and give a warning message if the value < 25 but >= 5. (Default: 25)

--tag <string>
The name of the optional field used in the SAM input for identifying a read with too many valid alignments. The field should have the format <tagName>:i:<value>, where a <value> bigger than 0 indicates a read with too many alignments. (Default: "")

--bowtie-path <path>
The path to the Bowtie executables. (Default: the path to the Bowtie executables is assumed to be in the user's PATH environment variable)

--bowtie-n <int>
(Bowtie parameter) max # of mismatches in the seed. (Range: 0-3, Default: 2)

--bowtie-e <int>
(Bowtie parameter) max sum of mismatch quality scores across the alignment. (Default: 99999999)

--bowtie-m <int>
(Bowtie parameter) suppress all alignments for a read if > <int> valid alignments exist. (Default: 200)

--bowtie-chunkmbs <int>
(Bowtie parameter) memory allocated for best first alignment calculation (Default: 0 - use Bowtie's default)

--phred33-quals
Input quality scores are encoded as Phred+33. (Default: on)

--phred64-quals
Input quality scores are encoded as Phred+64 (default for GA Pipeline ver. >= 1.3). (Default: off)

--solexa-quals
Input quality scores are solexa encoded (from GA Pipeline ver. < 1.3). (Default: off)

--bowtie2-path <path>
(Bowtie 2 parameter) The path to the Bowtie 2 executables. (Default: the path to the Bowtie 2 executables is assumed to be in the user's PATH environment variable)

--bowtie2-mismatch-rate <double>
(Bowtie 2 parameter) The maximum mismatch rate allowed. (Default: 0.1)

--bowtie2-k <int>
(Bowtie 2 parameter) Find up to <int> alignments per read. (Default: 200)

--bowtie2-sensitivity-level <string>
(Bowtie 2 parameter) Set Bowtie 2's preset options in --end-to-end mode. This option controls how hard Bowtie 2 tries to find alignments. <string> must be one of "very_fast", "fast", "sensitive" and "very_sensitive". The four candidates correspond to Bowtie 2's "--very-fast", "--fast", "--sensitive" and "--very-sensitive" options. (Default: "sensitive" - use Bowtie 2's default)

--gzipped-read-file
Input read file(s) is compressed by gzip. This option can be only used when aligning reads by STAR, i.e. --star-genome-path <path> is defined (Default: off)

--bzipped-read-file
Input read file(s) is compressed by bzip2. This option can be only used when aligning reads by STAR, i.e. --star-genome-path <path> is defined (Default: off)

--output-star-genome-bam
Save the BAM file from STAR alignment under genomic coordinate to 'sample_name.STAR.genome.bam'. This file is NOT sorted by genomic coordinate. In this file, according to STAR's manual, 'paired ends of an alignment are always adjacent, and multiple alignments of a read are adjacent as well'. (Default: off)

--sort-bam-by-read-name
Sort BAM file aligned under transcript coordidate by read name. Setting this option on will produce determinstic maximum likelihood estimations from independet runs. Note that sorting will take long time and lots of memory. (Default: off)

--sort-bam-buffer-size <string>
Size for main memeory buffer when sorting BAM file. It can be any string acceptable to GNU sort's '-S' option. See "sort --help" for details. (Default: '60G')

--forward-prob <double>
Probability of generating a read from the forward strand of a transcript. Set to 1 for a strand-specific protocol where all (upstream) reads are derived from the forward strand, 0 for a strand-specific protocol where all (upstream) read are derived from the reverse strand, or 0.5 for a non-strand-specific protocol. (Default: 0.5)

--fragment-length-min <int>
Minimum read/insert length allowed. This is also the value for the Bowtie/Bowtie2 -I option. (Default: 1)

--fragment-length-max <int>
Maximum read/insert length allowed. This is also the value for the Bowtie/Bowtie 2 -X option. (Default: 1000)

--fragment-length-mean <double>
(single-end data only) The mean of the fragment length distribution, which is assumed to be a Gaussian. (Default: -1, which disables use of the fragment length distribution)

--fragment-length-sd <double>
(single-end data only) The standard deviation of the fragment length distribution, which is assumed to be a Gaussian. (Default: 0, which assumes that all fragments are of the same length, given by the rounded value of --fragment-length-mean)

--estimate-rspd
Set this option if you want to estimate the read start position distribution (RSPD) from data. Otherwise, RSEM will use a uniform RSPD. (Default: off)

--num-rspd-bins <int>
Number of bins in the RSPD. Only relevant when '--estimate-rspd' is specified. Use of the default setting is recommended. (Default: 20)

--gibbs-burnin <int>
The number of burn-in rounds for RSEM's Gibbs sampler. Each round passes over the entire data set once. If RSEM can use multiple threads, multiple Gibbs samplers will start at the same time and all samplers share the same burn-in number. (Default: 200)

--gibbs-number-of-samples <int>
The total number of count vectors RSEM will collect from its Gibbs samplers. (Default: 1000)

--gibbs-sampling-gap <int>
The number of rounds between two succinct count vectors RSEM collects. If the count vector after round N is collected, the count vector after round N + <int> will also be collected. (Default: 1)

--ci-credibility-level <double>
The credibility level for credibility intervals. (Default: 0.95)

--ci-memory <int>
Maximum size (in memory, MB) of the auxiliary buffer used for computing credibility intervals (CI). Set it larger for a faster CI calculation. However, leaving 2 GB memory free for other usage is recommended. (Default: 1024)

--ci-number-of-samples-per-count-vector <int>
The number of read generating probability vectors sampled per sampled count vector. The crebility intervals are calculated by first sampling P(C | D) and then sampling P(Theta | C) for each sampled count vector. This option controls how many Theta vectors are sampled per sampled count vector. (Default: 50)

--samtools-sort-mem <string>
Set the maximum memory per thread that can be used by 'samtools sort'. <string> represents the memory and accepts suffices 'K/M/G'. RSEM will pass <string> to the '-m' option of 'samtools sort'. Please note that the default used here is different from the default used by samtools. (Default: 1G)

--keep-intermediate-files
Keep temporary files generated by RSEM. RSEM creates a temporary directory, 'sample_name.temp', into which it puts all intermediate output files. If this directory already exists, RSEM overwrites all files generated by previous RSEM runs inside of it. By default, after RSEM finishes, the temporary directory is deleted. Set this option to prevent the deletion of this directory and the intermediate files inside of it. (Default: off)

--temporary-folder <string>
Set where to put the temporary files generated by RSEM. If the folder specified does not exist, RSEM will try to create it. (Default: sample_name.temp)

--time
Output time consumed by each step of RSEM to 'sample_name.time'. (Default: off)

DESCRIPTION
In its default mode, this program aligns input reads against a reference transcriptome with Bowtie and calculates expression values using the alignments. RSEM assumes the data are single-end reads with quality scores, unless the '--paired-end' or '--no-qualities' options are specified. Alternatively, users can use STAR to align reads using the '--star' option. RSEM has provided options in 'rsem-prepare-reference' to prepare STAR's genome indices. Users may use an alternative aligner by specifying one of the --sam and --bam options, and providing an alignment file in the specified format. However, users should make sure that they align against the indices generated by 'rsem-prepare-reference' and the alignment file satisfies the requirements mentioned in ARGUMENTS section.

One simple way to make the alignment file satisfying RSEM's requirements (assuming the aligner used put mates in a paired-end read adjacent) is to use 'convert-sam-for-rsem' script. This script only accept SAM format files as input. If a BAM format file is obtained, please use samtools to convert it to a SAM file first. For example, if '/ref/mouse_125' is the 'reference_name' and the SAM file is named 'input.sam', you can run the following command:

```
  convert-sam-for-rsem /ref/mouse_125 input.sam -o input_for_rsem.sam
```

For details, please refer to 'convert-sam-for-rsem's documentation page.

The SAM/BAM format RSEM uses is v1.4. However, it is compatible with old SAM/BAM format. However, RSEM cannot recognize 0x100 in the FLAG field. In addition, RSEM requires SEQ and QUAL are not '*'.

The user must run 'rsem-prepare-reference' with the appropriate reference before using this program.

For single-end data, it is strongly recommended that the user provide the fragment length distribution parameters (--fragment-length-mean and --fragment-length-sd). For paired-end data, RSEM will automatically learn a fragment length distribution from the data.

Please note that some of the default values for the Bowtie parameters are not the same as those defined for Bowtie itself.

The temporary directory and all intermediate files will be removed when RSEM finishes unless '--keep-intermediate-files' is specified.

With the '--calc-pme' option, posterior mean estimates will be calculated in addition to maximum likelihood estimates.

With the '--calc-ci' option, 95% credibility intervals and posterior mean estimates will be calculated in addition to maximum likelihood estimates.

OUTPUTS

sample_name.isoforms.results

File containing isoform level expression estimates. The first line contains column names separated by the tab character. The format of each line in the rest of this file is:

- transcript_id
- gene_id
- length
- effective_length
- expected_count
- TPM FPKM IsoPct [posterior_mean_count, posterior_standard_deviation_of_count, pme_TPM pme_FPKM IsoPct_from_pme_TPM, TPM_ci_lower_bound, TPM_ci_upper_bound, FPKM_ci_lower_bound, FPKM_ci_upper_bound]

Fields are separated by the tab character. Fields within "[]" are optional. They will not be presented if neither '--calc-pme' nor '--calc-ci' is set.

'transcript_id' is the transcript name of this transcript. 'gene_id' is the gene name of the gene which this transcript belongs to (denote this gene as its parent gene). If no gene information is provided, 'gene_id' and 'transcript_id' are the same.

'length' is this transcript's sequence length (poly(A) tail is not counted). 'effective_length' counts only the positions that can generate a valid fragment. If no poly(A) tail is added, 'effective_length' is equal to transcript length - mean fragment length + 1. If one transcript's effective length is less than 1, this transcript's both effective length and abundance estimates are set to 0.

'expected_count' is the sum of the posterior probability of each read comes from this transcript over all reads. Because 1) each read aligning to this transcript has a probability of being generated from background noise; 2) RSEM may filter some alignable low quality reads, the sum of expected counts for all transcript are generally less than the total number of reads aligned.

'TPM' stands for Transcripts Per Million. It is a relative measure of transcript abundance. The sum of all transcripts' TPM is 1 million. 'FPKM' stands for Fragments Per Kilobase of transcript per Million mapped reads. It is another relative measure of transcript abundance. If we define l_bar be the mean transcript length in a sample, which can be calculated as

l_bar = \sum_i TPM_i / 10^6 * effective_length_i (i goes through every transcript),

the following equation is hold:

FPKM_i = 10^3 / l_bar * TPM_i.

We can see that the sum of FPKM is not a constant across samples.

'IsoPct' stands for isoform percentage. It is the percentage of this transcript's abandunce over its parent gene's abandunce. If its parent gene has only one isoform or the gene information is not provided, this field will be set to 100.

'posterior_mean_count', 'pme_TPM', 'pme_FPKM' are posterior mean estimates calculated by RSEM's Gibbs sampler. 'posterior_standard_deviation_of_count' is the posterior standard deviation of counts. 'IsoPct_from_pme_TPM' is the isoform percentage calculated from 'pme_TPM' values.

'TPM_ci_lower_bound', 'TPM_ci_upper_bound', 'FPKM_ci_lower_bound' and 'FPKM_ci_upper_bound' are lower(l) and upper(u) bounds of 95% credibility intervals for TPM and FPKM values. The bounds are inclusive (i.e. [l, u]).

sample_name.genes.results
File containing gene level expression estimates. The first line contains column names separated by the tab character. The format of each line in the rest of this file is:

gene_id transcript_id(s) length effective_length expected_count TPM FPKM [posterior_mean_count posterior_standard_deviation_of_count pme_TPM pme_FPKM TPM_ci_lower_bound TPM_ci_upper_bound FPKM_ci_lower_bound FPKM_ci_upper_bound]

Fields are separated by the tab character. Fields within "[]" are optional. They will not be presented if neither '--calc-pme' nor '--calc-ci' is set.

'transcript_id(s)' is a comma-separated list of transcript_ids belonging to this gene. If no gene information is provided, 'gene_id' and 'transcript_id(s)' are identical (the 'transcript_id').

A gene's 'length' and 'effective_length' are defined as the weighted average of its transcripts' lengths and effective lengths (weighted by 'IsoPct'). A gene's abundance estimates are just the sum of its transcripts' abundance estimates.

sample_name.alleles.results
Only generated when the RSEM references are built with allele-specific transcripts.

This file contains allele level expression estimates for allele-specific expression calculation. The first line contains column names separated by the tab character. The format of each line in the rest of this file is:

allele_id transcript_id gene_id length effective_length expected_count TPM FPKM AlleleIsoPct AlleleGenePct [posterior_mean_count posterior_standard_deviation_of_count pme_TPM pme_FPKM AlleleIsoPct_from_pme_TPM AlleleGenePct_from_pme_TPM TPM_ci_lower_bound TPM_ci_upper_bound FPKM_ci_lower_bound FPKM_ci_upper_bound]

Fields are separated by the tab character. Fields within "[]" are optional. They will not be presented if neither '--calc-pme' nor '--calc-ci' is set.

'allele_id' is the allele-specific name of this allele-specific transcript.

'AlleleIsoPct' stands for allele-specific percentage on isoform level. It is the percentage of this allele-specific transcript's abundance over its parent transcript's abundance. If its parent transcript has only one allele variant form, this field will be set to 100.

'AlleleGenePct' stands for allele-specific percentage on gene level. It is the percentage of this allele-specific transcript's abundance over its parent gene's abundance.

'AlleleIsoPct_from_pme_TPM' and 'AlleleGenePct_from_pme_TPM' have similar meanings. They are calculated based on posterior mean estimates.

Please note that if this file is present, the fields 'length' and 'effective_length' in 'sample_name.isoforms.results' should be interpreted similarly as the corresponding definitions in 'sample_name.genes.results'.

sample_name.transcript.bam, sample_name.transcript.sorted.bam and sample_name.transcript.sorted.bam.bai
Only generated when --no-bam-output is not specified.

'sample_name.transcript.bam' is a BAM-formatted file of read alignments in transcript coordinates. The MAPQ field of each alignment is set to min(100, floor(-10 * log10(1.0 - w) + 0.5)), where w is the posterior probability of that alignment being the true mapping of a read. In addition, RSEM pads a new tag ZW:f:value, where value is a single precision floating number representing the posterior probability. Because this file contains all alignment lines produced by bowtie or user-specified aligners, it can also be used as a replacement of the aligner generated BAM/SAM file. For paired-end reads, if one mate has alignments but the other does not, this file marks the alignable mate as "unmappable" (flag bit 0x4) and appends an optional field "Z0:A:!".

'sample_name.transcript.sorted.bam' and 'sample_name.transcript.sorted.bam.bai' are the sorted BAM file and indices generated by samtools (included in RSEM package).

sample_name.genome.bam, sample_name.genome.sorted.bam and sample_name.genome.sorted.bam.bai
Only generated when --no-bam-output is not specified and --output-genome-bam is specified.

'sample_name.genome.bam' is a BAM-formatted file of read alignments in genomic coordinates. Alignments of reads that have identical genomic coordinates (i.e., alignments to different isoforms that share the same genomic region) are collapsed into one alignment. The MAPQ field of each alignment is set to min(100, floor(-10 * log10(1.0 - w) + 0.5)), where w is the posterior probability of that alignment being the true mapping of a read. In addition, RSEM pads a new tag ZW:f:value, where value is a single precision floating number representing the posterior probability. If an alignment is spliced, a XS:A:value tag is also added, where value is either '+' or '-' indicating the strand of the transcript it aligns to.

'sample_name.genome.sorted.bam' and 'sample_name.genome.sorted.bam.bai' are the sorted BAM file and indices generated by samtools (included in RSEM package).

sample_name.time
Only generated when --time is specified.

It contains time (in seconds) consumed by aligning reads, estimating expression levels and calculating credibility intervals.

sample_name.stat
This is a folder instead of a file. All model related statistics are stored in this folder. Use 'rsem-plot-model' can generate plots using this folder.

'sample_name.stat/sample_name.cnt' contains alignment statistics. The format and meanings of each field are described in 'cnt_file_description.txt' under RSEM directory.

'sample_name.stat/sample_name.model' stores RNA-Seq model parameters learned from the data. The format and meanings of each filed of this file are described in 'model_file_description.txt' under RSEM directory.

EXAMPLES
Assume the path to the bowtie executables is in the user's PATH environment variable. Reference files are under '/ref' with name 'mouse_125'.

1) '/data/mmliver.fq', single-end reads with quality scores. Quality scores are encoded as for 'GA pipeline version >= 1.3'. We want to use 8 threads and generate a genome BAM file:

```
 rsem-calculate-expression --phred64-quals \
                   -p 8 \
                   --output-genome-bam \
                   /data/mmliver.fq \
                   /ref/mouse_125 \
                   mmliver_single_quals
```

2) '/data/mmliver_1.fq' and '/data/mmliver_2.fq', paired-end reads with quality scores. Quality scores are in SANGER format. We want to use 8 threads and do not generate a genome BAM file:

```
 rsem-calculate-expression -p 8 \
                   --paired-end \
                   /data/mmliver_1.fq \
                   /data/mmliver_2.fq \
                   /ref/mouse_125 \
                   mmliver_paired_end_quals
```

3) '/data/mmliver.fa', single-end reads without quality scores. We want to use 8 threads:

```
 rsem-calculate-expression -p 8 \
                    --no-qualities \
                    /data/mmliver.fa \
                    /ref/mouse_125 \
                    mmliver_single_without_quals
```

4) Data are the same as 1). This time we assume the bowtie executables are under '/sw/bowtie'. We want to take a fragment length distribution into consideration. We set the fragment length mean to 150 and the standard deviation to 35. In addition to a BAM file, we also want to generate credibility intervals. We allow RSEM to use 1GB of memory for CI calculation:

```
 rsem-calculate-expression --bowtie-path /sw/bowtie \
                    --phred64-quals \
                    --fragment-length-mean 150.0 \
                    --fragment-length-sd 35.0 \
                    -p 8 \
                    --output-genome-bam \
                    --calc-ci \
                    --ci-memory 1024 \
                    /data/mmliver.fq \
                    /ref/mouse_125 \
                    mmliver_single_quals
```

5) '/data/mmliver_paired_end_quals.bam', paired-end reads with quality scores. We want to use 8 threads:

```
 rsem-calculate-expression --paired-end \
                    --bam \
                    -p 8 \
                    /data/mmliver_paired_end_quals.bam \
                    /ref/mouse_125 \
                    mmliver_paired_end_quals
```

6) '/data/mmliver_1.fq.gz' and '/data/mmliver_2.fq.gz', paired-end reads with quality scores and read files are compressed by gzip. We want to use STAR to aligned reads and assume STAR executable is '/sw/STAR'. Suppose we want to use 8 threads and do not generate a genome BAM file:

```
 rsem-calculate-expression --star \
                    --star-path /sw/STAR \
                    --gzipped-read-file \
                    -p 8 \
                    /data/mmliver_1.fq.gz \
                    /data/mmliver_2.fq.gz \
                    /ref/mouse_125 \
                    mmliver_paired_end_quals
```