

RNA-Seq Analysis in KBase

Sunita Kumari*¹ (kumari@cshl.edu), Vivek Kumar¹, Priya Ranjan², Doreen Ware¹, **Bob Cottingham**², **Chris Henry**³ **Adam P. Arkin**⁴, and the KBase Team at the following institutions

¹ Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; ²Oak Ridge National Laboratory, Oak Ridge, TN; ³Argonne National Laboratory, Argonne, IL; ⁴Lawrence Berkeley National Laboratory, Berkeley, CA; ⁵Brookhaven National Laboratory, Upton, NY;

<http://kbase.us>

Project Goals: The DOE Systems Biology Knowledgebase (KBase) is a free, open-source software and data platform that enables researchers to collaboratively generate, test, compare, and share hypotheses about biological functions; analyze their own data along with public and collaborator data; and combine experimental evidence and conclusions to model plant and microbial physiology and community dynamics. KBase's analytical capabilities currently include (meta)genome assembly, annotation, comparative genomics, transcriptomics, and metabolic modeling. Its web-based user interface supports building, sharing, and publishing reproducible, annotated analysis workflows with integrated data. Additionally, KBase has a software development kit that enables the community to add functionality to the system.

RNA-seq analysis is emerging as one of the most powerful approaches for assessing differential gene expression. RNA-seq uses next-generation sequencing to account for all the transcripts in one or more biological samples at a particular time. It can be used for a variety of applications such as transcriptome assembly, gene discovery/annotation, and detection of differential transcript abundances between tissues, developmental stages, genetic backgrounds and environmental conditions. The overarching goal of the RNA-seq pipeline in KBase is to create differential expression estimates and use these to inform metabolic models and to perform functional analysis of genes with similar expression patterns.

RNA-seq analysis typically consists of (i) mapping short sequence reads to the reference genome; (ii) assembling the transcripts into full-length transcripts and expression quantification; and (iii) differential analysis of the gene expression. KBase provides a set of apps that allow users to run the tools from the popular Tuxedo RNA-seq suites to generate the normalized full and differential expression matrix of the reads obtained from Illumina sequencing platforms using the reference prokaryotic and eukaryotic genome. The RNA-seq apps in KBase can be combined into multiple workflows, allowing users to select their choice of reads aligner and assembler for the differential gene expression analysis. For alignment of the reads to the reference genome, transcriptome profiling, and identification of differentially expressed genes, the original Tuxedo suite uses the tools TopHat2, Cufflinks, and Cuffdiff respectively; the new Tuxedo suite uses HISAT2, StringTie and Ballgown.

All of these tools are available as KBase apps; detailed usage instructions can be found at <http://kbase.us/transcriptomics-and-expression-analysis/>.

There are three Narrative tutorials that demonstrate how to use the KBase RNA-seq pipeline end-to-end on plant and microbial reads. You can copy and re-run them to become acquainted with building RNA-seq analysis workflows in KBase.

- Arabidopsis RNA-seq Analysis using Original Tuxedo Suite:
<https://narrative.kbase.us/narrative/ws.19393.obj.1>
- Arabidopsis RNA-seq Analysis using New Tuxedo Suite:
<https://narrative.kbase.us/narrative/ws.19391.obj.1>
- E. coli RNA-seq Analysis
<https://narrative.kbase.us/narrative/ws.19340.obj.16>

KBase is funded by the Genomic Science program within the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under award numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.