illumina®

TECHNICAL NOTE

RNA-Seq data processing for Correlation Engine

An overview of the RNA sequencing pipeline for preparation of data for use in Correlation Engine

Automated multi-step RNA-Seq processing

Unifies sequence retrieval, quality control, alignment, and expression quantification in a single pipeline

High-speed alignment with DRAGEN™ RNA pipeline

Hardware-accelerated processing delivers fast, accurate mapping of high-throughput RNA-Seq data sets

Ready-to-import biosets for Correlation Engine

Includes fold change, FPKM, and q-values from differential expression analysis for biological context investigaton

Introduction

The RNA sequencing (RNA-Seq) pipeline for Correlation Engine processes next-generation sequencing (NGS) data from mRNA to estimate transcript abundance and identify differentially expressed transcripts across samples. Correlation Engine enables comparative analysis of gene expression across studies, conditions, and biological contexts, allowing researchers to explore molecular patterns and generate hypotheses from large-scale public and private data sets.

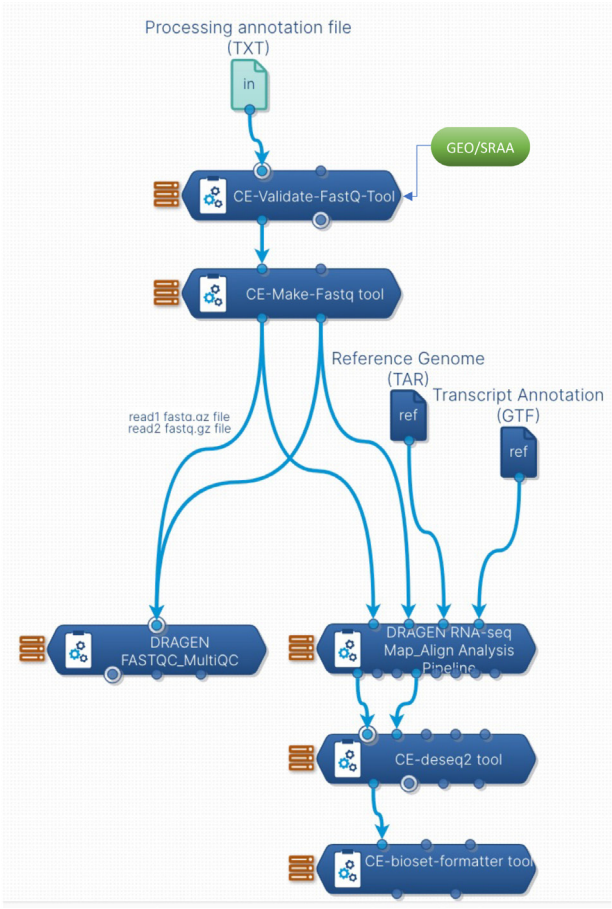

Figure 1: RNA-Seq workflow in Illumina Connected Analytics
The workflow, from raw FASTQ file input through alignment, quantification, and differential expression, provides a scalable, standardized method to prepare data for analysis in Correlation Engine.

RNA-Seq data from public sources are pulled into the Illumina Connected Analytics workflow and processed using DRAGEN analysis tools. This technical note provides details of the individual steps in the RNA-Seq pipeline workflow (Figure 1).

Extraction of raw sequences

The RNA-Seq pipeline supports the input of raw NGS data in FASTQ format. Many NGS studies published in the Gene Expression Omnibus (GEO)[1] provide direct links to raw sequence data stored at the Sequence Read Archive (SRA).[2] Sequence data from the SRA typically requires decompression and, in some cases, splitting to generate usable FASTQ files. When the input FASTQ files are from private sources, sequences need to be trimmed to remove adapter sequences and low-quality tails.

The RNA-Seq pipeline accepts a processing annotation file that identifies which samples to retrieve from GEO/SRA, the sample groupings, and the experimental design of the downstream comparative analysis. The initial step validates that the files are available for retrieval, then each SRA format file is downloaded and converted to FASTQ format using the standard SRA Toolkit provided by the National Center for Biotechnology Information (NCBI).

Sequence alignment

Input NGS sequences are aligned against a reference genome using the Connected Analytics DRAGEN RNA-Seq MAP Align Analysis pipeline. The DRAGEN RNA pipeline uses a spliced aligner that maps short seed sequences from RNA-Seq reads using methods similar to those for DNA alignment. It then incorporates splice junctions, boundaries where exons are joined during RNA splicing, into the final read alignments.[3] A branch in the workflow evaluates sample quality parameters such as degradation, ribonucleotide composition, insert size, and percentage of mapped reads.

The workflow enables rapid alignment and quantification of RNA-Seq data for integration into Correlation Engine, using a standardized interface that supports efficient data processing across multiple input types (Figure 2). The pipeline uses hardware-accelerated algorithms to map and align RNA-Seq reads with high speed and accuracy, outperforming many commonly used software tools. For example, it can align 100 million paired-end RNA-Seq reads in about three minutes. For Correlation Engine, the custom pipeline focuses on transcript

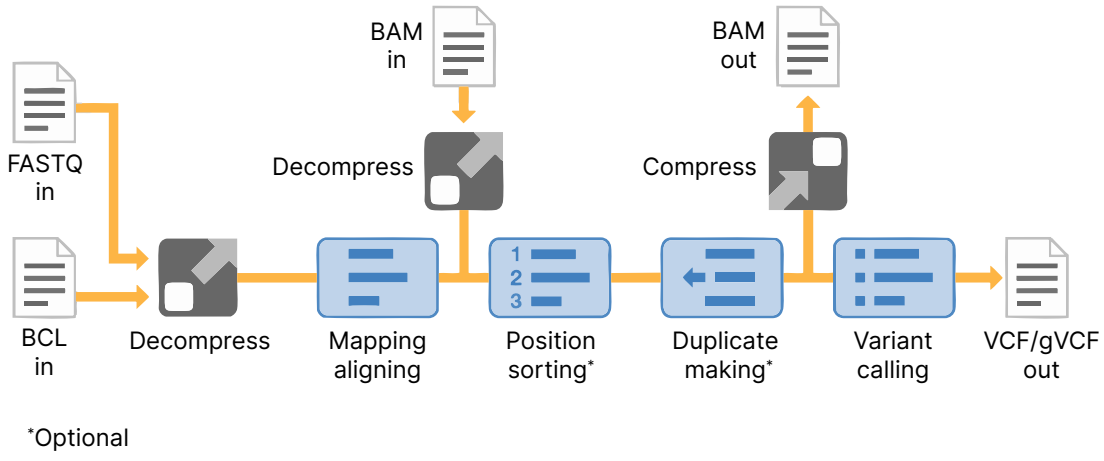

Figure 2: Data analysis using the DRAGEN RNA pipeline
The interface enables rapid and accurate alignment and quantification of RNA-Seq data for integration into Correlation Engine.

quantification for downstream differential signature generation using DESeq2. For a full description, refer to the DRAGEN RNA pipeline web page.

Transcript abundance estimation

The DRAGEN RNA pipeline contains a gene expression quantification module that estimates the expression of each transcript and gene in an RNA-Seq data set. First, it internally translates the genomic mapping of each read (read pair) to the corresponding transcript mappings. Then it uses an Expectation-Maximization (EM) algorithm to infer the transcript expression values that best match all the observed reads. The EM algorithm can also model GC-bias and correct for it in the reported quantification results.

Read counts are used as input for differential expression analysis between test and control groups using R and DESeq2. The base-mean read count, fold change, p-value, and q-value (Benjamini–Hochberg adjusted) are derived from this analysis. The median value of fragments per kilobase of transcript per million mapped reads (FPKM) per group are calculated separately based on normalized read counts, number of aligned reads, and the full gene length.

Bioset generation

The output of the differential expression analysis is consolidated into a single text file and converted into a bioset file containing processing metadata and expression data organized into the following columns:

- Transcript ID
- Fold change
- Test expression (FPKM)
- Control expression (FPKM)
- p-value
- q-value

Filters are applied to the bioset file to pass gene features with at least ±1.2-fold change values and q-values less than or equal to 0.05. The final bioset file is imported directly into Correlation Engine.

Summary

The RNA-Seq pipeline for Correlation Engine uses the DRAGEN RNA pipeline for rapid alignment and expression analysis, producing standardized expression data for downstream investigation. The workflow enables efficient integration of public and private RNA-Seq data sets.

Learn more →

Illumina RNA Sequencing

DRAGEN RNA pipeline

Correlation Engine

References

1. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nuc Acids Res*. 2002;30(1):207-210. doi:10.1093/nar/30.1.207
2. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. *Nuc Acids Res*. 2011;39(Database issue):D19-D21. doi:10.1093/nar/gkq1019
3. Illumina. Illumina DRAGEN Bio-IT Platform Support. support.illumina.com/sequencing/sequencing_software/dragen-bio-it-platform.html. Published 2025. Accessed May 7, 2025.

illumina®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-GL-02154 v2.0