RNA-Seq pipeline benchmark with Dell EMC Ready Bundle for HPC Life Sciences

Deferentially Expressed Gene (DEG) Analysis with Tuxedo Pipeline

Overview

Gene expression analysis is as important as identifying Single Nucleotide Polymorphism, indel or chromosomal restructuring. Eventually, the entire physiological and biochemical events depend on the final gene expression products, proteins. Many quantitative scientists, non-biologists tend to oversimplify the flow of genetic information and forget about what the actual roles of proteins are. Simplification is the beginning of most science fields; however, it is too optimistic to think that this practice also works for biology. Although all the human organs contain the identical genomic composition, the actual protein expressed in various organs are completely different. Ideally, a technology enables to quantify the entire proteins in a cell could excel the progress of Life Science significantly; however, we are far from to achieving it. Here, in this blog, we test one popular RNA-Seq data analysis pipeline known as Tuxedo pipeline. The Tuxedo suite offers a set of tools for analyzing a variety of RNA-Seq data, including short-read mapping, identification of splice junctions, transcript and isoform detection, differential expression, visualizations, and quality control metrics.

A typical RNA-Seq data set consists of multiple samples as shown in **Figure 1**. Although the number of sample sets depends on the biological experimental designs, two sets of samples are used to make comparisons between normal vs. cancer samples or untreated vs. treated samples, for examples.

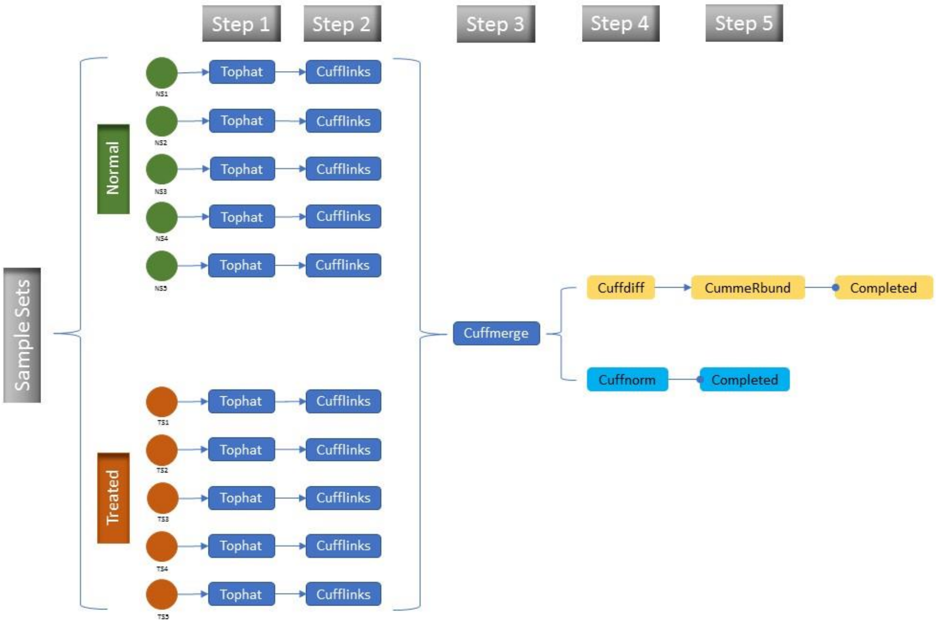

Figure 1 Tested Tuxedo pipeline workflow

All the samples are aligned individually in Step 1. In this pipeline, the Tophat process uses Bowtie 2 version 2.3.1 as an underlying short sequence read aligner. Step 3, Cuffmerge job has a dependency from all the previous jobs in Step 2. The results from Cufflinks jobs are collected at this step to merge together multiple Cufflinks assemblies which is required for Cuffdiff step. Cuffmerge also runs Cuffcompare in the background and automatically filters out transcription artifacts. Cuffnorm generates tables of expression values that are properly normalized for library size, and these tables can be used for other statistical analysis instead of CummeRbund. At Step 5, CummeRbund step is set to generate three plots, gene density, gene box and volcano plots by using R script.

A performance study of RNA-Seq pipeline is not trivial because the nature workflow requires non-identical input files. 185 RNA-Seq paired-end read data are collected from public data repositories. All the read data files contain around 25 Million Fragments (MF)[i] and have similar read lengths. The samples for a test randomly selected from the pool of 185 paired-end read files. Although these randomly selected data will not have any biological meaning, certainly these data will put the tests on the worst-case scenario with very high level of noise.

The test cluster configurations are summarized in **Table 1**.

Table 1 Test Cluster Configurations

	8x Dell EMC PowerEdge C6420
CPU	2x Xeon® Gold 6148 20c 2.4GHz (Skylake)
RAM	12x 16GB @2666 MHz
OS	RHEL 7.4
Interconnect	Intel® Omni-Path
BIOS System Profile	Performance Optimized
Logical Processor	Disabled
Virtualization Technology	Disabled

The test clusters and H600 storage system was connected via 4 x 100GbE links between two Dell Networking Z9100-ON switches. Each compute node was connected to the test cluster side Dell Networking Z9100-ON switch via single 10GbE. Four storage nodes in Dell EMC Isilon H600 was connected to the other switch via 4x 40GbE links. The configuration of the storage is listed in Table 2.

Table 2 Storage Configurations

	Dell EMC Isilon H600
Number of nodes	4
CPU per node	Intel® Xeon™ CPU E5-2680 v4 @2.40GHz
Memory per node	256GB
Storage Capacity	Total usable space: 126.8 TB, 31.7 TB per node
SSD L3 Cache	2.9 TB per node
Network	Front end network: 40GbE Back end network: IB QDR
OS	Isilon OneFS v8.1.0.0 B_8_1_0_011

Performance Evaluation

Two sample Test – Bare-minimum

DEG analysis requires at least two samples. In **Figure 2**, each step described in **Figure 1** is submitted to Slurm job scheduler with proper dependencies. For example, Cuffmerge step must wait for all the Cufflinks jobs are completed. Two samples, let's imagine one normal and one treated sample, begin with Tophat step individually and followed by Cufflinks step. Upon the completion of all the Cufflinks steps, Cuffmerge aggregates gene expressions in the entire samples provided. Then, subsequent steps, Cuffdiff and Cuffnorm begin. The output of Cuffnorm can be used for other statistical analysis. Cuffdiff steps generates gene expression differences at the gene level as well as isoformer level. CummeRbund step uses R-package CummeRbund to visualize the results as shown in **Figure 3**. The total runtime[ii] with 38 cores and two PowerEdge C6420s is 3.15 hours.

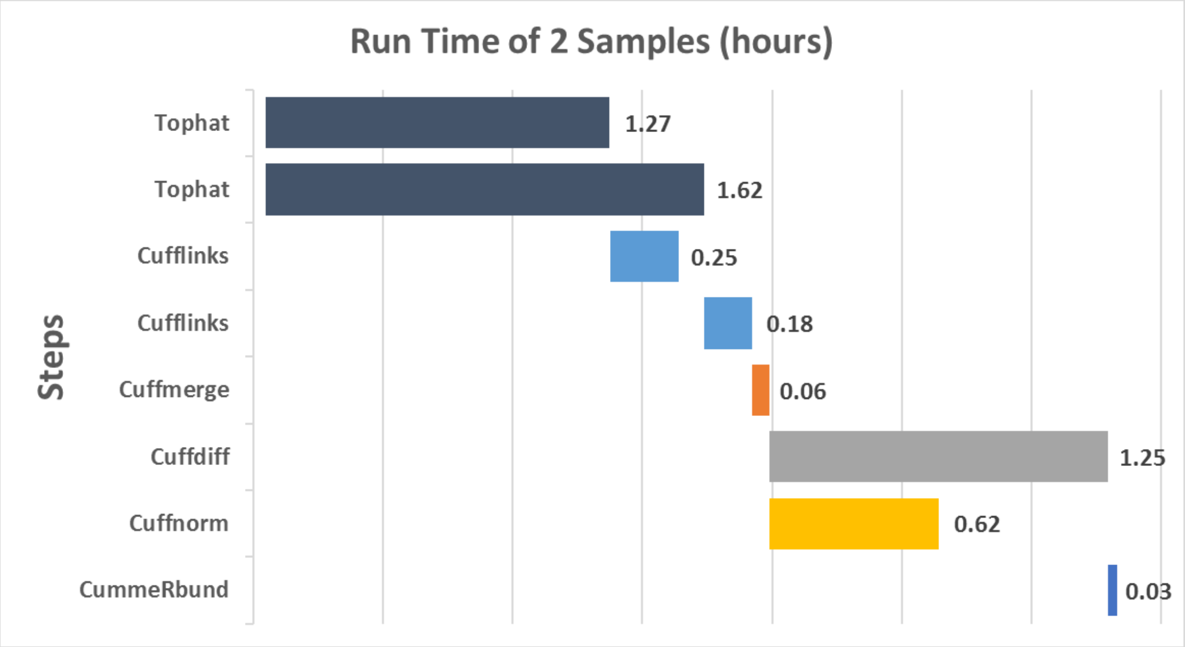

Figure 2 Tuxedo pipeline with two samples

Figure 3 shows differentially expressed genes in red with significantly lower p-values (Y-axis) compared to other gene expressions illustrated in black. X-axis is fold changes in log base of 2, and these fold changes of each genes are plotted against p-values. More samples will bring a better gene expression estimation. The right upper plot are gene expressions in sample 2 in comparisons with sample 1 whereas the left lower plot are gene expressions in sample 1 compared to sample 2. Gene expressions in black dots are not significantly different in both samples.

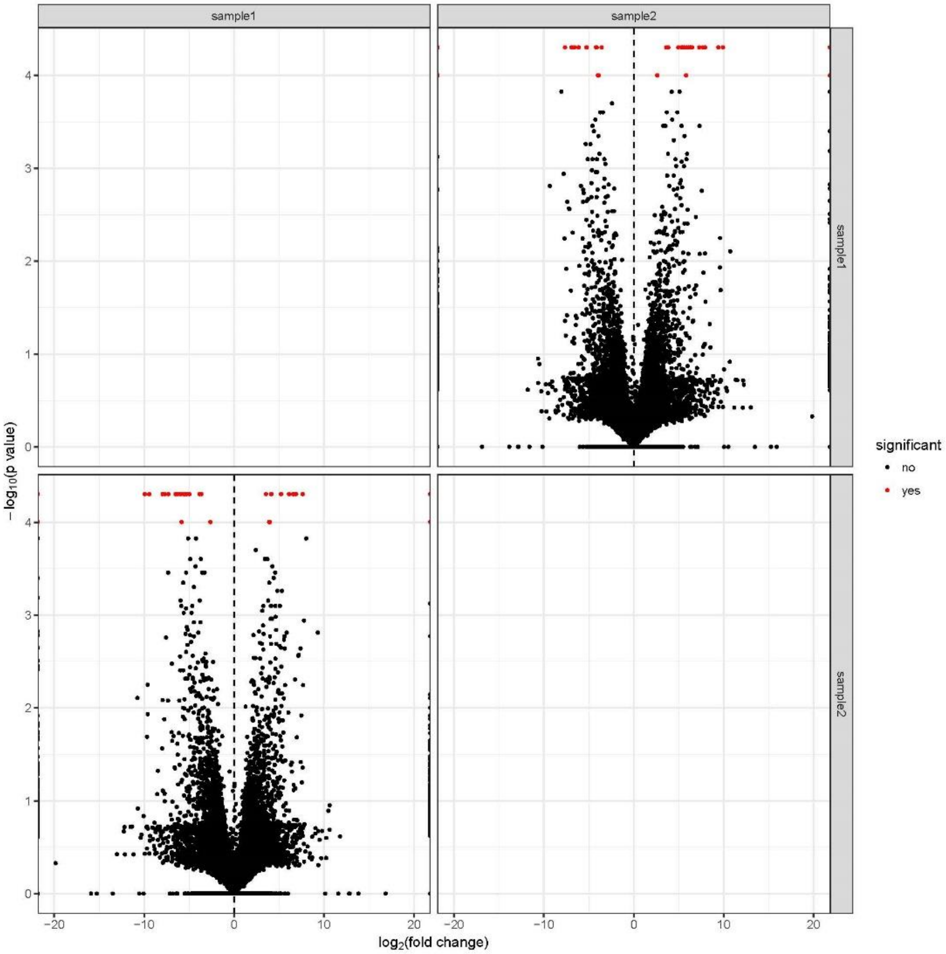

Figure 3 Volcano plot of the Cuffdiff results

Throughput Test – Single pipeline with more than two samples - Biological / Technical replicates

Typical RNA-Seq studies consist of multiple samples, sometime 100s of different samples, normal versus disease or untreated versus treated samples. These samples tend to have high level of noisy due to their biological reasons; hence, the analysis requires vigorous data preprocessing procedure.

Here, we tested various numbers of samples (all different RNA-Seq data selected from 185 paired-end reads data set) to see how much data can be processed by 8 nodes PowerEdge C6420 cluster. As shown in **Figure 4**, the runtimes with 2, 4, 8, 16, 32 and 64 samples grow exponentially when the number of samples increases. Cuffmerge step does not slow down as the number of samples grows while Cuffdiff and Cuffnorm steps slow down significantly. Especially, Cuffdiff step becomes a bottle-neck for the pipeline since the running time grows exponentially (**Figure 5**). Although Cuffnorm's runtime increases exponentially like Cuffdiff, it is ignorable since Cuffnorm's runtime is bounded by Cuffdiff's runtime.

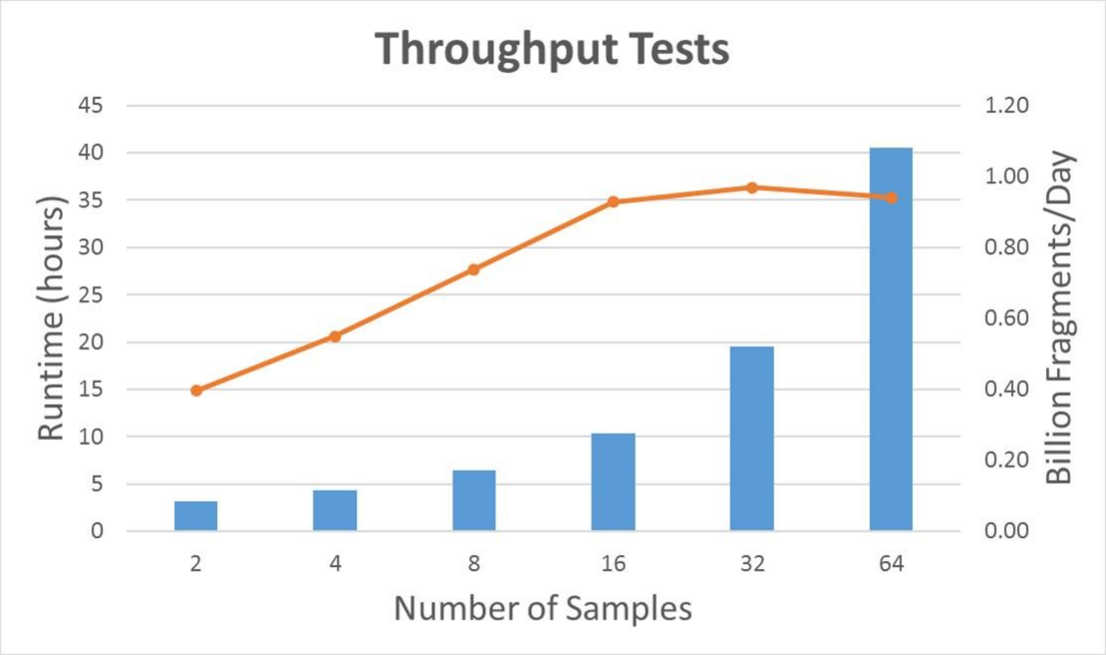

Figure 4 Runtime and throughput results

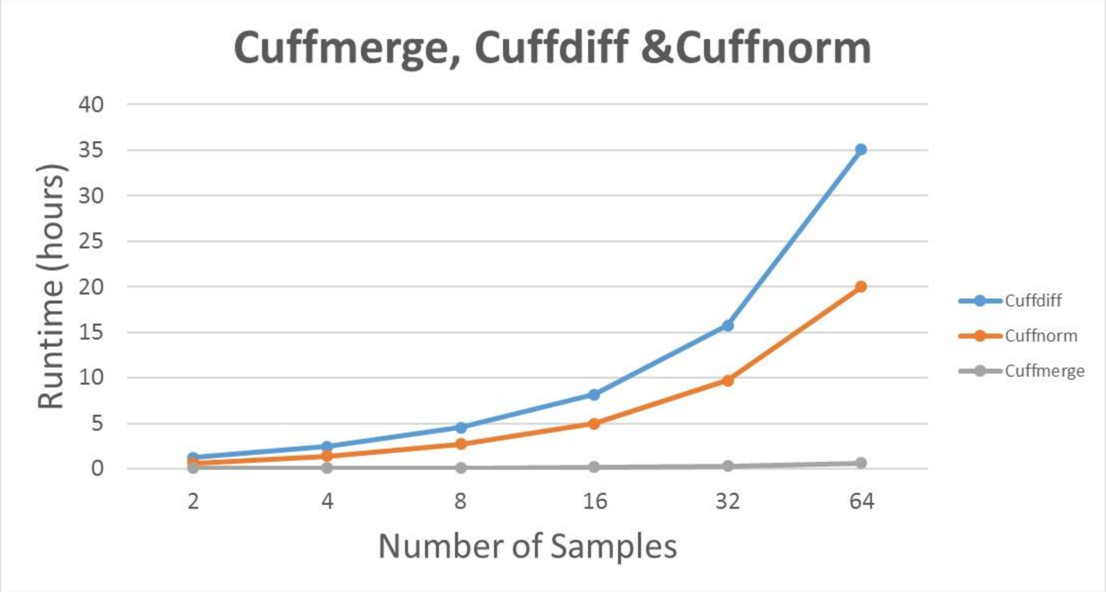

Figure 5 Behaviors of Cuffmerge, Cuffdiff and Cuffnorm

Conclusion

The throughput test results show that 8 node PowerEdge C6420s with Isilon H600 can process roughly 1 Billion Fragments which is little more than 32 samples with ~50 million paired reads each (25 MF) through Tuxedo pipeline illustrated in **Figure 1**.

Since Tuxedo pipeline is relatively faster than other popular pipelines, it is hard to generalize or utilize these results for sizing a HPC system. However, this provides a good reference point to help designing a right size HPC system.

Resources

Internal web page

External web page

Contacts

Americas

Kihoon Yoon

Sr. Principal Systems Dev Eng

Kihoon.Yoon@dell.com

+1 512 728 4191

[1] Footnote copy goes here.

[i] "For RNA sequencing, determining coverage is complicated by the fact that different transcripts are expressed at different levels. This means that more reads will be captured from highly expressed genes, and few reads will be captured from genes expressed at low levels. When planning RNA sequencing experiments, researchers usually think in terms of numbers of millions of reads to be sampled." – cited from https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf

[ii] Runtime refers Wall-Clock time throughout the blog.