Large Scale RNA Expression Analysis by leveraging Tophat, Cufflinks and Jnomics

Fall 2012

Piyush Kansal
Stony Brook University

Adviser:
Prof. Michael Schatz
Cold Spring Harbor Laboratory

- Motivation

Tophat and Cufflinks are used together to do RNA expression analysis. When run on a large experiment (~1GB) on only one of the machines on the cluster, these programs take upto 52 minutes to complete, which becomes a serious limitation when they are run on a large set of experiments (50) because in that case, a sequential run will take almost 45 hours to complete.

So, the motivation is to leverage the capabilities of Tophat and Cufflinks by enhancing these programs to work with Jnomics so as to do large scale RNA expression analysis in shorter amount of time.

- Background

- Tophat
  A fast splice junction mapper for RNA-Seq reads. It aligns these reads to genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons *[1]*

- Cufflinks
  It assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols *[2]*

- Jnomics
  A cloud-scale sequence analysis suite designed to help meet the computational challenges presented by the continuing revolution in massively parallel DNA sequencing technologies. In total, current worldwide second-generation sequencing capacity exceeds 13 Pbp/year, and continues to increase annually by a factor of five.
  The storage and analysis of such massive volumes of genomic data represents the primary challenge in computational biology today. Jnomics attempts to address these problems by applying recent innovations in distributed computing to the challenge of large-scale genomic storage and analysis. It is based on Apache Hadoop, an open-source implementation of Google's MapReduce framework *[3]*

- Source Files

- RNAExpAnalysisUsingTophatCufflinks.py (new)
  Main file responsible for processing all the experiments

- tophat_jnomics (new)
  Tophat code with hooks placed to do processing using Hadoop

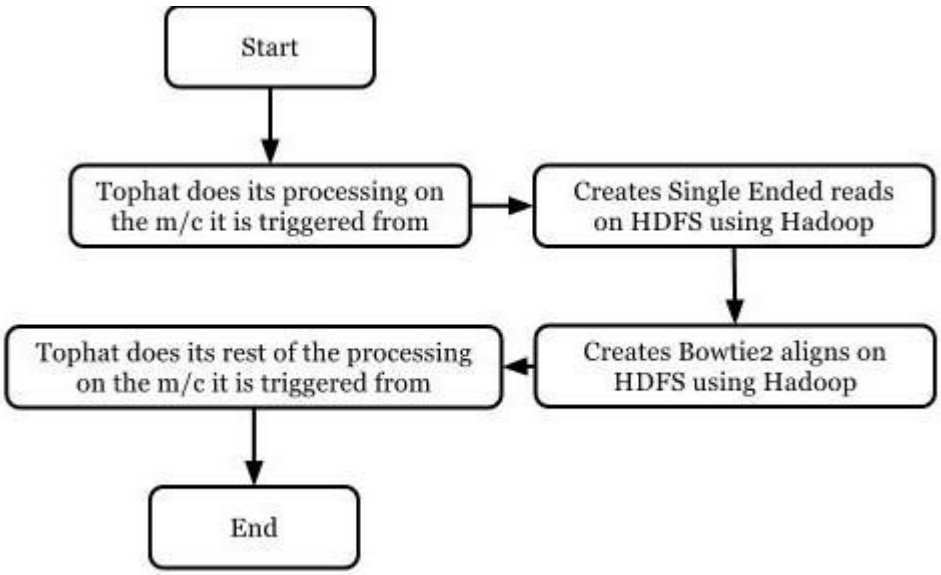
- tophatjnomicsext.py (new)
  File containing function definitions of the hooks placed in tophat_jnomics

- SingleEndLoader.java (new)
  File responsible to create single ended reads on HDFS

- CufflinksMap.java (new)
  File responsible for processing Tophat output using Cufflinks

- genPlotsBetweenCoverageAndFPKM.py (new)
  File responsible for creating graphs between Coverage(output of regular expression analysis) and FPKM(output of expression analysis using Tophat and Cufflinks) values

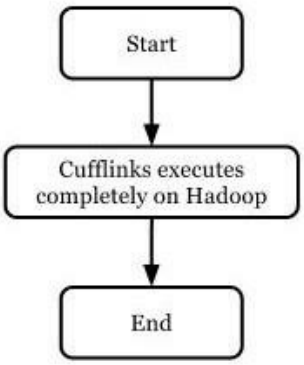
- Program Workflow of Tophat



- The single ended reads are created using the existing class in Jnomics, FastqParser. The records in this file are iterated till the end and then written to the output file on HDFS in compressed format using Gzip as a compression algorithm. The class responsible for this functionality is SingleEndLoader and *implements* existing class ManagerTask

- The bowtie2 alignments are created using the already existing class in Jnomics, Bowtie2Map. The mapper of this class uses Java Runtime to invoke the bowtie2-align command and creates the required alignments on HDFS

- Program Workflow of Cufflinks



- The Cufflinks program on Hadoop is run using Java Runtime, similar to Bowtie2Map. The command is invoked from the mapper class and the output of the Cufflinks is created in the temporary directory of a mapper which is then copied back to HDFS before the mapper finishes. The class responsible for this functionality is CufflinksMap and extends existing class AlignmentBaseMap

- Output

Output of running these programs are set of files on HDFS. There are two set of files:

- Tophat output files in BAM format. These files acts as an input to the class CufflinksMap
- Cufflinks output files in regular text format. These files contain the FPKM values for the genes and in our case, are further used to generate plots

- Installation

- Install latest code of Jnomics
- cd jnomics/code
- ant jar (this will build-up jnomics.jar)
- cd src/tools/edu/cshl/schatz/jnomics/tools
- sudo cp RNAExpAnalysisUsingTophatCufflinks.py /usr/local/bin/
- sudo cp tophat_jnomics /usr/local/bin/
- sudo cp tophatjnomicsext.py /usr/local/bin/
- sudo cp genPlotsBetweenCoverageAndFPKM.py /usr/local/bin/
- To generate graphs, install rpy2 as well

- Steps to run:

RNAExpAnalysisUsingTophatCufflinks.py <fasta-file> <number-of-experiments> <jnomics-jar> <binaries-location> <tophat-op-files-path-on-hdfs> <cufflinks-op-files-path-on-hdfs>

Eg, RNAExpAnalysisUsingTophatCufflinks.py ecoli.fa 50 ../code/bin/jnomics.jar . tophat_output cufflinks_ouput

- Test Results

- Data Set
  Consists of 50 experiments with each experiment containing paired-end reads of size ~1GB. So, total amount of data to be processed is almost 50GB. Total number of genes to be analysed are 1047

- Serial Execution Time
  Total time is *45 hours*

- Parallel Execution Time
  Total time is *2.15 hours*

- References

[1] http://tophat.cbcb.umd.edu/
[2] http://cufflinks.cbcb.umd.edu/
[3] http://sourceforge.net/apps/mediawiki/jnomics/?source=navbar