BC-DNA: A GUI BASED PIPELINE FORNGS DATA ANALYSIS

Amratha c[1] , Chaya kharvi[2] , Swathi. [3] .
[1] M.sc ,lecturer -Bhandarkars'college. amrathac2000@gmail.com
[2]M.Tech,lecturer -Bhandarkars'college. kharvichaya87@gmail.com
[3]M.sc ,lecturer -Bhandarkars'college. swathiskaranth@gmail.com

1. INTRODUCTION

Next Generation Sequencing (NGS) technology has evolved rapidly in the last five years, leading to the generation of hundreds of millions of sequences (reads) in a single run. The number of generated reads varies between 1 million for long reads generated ≈400 base pairs (bps) and 2.4 billion for short reads generated ≈75 bps. The invention of the high-throughput sequencers has led to a significant cost reduction of sequencing. NGS generally produces short reads or short read pairs meaning short sequences of <~200 bases. To compare DNA of the sequenced sample to its reference sequence - a sequence to which the subject is to be compared, there is a need to find the corresponding part of that sequence for each read in sequencing data. This is called aligning or mapping the reads against the reference sequence.

Alignment, also called mapping of reads (short DNA sequence), is an essential step in re-sequencing. Re-sequencing refers to acompletesequencing of the genome of DNA. An alignment of data from these re-sequenced organisms is a relatively simple method of detecting variation in samples. Genome sequence alignment and DNA/ RNA Sequence analysis help us to understand genetic variations, understanding various diseases, identification of mutations linked to different forms of cancer etc. The process of aligning these reads to a reference genome is time consuming and demands the development of fast and accurate alignment tools. However, the current available tools make different compromises between the accuracy and the speed of mapping.

Raw short DNA/RNA reads often come in a file format called FASTQ - a plain text format where each single read occupies four consecutive lines. For each of the short reads in the FASTQ file, a corresponding location in the reference sequence needs to be determined. This is achieved by comparing the sequence of the read to that of the reference sequence. A mapping algorithm will locate a location in the reference sequence that matches the DNA / RNA read while tolerating a certain amount of mismatch to allow subsequence variation detection.

RNA sequencing (RNA-Seq), a highly sensitive and accurate technique for measuring expression across the transcriptome is revolutionizing the study of the transcriptome. It is a powerful method for discovering, profiling, and quantifying RNA transcriptsthat provides visibility to previously undetected changes occurring in disease states in response to therapeutics under different environmental conditions and across a broad range of other study designs.

There are many software tools such as BBMap, BLAT, Mosaik, TMAP, NextGenMap, Bowtie etc, for short-read alignment and TopHat, Cufflinks, HISAT, DiffBind, Sailfish etc, for RNA-Seqanalysis tools. Bowtie - it is a reliable, convenient and fast tool compare to other short-read alignment tools and it supports running in parallel mode. Cufflink- is good for compute expression values because it contains a sophisticated algorithm for this calculation, which is far more accurate than other RNA-Seq analysis tools. This is the one of reason to include these modules in our project.

Bowtie is a fast short aligner based on the Burrows-Wheeler transform [1] and the FM (Full-text index in Minute Space) index [2]. An FM-index is a compressed full-text substring index based on the Burrows-Wheeler transform. Bowtie tolerates a small number of mismatches and works best when

aligning short reads to large genomes, though it supports arbitrarily small reference sequences and reads as long as 1024 bases. It is designed to be extremely fast for sets of short reads in the following situations:

(a) Many of the reads have at least one good, valid alignment.

(b) Many of the reads are relatively high-quality.

(c) The number of alignments reported per read is small (close to 1).

Cufflinks accepts the output of Bowtie Module in the SAM format in order to rebuild transcripts.

Cufflinks 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.Cufflinks includes other modules such as Cuffdiff, Cuffnorm and Cuffcompare, each module discussed below:

- **Cuffdiff** is a highly accurate tool for performing comparisons of expression levels of genes and transcripts in RNA-Seq experiments. It not only tell us which genes are up- or down-regulated between two or more conditions and also tells which genes are differentially spliced or undergoing other types of isoform-level regulation.
- **Cuffcompare**helps us to compare assembled transcripts to a reference annotation so it will take the transcripts and reference annotation as an input. It tracks the cufflinks transcripts across multiple experiments.
- **Cuffnorm** is helpful to generate tables of expression values that are properly normalized for library size. Expression levels reported by Cufflinks in FPKM units are usually comparable between samples, but in certain situations applying an extra level of normalization can remove sources of bias in the data. Cuffnorm normalizes a set of samples to be on as similar scales as possible, which can improve the results we obtain with other downstream tools.

Cuffdiff, Cuffcompare, Cuffnorm takes the output of Cufflinks module as input for performing various RNA analysis.

All the above discussed modules are dependent on one another in terms of input/output (i.e., output of one module need to be given as input to another module) which has to be handled manually. However, the execution of each module is done separately based on users' interest. All the tools are Linux based i.e., they can be run on Linux only. Running each module needs installing all the required software tools and selecting necessary parameters manually. Lot of human intervention is required to passdata to each module and collect the output. All this processes comes with lot of difficulty for biologists who are not much familiar with Linux environment and execution of each module. In addition, biologists with no programming experience often find it difficult to perform parameter settings and convertdata format to the required format.

To overcome all these challenges, there is a need to develop a user friendly pipeline(tool) integrating all these modules to perform short read DNA sequence alignment and RNAsequence analysis facilitating large scale data analysis with minimum human intervention.

In this work, we have designed a Linux based pipeline(tool) with a user friendly GUI for short read alignment and RNAsequence analysis, connecting Bowtie, Cufflinks and other sub modules of Cufflinks such as Cuffdiff, Cuffcompare and Cuffnorm. This system facilitates to provide all the data and necessary parameters at once and allows user to go with the execution to get the final output thus significantly reducing human intervention. This system is designed with several functional GUI submodules which are flexible to be used in any combination or even can be used separately based on the user requirements and availability of input data. This user friendly tool is very easy to use and will allow biologists to focus more on data analysis without worrying to learn the execution of the tool.

2. LITERATURE SURVEY

Various studies have been conducted on DNA, RNA by using various data analysis tools such as Bowtie [1], TopHat and Cufflinks(Cuffdiff, Cuffcompare and Cuffnorm)[3] that allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under various conditions etc.

Bowtie uses a different and novel indexing strategy to create an ultrafast, memory-efficient short read aligner geared toward mammalian re-sequencing. Bowtie aligns 35-base pair (bp) reads at a rate of more than 25 million reads per CPU-hour, which is faster than existing systems. Bowtie employs a Burrows-Wheeler index based on the full-text minute-space (FM) index, which has a memory footprint of only about 1.3 gigabytes (GB) for the human genome. The small footprint allows Bowtie to run on a typical desktop computer with 2 GB of RAM. The index is small enough to be distributed over the internet and to be stored on disk and re-used. Multiple processor cores can be used simultaneously to achieve even greater alignment speed.

The BWT (Burrows Wheeler Transformation) is a reversible permutation of the characters in a text. Although originally developed within the context of data compression, BWT-based indexing allows large texts to be searched efficiently in a small memory footprint. It has been applied to bioinformatics applications, including whole-genome alignment, tiling microarray probe design, and Smith-Waterman alignment to a human-sized reference. The Burrows-Wheeler transformation of a text T, BWT (T), is constructed as follows. The character $ is appended to T, where $ is not in T and is lexicographically less than all characters in T. The Burrows-Wheeler matrix of T is constructed as the matrix whose rows comprise all cyclic rotations of T$. The rows are then sorted lexicographically. BWT (T) is the sequence of characters in the rightmost column of the Burrows-Wheeler matrix (Figure 1a). BWT (T) has the same length as the original text T.

The FM Index

In 2000, six years after the BWT was published, Paolo Ferragina and Giovanni Manzini published a paper [2] describing how the BWT, together with some small auxilliary data structures, can be used as a space-efficient index of. It generally uses less than half the space required to store m5 integers. They named it the FM Index.

Cole Trapnell, Adam Roberts et al [3] described a protocol which explains about two software tools TopHat and Cufflinks for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. These tools allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. This protocol describes in detail how to use TopHat and Cufflinks and other sub modules in Cufflinks such as Cuffdiff, Cuffcompare and Cuffnorm to perform such analyses. It also covers several accessory tools and utilities that aid in managing data, including CummeRbund, a tool for visualizing RNA-seq analysis results.

They developed two popular tools that together serve all three roles, as well as a newer tool for visualizing analysis results. TopHat aligns reads to the genome and discovers transcript splice sites. Cufflinks uses this map against the genome to assemble the reads into transcripts. Cuffdiff, a part of the Cufflinks package, takes the aligned reads from two or more conditions and reports genes and transcripts that are differentially expressed using a rigorous statistical analysis. These tools are gaining wide acceptance and have been used in a number of recent high-resolution transcriptome studies. CummeRbund renders Cuffdiff output in publication-ready figures and plots. Cuffcompare and Cuffnorm is a part of Cufflinks.Cuffcompare tracks the cufflinks transcripts across multiple experiments and Cuffnorm normalizes a set of samples to be on as similar scales as possible which can improve the results obtained with other tools.

Galaxy[4] is an open source, web-based platform for data intensive biomedical research. It is a data integration, data and analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming experience. Although it was initially developed for genomics research which is largely domain agnostic and is now used as a

general bioinformatics workflow management system. It typically provide a graphical user interface for specifying what data to operate on, what steps to take, and what order to do them in. Galaxy is now also used for gene expression, genome assembly, proteomics, epigenomics, transcriptomics and host of other disciplines in the life sciences. It provides several tools for DNA/RNA sequence analysis like Bowtie, Cufflinks and other sub-modules in Cufflinks like Cuffcompare, Cuffdiff, Cuffnorm. The availability of various types of data analysis tools in Galaxy does not overcome the difficulty of running all the required tools in a user friendly and pipelined manner.

METHODOLOGY

The proposed system builds a pipeline (tool) based on most widely used DNA and RNA sequence analysis tools namely, Bowtie, Tophat and Cufflinks(various sub modules of cufflinks i.e,cuffdiff, cuffcompare, cuffnorm).

The proposed system consists of three main steps:

1. Building an index file of the user supplied reference genome.
2. Alignment of short DNA sequences with the reference genome.
3. Transcriptome assembly and isoform quantification from RNA-seq reads

The system works in the form of pipeline to accept inputs for all the independent modules connected and thus overcomes the difficulty of running each tool independently. In addition there is also an option to start the pipeline from any module, in case of availability of the input of a particular module.

Following figure shows the workflow of the pipeline, connecting different tools :

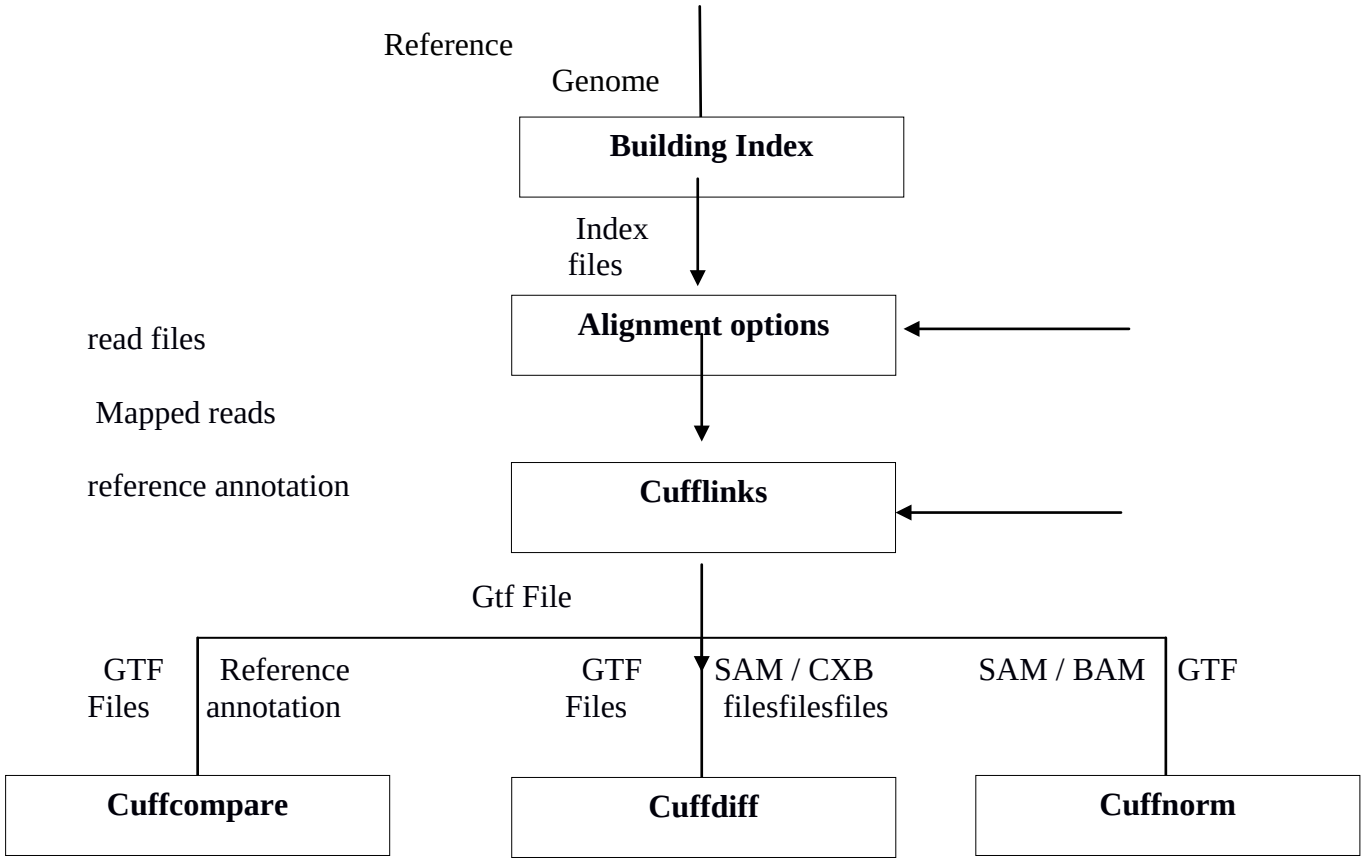

Figure 1: Describes the workflow of the pipeline.

The workflow shown in Figure 1 is briefly discussed below.

Building index:

To build index file, the system takes reference genome as input, and a binary file is produced as output. This index file is later used for alignment process either by bowtie or tophat based on user interest.

Alignment options:

For aligning the input reads, If the user is interested to perform alignment with bowtie, then all the option (parameters) related to bowtie will be displayed.Uponselecting necessary parameters user can proceed with the execution of the program. User can provide either user defined index file or already built in index files and a read file as a input. The accepted reads (Aligned reads) are given as output in SAM format.

Cufflinks:

This option is provided to assemble transcripts, estimation of their abundances, and tests for differential expression and regulation in RNA-Seq samples. The module takes sorted SAM file or a bam file as an input. Along with SAM file cufflinks also take the reference annotation file as an input and will produced the output in the form of GTF files.

Cuffcompare:

This module takes transcripts and reference annotation as an input to compare assembled transcripts to a reference annotation. Cuffcompare tracks the cufflinks transcripts across multiple experiments to trace any variation in the transcripts. The output is produced in the form of .stats, .refmap, .gtfetc

Cuffdiff:

Cuffdiff is to find significant changes in transcript expression, splicing, and promoter use. it will takes a transcript annotation file produced by cufflinks, cuffcompare, or other source as a input along with a SAM file of aligned RNA-Seq reads. If more than two files are provided, Cuffdiff tests for differential expression and regulation between all pairs of samples.

Cuffnorm:

Cuffnorm is to generate tables of expression values that are properly normalized for library size. It will take a transcript annotation file produced by cufflinks, cuffcompare, or other source as a input along with a SAM file of aligned RNA-Seq reads.

Algorithms used in the above modules are described below:

1. Algorithms used to build index file.

- **Building Index:**

  Input: Reference genome

  Output: Index files

  Step 1: Select either built-in index file or user defined reference genome FASTA format.

  Step 2: If built-in index file:

  Do nothing

  Else:

  Provide the file name for creating index file for reference genome.

  Step 3: Create the index for the reference genome of that name.

2. Alignment of short DNA Sequence

Input:Index files

Output:Mapped reads

Step 1: Select the index file of reference genome.

Step 2: Select necessary options available in Bowtie.

Step 3: Select the read File in the FASTA or FASTQ format.

Step 4: Pass the read file which is to be either single end or Paired-end.

Step 5:Generate the output in the SAM format.

Step 6: Sort the output if needed.

3. Cufflinks:

Input:Mapped reads

Output:Transfrags in GTF

Step 1: Choose the SAM /BAM file generated by the Bowtie tool.

Step 2: If it is not in the sorted order:

Do Sorting using “sort”

Step 3: If needed pass the reference annotation along with the SAM files.

Step 4: Output in the form of GTF files

4. Cuffcompare or Cuffdiff or Cufffnorm

Input:Transcripts.gtf

Output: Comparition, Differencial expression and Normalization of transfrags.

Step 1: If Cuffcompare:

Pass output of Cufflinks along with reference annotation if needed.

Else if Cuffdiff:

Pass output of cufflinks along with minimum two SAM files.

Else if Cuffnorm:

Pass output of cufflinks along with minimum two SAM files.

Step 2: Generate be produced by different modules.

EXPERIMENTS AND RESULTS

4.1.Experiments:

Python version 3.4 and pycharm community editor (IDE) are used to implement the pipeline with the user friendly GUI.

Python is a widely used high-level, general-purpose, interpreted, dynamic programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of codethan possible in languages such as C++ or Java. The language provides constructs intended to enable clear programs on both a small and large scale. Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library.

Python interpreters are available for many operating systems, allowing Python code to run on a wide variety of systems. Using third-party tools, such as Py2exe or Pyinstaller, Python code can be packaged into stand-alone executable programs for some of the most popular operating systems, so Python-based software can be distributed to, and used on, those environments with no need to install a Python interpreter.

PyCharm is an Integrated Development Environment (IDE) used for programming in Python. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django. PyCharm is developed by the Czech company JetBrains. It is cross-platform working on Windows, Mac OS X and Linux. PyCharm has a Professional Edition, released under a proprietary license and a Community Edition released under the Apache License. PyCharm Community Edition is less extensive than the Professional Edition.

We tried this system to create the index file of E-coli bacteria (approx 5 million bases long) and then tested the alignment of reads (approx 1000 reads each with length nearly 40 bases, both paired and single end) of the same organism . On matching the read with the reference genome, the output will be written in the file with the header information for each read such as gene name, version, ID, exon number and transcript ID. The sample file for E-coli reference genome is shown in the Figure E1.

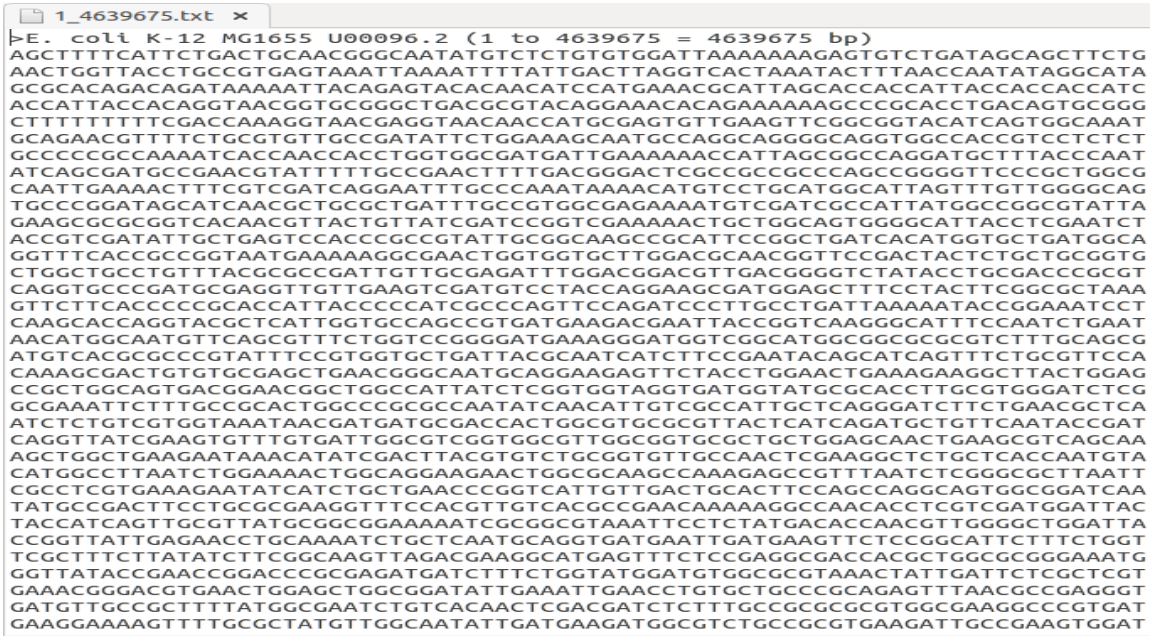

Figure E1: E-coli reference genome.

Read File

This file contains many short DNA sequences, typically an output by a sequencing instrument.

Reads can be in the FASTA format, FASTQ format, or in a raw one-sequence-per-line format. The reads used in this example are a set of simulated reads included in the Bowtie package. Figure E2 shows read file in the FASTQ format.

The read files have the following criteria.

- Single reads must be submitted using a single Fastq file and can be submitted with or without read names.
- Paired reads must split and submitted using either one or two Fastq files. The read names must have a suffix identifying the first and second read from the pair
- The first line for each read must start with '@'.
- The base calls and quality scores must be separated by a line starting with '+'.

Figure E2: read file in the FASTQ format

4.2 Validation:

In order to guide the user to use this system efficiently and generated necessary error messages whenever required, we have performed validation at different levels. Some of the validation screen shorts and corresponding error messages are shown below.

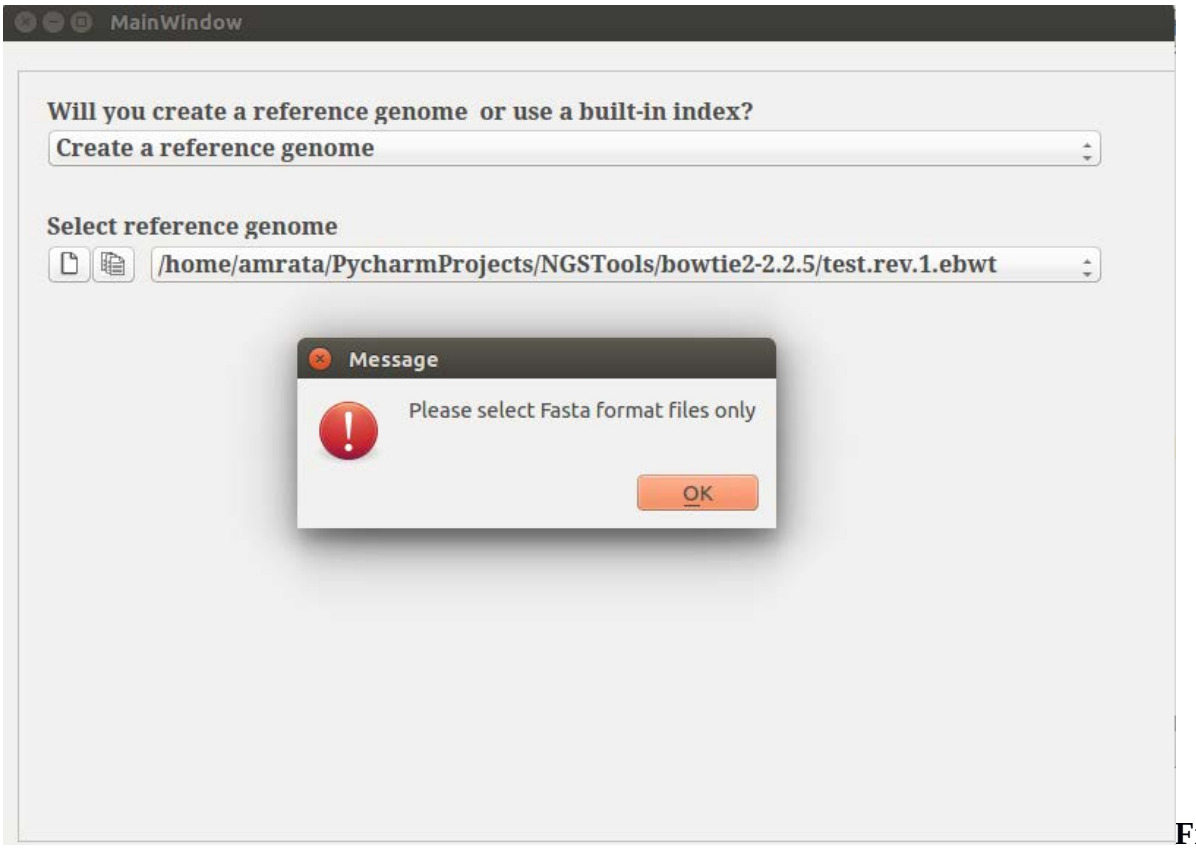

Figure V1: This screen generates an error message "please select fasta format file only" if the reference genome file selected is other than FASTA format.

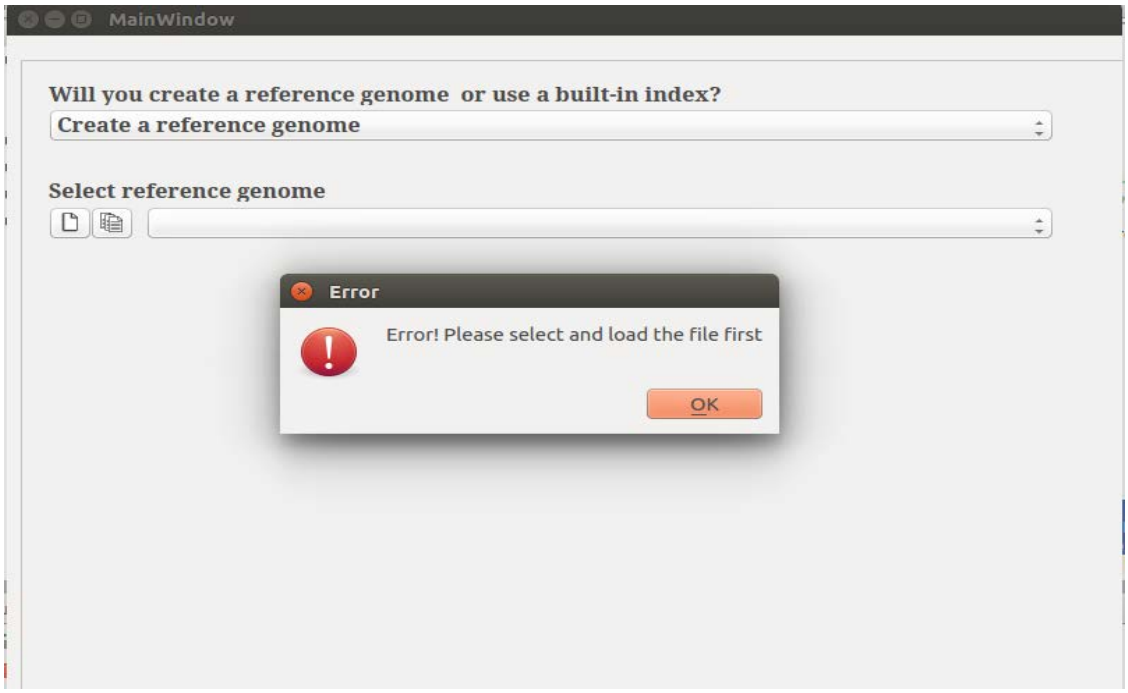

Figure V2: An error is raised if the user doesn't select any reference genome file.

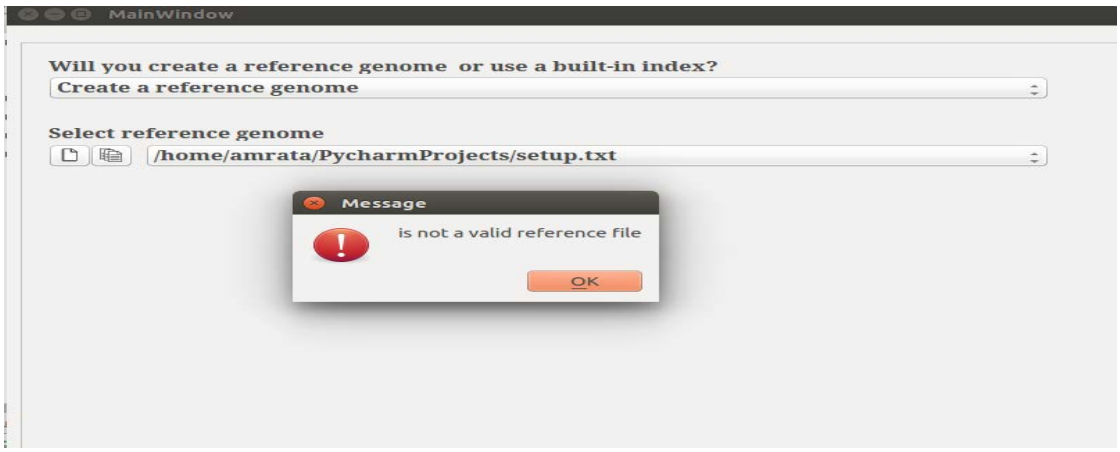

Figure V3: An error is raised in case user supplies an invalid FASTA file.

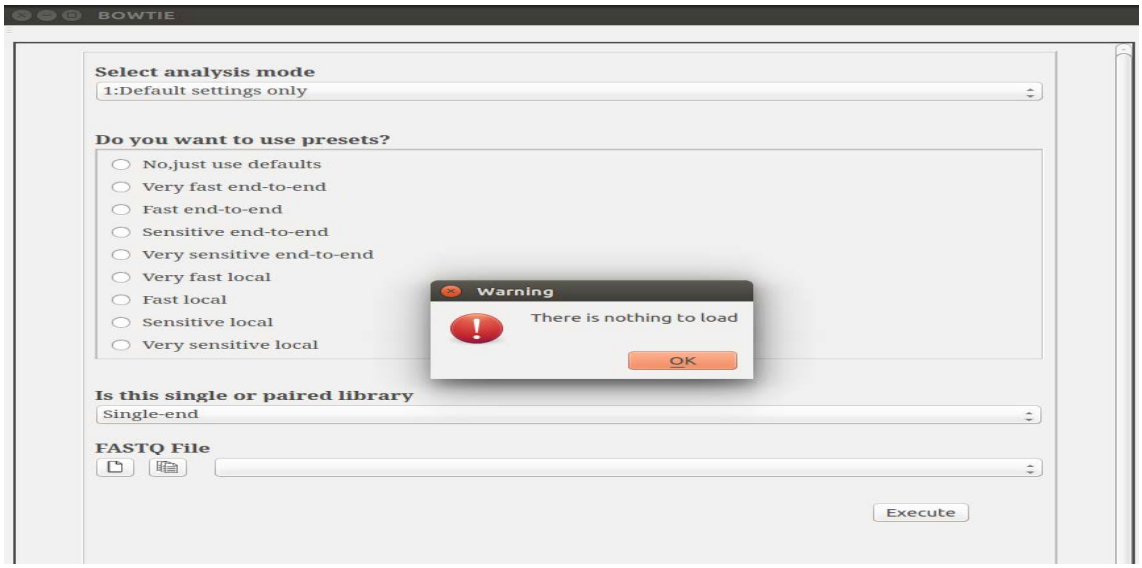

Figure V4: This error is generated in Bowtie module if user does not supply any read file for alignment.

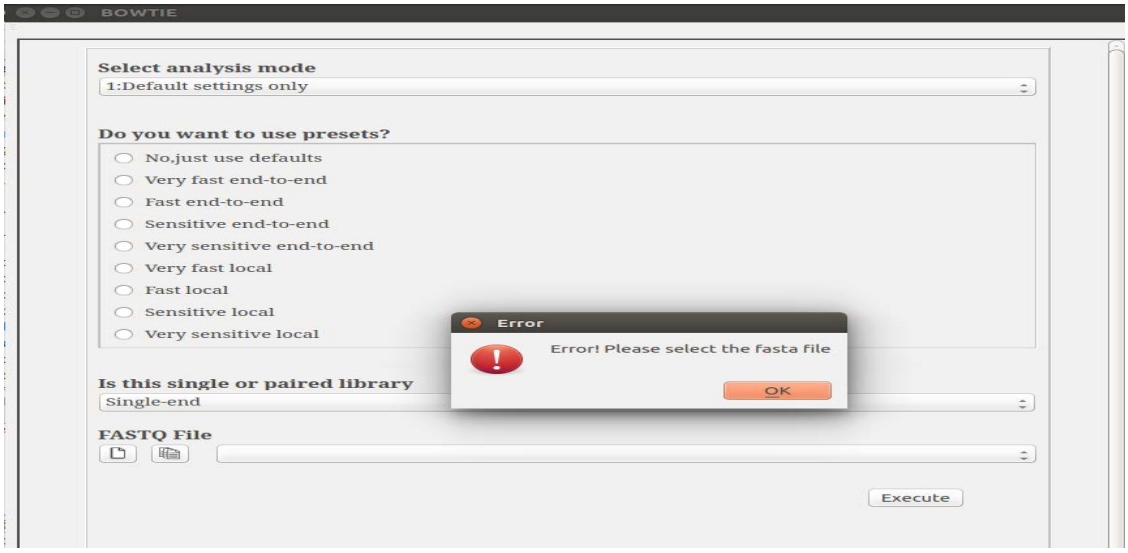

Figure V5: This error is generated in Bowtie module when the user select read file other than the FASTQ format.

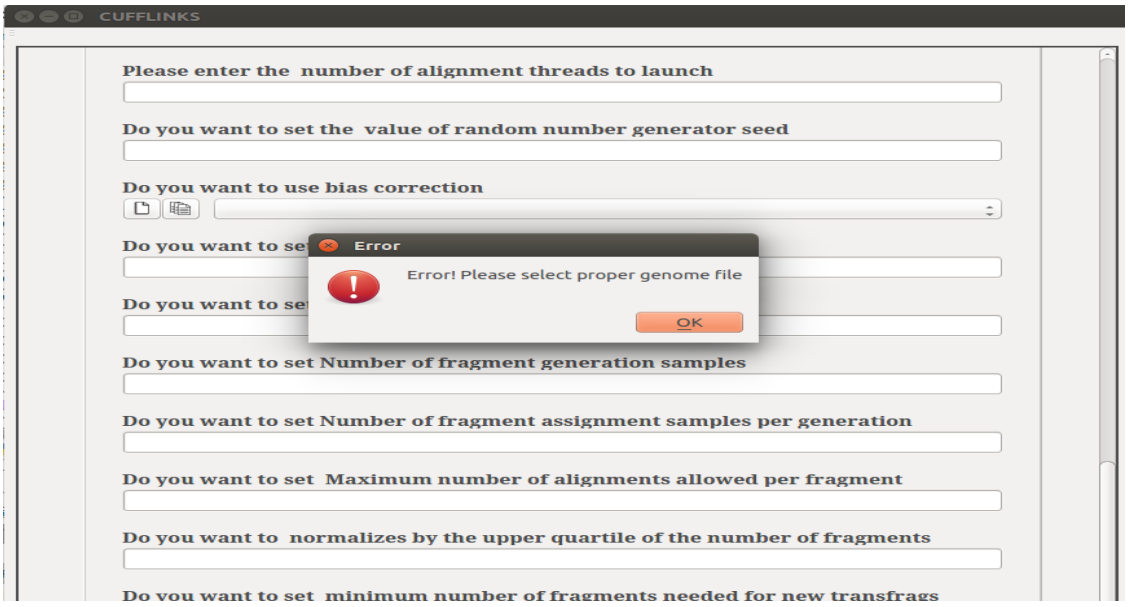

Figure V6: This error raised in the Cufflink module when the user select reference genome file other than FASTA format.

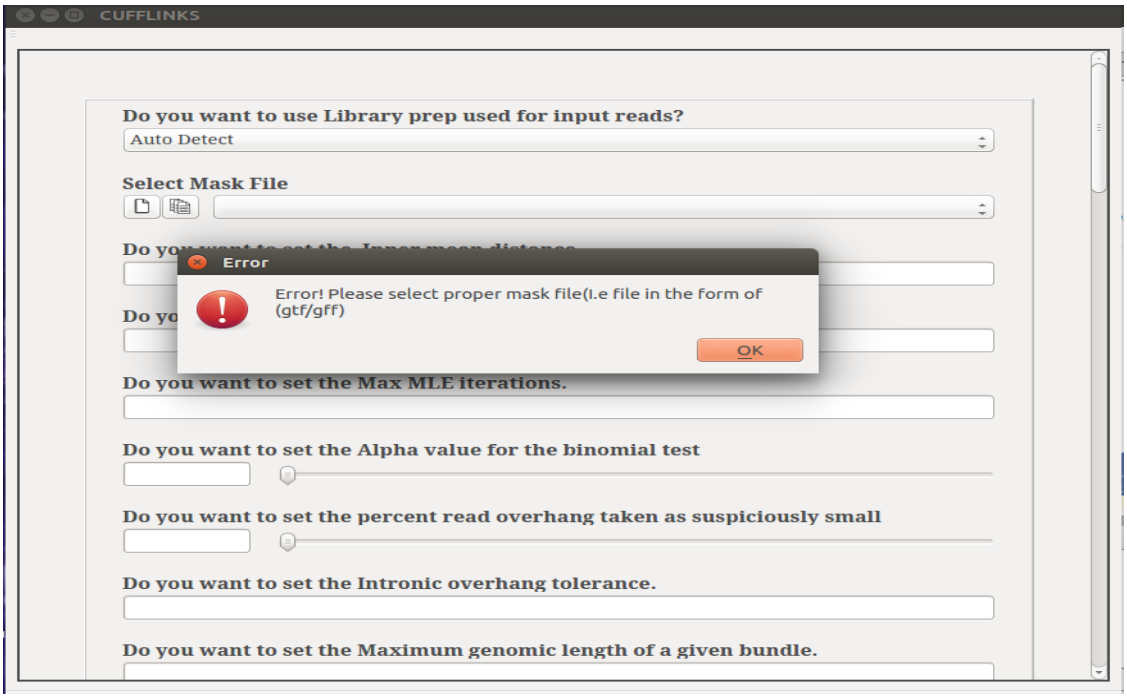

Figure V7: This error raised in the Cufflink module when the user select the Mask file other than GTF format.

4.3 Results and snapshots

1. Building index file:

A set of index files containing the index of the reference genome is produced as output . The index file format is unique to Bowtie, and FASTA formats are converted to this format using the Bowtie2-build tool which takes a collection of FASTA files for a reference genome and generate a collection of index files. Index files can then be used by bowtie to align reads to the reference genome. The same set of index files can be used across multiple runs of bowtie. The screen shot shown in Figure E3 shows the available options to build the index file.

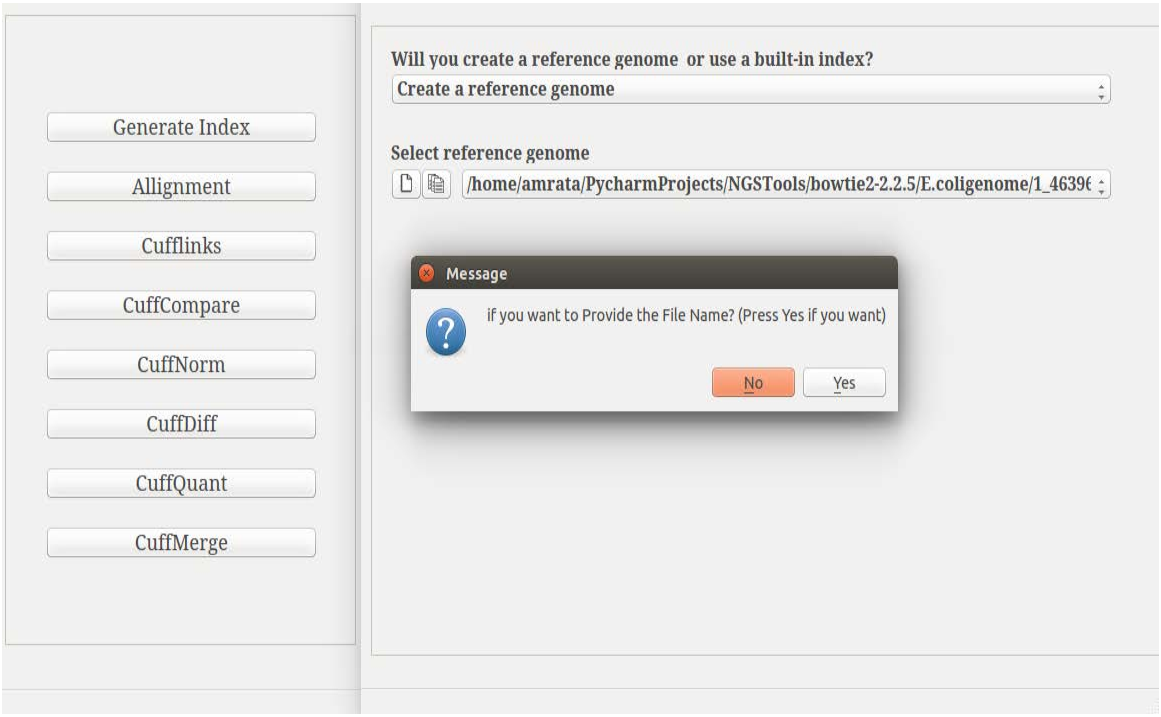

Figure E3: This screen shot shows set of options under Building index file.

2. Bowtie Module:

When invoking Bowtie, the user specifies an index file "basename," i.e. the prefix shared by all the index files, and the path(s) to the FASTQ read files to be aligned , and some additional parameters necessary to execute the module. Screen shot shown in Figure E4 shows the available options for alignment.

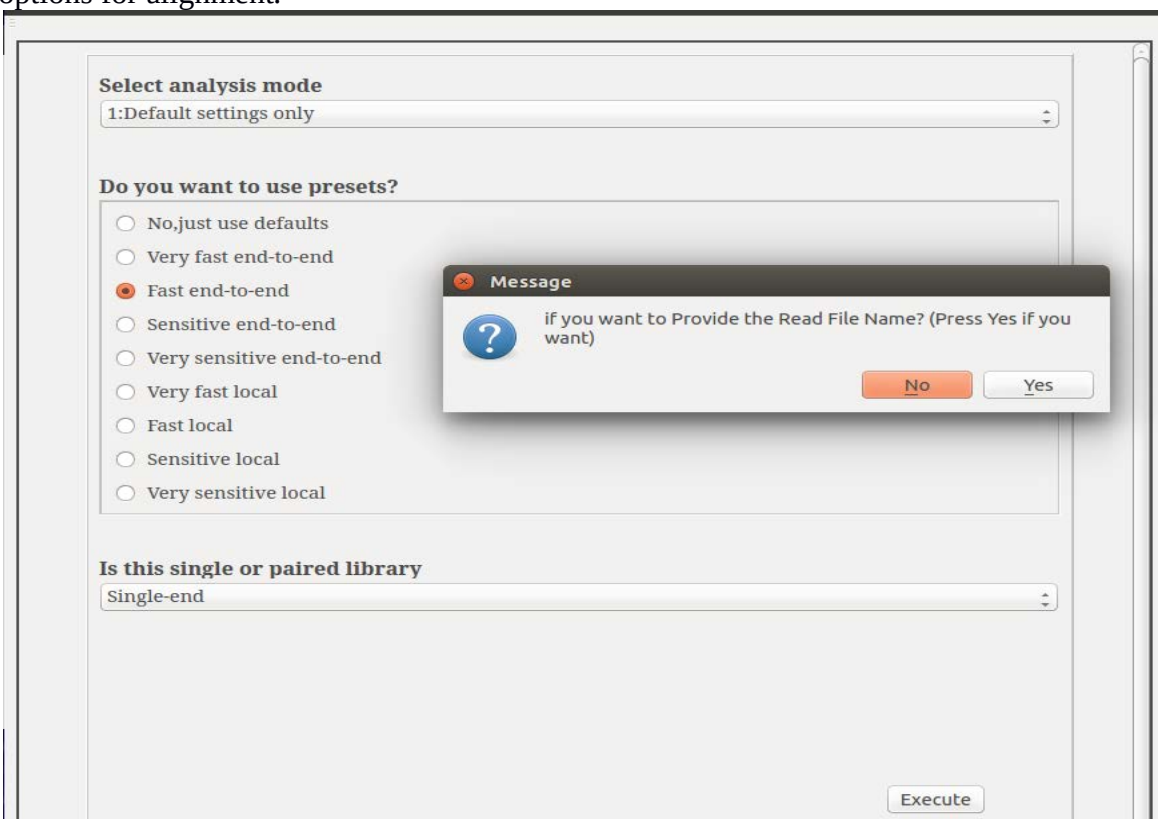

Figure E4: This screen shot shows set of options available under Bowtie.

3. Bowtie output:

The output file for an alignment consists of 8 fields, separated by tabs. The fields from left to right are:

- Read name
- Reference strand aligned to ("+" denotes forward strand, "−" denotes reverse strand)
- Name of reference sequence aligned to
- Offset of the leftmost position on the forward reference strand covered by the alignment
- Read sequence aligned (or its reverse complement, if the read aligned to the reverse strand)
- Quality sequence aligned (or its reverse, if the read aligned to the reverse strand)

Figure E5: This screen shot shows the Mapped reads of E_coli.

The output consists of two sets of lines those printed to the "standard out" file handle and those printed to the "standard error" file handle. Each line printed to "standard out" is a valid reportable alignment found by Bowtie. A valid alignment is one that satisfies the alignment policy specified in the−n, −l, −eor−voptions. A reportable alignment is one that is not suppressed by any other option such as the−k, −m,or−Moptions. Lines printed to "standard error" contain summary information about the entire alignment run,

6) Cufflinks:

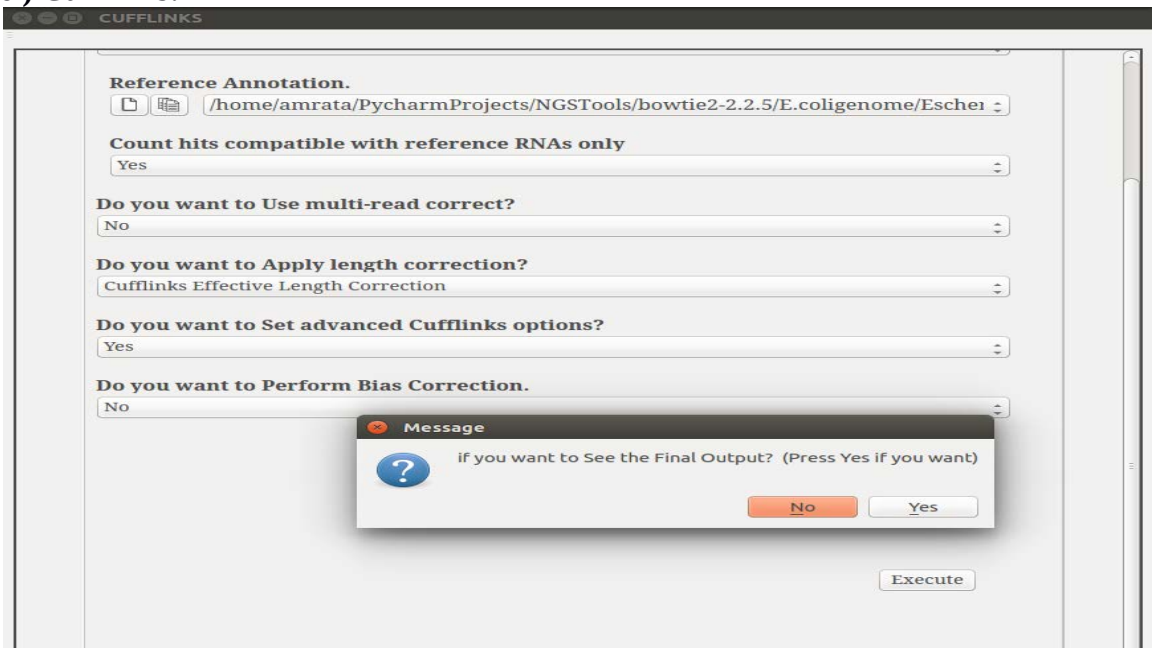

Figure E6: This screen shot shows set of options available under Cufflinks.

File of RNA-seq read alignments in SAM/BAM format is accepted as input.
As shown in Figure E6, reference genome annotation file can be submitted along with the SAM/BAM file. On selection of reference genome annotation file, Cufflinks will use this file to estimate isoform expression and will not assemble novel transcripts. The program will ignore alignments not structurally compatible with any reference transcript. The file can also be sent to the GTF guideparameter to enable Cufflinks to use the reference annotation based transcript (RABT) assembly algorithm.

7) Cufflinks output:

Figure E7 shows output in the format of GTF files and shows the label for the transcribed fragments (transfrags)

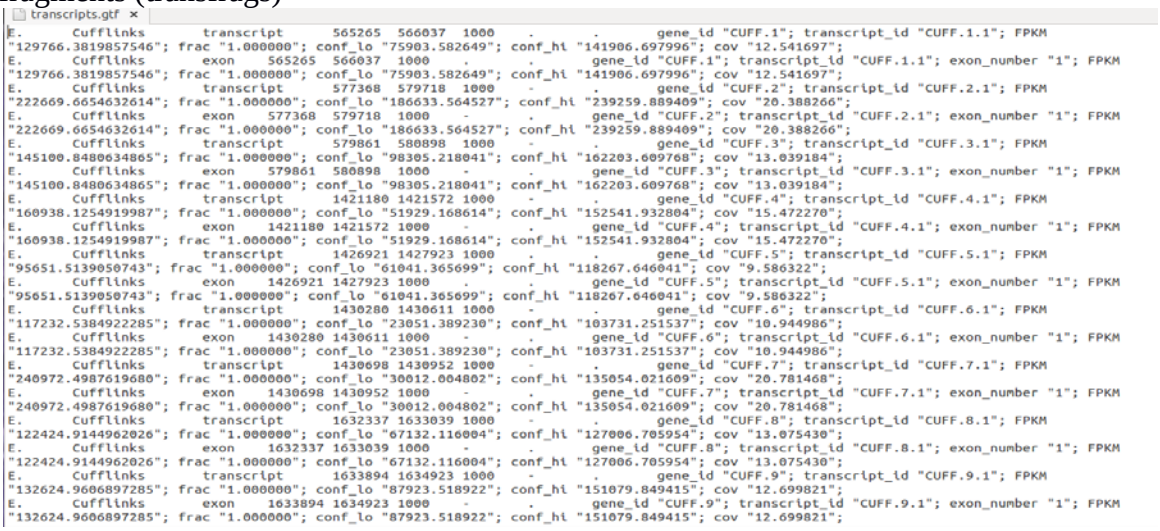

Figure E7 :This screen shot shows theoutput of Cufflink module.

This GTF file contains Cufflinks' assembled isoforms. The first 7 columns are standard GTF, and the last column contains attributes some of which are also standardized ("gene_id" and "transcript_id"). There is one GTF record per row, and each record represents either a transcript or an exon within a transcript.

8) CuffCompare:

Figure E8: Shows the options available under Cuffcompare for comparitions

cuffcompare requires at least one Cufflinks' GTF output file as input, and optionally can also take a "reference" annotation GTF/GFF file. Each sample is matched against this file, and sample isoforms are tagged as overlapping, matching, or novel where appropriate.

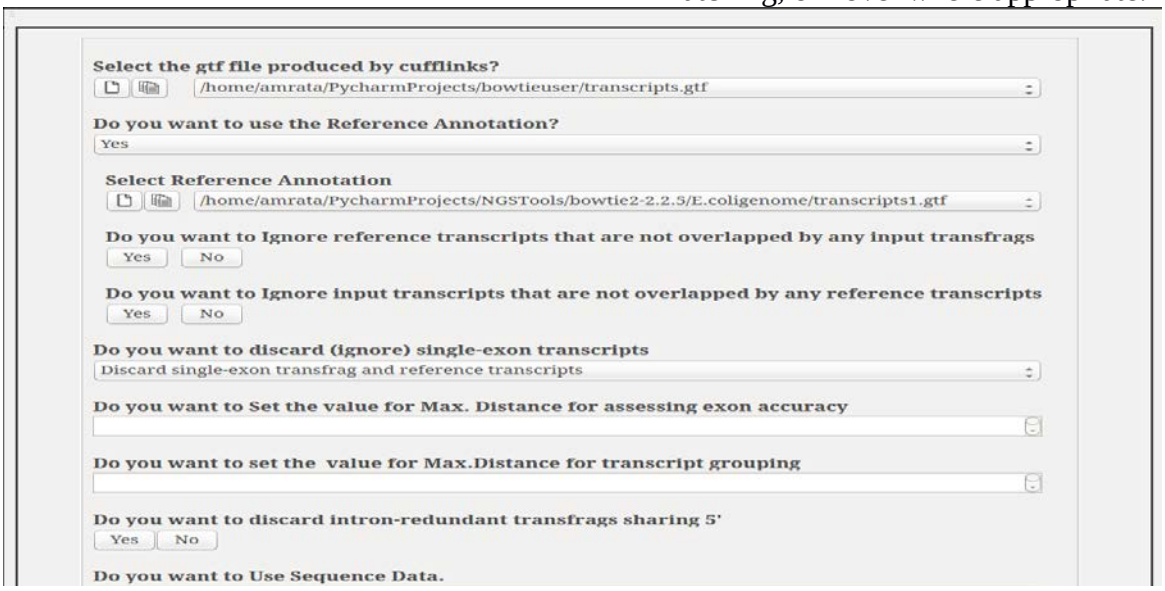

Figure E8:shows set of options available under Cuffcompare.

Some of the output file produced by Figure E8 will be described below:

1. <output.prefix>.stats
   this file contain various statistics related to the accuracy of the transcripts in each sample when compared to the reference annotation data.
2. <output.prefix>.combined.gtf
   Cufflinks.cuffcompare reports a GTF file containing the "union" of all transfrags in each sample. If a transfrag is present in both samples, it is thus reported once in the combined GTF.
3. *.tmap
   These tab-delimited files list the most closely matching reference transcript for each Cufflinks transcript. There is one row per Cufflinks transcript.Etc.

4. CONCLUSION

The system is implemented to provide a tool with a user friendly GUI for DNA and RNA- seq analysis. The system overcomes the difficulty of running each module(tool) separately, eventually reducing human intervention at significant level. The system provides an easy access to submit data and set all the necessary parameters at once, with an additional advantage of running the system in a pipelined manner or a module separately depending on the availability of input data. The system is easy to use and require no prior knowledge of programming and thus allow users to perform data analysis in an easy and user friendly manner with no in-between human intervention while running the system in pipeline manner and minimum intervention while executing sub-modules. The system is currently under testing at Institute of Bioinformatics and Applied Biotechnology(IBAB), Bangalore.

References:

1. B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-ef+-ficient alignment of short DNA sequences to the human genome.,” *Genome Biol.*, vol. 10, no. 3, p. R25, 2009.
2. FM Index, PaoloFerragina and Giovanni Manzin,2000
3. Cole Trapnell, Brian Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Jeltje van Baren, Steven Salzberg, Barbara Wold, LiorPachter.
4. https://usegalaxy.org/ - Galaxy tool

https://bowtie-bio.sourceforge.net – Bowtie manual

https://cole-trapnell-lab.github.io – Cufflinks manual

- For dataset:
  https://bowtie-bio.sourceforge.net
  http://www.ensembl.org
- For software development:
  www.wikipedia.com