ORIGINAL RESEARCH PAPER | Biological Science | Volume - 7 | Issue - 1 | January - 2017 | ISSN - 2249-555X | IF : 3.919 | IC Value : 79.96

A fundamental and efficient method for transcriptome analysis.

KEYWORDS: *Transcriptome*, NGS Data, Analysis

Varsha N
Department of Biotechnology, R V College of engineering, Bangalore, Karnataka, India

Sanjay Deshpande
Department of Biotechnology, R V College of engineering, Bangalore, Karnataka, India

Dr. Vidya Niranjan
Department of Biotechnology, R V College of engineering, Bangalore, Karnataka, India

ABSTRACT *Transcriptome analysis and investigations can portray all transcriptional movement (coding and non-coding), concentrate on a subset of pertinent target qualities and transcripts, or profile a great many qualities on the double to make a worldwide picture of cell capacity. Quality expression investigation studies can give a preview of snapshot of actively expressed genes and transcripts under different conditions.*

INTRODUCTION

The Transcriptome is the arrangement of all RNA molecules in one cell or a populace of cells. It varies from the exome in that it incorporates just those RNA molecules found in a predetermined cell populace, and generally incorporates the sum or convergence of every RNA particle notwithstanding the sub-atomic characters. There are two general techniques for construing Transcriptome. One methodology maps succession peruses onto a reference genome, both of the life form itself (whose Transcriptome is being concentrated on) or of a firmly related animal types. The other methodology, once more Transcriptome get together, utilizes programming to derive transcripts straightforwardly from short arrangement peruses.

Entire Transcriptome investigation is of developing significance in seeing how changed articulation of hereditary variations adds to complex ailments, for example, growth, diabetes, and coronary illness. Examination of far reaching differential RNA expression gives scientists more prominent bits of knowledge into natural pathways and atomic systems that direct cell destiny, advancement, and sickness movement.

SURVEY OF TOOLS

Table 1: Tools used in Transcriptome analysis

Sl. No	Process		Tools	Description
1	Data Processing	1.1	SRA Toolkit	Conversion of Raw data to readable format.
2	Quality Check	2.1	FastQC	Checking the quality of the Data
3	Sequence Processing	3.1	Cutadapt	Removes unwanted sequences from NGS Data
4	Alignment	4.1	Bowtie 1 & Bowtie2	Reference indexing and Alignment
		4.2	TopHat	Alignment of RNA seq reads with Reference genome and transcriptome
		4.3	Qualimap	Transcriptome alignment quality check
5	Transcript generation and Annotation	5.1	Cufflink	Cufflinks the program assembles transcriptomes from RNA-Seq data and quantifies their expression
		5.2	cuffmerge	cuffmerge performs merging of transcript assemblies from the RNA seq Data into a master trasnscript assembly
		5.3	cuffcompare	Cuffcompare helps you perform these comparisons and assess the quality of your assembly.
		5.4	Blast	The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences.
		5.5	Blast2go	Blast2GO is a bioinformatics platform for high-quality functional annotation and analysis of genomic datasets.
6	Differential expression	6.1	cuffdiff	Cuffdiff is a highly accurate tool for performing these comparisons, and can tell you not only which genes are up- or down-regulated between two or more conditions, but also which genes are differentially spliced or are undergoing other types of isoform-level regulation.
7	Visualization	7.1	R and Bioconductor	CummeRbund is an R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output.

METHODOLOGY

Data Processing

SRA Toolkit

The desired data is obtained from NCBI-SRA website. Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Bio systems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Quality Check

FastQC

Once the file is converted to FASTQ format. The file is then subjected to quality check. This is done using the FASTQC toolkit. FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

Sequence processing

Cutadapt: Once the quality is being checked the next step is to

Volume - 7 | Issue - 1 | January - 2017 | ISSN - 2249-555X | IF : 3.919 | IC Value : 79.96

remove adapter. The tool used is cutadapt. Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing read. Cleaning your data in this way is often required: Reads from small-RNA sequencing contain the 3' sequencing adapter because the read is longer than the molecule that is sequenced. Amplicon reads start with a primer sequence. Poly-A tails are useful for pulling out RNA from your sample, but often you don't want them to be in your reads.

Cutadapt helps with these trimming tasks by finding the adapter or primer sequences in an error-tolerant way. It can also modify and filter reads in various ways. Adapter sequences can contain IUPAC wildcard characters. Also, paired-end reads and even colour space data is supported. If you want, you can also just demultiplex the input data, without removing adapter sequences at all.

Alignment

Bowtie 1 & Bowtie 2

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index (based on the Burrows-Wheeler Transform or BWT) to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 gigabytes of RAM. Bowtie 2 supports gapped, local, and paired-end alignment modes. Multiple processors can be used simultaneously to achieve greater alignment speed. Bowtie 2 outputs alignments in SAM format, enabling interoperation with a large number of other tools (e.g. SAMtools, GATK) that use SAM. Bowtie 2 is distributed under the GPLv3 license, and it runs on the command line under Windows, Mac OS X and Linux.

TopHat

TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splice junctions. It is built on the ultrafast short read mapping program Bowtie. TopHat runs on **Linux** and **OS X.**

TopHat can find splice junctions without a reference annotation. By first mapping RNA-Seq reads to the genome, TopHat identifies potential exons, since many RNA-Seq reads will contiguously align to the genome. Using this initial mapping information, TopHat builds a database of possible splice junctions and then maps the reads against these junctions to confirm them.

Differential expression analyis

CuffDiff:

Comparing expression levels of genes and transcripts in RNA-Seq experiments is a hard problem. Cuffdiff is a highly accurate tool for performing these comparisons, and can tell you not only which genes are up- or down-regulated between two or more conditions, but also which genes are differentially spliced or are undergoing other types of isoform-level regulation.

Visualization of results

R and Bioconductor.

cummeRbund : cummeRbund is a visualization package for Cufflinks high-throughput sequencing data. It is designed to help you navigate through the large amount of data produced from a Cuffdiff RNA-Seq differential expression analysis. The results of this analysis are typically a large number of inter-related files that are not terribly intuitive to navigate through. cummeRbund helps promote rapid analysis of RNA-Seq data by aggregating, indexing, and allowing you easily visualize and create publication-ready figures of your RNA-Seq data while maintaining appropriate relationships between connected data points. CummeRbund is a multifaceted suite for streamlined analysis and visualization of massively parallel RNA differential expression data sequencing data.

DISCUSSION

Since initially presented in 2005, NGS advances have been connected to a assortment of fields in genomic research. Underneath we quickly present a couple of extra illustrations of intriguing applications, past Transcriptome, encouraged by NGS technologies. NGS innovation has advanced examination on customized pharmaceutical. Quality combination occasions and join variations can likewise be distinguished by NGS (e.g. RNA-Seq or genomic DNA sequencing) information. These sequencing variety thinks about would comprehend the relationship between human hereditary variety furthermore, wellbeing/maladies. There are numerous different utilizations of NGS. Case in point, joined use of exome and Transcriptome sequencing can be utilized to hunt down qualities with loss of heterozygosity and allele-particular expression; noncoding RNA (ncRNA) expression profiling reveals insight onto human disease Transcriptome. In general, NGS brings various trust and energy to genomic research, however challenges with respect to information preparing analyses still exist.

A Transcriptome analysis using cutting edge RNA sequencing (RNA-Seq) has been a standout amongst the most appealing themes among late research exercises. The center system of RNA-Seq is to utilize NGS advancements to grouping cDNAs to get data around a specimen's RNA content .The RNA-Seq measure, with profound scope and base level determination, has given a perspective of eukaryotic Transcriptome of remarkable subtle element and clarity. Since created, RNA-Seq has been generally used to uncover the mind boggling scene and progress of Transcriptome for various species. Contrasted with past high-throughput advancements, for example, microarray, RNA-Seq does not require tests or groundworks, and in this way on a basic level is conceivable to do applications that are unrealistic utilizing conventional microarray-based strategies. Case in point, RNA-Seq can be utilized to recognize novel qualities/ isoforms/exons, elective graft locales, allele-particular expression, and uncommon transcripts in a solitary investigation.

Acknowledgment

I am grateful to Department of Biotechnology, R V College of Engineering and would like to thank Mr. Sanjay Deshpande for his valuable inputs and timely technical guidance.

References

1. L. A. Goff, C. Trapnell, and D. Kelley, "CummeRbund : Visualization and Exploration of Cufflinks High-throughput Sequencing Data," 2014.
2. B. J. Haas and M. C. Zody, "news and views Advancing RNA-Seq analysis," Nat. Publ. Gr., vol. 28, no. 5, pp. 421–423, 2010.
3. R. K. Patel and M. Jain, "NGS QC Toolkit : A Toolkit for Quality Control of Next Generation Sequencing Data," vol. 7, no. 2, 2012.
4. J. J. Li, H. Huang, M. Qian, and X. Zhang, "NEXT-GENERATION SEQUENCING," 2008.
5. J. B. W. Wolf, "Principles of transcriptome analysis and gene expression quantification : an RNA-seq tutorial," pp. 559–572, 2013.
6. R. Leinonen, H. Sugawara, and M. Shumway, "The sequence read archive," vol. 454, pp. 1–3, 2010.
7. U. Nagalakshmi, K. Waern, and M. Snyder, "RNA-Seq : A Method for Comprehensive Transcriptome Analysis," no. January, pp. 1–13, 2010.
8. F. Tang, C. Barbacioru, E. Nordman, B. Li, N. Xu, V. I. Bashkirov, K. Lao, and M. A. Surani, "RNA-Seq analysis to capture the transcriptome landscape of a single cell," Nat. Protoc., vol. 5, no. 3, pp. 516–535, 2010.
9. C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," Nat. Protoc., vol. 7, no. 3, pp. 562–578, 2012.