Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 5, Issue, 07, pp.1752-1755, July, 2013

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

BIOINFORMATICS ANALYSIS OF RNA SEQUENCE DATA FOR POTENTIAL BIOMARKER DISCOVERY

***Supreetha, K. V., Shambhu, M. G. and Kusum Paul**

Department of Biotechnology, The Oxford College of Engineering, Bommanahalli, Hosur road, Bangalore – 560 068, Karnataka state, India

ARTICLE INFO

Article History:
Received 23rd April, 2013
Received in revised form
26th May, 2013
Accepted 18th June, 2013
Published online 18th July, 2013

Key words:
RNASeq., Biomarker,
NGS data, TopHat,
Cufflinks, FPKM value.

ABSTRACT

Normal and disease-specific genes can be identified by making use of the existing microarray and RNA (NGS) sequencing data. The study attempts to identify potential biomarkers for the mammalian testis. Available NGS analysis database and software were first compiled and compared. NGS data corresponding to normal human and mouse testes have been collected after careful selection. The read data were aligned using Top Hat tool and 'Cufflinks' & 'Cuffcompare' tools were then used to obtain the transcript-specific FPKM values. The processed data were compared across different reports, and a score has been assigned to identify the consistent transcripts. The final transcript list was then compared to the microarray data from MgEx-Tdb. The top-scoring transcripts have been identified as the potential biomarkers for the normal human testis. In addition, potential marker transcripts have also been identified for specific testis cell types in the mouse. These in turn could help in detecting any abnormality, and also assist in research towards male contraception. This study has identified four genes which are ACTB, SMC6, PCMT1, SON along with their transcripts as potential biomarkers for normal human and mouse testis. Only certain transcripts show higher expression in normal testis condition and hence genes-transcript pairs can be good biomarkers.

INTRODUCTION

RNA-sequencing refers to the use of high-throughput sequencing technologies to investigate the RNA content from a sample via the sequencing of cDNA. Next Generation Sequencing (RNASeq) provides detailed information about differential gene expression, alternatively spliced transcripts of RNA with deep coverage and base- level resolution. The most used high throughput technologies for RNA sequencing are 454 Sequencing, Illumina and Solid (Michael L. Metzker, 2010). There are also several databases such as Gene Expression Omnibus, Sequence Read Archive, which provides RNASeq data. A huge amount of tools are available across internet for processing/analysis of RNA sequencing data, which are scattered across internet (Naomi D. Elkin, 2010). Advances in technology (genomics, proteomics) have offered the promise of personalized medicine, where therapies and medical decision making can be finely tailored to patients. Potential benefits include improving clinical efficacy and decreasing toxicity by better treatment selection and patient selection. Biomarkers play a very important role in personalized medicine to understand and make decisions about the diseases and treatment. A biomarker is usually a molecule, in terms of its quantitative expression, which has a unique association with a specific tissue and condition. Biomarkers are crucial for basic as well as applied studies in terms of molecular characterization of the tissues, cells under specific conditions. Testis is the male gonad in animals. They are components of both the reproductive system and the endocrine system. The primary functions of the testes are to produce sperm and to produce androgens. It is a common organ for adverse drug effects leading to attrition of potential compounds. Data access and data analysis is an orphan area where more attention has to be shown. There are a huge number of databases and data analysis tools are present, but they are scattered across internet. Hence there is a need to do careful selection and analysis.

Corresponding author: Supreetha, K. V.
Department of Biotechnology, The Oxford College of Engineering, Bommanahalli, Hosur road, Bangalore – 560 068, Karnataka state, India.

METHODS

Almost all the RNA-Sequencing tools and data resources are listed in www.startbioinfo.com web portal. The corresponding rank by usage frequency is also given in the portal. Collect all the tools/resources and prepare a list. Tools/resources are compared based on rank by usage frequency and by selecting certain parameters such as number of samples available for testicular RNASeq. data.etc.

Collect information of RNASEQUENCE data for mammalian testis tissue

The testicular RNASeq data is retrieved from SRA database using different queries. Query for testis related studies: (testis OR testes OR sperm OR sperms OR spermatid OR spermatids OR spermatocyte OR spermatocytes OR spermatogonia OR spermatogenesis OR spermatozoa OR spermiogenesis OR leydig OR sertoli OR testicle OR testicles OR testicular OR "germ cell" OR "germ cells")

Download RNASEQ. data of mammalian testis tissue from the data resources

Since the size of the sample is big, downloading takes considerable amount of time. If internet connection is interrupted while downloading, again the downloading starts from the beginning. To overcome this problem a command: wget –c "link to be downloaded" is used. It resumes downloading the files whenever we run the command.

Conversion of data format: from SRA format to FASTQ format

The files downloaded from SRA database will be in SRA data format. To process the raw sample data it has to be in FASTQ data format. By using the command ./fastq-dump –gzip input_file –O output_file, SRA data files are converted to FASTQ files. The output of the

command will be in zipped format and it has to be unzipped after format conversion (Rasko Leinonen *et al*., 2011).

Quality check using FASTQC tool

Most high throughput sequence generators produce output in FastQ format. This format combines the base calls for the sequence which was generated with an encoded quality value for each base which says how confident the sequencer was that the base call generated was correct.Before proceeding with the analysis of a sequence data set, it is a good idea to do some basic quality control checks on the raw data to ensure that there are no hidden problems which might be more difficult to detect at a later stage.

Align RNA-SEQ reads to mammalian-sized genomes using TOPHAT

The sample sequence is aligned against the reference genome (Cole Trapnell *et al*., 2012).

Assemble transcripts, estimate their abundances using cufflinks

The transcripts obtained from the TOPHAT results are assembled and their level of expression has been estimated by using cufflinks and Cuffcompare (Cole Trapnell *et al*., 2009).

Class codes

Class codes tell the type of match between the Cufflinks transcript and the reference transcript. Hence sometimes single transcript can have multiple FPKM values based on its class code. Class codes along with their meaning have been provided in Table-1.

Table 1. Class codes and their meaning

Priority	Code	Description
1	=	Complete match of intron chain.
2	c	Contained.
3	j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript.
4		Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
5	i	A transfrag falling entirely within a reference intron.
6	o	Generic exonic overlap with a reference transcript.
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript).
8	r	Repeat.
9	u	Unknown, intergenic transcript.
10	x	Exonic overlap with reference on the opposite strand
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors).
12	0	Tracking file only.

Class codes tell the type of match between the Cufflinks transcript and the reference transcript.

FPKM value

FPKM (Fragments Per Kilo base of exon model per Million mapped fragments): Counts divided by transcript length (kb) times the total number of millions of mapped reads. FPKM values provides the estimated level of expression for each transcripts.

$$RPKM = \frac{\text{Number or reads in the Region}}{\frac{\text{Total Reads}}{1000000} \times \frac{\text{Region Length}}{1000}}$$

-------equation(1)

Conversion of ensemble id to gene symbol

Cuffcompare gives a list of expressed transcript ids along with other details such as class codes, cuffdiff_ids, FMI values, FPKM values, coverage values, etc. Here the transcript id will be in the form of ensemble id. Hence there is a need to convert these ids to gene symbols. Ensemble Genome Browser has a tool called BioMart which converts ENSEMBLE id to corresponding gene symbol. The ENSEMBLE ID converted to mgi_symbol is tabulated as shown in Table-2.

Table 2. ENSEMBLE ID along with corresponding gene symbols

	Ensembl_transcript_id	MGI_symbol
1	ENSMUST00000000001	Gnai3
2	ENSMUST00000000010	Hoxb9
3	ENSMUST00000000033	Igf2
4	ENSMUST00000000058	Cav2

Addition of FPKM values of multiple transcripts of the

Same gene

A single gene can have multiple transcripts due to alternative splicing. It can be obtained from the class codes. A single transcript can have multiple FPKM values due to corresponding class codes. To obtain actual level of expression, add the FPKM values of multiple transcripts of the same gene. A Perl code has been written to perform the addition of similar class code fpkm values.

Confidence score calculation

To calculate the confidence of genes/transcripts to be transcribed a scoring system is developed based on the expression values of the transcripts. The idea of scoring is, the level of expression i.e. average FPKM value should be high and the standard deviation between the levels of expression should be less (Table-3).

Table 3. A confidence scoring system.

Confidence Score = (((A)*1) + ((B)*5) + ((C)*10))* (D) + ((E) * 10)
Where,
A= count of FPKM values with a range 1-9
B= count of FPKM values with a range 10-99
C= count of FPKM values >100
D=AVG of corresponding FPKM values
E=1000-STDEV between corresponding FPKM values

Compare with microarray study (MgEx-Tdb) to identify potential biomarker(s)

There is a Mammalian gene Expression- Testis database which measures the extent of agreement or contradictions for each gene's expression status by using a consistency score/reliability score (Kshitish K Acharya *et al*., 2010). The scores obtained from the new scoring system have been compared with the MgEx-Tdb scores. The genes/transcripts with highest confidence score present in both the scoring methods i.e. MgEx-Tdb score and RNASeq score are highly transcribed genes/transcripts and are selected as biomarkers for the respective states (cell-type/ condition).

RESULTS AND DISCUSSIONS

From the comparison it has been noticed that SRA, DRA, ENA and GEO databases are inter-related. SRA database is selected as best for RNASeq. /NGS data analysis. Based on rank by usage frequency and based on the literature survey, TopHat and Cufflinks tools are selected as best tools for processing raw RNASeq.NGS data (Acharya *et al*., 2008). Information on mammalian testis tissue has been collected from SRA

Table 4. Biomarkers for normal human testis tissue

MgEx-Tdb Score	RANK(out of 100)	Gene symbol	transcript id	Confidence score	RANK(out of 100)	TOTAL RANK
138.4327	100	ACTB	ENST00000425660	1917.08305	98	198
138.4327	100	ACTB	ENST00000484841	149.176379	97	197
138.4327	100	ACTB	ENST00000331789	10445.0014	100	200
138.4327	100	ACTB	ENST00000462494	153.707979	97	197
126.995	92	SMC6	ENST00000402989	9671.48646	100	192
126.995	92	SMC6	ENST00000448223	10002.8003	100	192
125.3341	91	PCMT1	ENST00000367384	9936.37305	100	191
125.3341	91	PCMT1	ENST00000367378	9858.71731	100	191
125.3341	91	PCMT1	ENST00000464889	9979.70773	100	191
124.6102	91	SON	ENST00000356577	21.227878	97	188
124.6102	91	SON	ENST00000436227	41.010897	97	188
124.6102	91	SON	ENST00000300278	20.001728	97	188
124.6102	91	SON	ENST00000381679	9667.96093	100	191
124.6102	91	SON	ENST00000290239	9798.77866	100	191
124.6102	91	SON	ENST00000467616	34.754637	97	188

The above table provides list of identified potential biomarkers genes and transcript pairs along with their MgEx-Tdb score and confidence score obtained from the new scoring method.

Table 5. Biomarker list for normal mouse testis cell-types

				testis/primitive type A spermatogonia (germ cell)		testis/primitive type B spermatogonia (germ cell)	
MgEx-Tdb Score	RANK (out of 100)	Gene symbol	transcript id	Confidence score	RANK (out of 100)	Confidence score	RANK (out of 100)
138.4327	100	Actb	ENSMUST00000100497	155876.636	1	158038.488	98
126.995	93	Smc6	ENSMUST00000020931	27331.1303	1	32052.7032	98
125.3341	92	Pcmt1	ENSMUST00000159917	10237.7216	1	10116.0385	98
125.3341	92	Pcmt1	ENSMUST00000159977	4.840998	1	0	
125.3341	92	Pcmt1	ENSMUST00000162606	10246.6519	1	10224.2689	98
125.3341	92	Pcmt1	ENSMUST00000162682	10003.2284	1	0	
125.3341	92	Pcmt1	ENSMUST00000163085	11609.0707	1	11365.0726	98
125.3341	92	Pcmt1	ENSMUST00000161123	990.336	1	0	
125.3341	92	Pcmt1	ENSMUST00000160250	0		0	
125.3341	92	Pcmt1	ENSMUST00000161428	0		0	
125.3341	92	Pcmt1	ENSMUST00000159835	0		0	
124.6102	92	Son	ENSMUST00000114036	33249.1099	1	1271.24276	98
124.6102	92	Son	ENSMUST00000117633	3087.34568	1	3347.04806	98
124.6102	92	Son	ENSMUST00000140312	12966.821	1	11353.2567	98
124.6102	92	Son	ENSMUST00000147891	41502.5568	1	12094.7797	98
124.6102	92	Son	ENSMUST00000119368	867.201906	1	954.024368	98
124.6102	92	Son	ENSMUST00000114037	0		761.552186	98
124.6102	92	Son	ENSMUST00000151503	0		0	

The above table provides list of identified potential biomarkers genes and transcript pairs along with their MgEx-Tdb score and confidence score obtained

database. A total of 50 human, 31 mouse and 4 rat normal samples are available from SRA database are freely downloadable. From the quality check observed that the cutoff value for good quality score should: Total number of passes = > 4 A scoring system has been developed to suggest the degree of reliability/confidence. Confidence score for genes/transcripts are compared among RNASeq. study and micro-array study i.e. MgEx-Tdb score. The genes having similar scores between both the studies have more reliability. The transcripts/genes having highest score for its transcribed state from both the methods i.e. MgEx-Tdb and RNASeq are selected as biomarker genes having a potential to help for clinical diagnosis of normal state. Table-4, Table-5 shows the biomarker genes for normal human and mouse testis along with their transcripts and scores for transcribed state respectively. The transcripts of ACTB, SMC6, PCMT1, SON genes are having highest score. From the analysis it is clear that these genes have very high expression compared with the other genes. Hence they can be having more potential to become a biomarker. The graph in Figure-1, Figure-2 shows the level of expression for each transcript. A single gene can have multiple transcripts due to alternative splicing (Jennifer Michalowski, 2001). But the level of expression varies between each transcript across different tissues. The above graph shows gene expression along with their transcripts for normal human testis tissue.

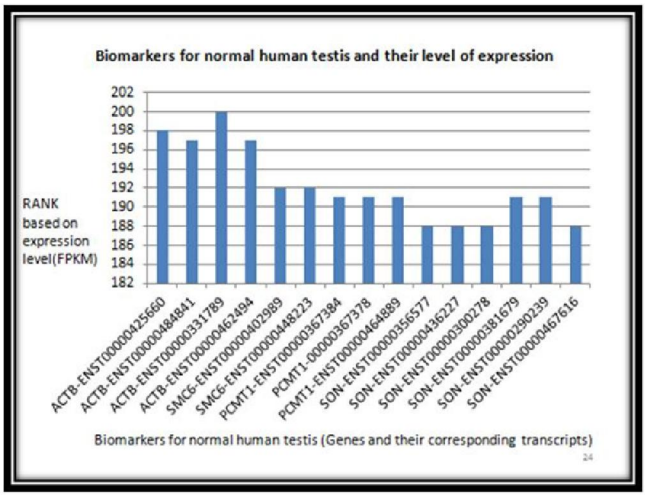

Fig.1. Transcript level expression of human biomarker genes/transcripts

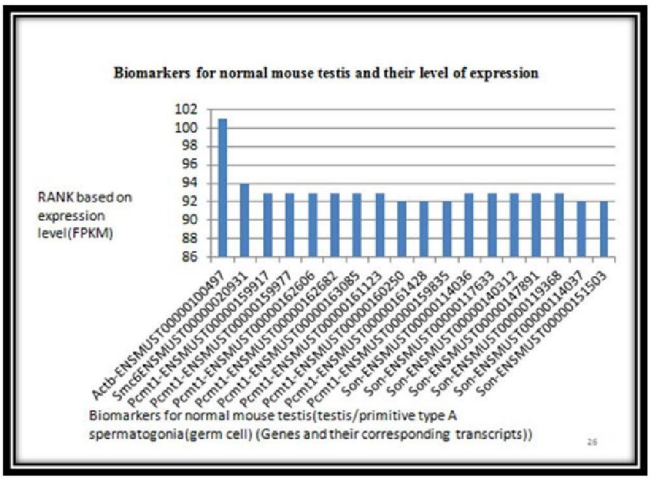

Fig.2. Transcript level expression (mouse-testis/primitive type-A spermatogonia (germ cell))

Conclusions

Analysis of mammalian normal testis tissue at the transcript level gives detailed information of gene expression. A gene can have multiple transcripts due to alternative splicing. Expression of each transcript under a single gene varies from each other. Due to ubiquitous expression of genes across multiple tissues, it is better to study transcript level gene expression compared to gene level expression. This study has identified four genes which are ACTB, SMC6, PCMT1, SON along with their transcripts as potential biomarkers for normal human and mouse testis. Pathway analysis has been done as a further analysis for the selected potential markers to give more proof on their expressions in different state/conditions. Variations in the expression of transcripts of the genes can disrupt normal gene expression and may cause disorders.

Only certain transcripts show higher expression in normal testis condition and hence genes-transcript pairs can be good biomarkers.

Acknowledgments

We extend our sincere thanks to the management of The Oxford College of Engineering for their support. We thank our principal Dr. Nagaraj and our Head of the Department, Department of Biotechnology for providing necessary resources and valuable suggestions.

REFERENCES

[1] Acharya K. K, Greta K. *et al.,* 2008, "A comparative analysis of 21 literature search engines", Nature precedings.

[2] Cole Trapnell, Adam Roberts *et al.,* 2012, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks".

[3] Cole Trapnell, Lior Pachter *et al.,* 2009 "TopHat: discovering splice junctions with RNA-Seq", bioinformatics/btp, 1105–1111. 120.

[4] Jennifer Michalowski, 2001, "ALTERNATIVE SPLICING". Jorge S Reis-Filho, 2009, "Next-generation sequencing", Breast Cancer Research.

[5] Kshitish K Acharya, Darshan S Chandrashekar *et al.,* 2010, "A novel tissue-specific meta-analysis approach for gene expression predictions, initiated with a mammalian gene expression testis database", BMC Genomics, 11:467.

[6] Michael L. Metzker, 2010, "Sequencing technologies - the next generation", Nature Reviews.

[7] Naomi D. Elkin, 2010, "Evaluation of Biomarkers for Testicular Toxicity".

[8] Rasko Leinonen, Hideaki Sugawara *et al.,* 2011, "The Sequence Read Archive", Nucleic Acids Research.

****