

## Identification of SNPs in Goats (*Capra hircus*) using RNA-Seq Analysis

<sup>1</sup>Upasna Sharma, <sup>2</sup>Priyanka Banerjee, <sup>2</sup>Jyoti Joshi and <sup>2</sup>Ramesh Kumar vjhh

<sup>1</sup>Singhania University, Pachheri Bari, Jhunjhunu-313515, Rajasthan, India

<sup>2</sup>National Bureau of Animal Genetic Resources (ICAR), Karnal-132001 India

**Abstract:** Single Nucleotide Polymorphisms (SNPs) have become the marker of choice for genome-wide association studies. In order to provide the best genome coverage for the analysis of performance and production traits, a large number of relatively evenly distributed SNPs are needed. Gene-associated SNPs may fulfill these requirements of large numbers and genome wide distribution. In addition, gene-associated SNPs could themselves be causative SNPs for traits. The main objective of our work was to identify large numbers of gene-associated SNPs using high-throughput next generation sequencing. Transcriptome sequencing was conducted on 2 tissues viz. liver and kidney for 5 breeds of goat (Kanniadu, Osmanabadi, Black Bengal, Changthangi and Sirohi) using Illumina next generation sequencing technology. Approximately 46.4 million reads for Black Bengal, 61.9 from Kanniadu, 58.2 from Changthangi, 47.3 from Osmanabadi, 73.2 from Sirohi were obtained by sequencing gene transcripts derived from kidney while 37, 27.2, 19.4, 56.9 and 80.7 million reads were obtained by gene transcripts derived from liver. The analysis of total number of SNPs in liver and kidney revealed that out of a total of 68597 SNPs in liver, the total number of transversions was 21300 and the number of transitions was 47297. A total of 1574 SNPs of liver were complex. Similarly for kidney the total number of 72047 SNPs were categorised into 22774 transversions and 49273 transitions. The total number of complex SNPs in kidney was 1597. The number of transitions is more than double the number of transversions in both the tissues. Further analysis of transversion revealed a preponderance of cytosine and guanine change compared to other nucleotides. 12863 and 11319 transversions out of 21300 and 22774 transversions respectively for liver and kidney revealed this bias. When multiple individuals with different genetic backgrounds were used, RNA-Seq was very effective for the identification of SNPs. The SNPs identified in this report provides a much needed resource for genetic studies in goat and shall contribute to the development of a high-density SNP array. Validation and testing of these SNPs using SNP arrays will form the material basis for genome association studies and whole genome-based selection in goats.

**Keywords:** *Capra hircus*, transcriptome, RNA-Seq, Single nucleotide polymorphism

### INTRODUCTION

The application of genomics-type technologies is beginning to have an impact, enhancing our understanding of animal responses to external factors abiotic stresses that disturb the homeostatic equilibrium. The term genomics is any technology that, preferably in a high-throughput genome-focused fashion, promises insights and answers on how genes in animals, proteins, protein activities and metabolite type and flux respond to external factors. Thermal stresses trigger a complex program of gene expression and biochemical adaptive responses (Fujita, 1999; Lindquist, 1986). Biologically, the ability to survive and adapt to thermal stress appears to be a fundamental requirement of cellular life, as cell stress responses are ubiquitous among livestock species as in case of goats which thrive in different climatic regions ranging from hot dry regions with a temperature of +45°C to semi humid, humid and to even cold dry regions of Leh and Himachal Pradesh with sub 0 temperatures. In addition,

the observation that induction of a cell stress response by one type of stressor (such as heat/cold) often leads to cross-protection to other stressors has raised the interesting possibility that the pathways involved in the cell stress response might present useful targets. Inevitably a high degree of understanding is required to exactly fish out the genes involved and the pathways that are shared by various stressors and of those that are unique to particular stressor. It is widely accepted that changes in gene and thereby in the protein sequences might be an integral part of the cellular response to thermal stress. Adaptability to various thermal stresses is likely to be the result of changed cellular responses, changed protein structure among different breeds of goats. The changed proteins may result from non-synonymous SNPs and adapted to the specific geographical location. A simpler and potentially more comprehensive way to measure transcriptome composition and to discover new exons or genes is by direct ultra-high-throughput sequencing (RNA-Seq approach). RNA-Seq is a powerful new method for

mapping and quantifying transcripts developed to analyze global gene expression in different tissues and also used to detect unannotated transcriptional activity, to differentiate between different transcriptional or splicing isoforms and to provide digital measurements at single base resolution. Recently, this technique has also been used as an efficient and cost-effective method to systematically identify SNPs in transcribed regions in different species (Chepelev *et al.*, 2009; Cirulli *et al.*, 2010; Cloonan *et al.*, 2008; Morin *et al.*, 2008a, b) as higher throughput expressed sequence reads are needed to increase coverage and depth and ensure sequence accuracy. RNA-Seq generates sequences on a very large scale at a fraction of the cost required for traditional Sanger sequencing, allowing the application of sequencing approaches to biological questions that would not have been economically or logistically practical before (Marguerat *et al.*, 2008). Taking this into account, we applied this novel approach to gene identification, polymorphism detection and transcript profiling of goat transcriptome. Our research work also provides the first report of high-throughput SNP detection and identification of genes and the pathways followed making them adaptable to various climatic regions in goats.

## MATERIALS AND METHODS

**Sample Collection and RNA isolation:** Liver and kidney tissue samples were obtained from Black Bengal (hot and humid region of West Bengal), Kanniadu (region with very small change in temperature during the year, South India, Tamil Nadu), Changthangi (Extreme cold region with temperature going to as low as minus 20 degree Celsius, cold desert region of northern part of India, Ladakh, Jammu and Kashmir), Osmanabadi (Central peninsular and semi arid regions of Maharashtra) and Sirohi goat (Arid regions of Rajasthan). The collected tissues were immediately frozen in dry ice and brought to lab and processed immediately for RNA extraction. The distribution of the animals (and also sample collection sites) in different climatic regions of India are shown below (Fig. 1).

RNA was isolated following the standard protocols of RNeasy Kit (Qiagen). The mRNA comprises only 1-3% of total RNA samples it was not readily detectable even with the most sensitive of methods. Quantification of RNA was done using Agilent 2100 Bioanalyzer (Agilent, Foster city, USA) which provided ng RNA/ $\mu$ l values. The clear 28S and 18S rRNA bands were indicative of intact RNA. The Bioanalyzer 2100 was used for all the ten isolated samples of RNA. The RNA samples were processed further if the RNA Integrity



Fig. 1: Map of India showing different breeds of goat in varied climatic zones

Number (RIN) was found to be greater than 8.5, for RNA-Seq library preparation.

### RNA-Seq library preparation and data generation:

Standard Illumina kit was utilised for DNA library preparation which facilitated reading both the forward and reverse template strands of each cluster during one paired-end read. In addition to sequence information, both reads contain long range positional information, allowing for highly precise alignment of reads. The paired-end sequencing assay utilized a combination of cBot (or the Cluster Station) and the paired-end module followed by paired-end sequencing on the Genome Analyzer Iix. The unique paired-end sequencing protocol allowed us the length of the insert (200–300 bp), generating high quality, align-able sequence data. A typical paired-end run could achieve  $2 \times 76$  bp reads and up to 40-60 million reads of the transcriptome data for each of the ten RNA samples (five each from kidney and liver). The liver and kidney tissues belonged to Osmanabadi, Changthangi, Sirohi, Black Bengal and Kanniadu breeds of goats. The image analysis, base calling and quality score calibration were processed using the Illumina Pipeline Software v1.4.1 according to the manufacturer's instructions. Reads were exported in the FASTQ format and used for further analysis.

**RNA-Seq data analysis:** The RNA-Seq data analysis was carried out using CLC Genomics Workbench (version 4.0.2; CLC bio). The quality filtered and

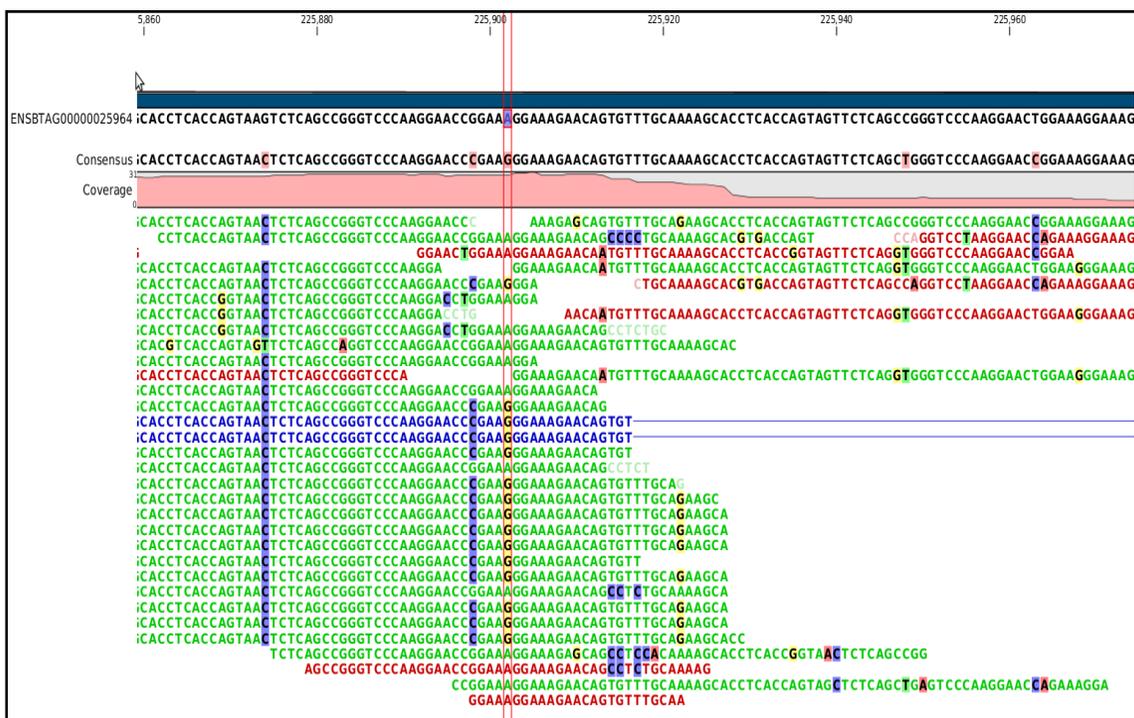


Fig. 2: Screenshot of CLC genomics workbench software showing SNP detection with *Bos taurus* as reference

Table 1: Summary of gene mapped and number of total reads utilised for two tissues

Goat breeds	Sequence length of data generated	Total reads (millions)		Total genes mapped	
		Kidney	Liver	Kidney	Liver
Black Bengal	2 X 76bp paired end reads	46.4	37	13604	12231
Kanniadu		61.9	27.2	7341	12519
Changthangi		58.2	19.4	4737	8126
Osmanabadi		47.3	56.9	12730	12199
Sirohi		73.2	80.7	12878	11949

trimmed reads were aligned to annotated *Bos taurus* genes downloaded from Ensembl Genome Browser (<http://www.ensembl.org/info/data/ftp/index.html>). For the genes downloaded, the default RNA-Seq settings were used to rank all potential matches, with mismatch cost of 2, deletion cost of 3 and insertion cost of 3. The highest scoring matches that shared  $\geq 80\%$  similarity with the reference sequence across  $\geq 50\%$  of their length were included in the alignment. This permissive alignment ensured that even reads derived from highly mutated orthodox between cattle and goat were not discarded. We utilised RNA-Seq tool of CLC Genomics Workbench and mapped the reads of two tissues from goats of five geographical locations (total of 10 tissues) separately and then merged alignment files of kidney and liver separately across the five geographical locations.

**SNP identification:** RNA-Seq alignment files for each of the 2 tissues of 5 different goats were scanned for

SNPs utilizing SNP detection module (Fig. 2) included in CLC Genomics Workbench (CLC bio, Aarhus, Denmark). The central base quality score of  $\geq 20$  and average surrounding base quality score of  $\geq 15$  were set to assess the quality of reads at positions for SNP detection. We utilized the criteria of depth of coverage of ten and the minor allele frequency of 2 out of 10 reads (minimum allele frequency of 20%) for the identification of SNPs. The complex SNPs were also identified. The SNPs were categorised into synonymous and non-synonymous SNPs and also transitions or transversions.

**Gene identification and ontology:** The genes with maximum number of SNPs were identified and classified on the basis of number of SNPs detected. The genes were mapped with the GO IDs. The characterization and manipulation of the gene lists resulting from mining high-throughput genomic data was done using g:Profiler (<http://biit.cs.ut.ee/gprofiler/>)

Table 2: Summary statistics of SNPs for two tissues- total number of SNPs, complex SNPs, non-synonymous SNPs, transitions, transversions.

S.no.	Tissue	Liver	Kidney
1	Total genes mapped using RNA-Seq	9575	10573
2	Total SNPs observed	68597	72047
3	Complex SNPs	1574	1597
4	Non-synonymous SNPs	13807	11779
8	Transitions	21300	49273
9	Tran versions	47297	22774
7	Functional annotation of genes showing SNPs	BP	224
		CC	87
		MF	72

Table 3: Number of genes with varying number of SNPs in the 2 tissues

S.No.	Tissue	Total number genes with SNPs in the range					
		11-15	16-20	21-25	25-30	31-40	Above 40
1	Liver	587	281	170	85	92	205
2	Kidney	805	359	195	129	113	158

for classification into Biological process, Cellular component and Molecular function. REVIGO web server (<http://revigo.irb.hr/>) was utilized for summarizing the long lists of Gene Ontology terms and then by removing redundant GO terms. The remaining terms were visualized in semantic similarity-based scatter plots, interactive graphs and tag clouds.

## RESULTS

### Generation and assembly of expressed short reads:

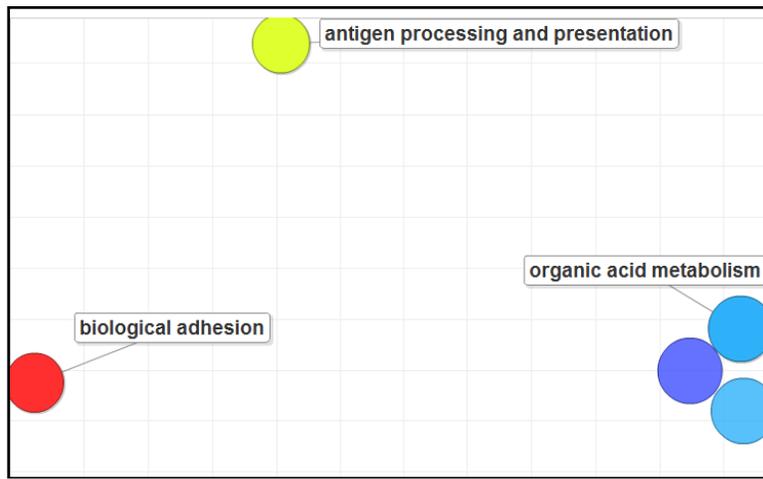
Illumina sequencing generated short sequence reads of expressed sequences. cDNA libraries were made from RNA samples prepared from 2 tissues of Black Bengal, Kanniadu, Changthangi, Osmanabadi and Sirohi goat. The mRNAs were sequenced with one lane for each tissue using Illumina GAIx and it generated 46.4, 61.9, 58.2, 47.3 and 73.2 million 2 X 76-bp paired-end reads in Kidney and 37, 27.2, 19.4, 56.9 and 80.7 million reads for liver in the above mentioned breeds respectively (Table 1). The reads were aligned against using annotated *Bos taurus* genes from Ensembl browser (<http://www.ensembl.org/info/data/ftp/index.html>) as a reference. In Black Bengal, 13604 genes mapped in kidney, while for Kanniadu maximum 12519 genes were mapped for liver. In Changthangi goats, 8126 genes mapped in liver, in Osmanabadi, maximum number of 12730 genes mapped in kidney and in Sirohi 12878 genes mapped in kidney (Table 1). The number of reads utilised and the number of genes mapped for each breed and tissue are summarised below.

**SNP identification:** From the RNA-Seq output, we also estimated the depth of each nucleotide (A, G, C, or T) at each base position with respect to gene sequence. For reducing false positive SNPs, we filtered potential SNPs using a stringent nucleotide depth cut off of 10. Large number of SNPs was detected in both the tissues. A total of 72047 SNPs were found in kidney and 68597

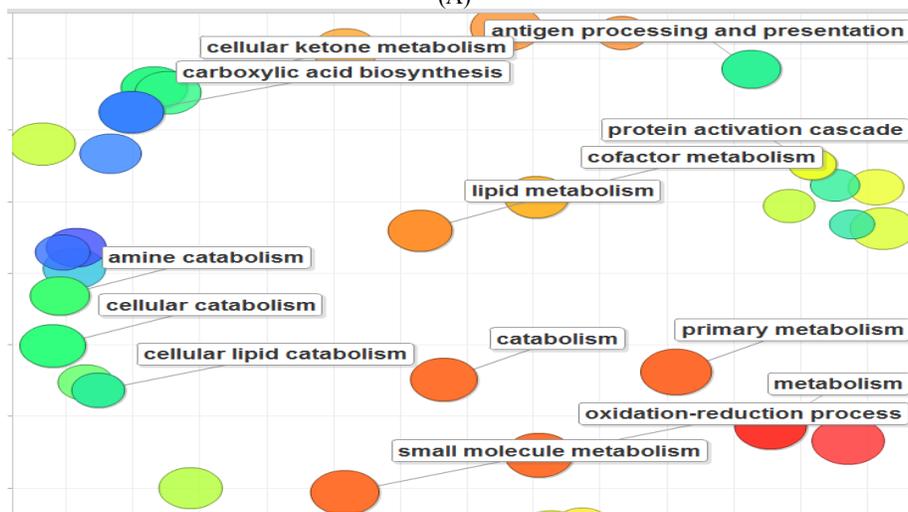
in liver. In depth analysis of the SNPs in kidney revealed 22774 transversions and 49273 were transition. The SNPs of liver revealed 21300 transversions and 47297 transitions (Table 2). Of these SNPs, 1597 were complex SNPs in kidney and 1574 in liver respectively. Allele frequencies for heterozygous SNPs were obtained for the tissue samples by counting the number of reads representing each allele. Summarizing the information revealed 3835 SNPs had an allele frequency of 60/40, 4144 SNPs had an allele frequency of 50/50 in kidney. The details for rest of the SNPs are given in Table 3.

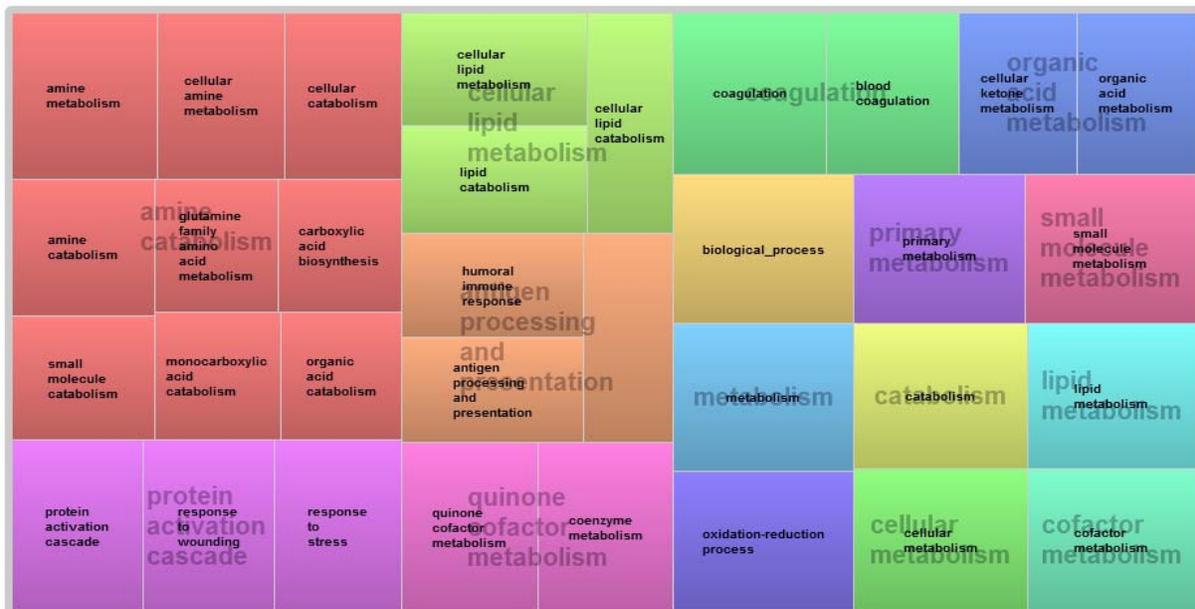
Among the SNPs recorded, maximum number of 570 SNPs was found in ENSBTAG00000026917 gene of Liver which shows serine- type end peptidase inhibitor activity and 250 SNPs was found in ENSBTAG0000002069 gene of kidney which is involved in the presentation of foreign antigens to the immune system. The genes with synonymous and non-synonymous SNPs were mapped on GO IDs using the web server g: Profiler and then these genes were summarised and visualised using REVIGO web server for removal of redundant GO terms.

In kidney, the genes with non-synonymous SNPs mapped on GOs associated with biological components (Fig. 3A) responsible for cell adhesion; antigen processing and presentation; organic acid, carboxylic acid and cellular ketone metabolism. The cellular components (Fig. 4A) were related to MHC protein complex and intracellular membrane bound organelle. The molecular function (Fig. 5A) was related to vitamin, ion, zinc- ion, cofactor and integrand binding; catalytic, oxidoreductase, GTPase regulator and hydrolase activity. The genes with synonymous SNPs were responsible for translational initiation, vesicle mediated transport, response to chemical stimulus, monocarboxylic acid metabolism (biological process) given in Fig. 6A; cell periphery, cell projection,



(A)



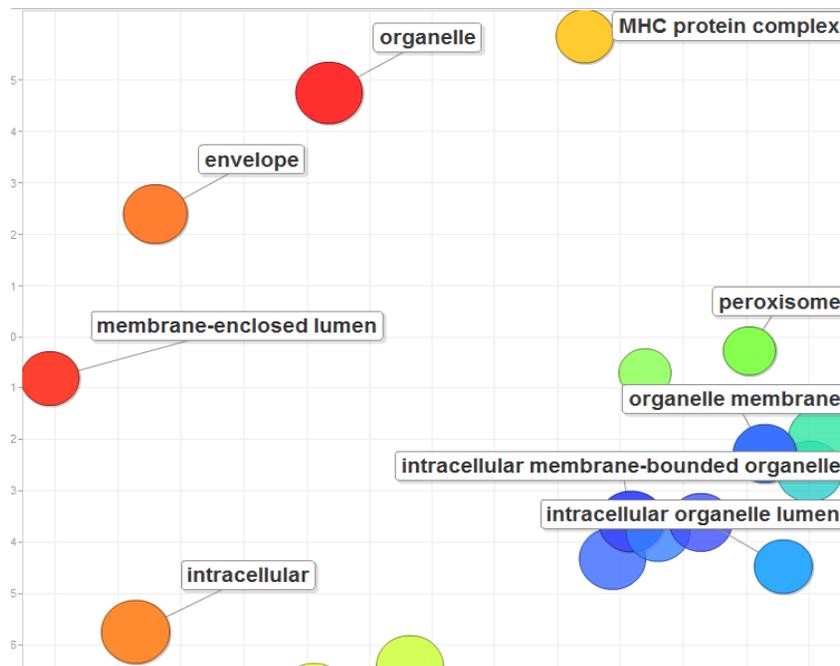


(B)

Fig. 3: Network and tree map graphs of biological process of genes with non-synonymous SNPs in A) Kidney and B) Liver tissue, respectively



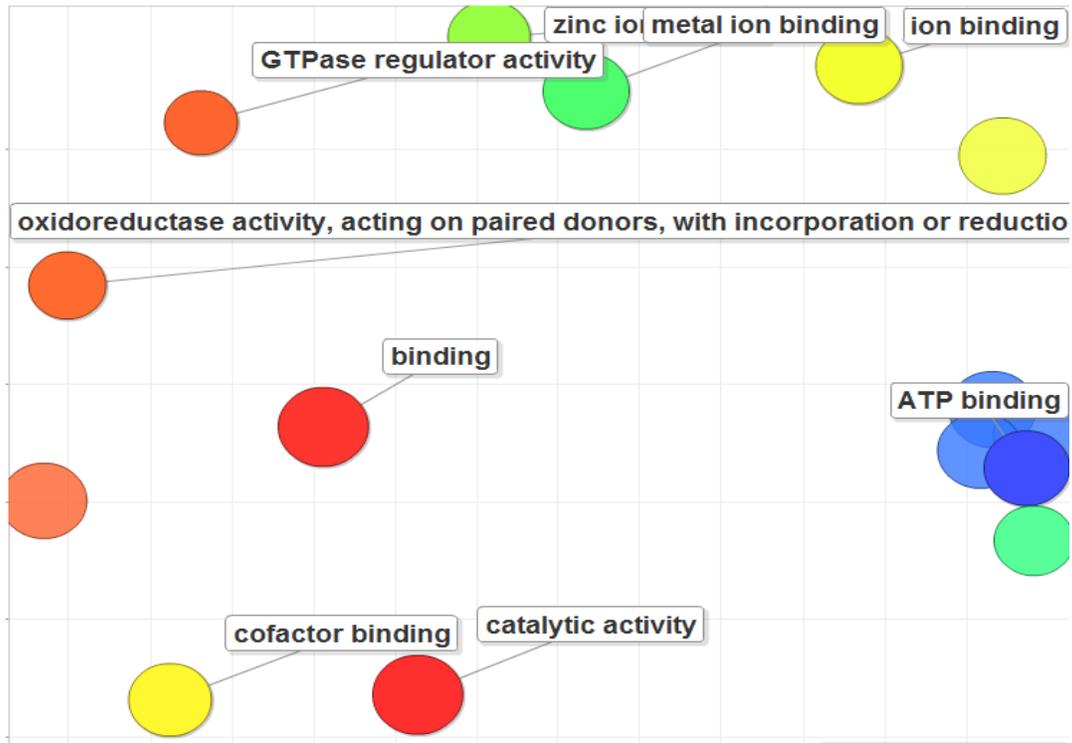
(A)



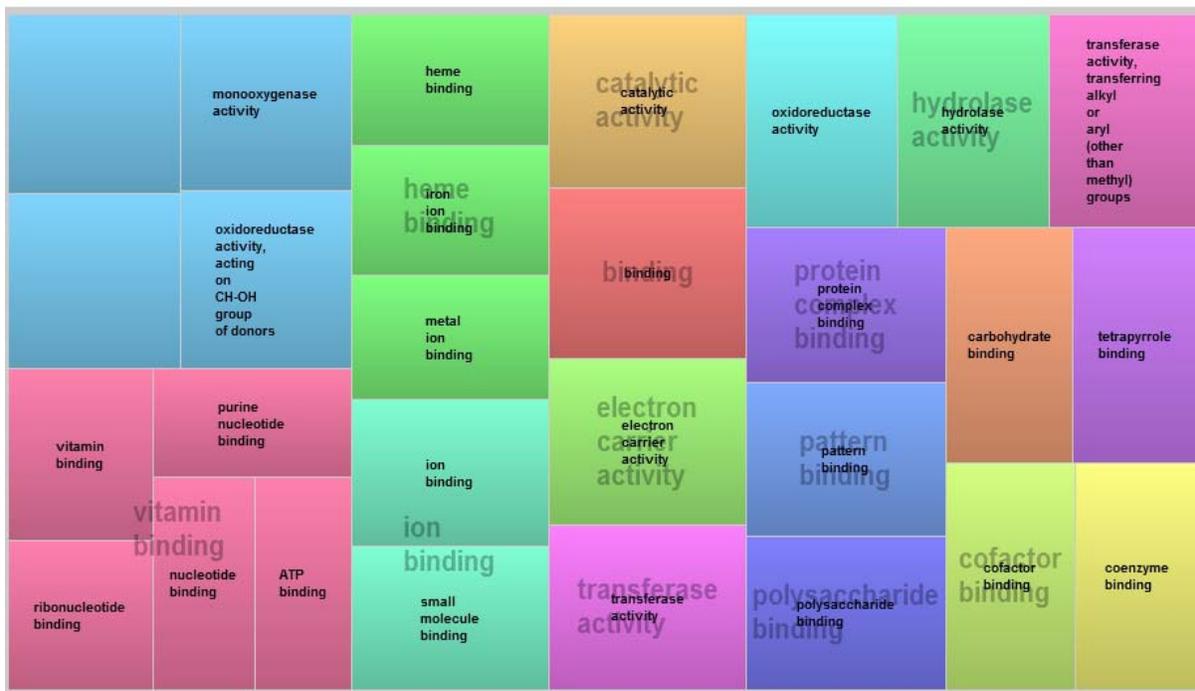
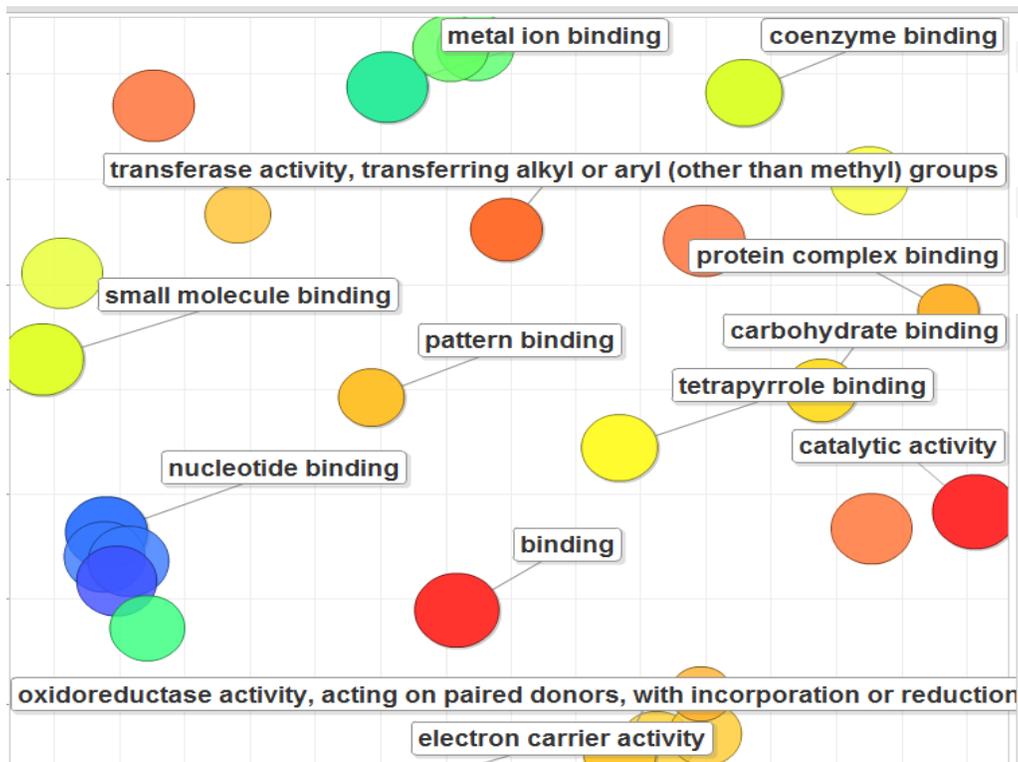
(B)

Fig. 4: Network and tree map graphs of cellular component of genes with non - synonymous SNPs in A) Kidney and B) Liver tissue, respectively

adherens junction (cellular process), nuclear hormone receptor binding, vitamin binding, ion binding etc (molecular function). In case of liver, the non-synonymous SNPs mapped on GOs associated with

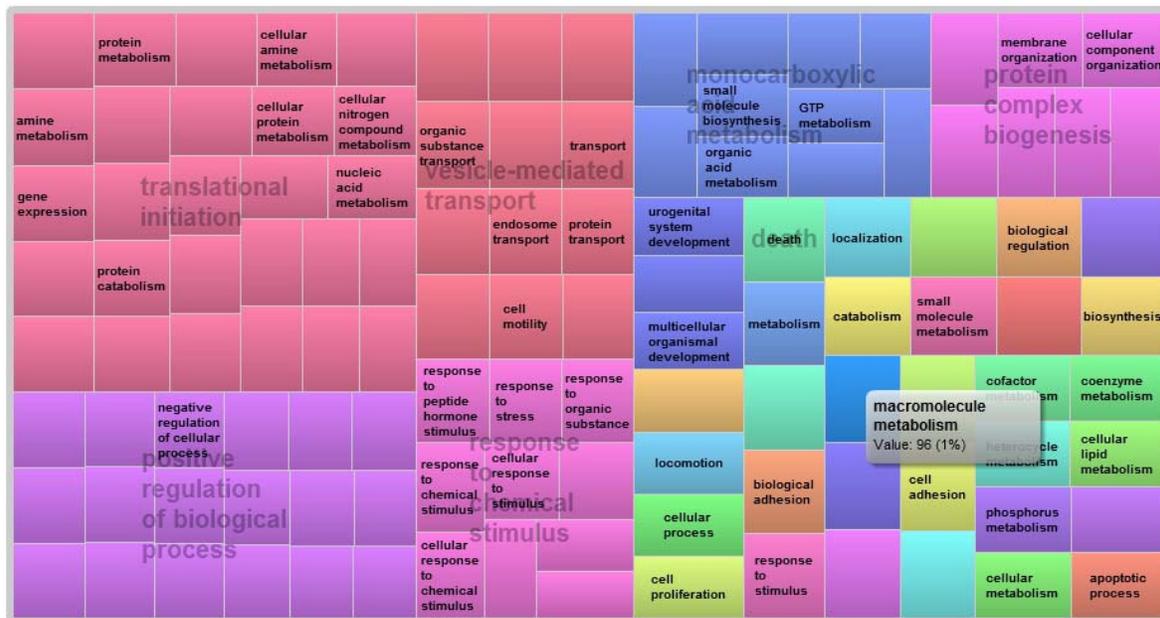


(A)

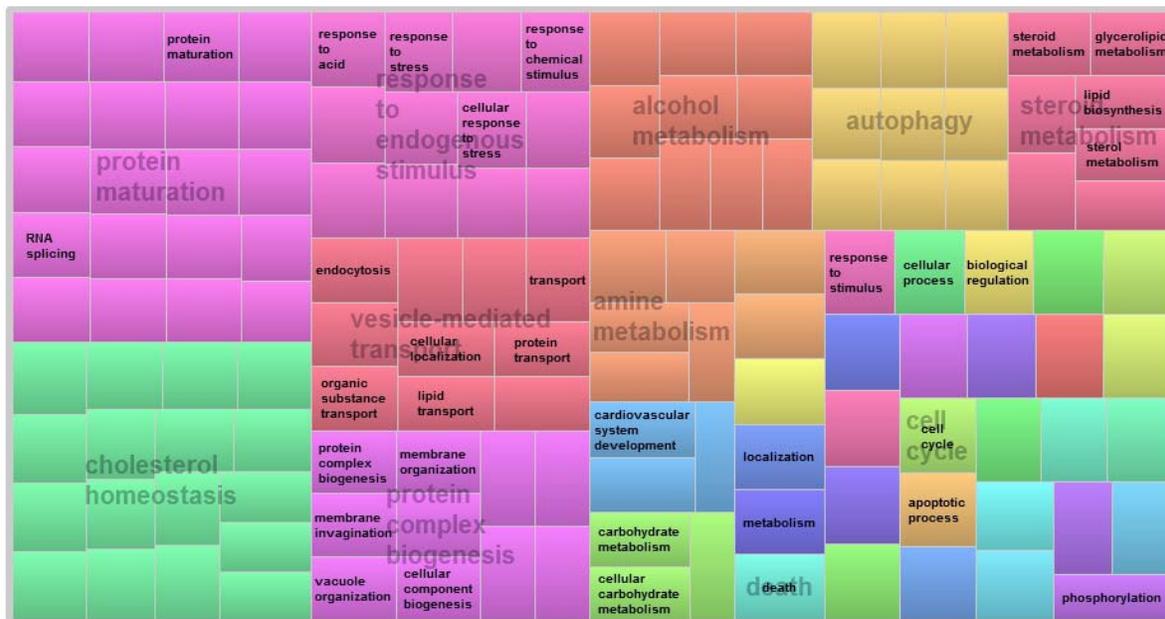


(B)

Fig. 5: Network and tree map graphs of molecular function of genes with non -synonymous SNPs in A) Kidney and B) Liver tissue, respectively



A) Kidney



B) Liver

Fig. 6: Tree map graphs of the genes with synonymous SNPs related to biological processes in Kidney and Liver, respectively

biological components (Fig. 3B) responsible for protein activation cascade, cellular lipid metabolism, quinone cofactor metabolism, organic acid metabolism. Cellular components (Fig. 4B) were related to peroxisomes, MHC protein complex while molecular function components (Fig. 5B) were related to transferase activity, catalytic activity, vitamin binding, hydrolase

activity polysaccharide binding etc. In synonymous SNPs of liver tissue, biological components (Fig. 6B) include protein maturation, cholesterol homeostasis, response to endogenous stimulus, autophagy, vesicle mediated transport; cellular components include adheren junctions, perinuclear regions of cytoplasm; molecular functions are related to identical protein

binding, hydrolase activity acting on ester bonds, lyase and ligase activity; protein, RNA and phospholipid binding etc.

## DISCUSSION

In this study, we carried out RNA-Seq analysis with RNA samples from multiple individuals of goats belonging to diverse geographical locations of the country. The analysis using RNAseq has been done with an aim of developing large numbers of high-quality SNPs. A total of 46.4 million reads for Black Bengal, 61.9 from Kanniadu, 58.2 from Changthangi, 47.3 from Osmanabadi, 73.2 from Sirohi from kidney while 37, 27.2, 19.4, 56.9 and 80.7 million reads were obtained by gene transcripts derived from liver, allowed generation of a total of approx 7-11 billion nucleotide bases (values vary in different tissues) of expressed sequences in kidney and approx 3-12 billion nucleotide base in liver. Our results demonstrate the efficiency and cost-effectiveness of next generation sequencing technologies in generating expressed sequences. One great challenge of using Illumina sequencing for transcriptome analysis is the short read length. In this study, we have used the Illumina GA-IIx sequencing platforms that generated read lengths of 2 X 76 base pair. A second challenge is the over representation of highly expressed gene tags in transcriptome analysis. In our present case since the mapping was done on the annotated genes of cattle our mappings are obviously limited to the annotated genes. The reads which do not belong to the reference genes shall not be mapped and discarded. This may also be primarily the reason for not detecting a large number of SNPs as only annotated genes were considered. Pooling of RNA samples was not exercised in our present study since the purpose was to analyse the transcriptome of goats from different geographical locations and comparing them. However we pooled the mappings of 2 tissues and carried out the SNP identification. Thus the merging of file of mapped genes actually tantamount to a situation of merging the tissues followed by next generation sequence data generation which is among the most efficient methods for SNP identification.

In this present study we have emphasised on the tissue specific SNP detection for kidney and liver. Large number of SNPs obtained in the present study belongs only to the annotated genes and their values are much higher than the expected values. The possible reasons are that we have taken unrelated and completely differentiated goat breeds of the country. These goats have adapted to harsh climatic regions of the country and have evolved independent to one another because of large differences in temperature and

humidity conditions. They may reflect the real SNPs between goat breeds. These may likely be affected by 1) RNA editing 2) they may likely belong to sequence errors in spite of sequence quality filtration 3) they may be derived from the mapping error, or they were simply mistaken by a reference sequencing error. We point out that, according to present data, we could not clearly determine how many SNPs are doubtless due to RNA editing. Using both genomic DNA and RNA from the same organism of the same variety for deep sequencing with higher base quality might be a better way to detect and analyze RNA editing. In case of goat the genome sequence is not available and we have utilised the orthologous genes of cattle to map goat transcripts.

## CONCLUSION

Transcriptome analysis using next generation sequencing is among the most efficient methods for SNP identification. This efficiency is even more obvious when considering filtered (high-quality) SNPs. Given the large numbers of SNPs generated that meet these minimal requirements, more stringent parameters can be applied in picking SNP sets for different applications. Average depth at putative SNP positions is greater than 100 sequences, providing high confidence in accuracy of identified SNPs. Resequencing or limited validation of these samples by low throughput SNP genotyping is costly and is unlikely to generate additional information. Ultimately, SNPs need to be validated by genotyping in a variety of reference mapping families and trait-selected populations using a high-density screening array. Genome-wide association studies of complex traits require a large number of SNPs. However, for research communities focused on non-model organisms, it is cost-prohibitive to genotype all SNPs in an association study with the throughput of current technologies. Selection of uniformly distributed SNPs across the genome for association studies is therefore, very important. The approach to sample animals of diverse genetic backgrounds and sequence to sufficient depth for reliable SNP identification allowed the ability to detect many common SNPs across the entire genome. We have demonstrated that transcriptome analysis RNA samples using RNA-Seq and then merging the alignment files from multiple individuals is both technically efficient and cost-effective for generating expressed sequences and identification of SNPs.

## ACKNOWLEDGMENT

The study was funded by Indian Council of Agricultural Research grant C 30033 under National

Agricultural Innovative Project Component IV, Basic and Strategic Research which is gratefully acknowledged.

#### REFERENCES

- Chepelev, I., G. Wei, Q. Tang, K. Zhao, 2009. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucl. Acid. Res.*, 37: 106.
- Cirulli, E.T., A. Singh, K.V., G.D.L. Shianna, J.P. Smith, J.M. Maia, E.L. Heinzen, J.J. Goedert, D.B. Goldstein, 2010. Immunology CHAV: Screening the human exome: A comparison of whole genome and whole transcriptome sequencing. *Genome Biol.*, 11(5): 57.
- Cloonan, N., A. Forrest, G. Kolle, B. Gardiner, G. Faulkner, 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods.*, 5: 613-619.
- Fujita, J., 1999. Cold shock response in mammalian cells. *J. Mol. Microbiol. Biotechnol.*, 1: 243-255.
- Lindquist, S., 1986. The heat-shock response. *Annu. Rev. Biochem.*, 55: 1151-1191.
- Marguerat, S., B.T. Wilhelm, J. Bahler, 2008. Next-generation sequencing: Applications beyond genomes. *Biochem. Soc. Trans.*, 36: 1091-1096.
- Morin, P.A., G. Luikart, R.K. Wayne and S.W. Grp, 2008a. SNPs in ecology, evolution and conservation. *Trends Ecol. Evol.*, 19(4): 208-216.
- Morin, R., M. O'Connor, M. Griffith, F. Kuchenbauer, A. Delaney, 2008b. Application of massively parallel sequencing to micro RNA profiling and discovery in human embryonic stem cells. *Genome. Res.*, 18: 610-621.