

RESEARCH PROJECT FORMULATION

**Transcriptome Analysis of growth traits in mithun
through RNA-seq approach and Whole
Genome Sequencing of Mithun
(*Bos frontalis*)**

(Inter-Institutional collaborative project)

Submitted By

**Dr. Sabyasachi Mukherjee
Sr. Scientist**

**Animal Genetics and Breeding
National Research Centre on Mithun
Jharnapani, Medziphema
Nagaland 797 106, INDIA**

PROFORMA FOR SUBMISSION OF RESEARCH PROJECT

Part-I: General Information

200 Project code

2001 Institute project code:

2002 ICAR project code: (PIMS-ICAR):

201 Name of the Institute and Division

2011 Name and address of Institute: **NRC ON MITHUN (ICAR)**

2012 Name of Division/Section:

2013 Location of the project: NRCM, Nagaland, INDIA

2014 Location of the project: CARI, Izatnagar (UP), INDIA

202 Project title:

2021 Sub project title: **Transcriptome analysis through RNA-seq approach and Whole Genome Sequencing of Mithun (*Bos frontalis*) (Inter-Institutional collaborative project)**

203 Priority area: Molecular Genetics

204 Research approach:

Applied research/ Basic research/ Process or Tech. Dev./TOT

01

02

03

04

Applied research (01)

205 Specific area: Molecular Characterization of Mithun

206 Duration of project: 3 Years

2061 Date of start of the project: 01.2.2013

2002 Likely date of completion of the project: 31.3.2016

207 Total cost of the project: **Rs. 91.00 lakhs**

2071 Foreign exchange component: Nil

208 Profile summary:

Gene expression differs markedly between muscle types [Cassar-Malek *et al.* 2005] due to the fact that the muscle tissue is a composition of many cell types (including myofibres, connective tissue fibroblasts and adipocytes) which differ in their proportions between individual muscles.

Hence, genomic studies may help to identify genes, especially those which show significant changes in expression in different environments, suggesting that their expression level depends mainly on genetic factors.

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition. Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease. The key aims of transcriptomics are to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs; to determine the transcriptional structure of genes, in terms of their start sites, 5' and 3' ends, splicing patterns and other post-transcriptional modifications; and to quantify the changing expression levels of each transcript during development and under different conditions.

The present project has been formulated to perform transcriptome analysis of mithuns by high throughput RNA-seq approach using latest sequencing platform for ***De novo assembly of transcriptome (genes and transcripts) of mithun reared under differential growth conditions*** as well as to discover the underlying variations with respect to important economic traits like growth of mithuns. We expect that this transcriptome analysis will be very informative to find out genomic variations in Mithun (*Bos frontalis*) belonging to different growth lines.

On the other hand, Whole Genome Sequencing of mithun will be a very important research work in the field of livestock genomics where a very unique livestock species will be explored with whole genome studies with immense possibilities for the future. Most of the Next-Gen Sequencing (NGS) and other related work may be outsourced through reputed vendors dealing with NGS works. There is tremendous advance in the field of high-throughput sequencing technique with the starting of Human Genome Project and subsequent sequencing of many important livestock species in recent years.

Collaboration with Dr. James Reecey will give an advantage to the mithun genome project to get one of the best technical expertise in the field of genomics and bioinformatics.

The breeding of farm animals has entered the genome era and mithun will be no exception. Despite some deficiencies of NGS, e.g. poor coverage of GC rich areas and the challenges in the assembly when a good reference genome is not available, NGS technologies (RNA-Seq, Chip-Seq, and Genome-resequencing) are still able to help animal scientists study individual genomes at a pace far quicker than previously could be achieved. By utilizing NGS approaches/tools, it will be possible to identify and further analyze individual genes controlling/affecting economic traits in mithuns, which will eventually benefit the mithun owners as well as researchers as well.

209 Key words: Mithun, Strains, transcriptome, RNA-seq, Whole Genome Analysis

Part -II: Investigator Profile

Main Centre

210 Principal investigator of the project

2101 Name: **Dr. Sabyasachi Mukherjee**
2102 Designation: Sr. Scientist
2103 Division/ Section: AGB
2104 Location: Nagaland
2105 Institute address: NRC Mithun, Nagaland

211 (i) Co-investigator

2111 Name: **Dr (Mrs) Anupama Mukherjee**
2112 Designation: Sr. Scientist
2113 Division/ Section: AGB
2114 Location: Nagaland
2115 Institute address: NRC Mithun, Nagaland

211(ii) Co-investigator

2111 Name: **Dr Kezhavituo**
2112 Designation: ACTO
2113 Division/ Section: Animal Physiology
2114 Location: Nagaland
2115 Institute address: NRC Mithun, Nagaland

211(iii) Co-investigator

2111 Name: **Dr Kobu Khate**
2112 Designation: ACTO
2113 Division/ Section: LPM
2114 Location: Nagaland
2115 Institute address: NRC Mithun, Nagaland

211 (iv) Co-investigator

2111 Name: **Dr. C. Rajkhowa**
2112 Designation: Director
2113 Division/ Section:
2114 Location: Nagaland
2115 Institute address: NRC Mithun, Nagaland

Co-operating Centre - I

210 Co-Investigator

2101 Name: **Dr. C. G. Joshi**
2102 Designation: Professor
2103 Division/ Section: Molecular Genetics
2104 Location: Dept of Biotechnology
2105 Institute address: Anand Agricultural University, Gujrat

Part-III: Technical Details

212 Introduction and objectives

2121 Origin of project (Problem identification)

Designing sustainable animal production systems that better balance productivity and adaptation to climate change is a major concern throughout the world. In order to address questions related to suitability to changing environmental conditions in mithuns, it is necessary to increase knowledge on its growth and meat traits, and to produce efficient tools dedicated to this species as mithun is mostly used for ceremonial meat purpose by the tribal population of remote North Eastern States of India.

Animal biodiversity must be an integrated part of climate change adaptation and mitigation efforts. Especially in marginal areas, climate change adaptation, biodiversity conservation and poverty alleviation should complement each other (FAO, 2008). Strategies thus need to be developed that strengthen livestock keepers' adaptation strategies, their ecological knowledge and local institutions. Mithun which is the "ceremonial ox" of the North Eastern Hills has immense potential as future meat animals providing wholesome 'organic meat' for local consumption as well as export.

Under the likely threat of future climate change scenario, how their growth and meat qualities are affected could be a very interesting study through the transcriptome analysis of mithuns belonging to different strains/population groups reared under different environment/altitude and the outcome could well guide us for devising suitable breeding as well as climate mitigation strategies of this unique bovine species, presently under threat of declining population and need proper conservation and genetic improvement strategies.

Mithun (*Bos frontalis*) which is often referred to as the 'ship of highland' and a good example of integration of agro-ecology, subsistence livelihood, culture and livestock rearing of the North Eastern part of India, is till date having limited information about its genome. Obviously, this species will be a very interesting candidate animal to explore their 'code of life' through transcriptome and Whole genome analysis.

Next-generation sequencing (NGS) technologies have been recently used for whole genome sequencing and for re-sequencing projects of different livestock species where the genomes of several specimens are sequenced to discover large numbers of single nucleotide polymorphisms (SNPs) for exploring within-species diversity, constructing haplotype maps and performing genome-wide association studies (GWAS).

The whole genome sequencing related work in other livestock viz. cattle, buffalo, sheep, goat, pig, camel, equines, poultry are either already completed in foreign countries or in the verge of completion.

Mithun is the only species of livestock which may not still under the scanner of any international consortium for genome sequencing and there is enormous opportunity to

explore the genome of this species to generate valuable information. NRC on Mithun (ICAR) can play a significant and leading role in this regard and our project will be the first of this kind for exploring the genome of this unique species of livestock.

2122 Definition of problem

- Mithun are adapted to very harsh terrains high-altitude environments in the North Eastern Hilly Region. However, mithun is not well characterized genetically till date and their adaptation includes not only heat tolerance but also to their ability to survive, grow and reproduce in the presence of poor seasonal nutrition as well as parasites and diseases. It is reported that breeds adapted to these areas will more likely be affected by natural resource degradation linked to climate change rather than temperature or precipitation change per se (Hoffman, 2010).
- This is the main emphasis of this present research proposal that studying only the heat stress related variations in the physiology of mithuns will not be enough indicator of their adaptability under climatic stress. Rather, a concerted effort must include the variation in the genomic level of mithuns to check variation in their growth and meat quality through the elaborate transcriptome analysis of muscle tissues.
- Comparative genomics/transcriptomics of Mithun belonging to different growth lines will be helpful to understand the underlying mechanism of growth and meat traits of mithuns. The outcome of this study will be very useful to devise suitable breeding and climate mitigation strategies of mithuns under the threat of changing climate scenario.
- What is genome structure of *Bos frontalis*?
- To what extent it differs from the genome sequence of *Bos taurus*, whose genome information is available?
- What are the numbers of genes as well novel genes present in the Mithun under study?
- Obviously, such studies will be immensely helpful in the long-run for conservation and propagation of mithun with higher genetic merit.

2123 Immediate objectives

- **Objective 1: *De novo* assembly of transcriptome (genes and transcripts) of mithun reared under differential growth conditions**
The transcriptome of mithuns reared under differential growth and environmental conditions will be assembled using various assemblers and a list of high quality

assembled mRNAs will be provided. List of assembled mRNA and expression (total number of reads) of each assembled mRNA will be analyzed. This will also include distribution pattern of expression value in FPKM (Fragment Per Kilo per Million reads), GC-content distribution of assembled mRNAs and analysis of length distribution of assembled mRNAs

- **Objective 2: Quantification of transcriptome (genes and transcripts) of mithun**

The assembled transcriptome will be annotated using various annotation techniques. This include generation of BLAST summary, Uniprot annotation of gene and protein name, gene description, protein id, protein review status, protein taxonomy report and GO annotation.

- **Objective 3: Identification of variants and differential gene expression for growth traits of mithuns**

Atleast 60-100 million reads of RNA sequencing data is expected from each muscle sample for characterization of the transcriptome of mithun and for identification of the genes/loci responsible for the traits (growth and meat quality) . The differential gene expression in various muscle tissues of mithuns reared under differential environmental factors will be studied through the transcripts and their variants that might be linked with traits will be identified and statistical analysis will be carried out.

- **Objective 4: Whole genome sequencing and *de novo* assembly of the mithun genome**

This whole genome sequencing work involves library preparation from Genomic DNA of mithun and will be validated on the Bio analyzer for quality. The sequencing will be carried out in high throughput Nextgen sequencing platform e.g. Illumina HiSeq for generation of short reads and Roche GS FLX+/PacBio/any other platform for longer reads. 2 x 100 bp high quality paired-end reads will be generated for approximately 50-60X coverage on Hiseq and long reads (450bp-1kb) reads on Roche GS FLX+/any other platform suitable.

2124 **Long term objectives**

Exploiting the genomes of mithun for further exploration and SNPs identification.

213 **Technical profile**

2131 Organisation of work element (for each objective)

Participating investigators giving man months involved and work to be done.

S.No.	Name of scientist	Man months	Work to be done
Main Centre			
1.	Dr. Sabyasachi Mukherjee	12.00 (50%)	<ul style="list-style-type: none"> • Planning of the project, literature review • Selection of animals (Mithun) • Recording of growth data of Mithuns • DNA/RNA isolation from Mithuns • Generation of RNA-seq and WGS data through outsourcing • Report writing • Analysis of molecular data
2.	Dr. Anupama Mukherjee	7.20 (30%)	<ul style="list-style-type: none"> • Planning and co-ordination • Generation of RNA-seq data • Analysis of molecular data
3.	Dr. Kezhavituo	2.40 (10%)	<ul style="list-style-type: none"> • Sample collection • Experimental design
4.	Dr. Kobu Khate	2.40 (10%)	<ul style="list-style-type: none"> • Sample collection • Experimental design
5.	Dr. C. G. Joshi	-	<ul style="list-style-type: none"> • Technical Advice
	Total	24.00 (100%)	

2132 Brief Methodology

The following methodology is proposed to accomplish the stated aim –

- Selection of Mithuns and collection of growth data.
- Collection of tissue and blood samples from Mithun and cattle (for reference).
- Isolation of DNA and RNA by standard protocol from tissues/blood.
- Preparation of cDNA as per standard protocol.
- Design and construction of RNA-seq libraries of Mithun.
- Ligation of the adapters.
- The products will be purified and enriched with PCR to create the final cDNA library.
- Bioanalyzer plots will be used at every step to assess mRNA quality, enrichment success, fragmentation sizes, and final library sizes. The size distribution of the sequencing library will be determined by gel electrophoresis. Both picogreen and qPCR will be used for measuring the quantity of the library before sequencing.

- High-throughput sequencing of RNA-seq and generation of quality transcriptome data.
- Validation of RNA-seq data through real-time PCR technique.
- Generation of whole genome sequencing data based on two different platforms (short reads and longer reads NGS technique) through outsourcing.
- Bioinformatic analysis of transcriptome/WGS data
- Interpretation of results and report writing.

2133 Estimated man months

(a) Scientific – 36.00 man-months (total 3 years)

214 Technical programme (Year wise)

Period of study	Achievable targets
6 -12 Months	<p>Administrative, Recruitment of Research fellow Recording and Selection of animals based on phenotypes, collection of growth data</p> <p>Isolation of genomic DNA, preparation of library for sequencing and sequencing run in Illumina/Roche 454 platform</p> <p>Isolation of RNA from tissue samples, quality checking by Bioanalyzer, preparation of cDNA library, and high throughput sequencing (Outsourcing)</p> <p>Curation and assembly of data. <i>De novo</i> assembly of transcriptome (genes and transcripts) of mithuns and quantification of transcriptome</p>
13-18 Months	<p>Continuation of mithun transcriptome work.</p> <p>Generation of WGS data of mithun through outsourcing (short reads)</p>
19-24 Months	<p>Functional annotation of expressed genes (BLASTX based protein similarity, Gene function, Gene Ontology, Metabolic and other pathway annotation)</p> <p>Generation of WGS data of mithun through outsourcing (long reads).</p>

25-36 Months	<p>Differential gene expression, calculation and identification of development related genes, transcripts and pathways</p> <p>Bioinformatic analysis of WGS data. Development of draft genome sequence of mithun.</p> <p>Submission of transcriptome data in Genbank. Submission of WGS data in Genbank.</p> <p>Finalization of analysis, interpretation and report writing</p>
--------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

215 Importance of proposed project (gap in knowledge/ product/ process/ technology) to the institute mandate.

The project has been formulated keeping in view of the mandate of the Institute and the priority area of research in the field of Animal Genetics and Breeding section.

The mandates of the Institute are

- Identification, evaluation and characterization of Mithun germplasm available in the country
- Conservation and improvement of Mithun for meat and milk
- To act as repository of germplasm and information centre on Mithun

The project is very important for the following reasons –

The proposed study is very relevant topic under Indian context, particularly for mithuns under the threat of impending climatic stress scenario. Mithun (*Bos frontalis*) is one of the most important and unique bovine species of the North Eastern part of India. However, information of this species particularly about its genome is scanty till date. Obviously, it will be a very important research study to explore the extent of genetic variability under different growth parameters as well as differential gene expression in the muscle of mithun through transcriptome sequencing.

This study will also be very important for comparative genome analysis among bovine speices, due to the fact that Mithun and local Tho tho cattle even through reared under same environment of Nagaland, still, mithun has better growth potential and mithun meat is more tender and palatable with higher marbling score than beef.

Part -IV: Budget estimate

216 Expected expenditure (Rs. in lakhs)

Particulars	1 st yr	2 nd yr	3 rd yr	4 th Yr	Total
Recurring expenditure					
A. Consumables					
(a) Chemicals/Kit	5.00	5.00	-	-	10.00
(b) Glassware	1.00	1.00	-	-	2.00
(c) Plastic ware	2.00	2.00	-	-	4.00
Sub total A	8.00	8.00	-	-	16.00
B. Contractual labour	-	-	-	-	-
C. Travel	1.00	2.00	-	-	3.00
D. Miscellaneous (outsourcing of sequencing work, bioinformatic analysis etc)	25.00	25.00	-	-	50.00
Sub total	34.00	35.00	-	-	69.00
Non-recurring expenditure					
E. Instruments	19.00	0		-	19.00
F. Software	1.00	0		-	1.00
Sub total	20.00	-	-	-	20.00
F. Purchase of animals	0	0		-	0
G. Renovation of lab / exptl. shed/building	0	0			0
H. Petty equipments	2.00	0		-	2.00
Sub total	22.00	-	-	-	22.00
OVER ALL TOTAL	56.00	35.00	-	-	91.00

217 Infrastructure facilities required

Most of the facilities are available in the Institute itself. However, a few specific instrument need to be procured as below –

- High capacity computing system for data analysis (mentioned in XII Plan EFC memo under Equipment Head) - approx cost Rs. 8.00 lakh
- Software for analysis of molecular data - Rs. 1.00 lakh
- Ultra-low deep freezer (-80°C) - Rs. 7.00 lakh
- Homogenizer - Rs. 2.00 lakh
- UPS (5 kv) with suitable batteries - Rs. 1.50 lakh
- Servo stabilizer (2) - Rs. 0.50 lakh

The technical assistance will be sought from the concerned scientific expert from Anand Agricultural University, Anand, Gujrat and ICAR HQ, New Delhi the collaborating Institutes. On the other hand, technical expertise of bioinformatic analysis will be sought from any other national/international agency/Institute as and when necessary.

This is to certify that

- the research work proposed in the project does not in any way duplicate the work already done or being carried out in the Institute on the subject.
- the same project has not been submitted to any other agency for financial support (if already submitted identify project & agency).
- the investigator/co-investigators have been fully consulted in the development of project and have fully undertaken their responsibility to carry out the programme as per the technical programme.

Signature of project leader - Main Centre/Institute

Dr. Sabyasachi Mukherjee_____



Co-investigators 1. Dr (Mrs) Anupama Mukherjee_____

2. Dr. Kezhavituo_____

3. Dr. Kobu Khate_____

Co-operating Centre - Principal Investigator

2. Dr. C J Joshi_____

Signature & comments of the Head of the Division/Institute

Co-operating Centre/Institute Veterinary Science Faculty Anand Agricultural University, Gujrat	Main Centre/Institute ICAR: NRC on Mithun, Nagaland

ANNUAL PROGRESS REPORT

R. P. F. II

Year: (01 April) 2014- (31 March) 2015

1. Institute Code No.

2. I.C.A.R. Code No. (**PIMS-ICAR**): **IXX10452**

3. Name & Address of Research Institute/Centre.

**National Research Centre on MITHUN (ICAR)
Medziphema, Nagaland 797 106 India**

4. Project Title: **Transcriptome analysis through RNA-seq approach and Whole Genome Sequencing of Mithun (*Bos frontalis*) (Inter-Institutional collaborative project)**

4. Name & Designation of Principal Investigator: **Dr. Sabyasachi Mukherjee, Principal Scientist**

5. Name (s) and Designation (s) of Project Associates and establishment (s) on which borne

(a). Whole time

(b). Part time (indicate proportion of time to be devoted and other area(s))

Dr. Anupama Mukherjee, Sr Scientist
Dr. Kezha Vituo, Technical Officer
Dr. Kobu Khate, Technical Officer
Dr. Chandan Rajkhowa, Director, NRCM

6. Location of Research Project with complete address (Division/Section/Sub-Centre) Department of Animal Breeding and Genetics
National Research Centre on Mithun (ICAR)
Medziphema, Nagaland 797 106 India

7. Date of Start: **01 February 2013**

8. Likely date of completion: **31 March 2016**

10. Annual Report on the Project:

- Please see Annexure - I

(A summary on the results not exceeding 2 pages precisely and concisely starting the work already carried out and the work contemplated.

- RNA was isolated from four tissue samples and transcriptome sequencing was outsourced from Scigenom lab.
- Bioinformatic analysis of transcriptome data was performed.
- DNA was isolated from one female mithun, quality was checked to be very high quality and whole genome sequencing was outsourced using Illumina Hiseq 2000 platform. Total 150 GB paired end data (2x 100 bp) was generated.

11. Progress of work in relation to the time targeted for completion of work, reasons for non-achievement of targets if any.

- The progress of the project was as per schedule
- Requirement of a suitable high performance computing system is felt urgently for bioinformatic analysis

12. Publications and Material:

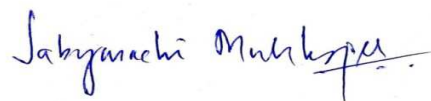
Research Papers One research abstract published as Invitation Lecture,
World DNA Day 2014 Conference, China

Popular articles

Reports

(Two copies each to be supplied with this proforma)

13. Signature of Project Principal Investigator



**14. Signature (with comments, if any)
of Head of Division/Section/Station.**

**15. Signature (with comments, if any)
of Director.**

Targets and milestones with duration

S. No	Targets	Duration
1.	Targets <ul style="list-style-type: none"> • RNA isolation from mithun samples, quality/quantity checking, and high-throughput sequencing for RNA-seq through outsourcing services. • Selection of mithun, isolation of DNA sample, quality/quantity checking and high-throughput sequencing for WGS through outsourcing service. 	April 2013 – March 2014

Achievements against the targets

- RNA was isolated from four tissue samples and transcriptome sequencing was outsourced from Scigenom lab.
- Bioinformatic analysis of transcriptome data was performed.
- DNA was isolated from one female mithun, quality was checked to be very high quality and whole genome sequencing was outsourced using Illumina Hiseq 2000 platform. Total 150 GB paired end data (2x 100 bp) was generated.

1. Bioinformatics analysis

The bioinformatics analysis pipeline for the *denovo transcriptome* analysis is shown below. Briefly the following analysis was performed -

- i. Fastq quality checking and filtering
- ii. *Denovo* transcriptome assembly
- iii. Expression estimation
- iv. Transcriptome annotation

2. Samples summary - Table 1

Species	<i>Bos frontalis</i>
Condition types	High & Low

Sequencing Platform	Illumina HiSeq 2000
Library type	Paired End

3. Sequence read quality checking

3.1 **Raw read summary** - Below is the summary of raw fastq files obtained from sequencer.

Table2: Raw read summary

	'Sample high 1'	'Sample high 2'	'Sample low1'	'Sample low 2'
Number of paired-end reads	84,564,192	82,641,184	94,834,328	72,549,304
Number of bases (Gb)	8.54	8.35	9.58	7.33
GC %	48	47	47	47
Read length (bp)	101*2	101*2	101*2	101*2

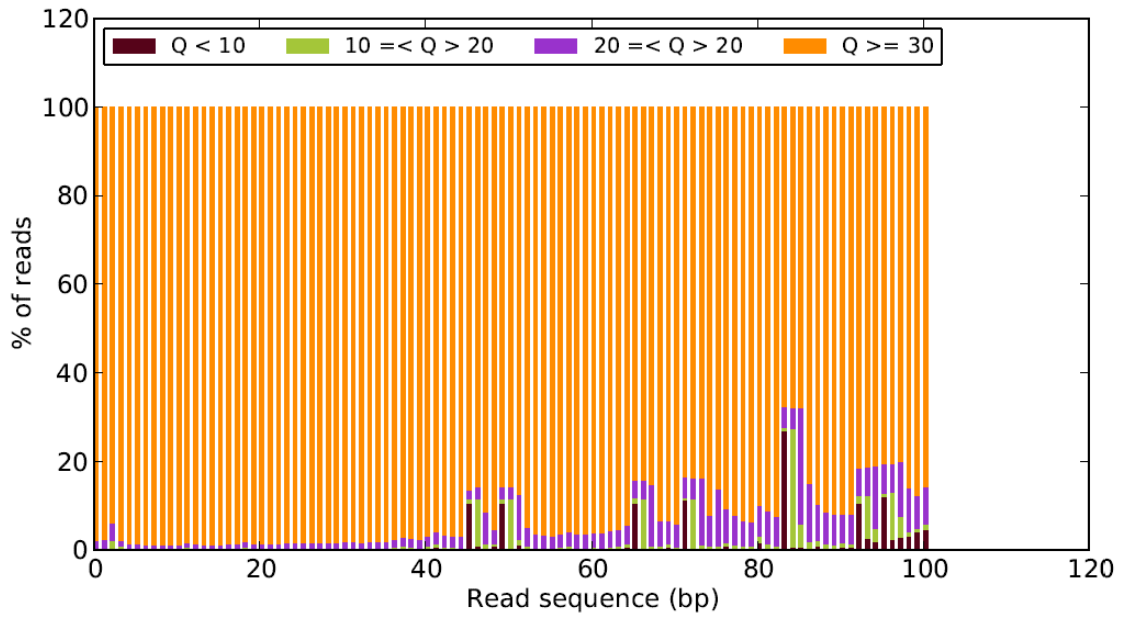
3.2 Fastq quality check

This step involves checking of quality parameters for the sequences obtained from sequencer. The following checks are performed for an input fastq file

- base quality score distributions
- average base content per read
- GC distribution in the reads

3.2.1 Base quality score distribution

The x-axis represents sequencing cycle and y-axis represents the Phred quality score of bases. The quality of left (also called R1) and right (also called R2) end of the paired-end read sequence is shown in Fig. 1(a) and Fig. 1(b), respectively. It can be clearly seen that the average base quality is above Q20 (error-probability ≥ 0.01) for majority of read cycle in R1.



Fi

g. 1(a): Base quality distribution of 'Sample high 1' left end of paired-end read

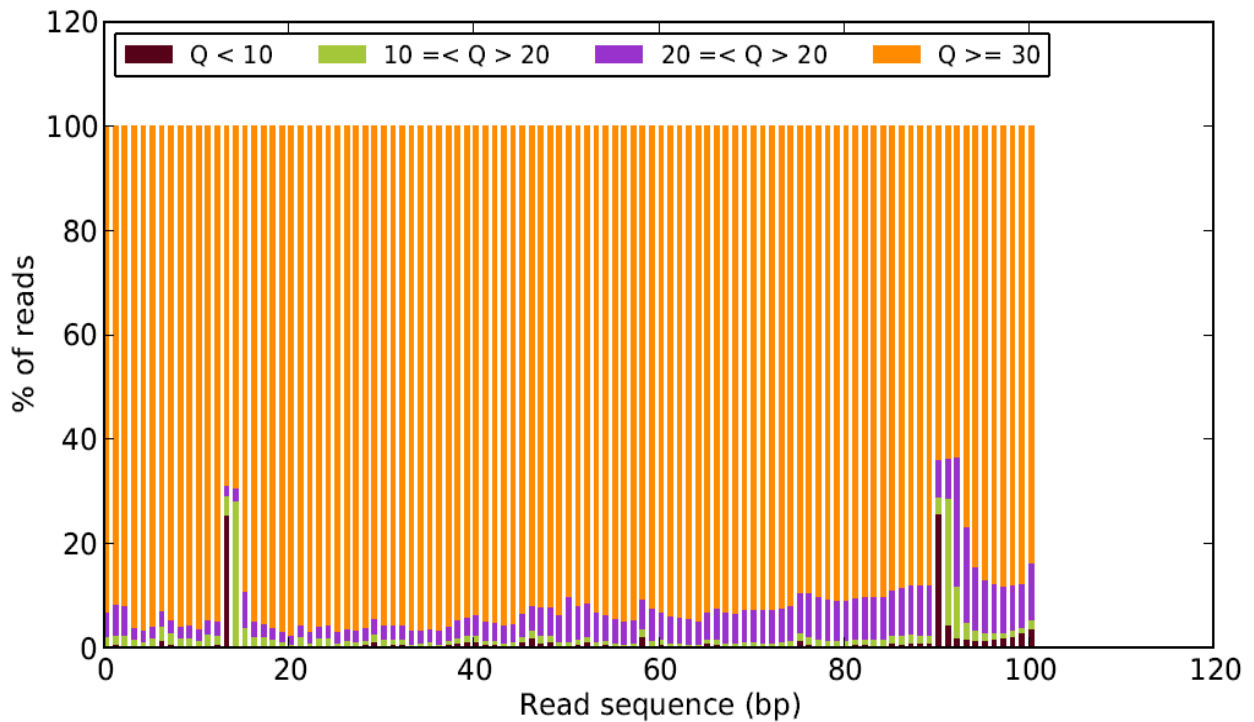


Fig. 1(b): Base quality distribution of 'Sample high 1' right end of paired-end read

3.2.2 Base composition distribution

The composition of nucleotides in the sequence read is shown in Fig. 2a - 2b.. The x-axis represents sequencing cycle and y-axis represents nucleotide percentage. The base composition of left and right end of the paired-end read sequence is shown in Fig.

2(a) and Fig. 2(b), respectively. A bias in first 18 cycles and last 19 cycles of R1 and first 18 cycles and last 11 cycles of R2 is observed in the across the samples. Biasing in sequence composition is in general observed in transcriptome experiments.

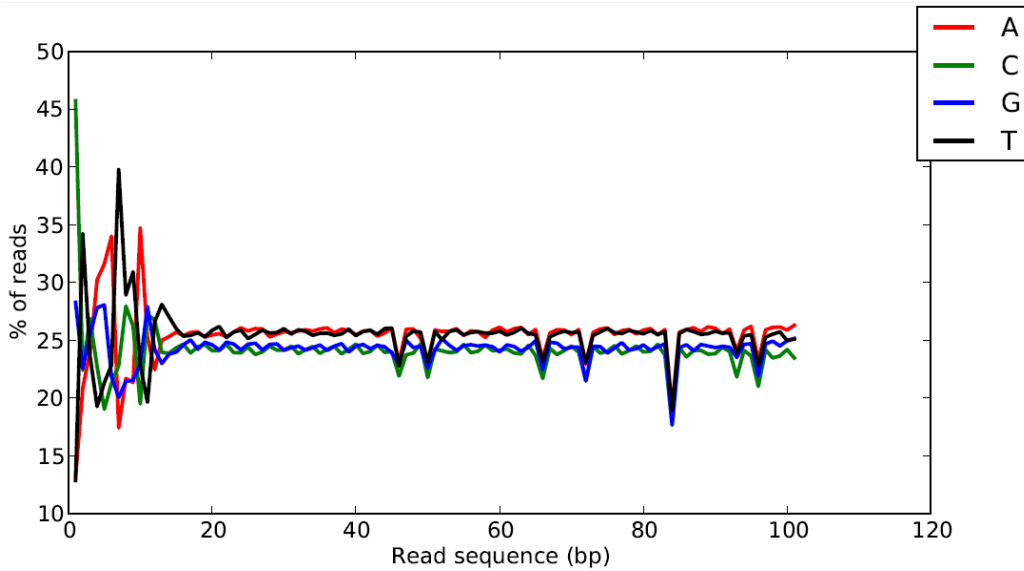


Fig.

2(a): Base composition in the left end of 'Sample high 1' paired-end read

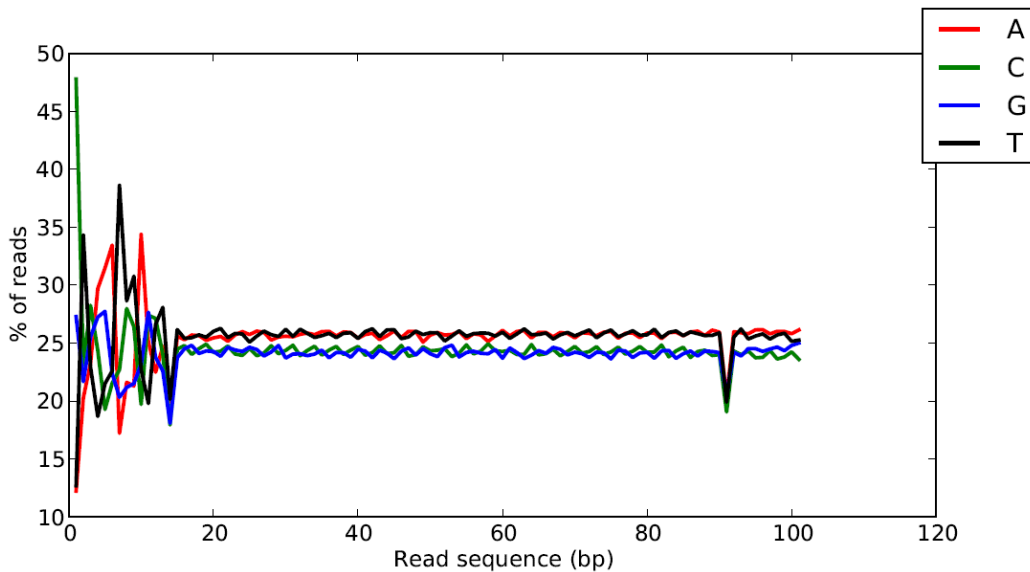


Fig. 2(b): Base composition in the right end of 'Sample high 1' paired-end read

3.2.3 GC distribution

The average GC content distribution in the sequenced read of the sample is shown in Fig. 3(a) & 3(b). The x-axis represents average GC content in the sequence and y-axis represents total percentage of reads. The average GC content of the reads in the sample follows a normal distribution.

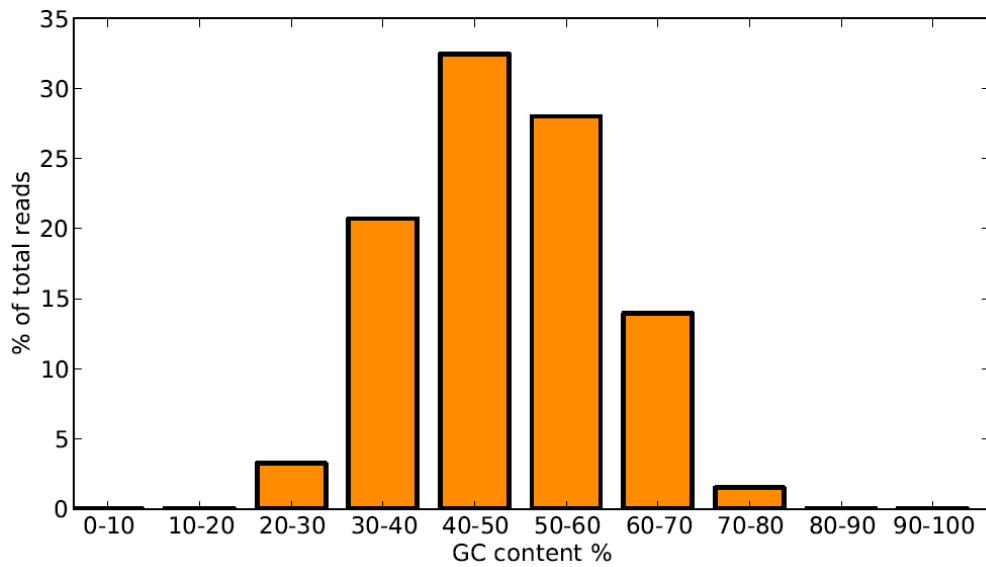


Fig. 3(a):

GC distribution over left end read sequence of 'Sample high 1' paired-end read

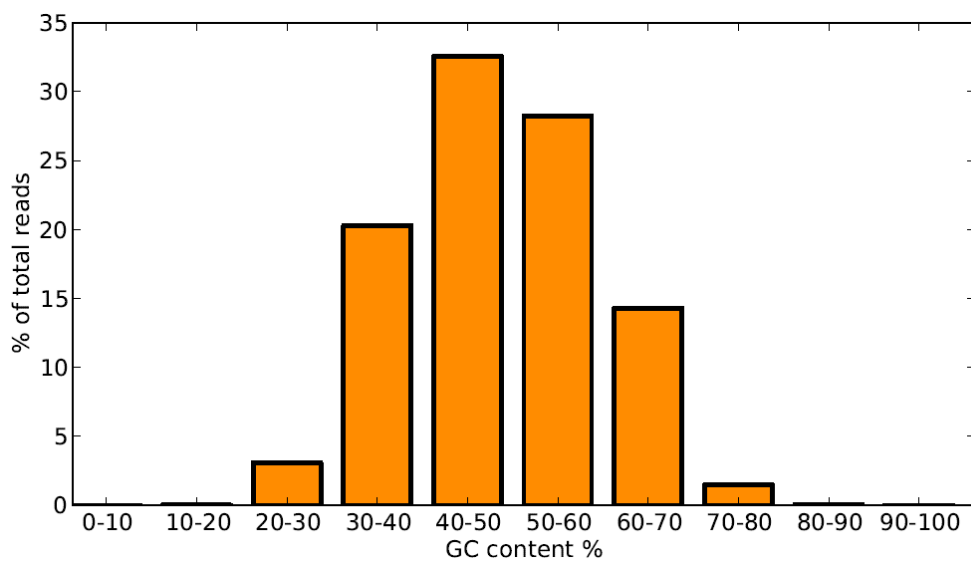


Fig. 3(b): GC distribution over right end read sequence of 'Sample high 1' paired-end read

4. *Denovo* transcriptome assembly

The fastq files were trimmed before performing assembly. First 18 bases and last 19 bases were removed from all R1 reads and first 18 bases and last 11 bases were trimmed from all R2 reads to avoid specific sequence bias and low quality bases. We also filtered out reads whose average quality score <20 in any of the paired end and reads contaminated with Illumina adapter. Summary of the paired end reads are provided in Table 3.

Table 3: Trimmed read summary

	'Sample high 1'	'Sample high 2'	'Sample low 1'	'Sample low 2'
Number of paired-end reads	42,275,516	41,314,410	47,409,867	36,268,355
Number of bases (Gb)	5.75	5.62	6.45	4.93
Read length	64bp (R1), 72bp (R2)	64bp (R1), 72bp (R2)	64bp (R1), 72bp (R2)	64bp (R1), 72bp (R2)

The trimmed reads were then assembled using SOAPdenovo31mer algorithm with default options. The transcriptome assembly result is summarized below in Table 4. The transcript length distribution for all assembled is shown in Fig. 4. Only 35% of total assembled transcripts are of length ≥ 150 bp. The GC content distribution of the all assembled transcripts is shown in Fig. 5. Main focus was on transcript of length ≥ 150 bp for transcript expression estimation and downstream annotations.

Table 4: Assembled transcript summary

	All assembled transcripts	Transcripts of length ≥ 150 bp
Number of assembled transcript	408,911	143,090
Longest transcript length (bp)	29,564	29,564
Mean GC % of transcripts	47.31	47.63

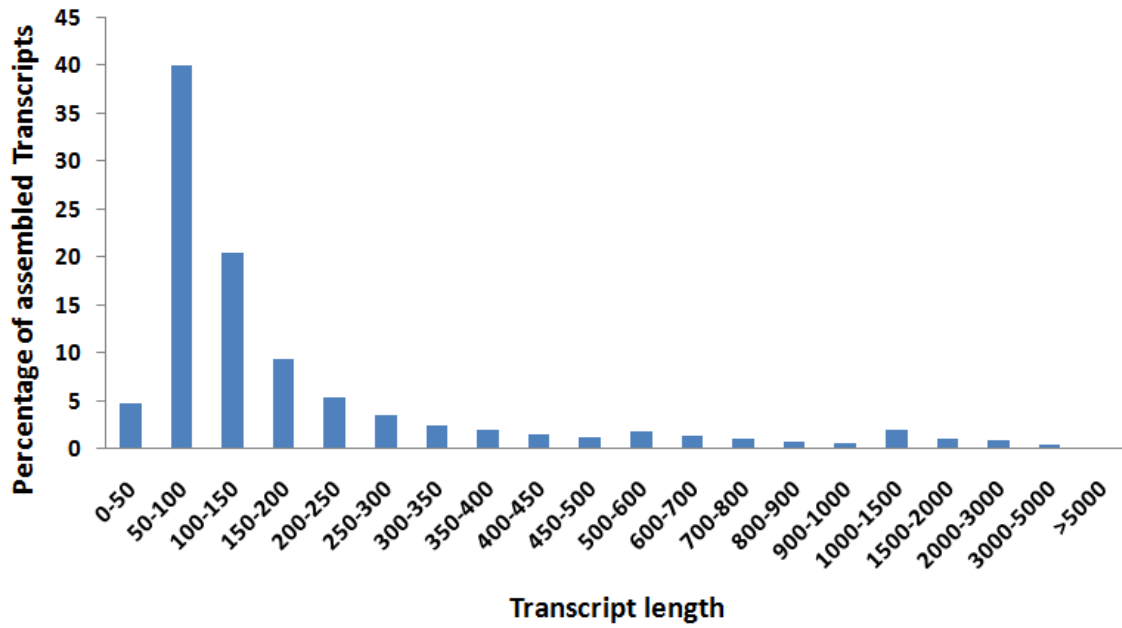


Fig. 4: Assembled transcript length distribution

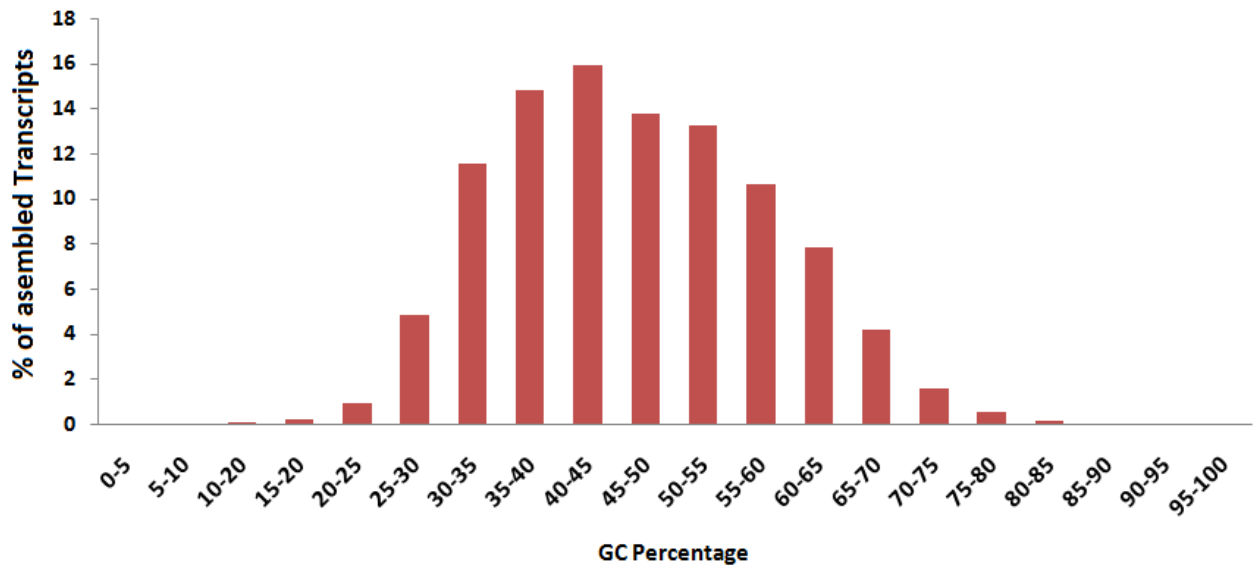
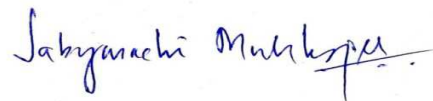


Fig. 5: GC content distribution of all assembled transcripts

This is to certify that

- * the research work proposed in the project does not in any way duplicate the work already done or being carried out in the Institute on the subject.
- * the same project has not been submitted to any other agency for financial support (if already submitted identify project & agency).
- * the investigator/co-investigators have been fully consulted in the development of project and have fully undertaken their responsibility to carry out the programme as per the technical programme.

Signature of project leader



Dr. Sabyasachi Mukherjee

Co-investigators 1. Dr (Mrs) Anupama Mukherjee

2. Dr. Kezhavituo

3. Dr. Kobu Khate

4. Dr. C. J. Joshi

Signature & comments of the Head of the Division



Signature & comments of the Director