ANNEXURE -VI

INDIAN COUNCIL OF AGRICULTURAL RESEARCH
CHECKLIST FOR SUBMISSION OF FINAL RESEARCH PROJECT REPORT (RPP-III)

(For Guidelines Refer ANNEXURE – XI (F))

1. **Institute Project Code** : **IXX10452**

2. **Investigators as approved in RPP-I, If any change attach IRC proceedings:**

Principal Investigator	CC-PI	Co-PI
Sabyasachi Mukherjee	-	Anupama Mukherjee Kezhavituo Vupru Kobu Khate

3. **Any change in objectives and activities** Yes/**No**
(If yes, attach IRC proceedings)

4.	Date of Start & Date of Completion (Actual). If any extension granted enclose IRC proceedings		Yes	No
5.	Whether all objectives met		Yes	No
6.	All activities completed		Yes	No
7.	Salient achievements/major recommendations included		Yes	No
8.	Annual Progress Reports (RPP-II) submitted	1st Year	Yes	No
		2nd Year	Yes	No
		3rd Year	Yes	No
		nth year	Yes	No
9.	Reprint of each of publication attached		Yes	No
10.	Action for further pursuit of obtained results indicated		Yes	No
11.	Report presented in Divisional seminar (enclose proceedings & action taken report)		Yes	No
12.	Report presented in Institute seminar (enclose proceedings & action taken report)		Yes	No
13.	IRC number in which the project was adopted		IRC No:	
14.	Any other Information			

15. Signature:

17/6/16 Sabyasachi Mukherjee Project Leader	Anupama Anupama Mukherjee Co-PI	Kezhavituo Vupru Co-PI	Kobu Khate Co-PI

HOD/PD/I/c.

ANNEXURE - VII

INDIAN COUNCIL OF AGRICULTURAL RESEARCH

FINAL RESEARCH PROJECT REPORT (RPP- III)

(For Guidelines Refer ANNEXURE – XI(G))

1. Institute Project Code **IXX10452**
2. Project Title **Transcriptome analysis through RNA-seq approach and Whole Genome Sequencing of Mithun (*Bos frontalis*) (Inter-Institutional collaborative project)**
3. Key Words Mithun, muscle, RNA-seq, transcriptome, whole genome, sequencing
4. (a) Name of the Lead Institute **ICAR-NRC on Mithun, Nagaland**

   (b) Name of Division/ Regional Center/ Section **Animal Genetics & Breeding**
5. (a) Name of the Collaborating Institute(s) **Gujrat Agricultural University, Anand**

   (b) Name of Division/ Regional Center/ Section of Collaborating Institute(s) Nil
6. Project Team(Name(s) and designation of PI, CC-PI and all project Co-PIs, with time spent)

S. No.	Name, designation and institute	Status in the project (PI/CC-PI/ Co-PI)	Time to be spent (%)	Work components assigned to individual scientist
1	Sabyasachi Mukherjee	PI	50	Planning, execution
2	Anupama Mukherjee	Co-PI	30	Planning, execution
3	Kezha Vituo	Co-PI	10	Sample collection
4	Kobu Khate	Co-PI	10	Management of mithuns

7. Priority Area Animal Genomics
8. Project Duration: Date of Start - 01.2.2013 Date of Completion – 31.3.2016
9. a. 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 mtihun 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.

b. Practical utility

10. Final Report on the Project (materials and methods used, results and discussion, objective wise achievements and conclusions) **Please see Annexure - A**

11. Financial Implications (₹ in Lakhs)

11.1 Expenditure on

(a) Manpower: -

(b) Research/Recurring Contingencies: Rs. 45.00 lakh

(c) Non-Recurring Cost (Including cost of equipment): High-speed computing system Rs. 8.50 lakh

(d) Any Other Expenditure Incurred - Rs.

11.2 Total Expenditure - approx. Rs. 53.50 lakh (excluding man power)

12. Cumulative Output

a. Special attainments/innovations :

- Generation of muscle transcriptome data of mithun and identification of genes responsible for higher growth
- Generation of whole genome sequence data ~250 gb

- Construction of a draft genome assembly of mithun with 90-100X coverage
- Estimation of mithun genome size to be ~ 3.2 gb

b. List of Publications (one copy each to be submitted if not already submitted)
   i. Research papers - 1 paper communicated
   ii. Reports/Manuals - Nil
   iii. Working and Concept Papers - Nil
   iv. Popular articles - Nil
   v. Books/Book Chapters - Nil
   vi. Extension Bulletins - Nil

c. Intellectual Property Generation Nil
(Patents - filed/obtained; Copyrights- filed/obtained; Designs- filed/obtained; Registration details of variety/germplasm/accession if any)

d. Presentation in Workshop/Seminars/Symposia/Conferences **Two numbers**
(relevant to the project in which scientists have participated)

❖ Sabyasachi Mukherjee, Anupama Mukherjee, S. Longkumer, Moonmoon Mech, Subhash Jakhesara, C. J. Joshi and C. Rajkhowa. 2014. Comparative Transcriptome Profiling of Muscle Tissue from Mithuns with Differential Growth Rates. Invited Presentation: BIT's 5th Annual World DNA and Genome Day-2014, Dalian, China, April 25-28, 2014.

❖ Sabyasachi Mukherjee, Anupama Mukherjee, S. Longkumer, Moonmoon Mech, Kobu Khate, Kezhavituo, C. J. Joshi, Sanjeev Kumar and C. Rajkhowa. 2015. Muscle Transcriptome Analysis with differential phenotypes in Mithun (Bos frontalis). XII Agricultural Science Congress. NDRI, Karnal, 3-6 February, 2015.

e. Details of technology developed Nil
(Crop-based; Animal-based, including vaccines; Biological – biofertilizer, biopesticide, etc; IT based – database, software; Any other – please specify)

f. Trainings/demonstrations organized **Nil**

g. Training received Nil

h. Any other relevant information

13. (a) Extent of achievement of objectives and outputs earmarked as per RPP-I

Objective wise	Activity	Envisaged output of monitorable target(s)	Output achieved	Extent of Achievement (%)
1: *De novo* assembly of transcriptome (genes and transcripts) of mithun reared under differential growth conditions	1. Isolation of RNA from muscle tissues of experimental mithuns and generation of RNA-seq data 2. Analysis and assembly of RNA-seq data	1. Muscle transcriptome data generation from 4 mithuns (two each of high and low growth groups) 2. Various bioinformatic analysis of RNA-seq/ transcriptome data	1. The muscle RNA-seq/transcriptome of mithuns reared under differential growth are generated using Illumina Hiseq 2000 2. Transcriptome data was analysed and assembled using various assemblers and a list of high quality assembled mRNAs was provided. 3. List of assembled mRNA and expression (total number of reads) of each assembled mRNA was analyzed including 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	100%
2. **Quantification of transcriptome (genes and transcripts) of mithun**	Annotation of transciptomes	Quantification and annotations of muscle transcriptomes of mithun	The assembled transcriptomes were 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.	90%
3: Identification of variants and differential gene expression for	Variant calling and gene expression	Identification of variations of the available mithun trancriptomes and genes responsible of high and low growth rates	More than 60 million reads of RNA sequencing data were generated from each muscle sample for characterization of the transcriptome of	90%

growth traits of mithuns			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 were studied through the transcripts and their variants that linked with traits identified and statistical analysis were carried out.	
• **4: Whole genome sequencing and *de novo* assembly of the mithun genome**	Generation of whole genome sequencing Bioinformatic analysis of whole genome data	Whole geneome sequencing of one female mithun by Illumina Hiseq2000 platform for short reads and Moleculo technique for long reads Construction of draft mithun genome assembly	The NGS sequencing was carried out in high throughput Illumina HiSeq 2000 for generation of short reads and Moleculo platform for longer reads. 2 x 100 bp high quality paired-end reads were generated for approximately 50-60X coverage and long reads on Moleculo other platform suitable. A draft mithun genome assembly was constructed through various stages of bioinformatic analysis.	90%

(b) Reasons of shortfall, if any The bioinformatic analysis work through outsourcing took longer time than expected due to very late submission of analysis report by the outsourcing company.

14. Efforts made for commercialization/ technology transfer

This is a basic research work for identification of genes and transcript in muscle tissues of mtihun and to construct a draft genome assembly of Mithun.

15. (a) How the output is proposed to be utilized?

- The genes identified as responsible for better growth of mithuns may be further studied individually in a large data set for use as candidate gene approach for selection of mtihuns with growth rate.
- The output of this research may be utilized to re-sequence few more Mithun genome to identify valuable SNPs for important economic traits.
- The initial draft genome of Mithun may be used as a baseline information and may be further refined through re-sequencing few more mithuns with low coverage and more gaps may be filled up using more sequence data.

(b) How will it help in knowledge creation

- Creation of Mithun whole genome data as base line information
- Creation of Mithun muscle transcriptome data as base line information for further research

16. Expected benefits and economic impact. (If any)

- Economic impact may not be immediate. However, whole genome sequencing and muscle transcriptome analysis of Mithun has given the advantage of exploring this unique animal more genetically.
- These new insights about tis genetic architecture will be very useful for devising suitable breeding and conservation policy of mtihun.

17. Future line of research work/other identifiable problems

- Transcriptome profiling of few more important organs of mtihun may be carried out for identification of genes affecting important metabolic and cellular pathways
- Refining the draft genome assembly of Mithun with more sequence data will be helpful for bridging the gaps of the scaffolds and the assembly.

18. Details on the research data (records) generated out of the project deposited with the institute for future use.

- The soft copy of muscle transcriptome and whole genome sequence data generated in the project is available with the In-charge Animal Genetics and Breeding section, NRC on mithun, Jharnapani as below:
- One hard disk from Nucleome contains all the genome sequence data and files
- One hard disk from Scigenome contains the muscle transcriptome data of Mithun and mtihun genome sequence data (Illumina data).

19. **Signature of PI, CC-PI(s), all Co-PIs**

Sabyasachi Mukherjee   17/6/16

Anupama Mukherjee   Anupama

Kezhavituo Vupru

Kobu Khate

20. **Signature of Head of Division**

21. **Observations of PME Cell based on Evaluation of Research Project after Completion**

The project was completed as per schedule and generated very important information regarding genetic architecture of mithun genome and transcriptome of muscle tissue for growth.

17/6/16

(Sabyasachi Mukherjee)
PS & I/C PME/RFD Cell

22. **Signature (with comments if any along with rating of the project in the scale of 1 to 10** □
on the overall quality of the work) of JD (R)/ Director

ANNEXURE - A

Final Project Report

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

Materials & Methods - Mithun transcriptome analysis

- RNA was isolated from four tissue samples and transcriptome sequencing was outsourced from Scigenom lab on Illumina Hiseq 2000 platform
- Bioinformatic analysis of transcriptome data was performed.

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

Samples summary - Table 1

Species	*Bos frontalis*
Condition types	High &Low
Sequencing Platform	Illumina HiSeq 2000
Library type	Paired End

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

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

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.

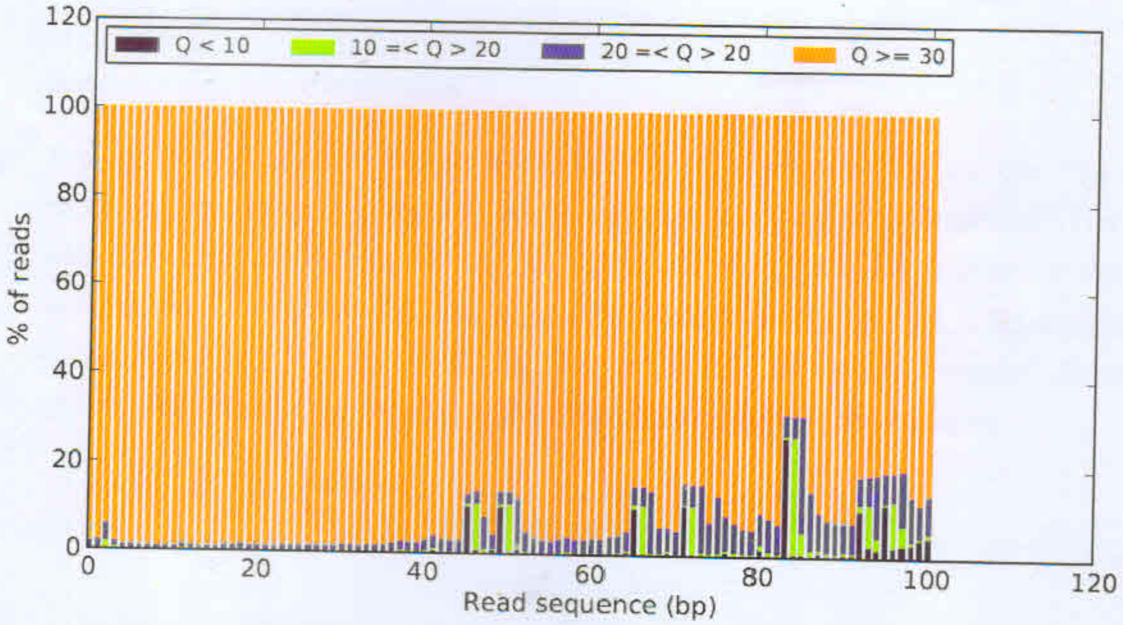


Fig. 1(a): Base quality distribution of 'Sample high 1' left end of paired-end read

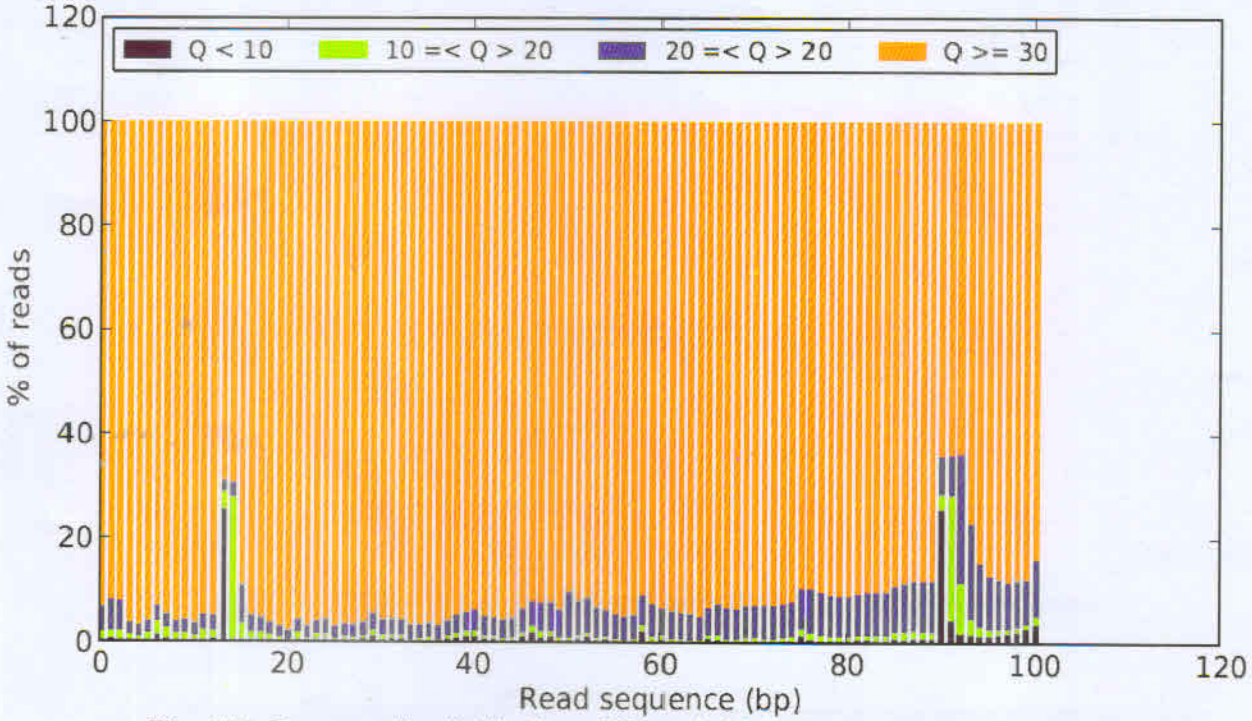


Fig. 1(b): Base quality distribution of 'Sample high 1'right end of paired-end read

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.

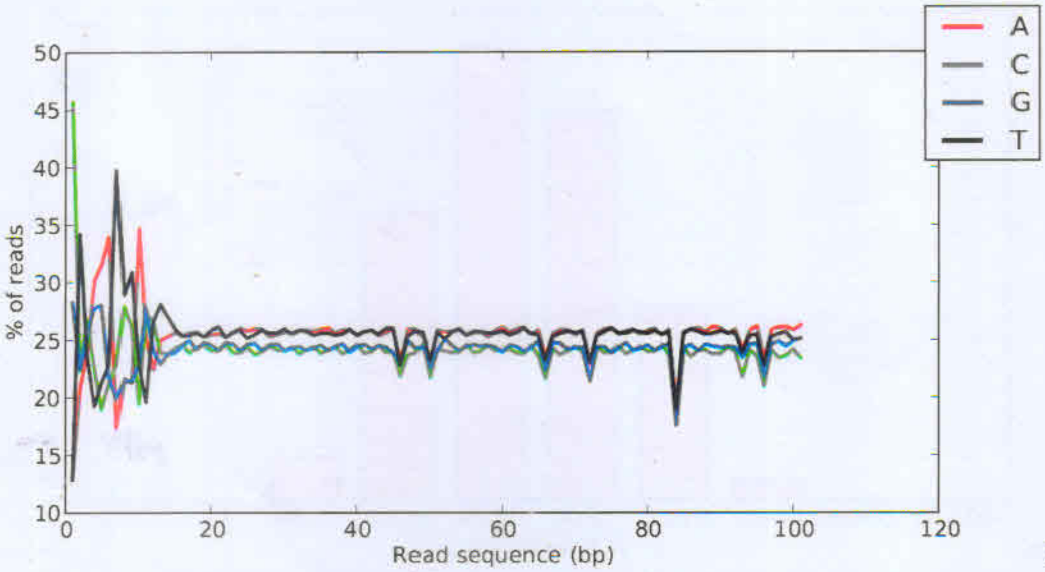


Fig. 2(a): Base composition in the left end of'Sample high 1' paired-end read

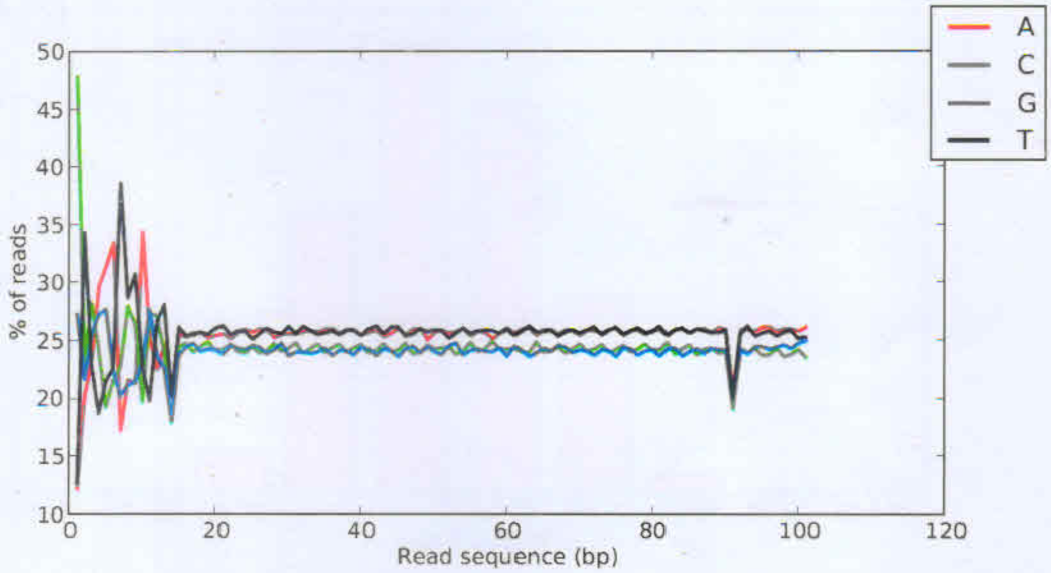


Fig. 2(b): Base composition in the right end of'Sample high 1' paired-end read

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.

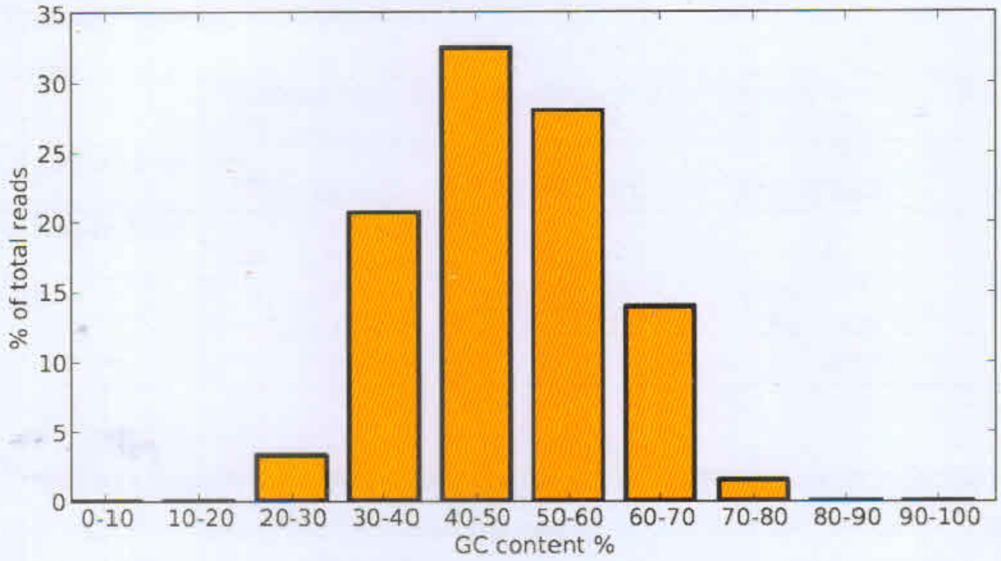


Fig. 3(a): GC distribution over left end read sequence of 'Sample high 1' paired-end read

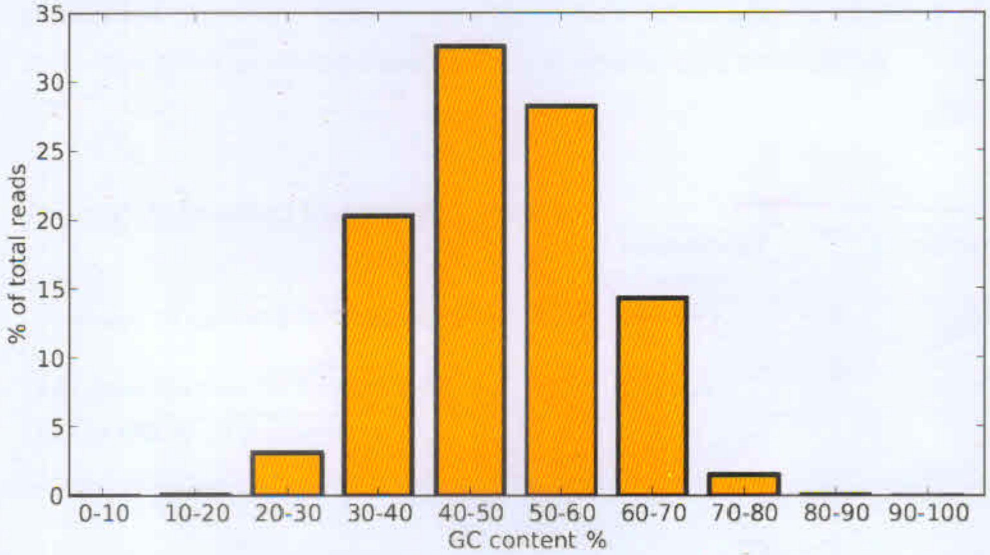


Fig. 3(b): GC distribution over right end read sequence of 'Sample high 1' paired-end read

Results and Discussion - Mithun transcriptome analysis

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 >= 150bp. 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

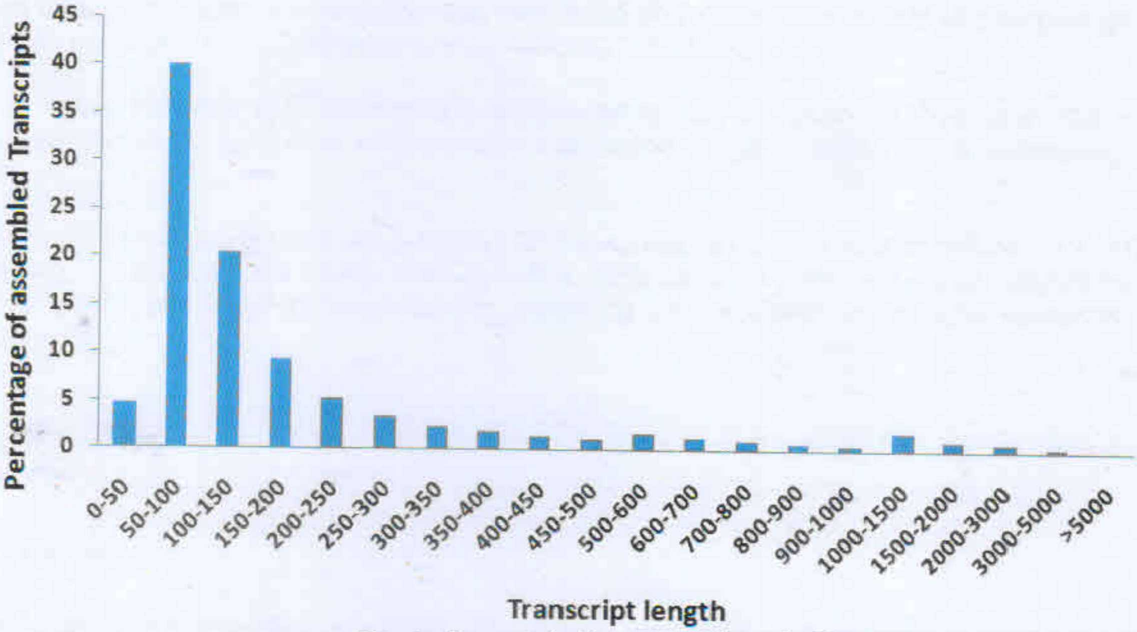


Fig. 4: Assembled transcript length distribution

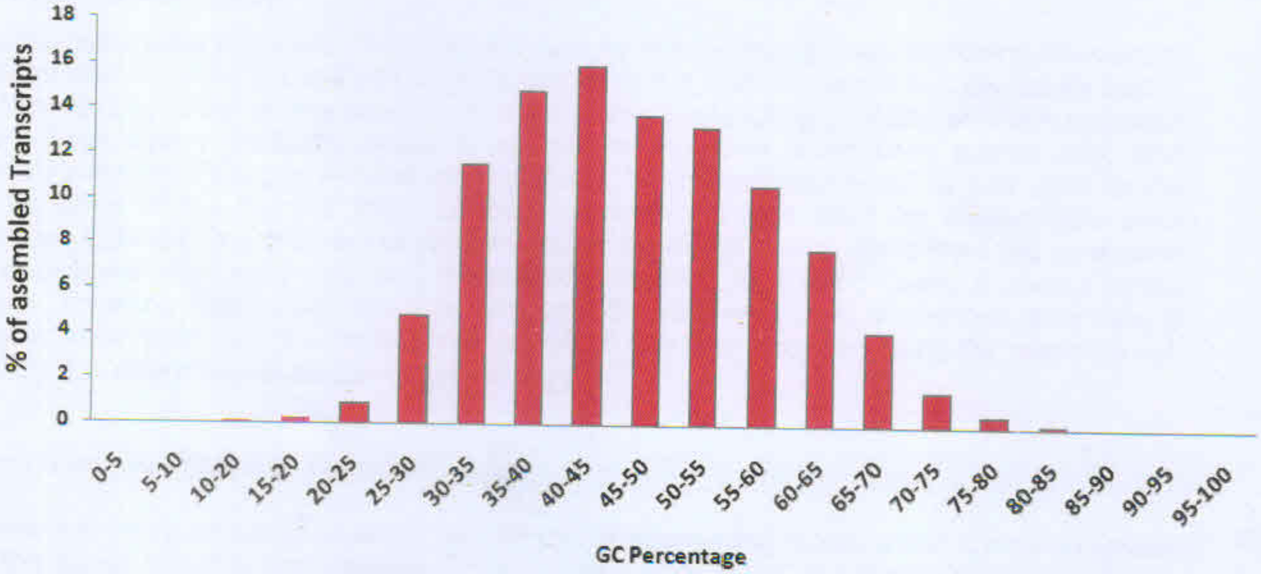


Fig. 5: GC content distribution of all assembled transcripts

Mithun muscle transcriptome analysis was completed and transcriptome data of divergent growth lines was submitted in NCBI Genbank (BioProject ID PRJNA307305).

Our analysis revealed 297 differentially expressed genes in which 173 up and 124 down regulated unigenes. Extensive conservation was found in genic region while comparing with cattle.

Analysis also revealed 57 pathways having 72 transcriptional factors and cofactors, 212 miRNA regulating 71 DEGs, 28963 SSRs, 104822 SNPs, 7288 indels, a gene regulatory network having 24 hub genes and transcriptional factors regulating cell proliferation, immune tolerance and myogenesis.

Results and Discussion - Mithun Whole Genome bioinformatic analysis and *De novo* draft assembly

Raw Data Generation

De novo genome sequencing of mithun was used to sequence uncharacterized genomes where there is no reference sequence available. *De novo* sequence reads were typically generated on the Illumina HiSeq 2000/2500 platforms, Moleculo Long-Read sequencing technology platform and PacBio RS II platform for Mithun whole genome sequencing project. Approximately 100x of data was generated on the above mentioned sequencing technology for this project.

Raw Data Preprocessing

Data pre-processing consisted in filtering the data to remove errors, thus facilitating the work of the assembler. In order to cope with lower quality data, it is common to remove low quality bases. Typically remove lower quality bases from the e.g. the 3' end using a sliding windows approach as per base quality gradually drops. In addition to removing lower base quality data, also removed adapters, PCR primers and other artifacts. Trimmomatic version-0.32 was used for the preprocessing of the Illumina data. LoRDEC version-0.4.1 was used for PacBio data error correction. LoRDEC is a program to correct sequencing errors in long reads from 3rd generation sequencing with high error rate, and is especially intended for PacBio reads. It uses a hybrid strategy, meaning that it uses two sets of reads: the reference read set, whose error rate is assumed to be small, and the PacBio read set, which was then corrected using the reference set. Typically, the reference set contains Illumina reads.

Mithun Genome Assembly Workflow

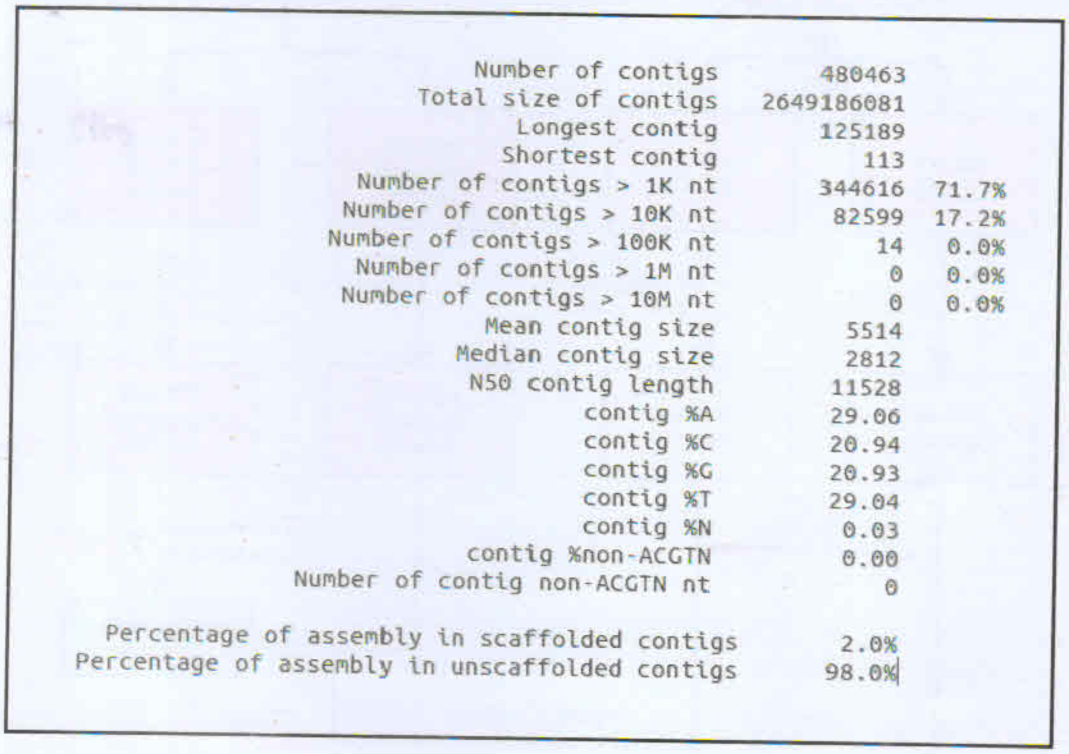
Genome assembly consisted of taking a collection of sequencing reads, which are much shorter than the actual genome, and creating a genome sequence which is a likely source of all these fragments. The genome must include as much of the input data as possible. The output of an assembler is generally decomposed into contigs, or contiguous regions of the genome which are nearly completely resolved, and scaffolds/super-scaffolds, or sets of contigs which are approximately placed and oriented with respect to each other. *De novo* whole genome assembly workflow used for Mithun Genome Project (MGP) was shown below:

The *De novo* whole genome assembly of Mithun was carried out in three different stages:

Stage-1

MaSuRCA is whole genome assembly software. It combines the efficiency of the de Bruijn graph and Overlap-Layout-Consensus (OLC) approaches. MaSuRCA can assemble data sets containing only short reads from Illumina sequencing, thus Illumina Paired-End clean data was used to assemble at this stage of assembly and following is the assembly statistics for Stage-1:

Table: Stage-1 Genome Assembly Statistics



```
                Number of contigs      480463
              Total size of contigs  2649186081
                   Longest contig      125189
                  Shortest contig         113
         Number of contigs > 1K nt     344616  71.7%
        Number of contigs > 10K nt      82599  17.2%
       Number of contigs > 100K nt         14   0.0%
         Number of contigs > 1M nt          0   0.0%
        Number of contigs > 10M nt          0   0.0%
                 Mean contig size        5514
               Median contig size        2812
                N50 contig length       11528
                        contig %A       29.06
                        contig %C       20.94
                        contig %G       20.93
                        contig %T       29.04
                        contig %N        0.03
                 contig %non-ACGTN        0.00
    Number of contig non-ACGTN nt           0

  Percentage of assembly in scaffolded contigs     2.0%
Percentage of assembly in unscaffolded contigs    98.0%
```

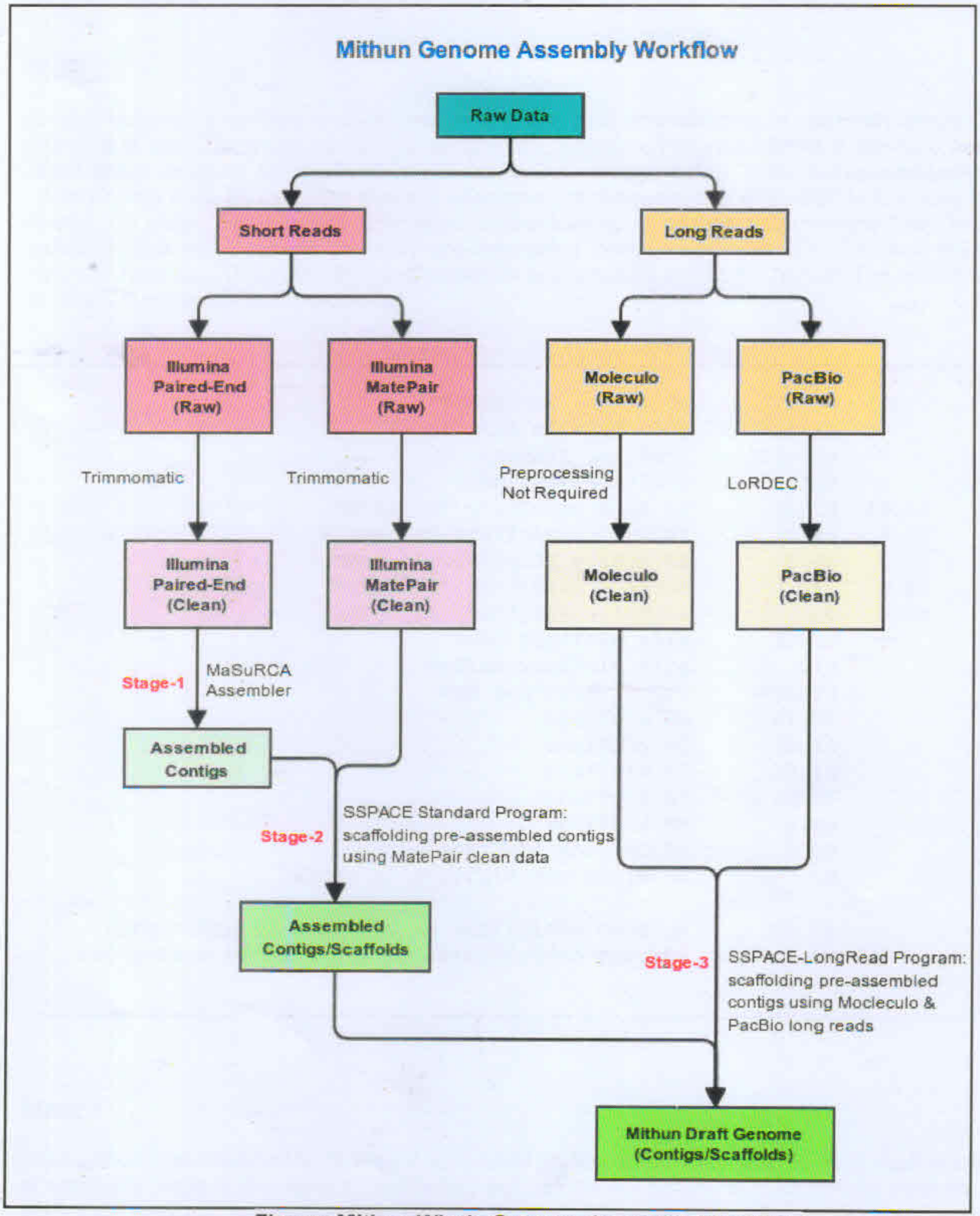
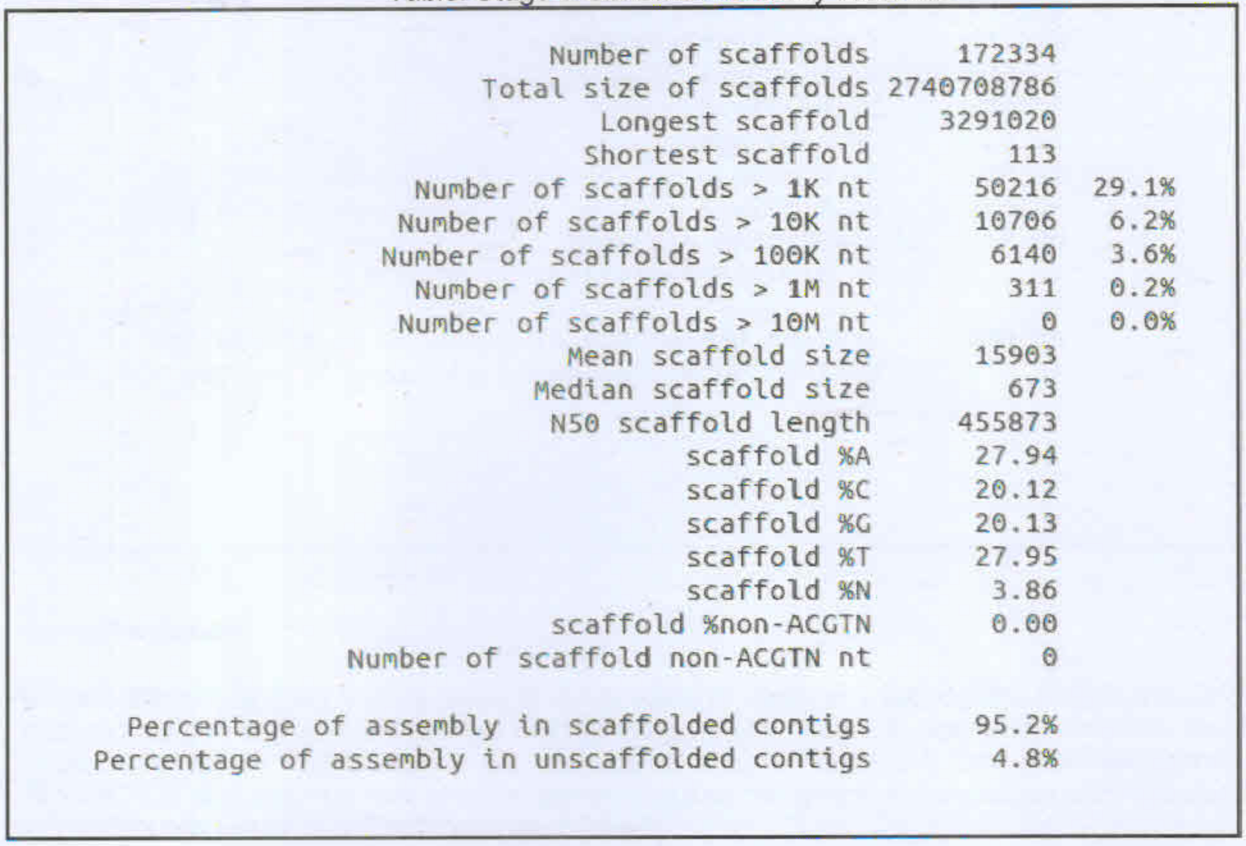

Mithun Genome Assembly Workflow
Raw Data
Short Reads
Long Reads
Illumina Paired-End (Raw)
Illumina MatePair (Raw)
Moleculo (Raw)
PacBio (Raw)
Trimmomatic
Trimmomatic
Preprocessing Not Required
LoRDEC
Illumina Paired-End (Clean)
Illumina MatePair (Clean)
Moleculo (Clean)
PacBio (Clean)
Stage-1
MaSuRCA Assembler
Assembled Contigs
Stage-2
SSPACE Standard Program: scaffolding pre-assembled contigs using MatePair clean data
Assembled Contigs/Scaffolds
Stage-3
SSPACE-LongRead Program: scaffolding pre-assembled contigs using Mocleculo & PacBio long reads
Mithun Draft Genome (Contigs/Scaffolds)

Figure: Mithun Whole Genome Assembly Workflow

Stage-2

Contigs obtained from Step-1 were used for the first level of scaffolding at assembly Stage-2, this level of scaffolding was carried out by SSPACE-Standard Program.SSPACE standard is a stand-alone program for scaffolding pre-assembled contigs using NGS paired-readand/or matepair data data. By using the distance information of matepairclean data, SSPACE is able to assess the order, distance and orientation of pre-assembled contigs and combine them into scaffolds. The input data is given by pre-assembled contig sequences (FASTA) and NGS matepair-read data (FASTQ). The final scaffolds are provided in FASTA format. The statistics for Stage-2 assembly is shown below:

Table: Stage-2 Genome Assembly Statistics



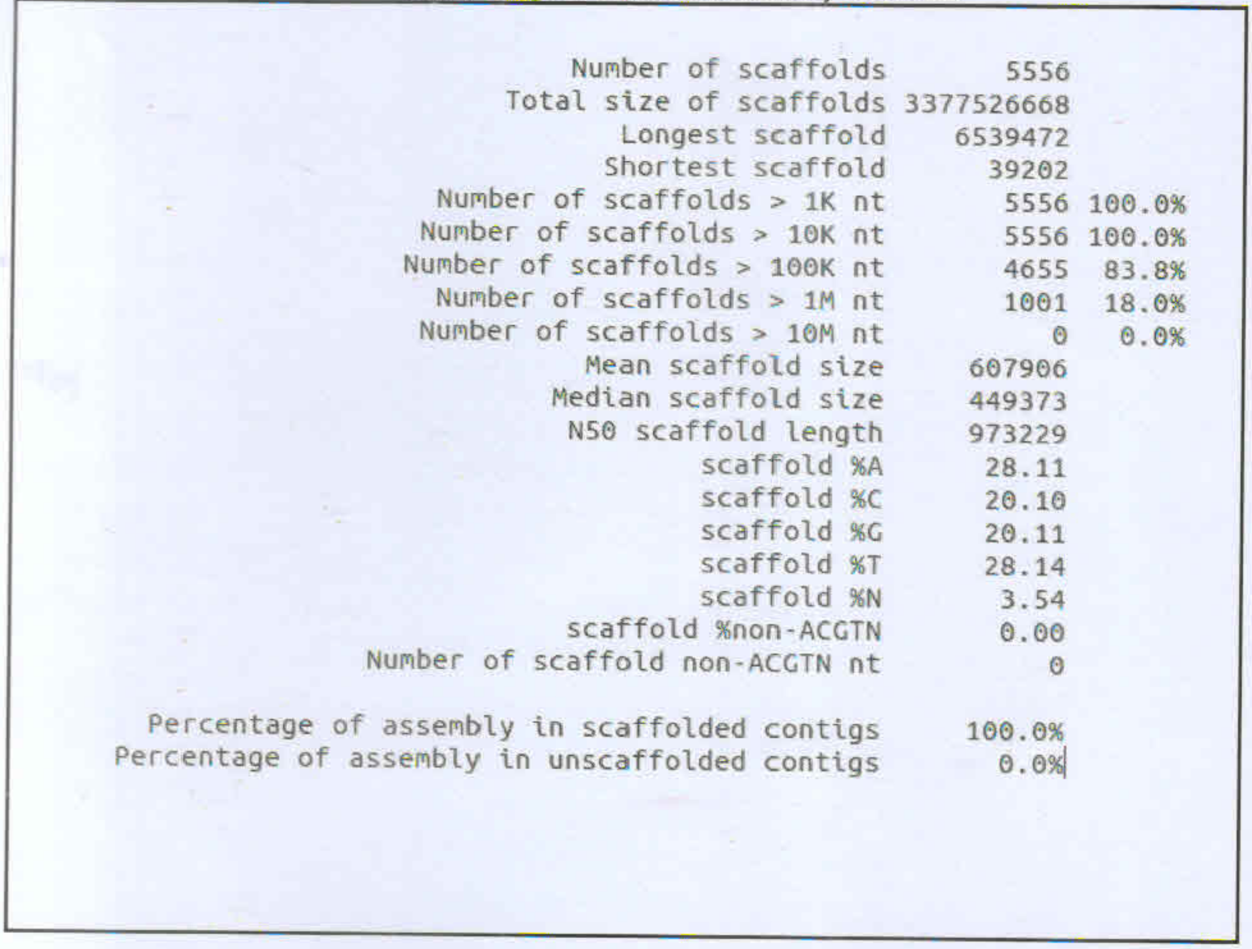
```
                    Number of scaffolds     172334
                Total size of scaffolds 2740708786
                       Longest scaffold    3291020
                      Shortest scaffold        113
            Number of scaffolds > 1K nt      50216  29.1%
           Number of scaffolds > 10K nt      10706   6.2%
          Number of scaffolds > 100K nt       6140   3.6%
            Number of scaffolds > 1M nt        311   0.2%
           Number of scaffolds > 10M nt          0   0.0%
                     Mean scaffold size      15903
                   Median scaffold size        673
                    N50 scaffold length     455873
                             scaffold %A     27.94
                             scaffold %C     20.12
                             scaffold %G     20.13
                             scaffold %T     27.95
                             scaffold %N      3.86
                    scaffold %non-ACGTN       0.00
         Number of scaffold non-ACGTN nt         0

  Percentage of assembly in scaffolded contigs     95.2%
Percentage of assembly in unscaffolded contigs      4.8%
```

Stage-3

Contigs/Scaffolds obtained from Step-2 were used for the second and final level of scaffolding at assembly Stage-3, this level of scaffolding was carried out by SSPACE-LongRead Program. SSPACE-LongRead is a stand-alone program for scaffolding pre-assembled contigs using long reads (e.g. PacBio RS reads,Moleculo reads). Using the long read information, contigs (or scaffolds) are placed in the right order and orientation in so-called super-scaffolds. The input data is given by contigs/scaffolds sequences (FASTA) from step-2 and PacBio RS II &Moleculo long readsdata (FASTA). The final super-scaffolds are provided in FASTA format. The statistics for Stage-3 assembly is shown below:

Table: Stage-3 Genome Assembly Statistics



```
                    Number of scaffolds       5556
              Total size of scaffolds 3377526668
                     Longest scaffold    6539472
                    Shortest scaffold      39202
          Number of scaffolds > 1K nt       5556 100.0%
         Number of scaffolds > 10K nt       5556 100.0%
        Number of scaffolds > 100K nt       4655  83.8%
          Number of scaffolds > 1M nt       1001  18.0%
         Number of scaffolds > 10M nt          0   0.0%
                   Mean scaffold size     607906
                 Median scaffold size     449373
                  N50 scaffold length     973229
                           scaffold %A     28.11
                           scaffold %C     20.10
                           scaffold %G     20.11
                           scaffold %T     28.14
                           scaffold %N      3.54
                   scaffold %non-ACGTN      0.00
       Number of scaffold non-ACGTN nt         0

  Percentage of assembly in scaffolded contigs    100.0%
Percentage of assembly in unscaffolded contigs      0.0%
```

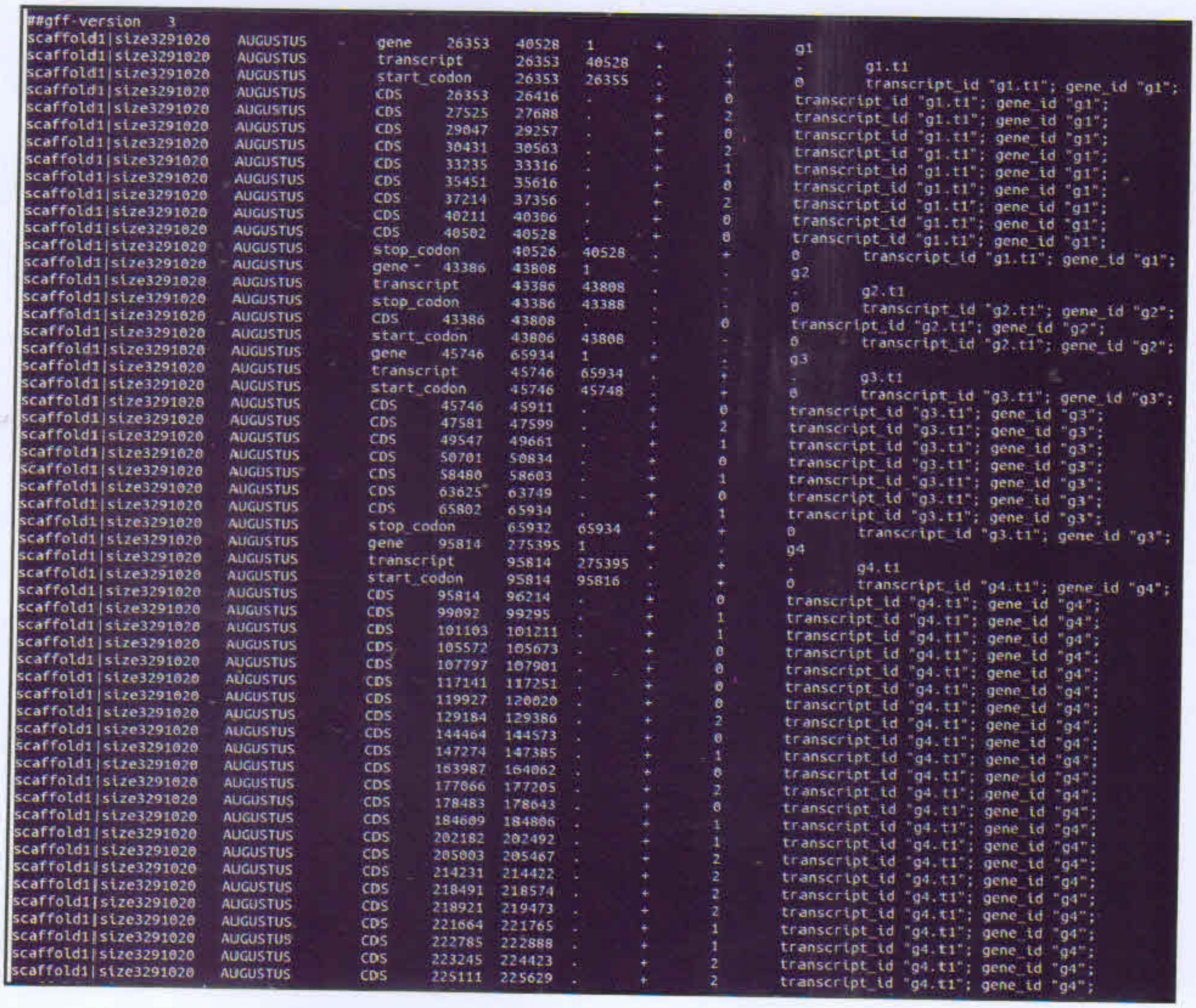
Gene Prediction

Gene prediction or gene finding refers to the process of identifying the regions of genomic DNA that encode genes in the assembled draft genome. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced. AUGUSTUS is a program that predicts genes in eukaryotic genomic sequences and was used to predict genes in Mithun Genome Project (MGP).

A total of 64,311 genes having 394,346 CDS sequences were predicted for Mithun draft genome. A snapshot of predicted genes was depicted in the figure shown below:

Table: Augustus Gene Prediction Statistics

S.No.	Predicted Features	Number
1	Genes	64311
2	Protein-Coding Genes	64311
3	mRNAs	64311
4	Exons	394346
5	CDSs	394346



```
##gff-version   3
scaffold1|size3291020   AUGUSTUS        gene    26353   40528   1       +       .       g1
scaffold1|size3291020   AUGUSTUS        transcript      26353   40528   .       +       .       g1.t1
scaffold1|size3291020   AUGUSTUS        start_codon     26353   26355   .       +       0       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        CDS     26353   26416   .       +       0       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        CDS     27525   27688   .       +       2       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        CDS     29047   29257   .       +       0       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        CDS     30431   30563   .       +       2       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        CDS     33235   33316   .       +       1       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        CDS     35451   35616   .       +       0       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        CDS     37214   37356   .       +       2       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        CDS     40211   40306   .       +       0       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        CDS     40502   40528   .       +       0       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        stop_codon      40526   40528   .       +       0       transcript_id "g1.t1"; gene_id "g1";
scaffold1|size3291020   AUGUSTUS        gene -  43386   43808   1       -       .       g2
scaffold1|size3291020   AUGUSTUS        transcript      43386   43808   .       -       .       g2.t1
scaffold1|size3291020   AUGUSTUS        stop_codon      43386   43388   .       -       0       transcript_id "g2.t1"; gene_id "g2";
scaffold1|size3291020   AUGUSTUS        CDS     43386   43808   .       -       0       transcript_id "g2.t1"; gene_id "g2";
scaffold1|size3291020   AUGUSTUS        start_codon     43806   43808   .       -       0       transcript_id "g2.t1"; gene_id "g2";
scaffold1|size3291020   AUGUSTUS        gene    45746   65934   1       +       .       g3
scaffold1|size3291020   AUGUSTUS        transcript      45746   65934   .       +       .       g3.t1
scaffold1|size3291020   AUGUSTUS        start_codon     45746   45748   .       +       0       transcript_id "g3.t1"; gene_id "g3";
scaffold1|size3291020   AUGUSTUS        CDS     45746   45911   .       +       0       transcript_id "g3.t1"; gene_id "g3";
scaffold1|size3291020   AUGUSTUS        CDS     47581   47599   .       +       2       transcript_id "g3.t1"; gene_id "g3";
scaffold1|size3291020   AUGUSTUS        CDS     49547   49661   .       +       1       transcript_id "g3.t1"; gene_id "g3";
scaffold1|size3291020   AUGUSTUS        CDS     50701   50834   .       +       0       transcript_id "g3.t1"; gene_id "g3";
scaffold1|size3291020   AUGUSTUS        CDS     58480   58603   .       +       1       transcript_id "g3.t1"; gene_id "g3";
scaffold1|size3291020   AUGUSTUS        CDS     63625   63749   .       +       0       transcript_id "g3.t1"; gene_id "g3";
scaffold1|size3291020   AUGUSTUS        CDS     65802   65934   .       +       1       transcript_id "g3.t1"; gene_id "g3";
scaffold1|size3291020   AUGUSTUS        stop_codon      65932   65934   .       +       0       transcript_id "g3.t1"; gene_id "g3";
scaffold1|size3291020   AUGUSTUS        gene    95814   275395  1       +       .       g4
scaffold1|size3291020   AUGUSTUS        transcript      95814   275395  .       +       .       g4.t1
scaffold1|size3291020   AUGUSTUS        start_codon     95814   95816   .       +       0       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     95814   96214   .       +       0       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     99092   99295   .       +       1       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     101103  101211  .       +       1       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     105572  105673  .       +       0       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     107797  107901  .       +       0       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     117141  117251  .       +       0       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     119927  120020  .       +       0       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     129184  129386  .       +       2       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     144464  144573  .       +       0       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     147274  147385  .       +       1       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     163987  164062  .       +       0       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     177066  177205  .       +       2       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     178483  178643  .       +       0       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     184609  184806  .       +       1       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     202182  202492  .       +       1       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     205003  205467  .       +       2       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     214231  214422  .       +       2       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     218491  218574  .       +       2       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     218921  219473  .       +       2       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     221664  221765  .       +       1       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     222785  222888  .       +       1       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     223245  224423  .       +       2       transcript_id "g4.t1"; gene_id "g4";
scaffold1|size3291020   AUGUSTUS        CDS     225111  225629  .       +       2       transcript_id "g4.t1"; gene_id "g4";
```

Figure: Snapshot of genes predicted using AUGUSTUS

ANNEXURE - VIII

INDIAN COUNCIL OF AGRICULTURAL RESEARCH

(For Guidelines Refer ANNEXURE – XI(H))

PROFORMA FOR RESEARCH PERFORMANCE EVALUATION OF INDIVIDUAL SCIENTIST

1. Institute Project Code * - **IXX10452**

2. Evaluation by PI on the contribution of the team in the project including self

S. No.	Name	Status in the project (PI/CC-PI/Co-PI)	*Rating in the scale of 1 to 10
1	Sabyasachi Mukherjee	PI	7
2	Anupama Mukherjee	Co-PI	7
3	Kezhavituo Vupru	Co-PI	7
4	Kobu Khate	Co-PI	7

3. Signature of PI 17/6/16

* Individual scientists participating in the project would be assessed for their performance through an appraisal system in a scale of 1 to 10 for each of the following attributes:

S. No.	Criteria	Marks
1.	Percentage of the assigned activity completed	40
2.	Quality of the completed activity	10
3.	Authenticity/reliability of the data generated	10
4.	Enthusiasm and sincerity to work	10
5.	Inferences made	10
6.	Collaboration and cooperation demonstrated in performing the task at hand	10
7.	Amenability to scientific/academic/laboratory discipline	10
	Total Score	100

ANNEXURE - IX

INDIAN COUNCIL OF AGRICULTURAL RESEARCH

(For Guidelines Refer ANNEXURE – XI(I))

PROFORMA FOR EVALUATION OF A RESEARCH PROJECT AFTER COMPLETION BY PI

1. Institute Project Code - **IXX10452**
2. Evaluation research project after completion by PI

S. No.	Criteria	Methodology	Marks (output)	Self Evaluation by PI
1.	Achievements Against approved and stipulated outputs under project	**Qualitative and quantitative assessment of objectives and stipulated outputs under the project will be carried out**	**75**	
		a) Activity Input /Projected Output/ Output Achieved	35	**34**
		b) Extent to which standard design methodology, experimental designs, test procedures, analytical methods followed	10	**9**
		c) Does the data justify the conclusions?	05	**04**
		d) Innovativeness and creating of new knowledge	10	**09**
		e) Additional outputs over those stipulated under the project	05	**04**
		f) Creation of linkages for commercialization of technology developed under the project	05	**02**
		g) Is scientific input commensurate to output (manpower, Financial input and time duration)?	05	**04**
2.	Publication/ awards	Assessment will be done in respect of: Research papers; Reports/Manuals; Working and Concept Papers; Books/Book Chapters/Bulletins. Quality of publication (s) and Awards /Scientific recognitions received	**10**	**02**
3.	Additional facilities created	Facilities created in terms of laboratory. Research set-up, instrumentation, etc. during the project.	**05**	**04**
4.	Human Resource Development (Scientific and Technical)	Scientist trained in different areas	**05**	**02**
5.	Revenue generated under the project/ avenues created for revenue generation	Resources and revenues generated	**05**	**00**

6.	Product/Process/Technology/ IPR / commercial value of the technology developed	Details to be provided on a) Products b) Process c) Technology d) IPR e) Registration of the varieties		**10**	**00**
7.	Quality of available documents of the project duly authenticated	Research Project Files, Data, Reports etc.		**05**	**05**
Total Marks				**115**	**79**
8.	Timelines of execution of the project	Marks will be deducted if extension sought over the approved project duration beyond recorded and officially granted extension with recorded reasons	Marks to be deducted		
		Up to 5%	01		
		Up to 10%	02		
		Up to 30 %	03		
		Beyond 30 %	05		
Net Score: Score obtained to be counted out of 100 to compensate for activities not relevant to the project				**100**	**68.69**

However, looking into the requirements of different research institutes and disciplines, IRC may modify the indicators, their weights and total scores. The time gap for assessment of different indicators may also be decided by IRC

3. Signature of PI 17/6/16

ANNEXURE - X

INDIAN COUNCIL OF AGRICULTURAL RESEARCH

(For Guidelines Refer ANNEXURE – XI (J))

PROFORMA FOR EVALUATION OF A RESEARCH PROJECT AFTER COMPLETION BY EVALUATION COMMITTEE

1. Institute Project Code - **IXX10452**
2. Evaluation research project after completion by Evaluation Committee

S. No.	Criteria	Methodology	Marks (output)	Evaluation by Evaluation Committee
1.	Achievements Against approved and stipulated outputs under project	**Qualitative and quantitative assessment of objectives and stipulated outputs under the project will be carried out**	**75**	
		a) Activity Input /Projected Output/ Output Achieved	35	
		b) Extent to which standard design methodology, experimental designs, test procedures, analytical methods followed	10	
		c) Does the data justify the conclusions?	05	
		d) Innovativeness and creating of new knowledge	10	
		e) Additional outputs over those stipulated under the project	05	
		f) Creation of linkages for commercialization of technology developed under the project	05	
		g) Is scientific input commensurate to output (manpower, Financial input and time duration)?	05	
2.	Publication/ awards	Assessment will be done in respect of: Research papers; Reports/Manuals; Working and Concept Papers; Books/Book Chapters/Bulletins. Quality of publication (s) and Awards /Scientific recognitions received	**10**	
3.	Additional facilities created	Facilities created in terms of laboratory. Research set-up, instrumentation, etc. during the project.	**05**	
4.	Human Resource Development (Scientific and Technical)	Scientist trained in different areas	**05**	
5.	Revenue generated under the project/ avenues created for	Resources and revenues generated	**05**	

	revenue generation				
6.	Product/Process/Technology/ IPR / commercial value of the technology developed	Details to be provided on a) Products b) Process c) Technology d) IPR e) Registration of the varieties		**10**	
7.	Quality of available documents of the project duly authenticated	Research Project Files, Data, Reports etc.		**05**	
Total Marks				**115**	
8.	Timelines of execution of the project	Marks will be deducted if extension sought over the approved project duration beyond recorded and officially granted extension with recorded reasons	Marks to be deducted		
		Up to 5%	01		
		Up to 10%	02		
		Up to 30 %	03		
		Beyond 30 %	05		
Net Score: Score obtained to be counted out of 100 to compensate for activities not relevant to the project				**100**	

4. Signature of Evaluation Committee