Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in
NGS-240592| Date: 10-12-25

Final Report of

Whole Transcriptome Sequencing

& Analysis

On Illumina Novaseq X Plus Platform

Submitted to:

Dr. Ankita Singh
Scientific Affairs
ICGA Foundation
India

Submitted by:
Unigenome
A life sciences division of Unipath Specialty Laboratory Ltd.
3rd Floor, PASL House | Beside Sahjanand College | Panjara pol, Ambawadi,
Ahmedabad – 380015 | Gujarat, India
Tel :- +91-79-66197701
genomics@unipath.in | www.unigenome.in

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

Table of Contents

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

1. Project Details

Service Type	**RNA Sequencing**
Platform	Illumina Novaseq X Plus
Read Length	2 X 150 PE
Data	~ 18-20 Gb/Sample

2. Sample Details

Type of Sample	**Frozen Tissue**
No. of Samples	8
Sample Name	ICGA_BC_NT_125, ICGA_BC_TT_125, ICGA_BC_NT_126, ICGA_BC_TT_126, ICGA_BC_NT_127, ICGA_BC_TT_127, ICGA_BC_NT_129, ICGA_BC_TT_129
Shipping condition	NA
No. of Libraries prepared	8

3. Methods

3.1. Isolation and Quantitative analysis of RNA:

RNA sample was extracted from tissue sample using QIAGEN RNeasy mini kit (CAT.NO:74106). RNA quantity was measured using Qubit® 4.0 fluorometer and quality were analyzed by using 1% agarose gel.

3.2. Preparation of Library

The paired-end sequencing library was prepared using NEBNext® Ultra™ RNA Library Prep Kit for Illumina. (NEB #E7770). The library preparation process was initiated with 100 ng input. Ribosomal RNA was removed using depletion was carried out using ribodepletion kit KAPA RNA HyperPrep Kit with RiboErase (HMR), Cat no: KK8561) as per user manual. Ribo-depleted RNA was subjected to fragmentation, first & second-strand cDNA synthesis, end-repair, 3′ adapter ligation, selective enrichment of adapter-ligated DNA fragments through PCR amplification, followed by validation of

Library on Agilent 4150 tape station. The final library was pooled with other samples, denatured & loaded on to flow cell. On the flowcell, cluster generation & sequencing was performed using Illumina Novaseq X plus platform to generate 2×150bp paired-end (PE) reads.

3.3. Quantity and quality check (QC) of library on Agilent Tape Station 4150

The amplified libraries were analyzed on TapeStation 4150 (Agilent Technologies) using High Sensitivity D1000 ScreenTape® as per manufacturer's instructions.

3.4. Cluster Generation and Sequencing

After obtaining the Qubit concentration for the library and the mean peak size from Tape Station profile, library will be loaded onto illumine Novaseq X Plus for cluster generation and sequencing. Paired-End sequencing allows the template fragments to be sequenced in both the forward and reverse directions. The library molecules will bind to complementary adapter oligos on paired-end flow cell. The adapters are designed to allow selective cleavage of the forward strands after re-synthesis of the reverse strand during sequencing. The copied reverse strand is then used to sequence from the opposite end of the fragment.

3.5. Results: RNA QC on 1% agarose gel

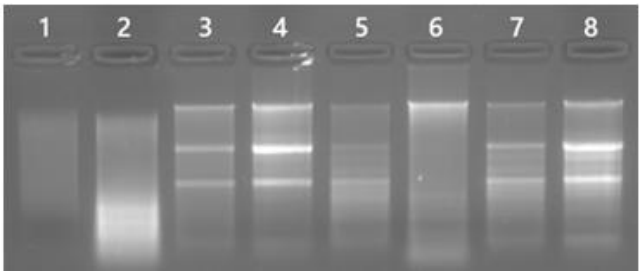

Figure 1: QC on 1.0% agarose gel.

Lane id	Sample name
1	ICGA_BC_NT_125
2	ICGA_BC_TT_125
3	ICGA_BC_NT_126
4	ICGA_BC_TT_126
5	ICGA_BC_NT_127
6	ICGA_BC_TT_127
7	ICGA_BC_NT_129
8	ICGA_BC_TT_129

3.6. Quantification using Qubit 4.0 Fluorometer

S. N	Sample ID	Concentration (ng/µl)	Volume (µl)	Yield (µg)	Remarks
1	ICGA_BC_NT_125	8.34	20	0.17	QC PASS
2	ICGA_BC_TT_125	8.8	20	0.18	QC PASS
3	ICGA_BC_NT_126	89.8	20	1.80	QC PASS
4	ICGA_BC_TT_126	88.6	20	1.77	QC PASS
5	ICGA_BC_NT_127	88.4	20	1.77	QC PASS
6	ICGA_BC_TT_127	32.8	20	0.66	QC PASS
7	ICGA_BC_NT_129	89	20	1.78	QC PASS
8	ICGA_BC_TT_129	90	20	1.80	QC PASS

4. Quantity and quality check (QC) of library on Agilent Tape Station 4150

4.1. TapeStation 4150 profiles of library loaded in Agilent High Sensitivity D1000ScreenTape®:

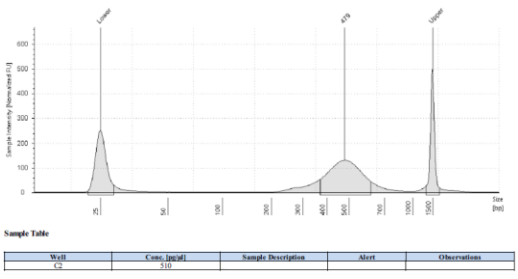

Figure 2: ICGA_BC_NT_125

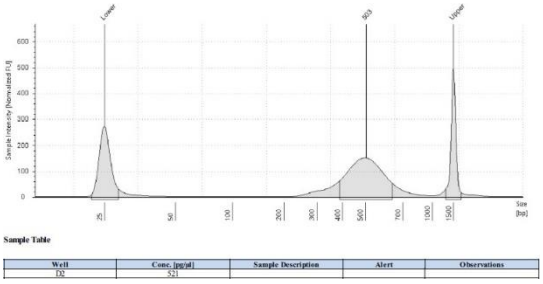

Figure 3: ICGA_BC_TT_125

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

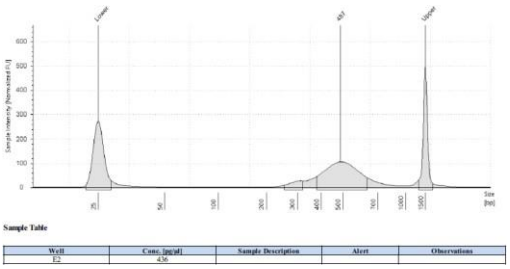

Figure 4: ICGA_BC_NT_126

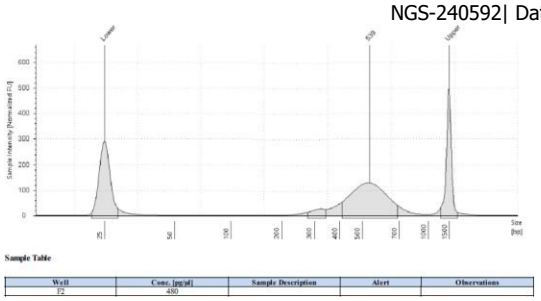

Figure 5: ICGA_BC_TT_126

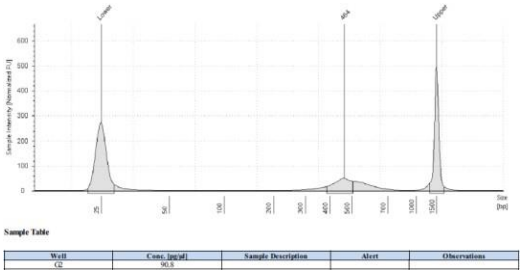

Figure 6: ICGA_BC_NT_127

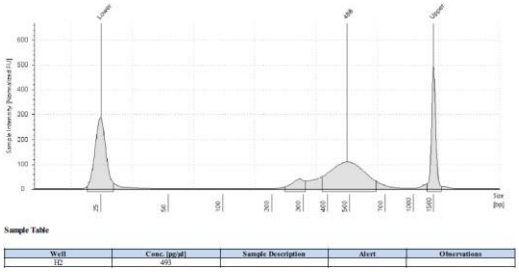

Figure 7: ICGA_BC_TT_127

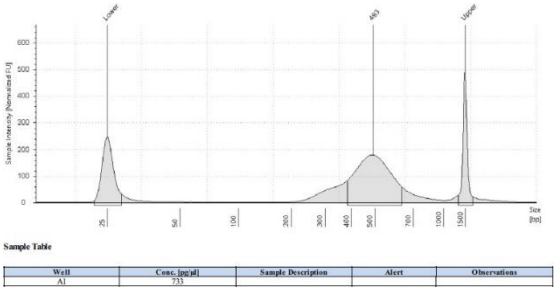

Figure 8: ICGA_BC_NT_129

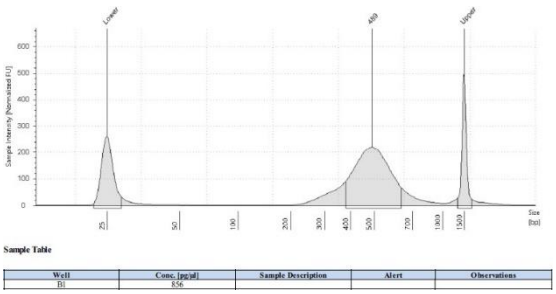

Figure 9: ICGA_BC_TT_12

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

4.2.Wet Lab Inferences

- The libraries were prepared from the samples by KAPA mRNA Hyper Prep Kit for Illumina (CAT #KK8581).
- The average size of libraries is 479bp, 503bp, 487bp, 539bp, 464bp, 488bp, 493bp, 489bp, for ICGA_BC_NT_125
- ICGA_BC_TT_125, ICGA_BC_NT_126, ICGA_BC_TT_126, ICGA_BC_NT_127, ICGA_BC_TT_127, ICGA_BC_NT_129 and ICGA_BC_TT_129.The libraries were sequenced on Illumina Novaseq X Plus platform using 2-x 150 bp chemistry to generate ~18-20 Gb/Sample.

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

5. Bioinformatics Analysis

5.1. Bioinformatics Workflow

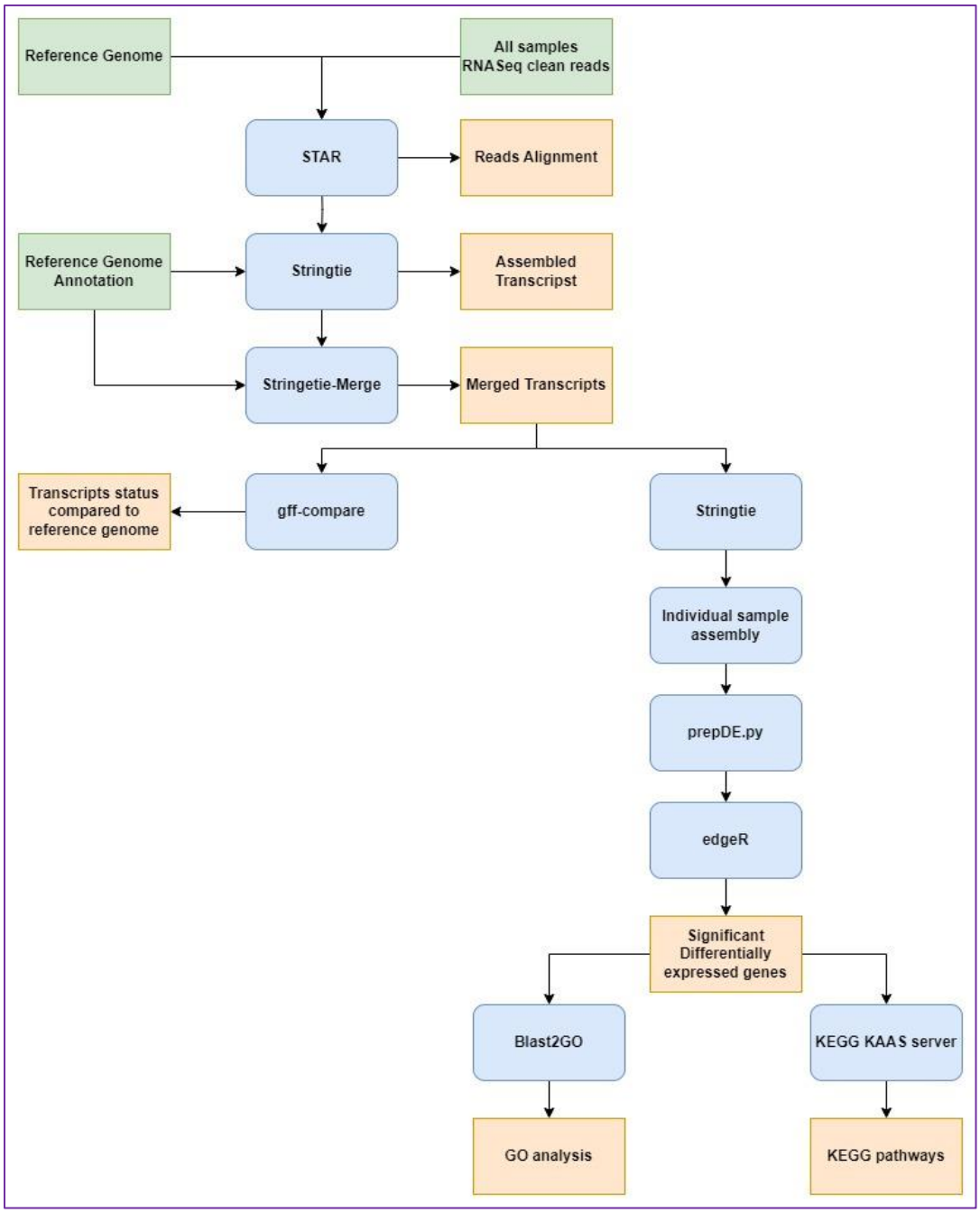

Figure 10: Bioinformatics Analysis Workflow

Workflow overview: Raw data was filtered to remove adapters and low-quality bases. Clean reads thus produced were used for analysis. Reference guided transcript assembly was performed for samples first by mapping clean reads on reference genome using STAR (v2.7.10a) aligner and then performing transcript assembly by StringTie (v 2.2.1). A consensus set of transcripts was obtained using SringTie merge function which merges together all the gene structures found in any of the samples. Transcript abundance was then estimated for individual sample using merged transcript consensus again using StringTie and read counts thus obtained for each transcript were taken as input for differential expression analysis using edgeR package. Gene Ontology and pathway analysis of the significantly differentially expressed transcripts were performed using Blast2go cli and KEGG-KAAS server respectively. Overall bioinformatics workflow is graphically represented in figure above and software version along with its use in table below.

5.2. Detail of Software used for analysis

Software	Version	Application
Trim Galore	0.6.4	Adapter and low-quality data removal
STAR	2.7.10	Mapping of reads to reference genome
Stringtie	2.2.1	Reference based assembly
gffcompare	0.12.6	Comparing reference genome gff with merged gff file generated from stringtie
gffread	0.12.7	Fetching fasta sequences corresponding to gff file generated after reference-based assembly by stringtie
Blastx	2.30.0+	Similarity search against NCBI NR database
Blast2go cli	1.5	GO mapping and annotation
edgeR	3.6.2	Differential expression analysis and its visualization
KEGG KAAS	Webserver	Pathway analysis
clusterProfiler	4.12.6	Over-representation (or enrichment) analysis

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

5.3.Data statistics

The next generation sequencing for these samples were performed on the Illumina Novaseq plus X platform. Statistics of raw data is shown in Table below:

Table 1: Data statistics

Sample	Reads in R1	Reads in R2	Total Reads (R1+R2)	Total Bases (R1+R2)	Total Data (GB)
ICGA-BC-NT-125	72417882	72417882	144835764	23028886476	23.03
ICGA-BC-NT-126	72545812	72545812	145091624	23069568216	23.07
ICGA-BC-NT-127	79142618	79142618	158285236	25167352524	25.17
ICGA-BC-NT-129	76856474	76856474	153712948	24440358732	24.44
ICGA-BC-TT-125	74210455	74210455	148420910	23598924690	23.6
ICGA-BC-TT-126	72602199	72602199	145204398	23087499282	23.09
ICGA-BC-TT-127	75547160	75547160	151094320	24023996880	24.02
ICGA-BC-TT-129	70056690	70056690	140113380	22278027420	22.28

The raw data has been provided in data deliverables folder entitled "***01_Raw_Data***". These reads were subjected to adapter filtration and low quality bases filtration before analysis to generate clean/high quality (HQ) reads.

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

5.4. Reference genome information

Based on information received from client, the reference genome of *Homo sapiens* (human) (GCF_000001405.40) and its corresponding GFF file was downloaded from (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.40_GRCh38.p14/).

Table 2: Reference genome stats

Description	Stats
Total Chromosomes/Scaffolds	705
Total genome length (bp)	3,298,430,636
Mean chromosome/scaffold size	4,678,625
Max chromosome/scaffold size (bp)	248,956,422
Min chromosome/scaffolds size (bp)	970

The reference GFF file contains genome annotation. It includes information (locus and description) regarding genes and proteins. The fasta sequence of the reference genome along with the corresponding GFF file is provided with the deliverables in the folder named **"02_reference_genome_and_gff"** for client's reference.

Table 3: Detail Information of GFF file

Genome elements	# Sequences
mRNA	144447
CDS	1836136
exon	2301289
gene	47876

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

5.5. Mapping to reference genome

Indexing of the reference genome was carried out using STAR genomeGenerate mode option. Then the input reads, in FASTQ format, along with the indexed reference genome generated in previous step was given to the STAR aligner. The analysis begins by mapping HQ reads against the reference genome to identify the positions from where the reads originated. This mapping information allows us to collect subsets of the reads corresponding to each gene, and then to assemble and quantify transcripts represented by those reads. Hence, the high-quality reads were mapped to reference genome using STAR aligner to create alignment in BAM format for each sample with default parameters. The mapping statistics are provided in the table below:

Table 4: Mapping Statistics

Sample Name	Total clean Reads	No. of Mapped Reads	% of mapped read	# uniquely mapped reads	% uniquely mapped reads
ICGA-BC-NT-125	72149275	37104104	51.43	33677874	46.68
ICGA-BC-NT-126	72286617	63654339	88.06	54524678	75.43
ICGA-BC-NT-127	78835961	70381220	89.28	60746952	77.05
ICGA-BC-NT-129	76538753	75296646	98.38	64338373	84.06
ICGA-BC-TT-125	73965951	49678232	67.16	15945694	21.56
ICGA-BC-TT-126	72221778	70916221	98.19	55033379	76.2
ICGA-BC-TT-127	75102744	69862954	93.02	46717111	62.2
ICGA-BC-TT-129	69702563	68143610	97.76	53526342	76.79

Note: Mapping for sample ICGA-BC-NT-125 and ICGA-BC-TT-125 was less than optimum which may indicate presence of reads originating from other organism.

5.6.Transcript assembly

StringTie assembles transcripts from RNA-seq reads that have been aligned to the genome, first grouping the reads into distinct gene loci and then assembling each locus into as many isoforms as are needed to explain the data. Following this, StringTie simultaneously assembles and quantify the final transcripts by using network flow algorithm and starting from most highly abundant transcripts. The reference genome GFF annotation files, containing exon structures of "known" genes, are then used to annotate the

assembled transcripts and quantify the expression of known genes as well derive clues if a novel transcript has been found in the sample.

After assembling each sample, the full set of assemblies is passed to StringTie's merge function, which merges together all the gene structures found in any of the samples. This step is required because transcripts in some of the samples might only be partially covered by reads, and as a consequence only partial versions of them will be assembled in the initial StringTie run. The merge step creates a set of transcripts that is consistent across all samples, so that the transcripts can be compared in subsequent steps in figure below:

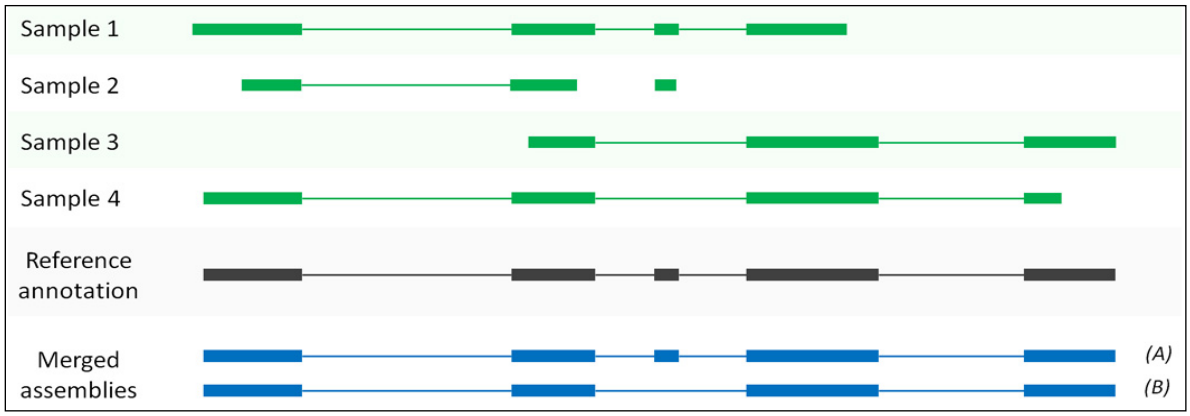

Figure 11: Explanation of merging transcript assemblies using StringTie's merge function. In this example, four partial assemblies from four different samples are merged into two transcripts A and B. Samples 1 and 2 are both consistent with the reference annotation, which is used here to merge and extend them to create transcript A. Samples 3 and 4 are consistent with each other but not with the annotation, and these are merged to create transcript B. Please note that the 4 samples shown here are for example purpose only (Ref: Micheal P et al., 2016).

The statistics of merged transcripts and individual transcript assembly is provided in the table below:

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

Table 5: Transcript assembly statistics

Sample Name	# assembled transcripts	Total assembled (bp)	Mean transcript size (bp)	Max transcript size (bp)
merged.fasta	6,53,846	98,30,19,883	1,503	1,55,532
ICGA-BC-NT-125	4,84,292	35,52,47,280	733.5	1,55,532
ICGA-BC-NT-126	3,09,175	61,03,86,499	1,974.20	1,55,532
ICGA-BC-NT-127	3,04,981	59,83,56,143	1,961.90	1,55,532
ICGA-BC-NT-129	3,21,143	61,65,35,962	1,919.80	1,55,532
ICGA-BC-TT-125	4,16,107	41,76,51,182	1,003.70	1,55,532
ICGA-BC-TT-126	2,55,210	56,41,16,378	2,210.40	1,55,532
ICGA-BC-TT-127	2,82,706	50,90,91,759	1,800.80	1,55,532
ICGA-BC-TT-129	2,69,530	57,79,75,701	2,144.40	1,55,532

The output consists of assembled gene/transcript GTF file for all samples and another GTF file resulting from the merge step as described above. These files have been provided along with the deliverables in the folder named **"03_transcript_assembly_gtf"**. In addition, the transcript sequences derived from these individual assembly is provided in folder **"04_transcript_sequences_fasta"**.

Note: "MSTRG" in output files generated from stringtie stands for "Merged StringTie Transcripts," followed by a unique number. This identifier is created by StringTie merged step to label transcript isoforms that were not previously annotated in the reference genome used.

5.7. Comparison of assembled transcripts with reference transcripts

For comparison of assembled transcripts with reference transcripts, Gffcompare utility was run taking the reference GTF and the string-tie merged GTF file. This produces an output file, which adds to each transcript a "class code" and the name of the transcript from the reference annotation file to check how the predicted transcripts relate to an annotation file. Meaning of each class code is depicted in figure below:

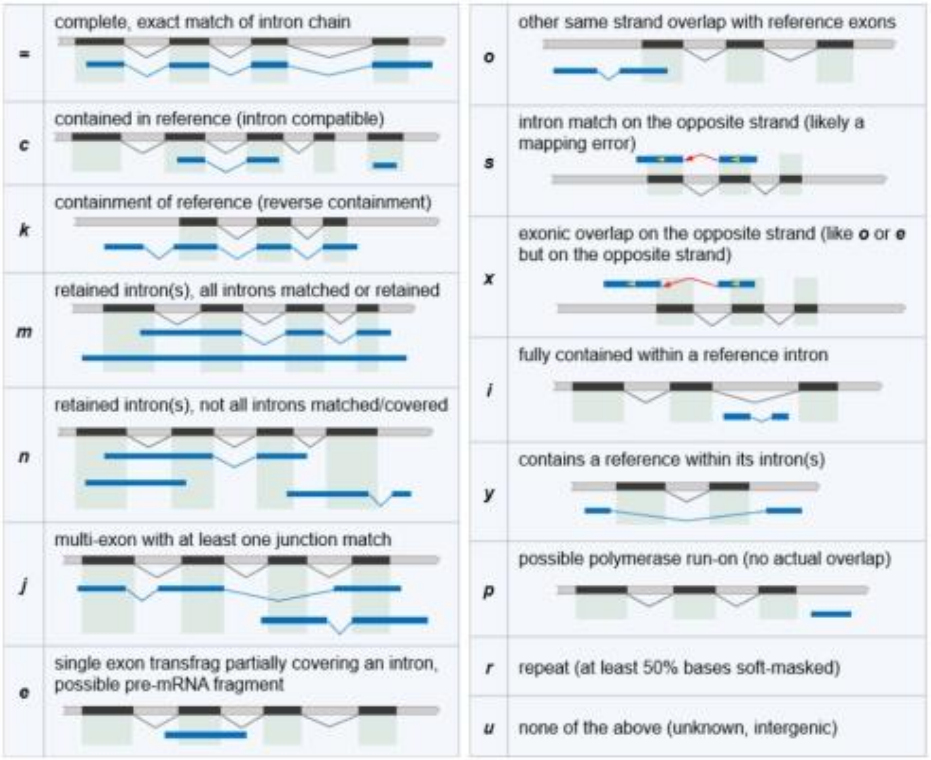

Figure 12: Different class codes and their descriptions

(https://ccb.jhu.edu/software/stringtie/gffcompare.shtml)

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

5.7.1. Identification of novel isoform transcripts

To identify novel isoform transcripts not present in the reference GTF file, gffcompare utility was ran taking the reference GTF and the string-tie merged GTF file. This produces an output file, which adds to each transcript a "class code" and the name of the transcript from the reference annotation file to check how the predicted transcripts relate to an annotation file. Class code "j" means that predicted transcript is a potential novel isoform that shares at least one splice junction with a reference transcript. A total of **3,096** novel isoforms with the class code "j" were extracted and are provided with the deliverables in a file named novel.isoforms.gtf in the folder **"03_transcript_assembly_gff"**. A brief description of the column names in the GTF file is provided in the same folder in the file named ReadMe.txt.

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

5.8. Differential expression analysis

Abundances of the transcripts in all samples were estimated using StringTie with the help of merged transcripts generated from merge stringtie step. A python program (prepDE.py) was used to extract the read count information directly from the files generated by StringTie. For differential expression analysis, sample comparison was made according to grouping information provided by client:

Comparison information

Comparison	Description
TT-125 v/s NT-125	Comparison 1 [TT-125 (ICGA-BC-TT-125) Vs NT-125 (ICGA-BC-NT-125)]
TT-126 v/s NT-126	Comparison 2 [TT-126 (ICGA-BC-TT-126) Vs NT-126 (ICGA-BC-NT-126)]
TT-127 v/s NT-127	Comparison 3 [TT-127 (ICGA-BC-TT-127) Vs NT-127 (ICGA-BC-NT-127)]
TT-129 v/s NT-129	Comparison 4 [TT-129 (ICGA-BC-TT-129) Vs NT-129 (ICGA-BC-NT-129)]
TT v/s NT	Comparison 5 [TT (ICGA-BC-TT-125+ICGA-BC-TT-126+ICGA-BC-TT-127+ICGA-BC-TT-129) Vs NT (ICGA-BC-NT-125+ICGA-BC-NT-126+ICGA-BC-NT-127+ICGA-BC-NT-129)]

Group wise sample information

Table 6: Sample description

Group Name	Sample Name
NT-125	ICGA-BC-NT-125
NT-126	ICGA-BC-NT-126
NT-127	ICGA-BC-NT-127
NT-129	ICGA-BC-NT-129
TT-125	ICGA-BC-TT-125
TT-126	ICGA-BC-TT-126
TT-127	ICGA-BC-TT-127
TT-129	ICGA-BC-TT-129
NT	ICGA-BC-NT-125

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

NT	ICGA-BC-NT-126
NT	ICGA-BC-NT-127
NT	ICGA-BC-NT-129
TT	ICGA-BC-TT-125
TT	ICGA-BC-TT-126
TT	ICGA-BC-TT-127
TT	ICGA-BC-TT-129

Differential gene expression was inferred between samples by applying the R package edgeR It is Bioconductor package based on negative binomial distribution method. The analysis provides tabular result along with the normalized i.e., Counts Per Million mapped reads (CPM) for each involved sample. The description of selected result columns has been given in below mentioned table:

Table 7: Column description of edgeR output

Column Name	Feature identifier
logFC	log2FoldChange: the logarithm (base 2) of the fold change (Test/Control)
logCPM	Log10 of CPM value
pval	pvalue for the statistical significance of this change
FDR	FDR adjusted pvalue (q-value)

The criterion used to identify upregulated and downregulated transcripts along with the significance is provided in the below table:

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

Table 8: Criteria used to identify upregulated and downregulated genes and assigning the significance

Condition	Status
log2FC > 0	Up regulated
log2FC < 0	Down regulated
log2FC > 0 and q-value <0.05	Significantly up regulated
log2FC < 0 and q-value <0.05	Significantly down regulated

Differential expression analysis statistics is provided in table below and the data deliverable are given in the folder ***"05_differential_expression_analysis"*** which contains the Differential gene expression (DGE) analysis results.

Table 9: Differential gene expression statistics

DGE Combination	Total DEGs	Down Regulated	Up Regulated	Significant Downregulated	Significant Upregulated	# Significant DGE
Comparison1	9160	4590	4570	493	768	1261
Comparison2	36826	18306	18520	2666	2481	5147
Comparison3	28797	14251	14546	3678	3923	7601
Comparison4	37177	18706	18471	2228	2467	4695
Comparison5	68563	30178	38385	127	61	188

Annotation for differentially expressed transcripts is provided based on blastX result obtained by similarity search against NCBI's NR database followed by gene ontology analysis using blast2GO.

Note: We recommend a minimum of triplicates per group in RNA-Seq experiments to obtain statistically significant results in differential expression analysis. However, as there are no replicates in this study, we performed DEG analysis using edgeR, which is capable of producing reasonable results in single-sample comparisons.

5.8.1. MA Plot for differentially expressed genes

The plot visualizes the differences between measurements taken in two samples, by transforming the data onto M (log ratio) and A (mean average) scales, then plotting these values. For illustration purpose, MA plot for Comparison1 has been depicted below:

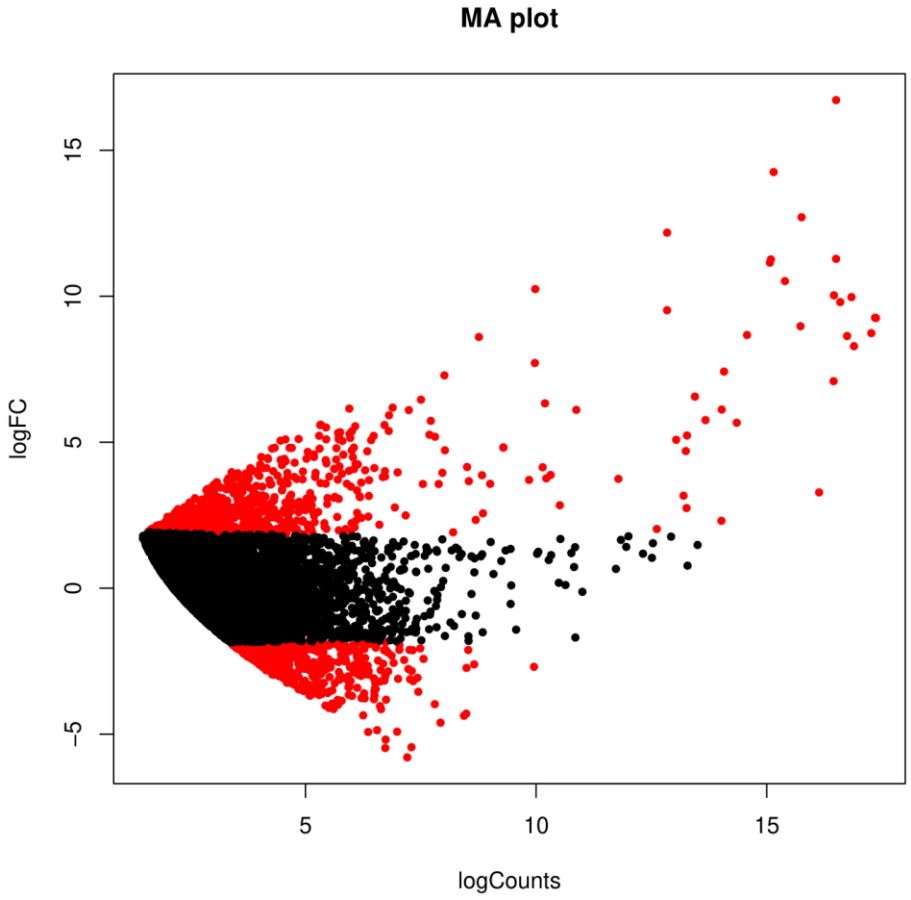

Figure 13: MA plot showing differentially expressed transcripts in Comparison-1. On X-axis normalized counts for all the samples and on Y-axis log2foldchange are plotted. Points colored are with red if the q-value is less than 0.05 and black if the q-value is greater than 0.05.

Note: MA plot for only one comparison has been shown above, however plots for all other combinations have been provided as a part of data deliverables "05_differential_expression_analysis".

5.8.2. Volcano plot for differentially expressed genes

The "volcano plot" arranges expressed genes along dimensions of biological as well as statistical significance. For illustration purpose, Volcano plot for Comparison1 has been depicted below:

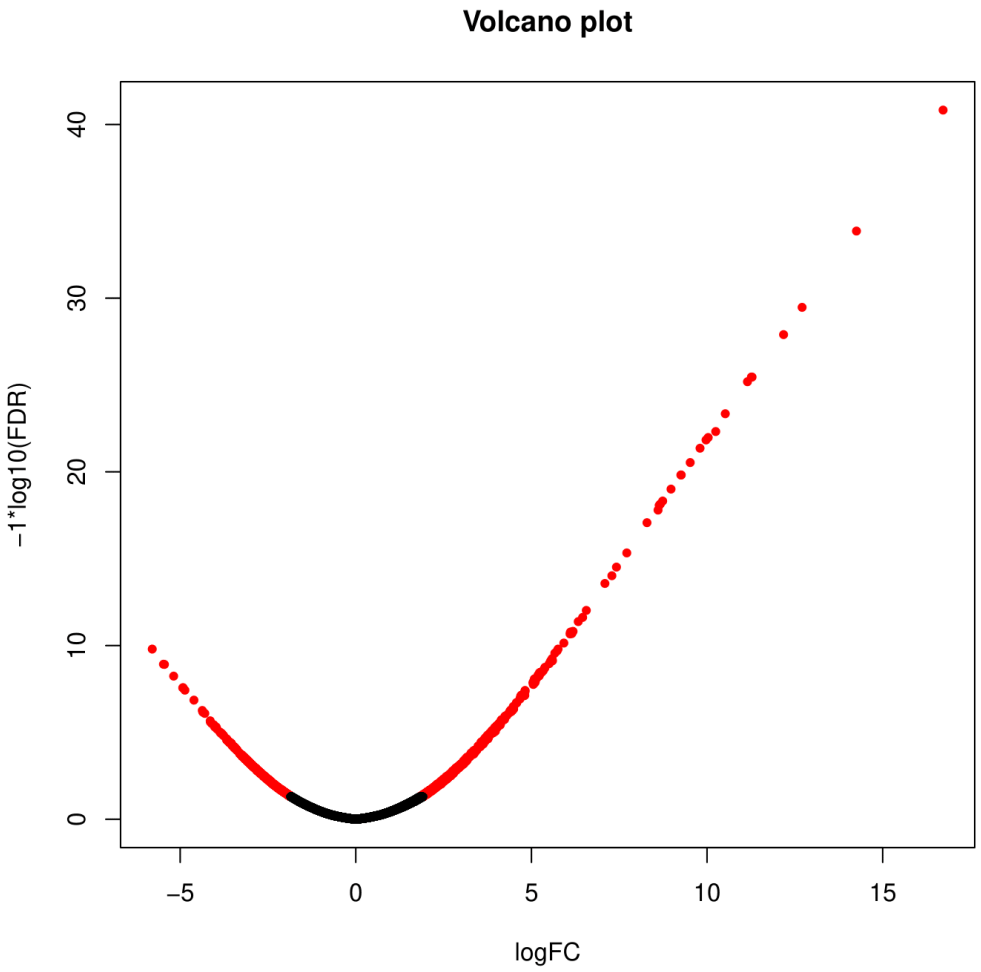

Figure 14: Volcano plot showing differentially expressed transcripts in Comparison-1. X-axis represents the log2fold change values and Y-axis represents log10 of q-value (FDR corrected p-value. Points are colored red if the q-value is less than 0.05 and black if the q-value is greater than 0.05.

Note: Volcano plot for only one comparison has been shown above, however plots for all other combinations has been provided as a part of data deliverables "*05_differential_expression analysis*".

5.8.3. Heatmap for significantly differential expressed genes

List of transcripts were selected for heatmap generation based on following criteria:

- 25 transcripts (downregulated) and 25 transcripts (upregulated) having lowest FDR/adjusted pvalue
- Transcripts must have proper annotations.

pheatmap package from R software was used for producing heatmap. The colour coding ranges from red to blue where shades of red represent high transcripts expression and shades of blue represents low transcript expression.

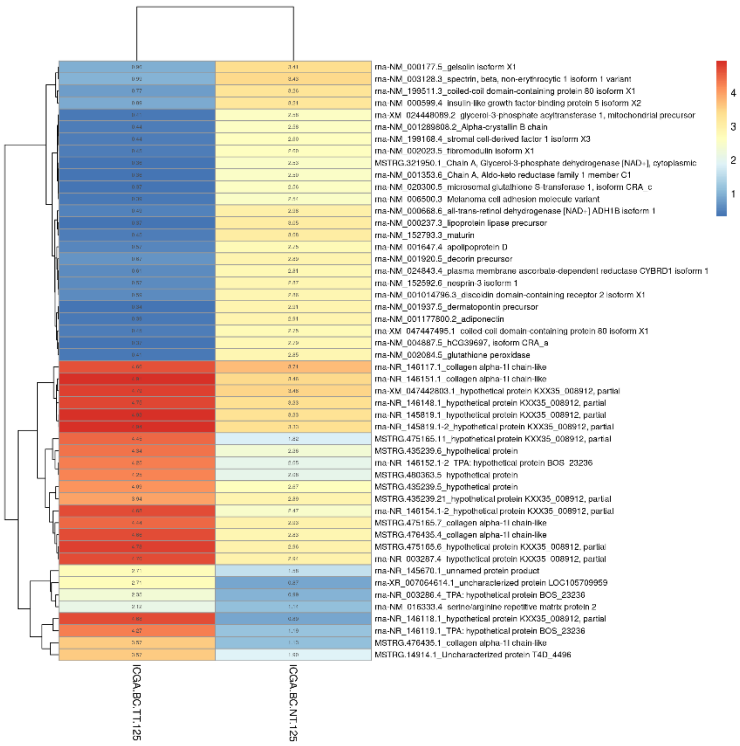

Figure 15: Heatmap representing most significant genes expressed in the samples was plotted using log10 of normalized read count values (CPM) for Comparison-1, where shades of blue represent downregulated genes and shades of red represents highly expressed genes.

Note: Heatmap plot for only one comparison has been shown above, however plots for all other combinations have been provided as a part of data deliverables "*05_differential_expression_analysis*".

5.8.4. Gene Ontology (GO) analysis for significant DEG

The Gene Ontology project provides controlled vocabularies of defined terms representing gene product properties. These cover three domains: Cellular Component, the parts of a cell or its extracellular environment; Molecular Function, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and Biological Process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms. GO was assigned to **significant differentially expressed transcripts using Blast2go cli**. Single gene can be assigned with multiple GO categories and hence multiple GO terms. The GO domain distribution is shown in table below:

Table 10: GO Distribution of Significantly Differential expression transcripts

Sample Name	Significant DGE	# Seq with GO	Biological Process	Cellular Component	Molecular Function
Comparison1	1261	504	445	463	450
Comparison2	5147	2291	2041	2127	2091
Comparison3	7601	3008	2644	2774	2677
Comparison4	4695	2137	1877	1975	1929
Comparison5	188	70	62	67	62

Assigned Gene Ontology are also provided in deliverables "***06_Significant_DGE_GO***".

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

5.8.5. Pathway analysis for significant DGE

Ortholog assignment and mapping of **Significant differentially expressed transcripts** to the biological pathways were performed using KEGG automatic annotation server (KAAS). Significant differentially expressed transcripts were compared against the KEGG database using BLASTX with threshold bit-score value of 60 (default). Pathway analysis was performed using all differentially expressed transcripts. The mapped transcripts represented metabolic pathways of major biomolecules such as carbohydrates, lipids, nucleotides, amino acids, glycans, etc. The mapped transcripts also represented the genes involved in metabolism, genetic information processing, and environmental information processing, cellular processes and organismal systems. Detail of the transcripts following in particular pathway along with the pathway images is given in data deliverables entitled "***07_Significant_DGE_pathways***". **Respective pathway images are also provided in the same folder.** Pathway statistics for one of the comparisons and one representative pathway image have been given below for illustration purpose.

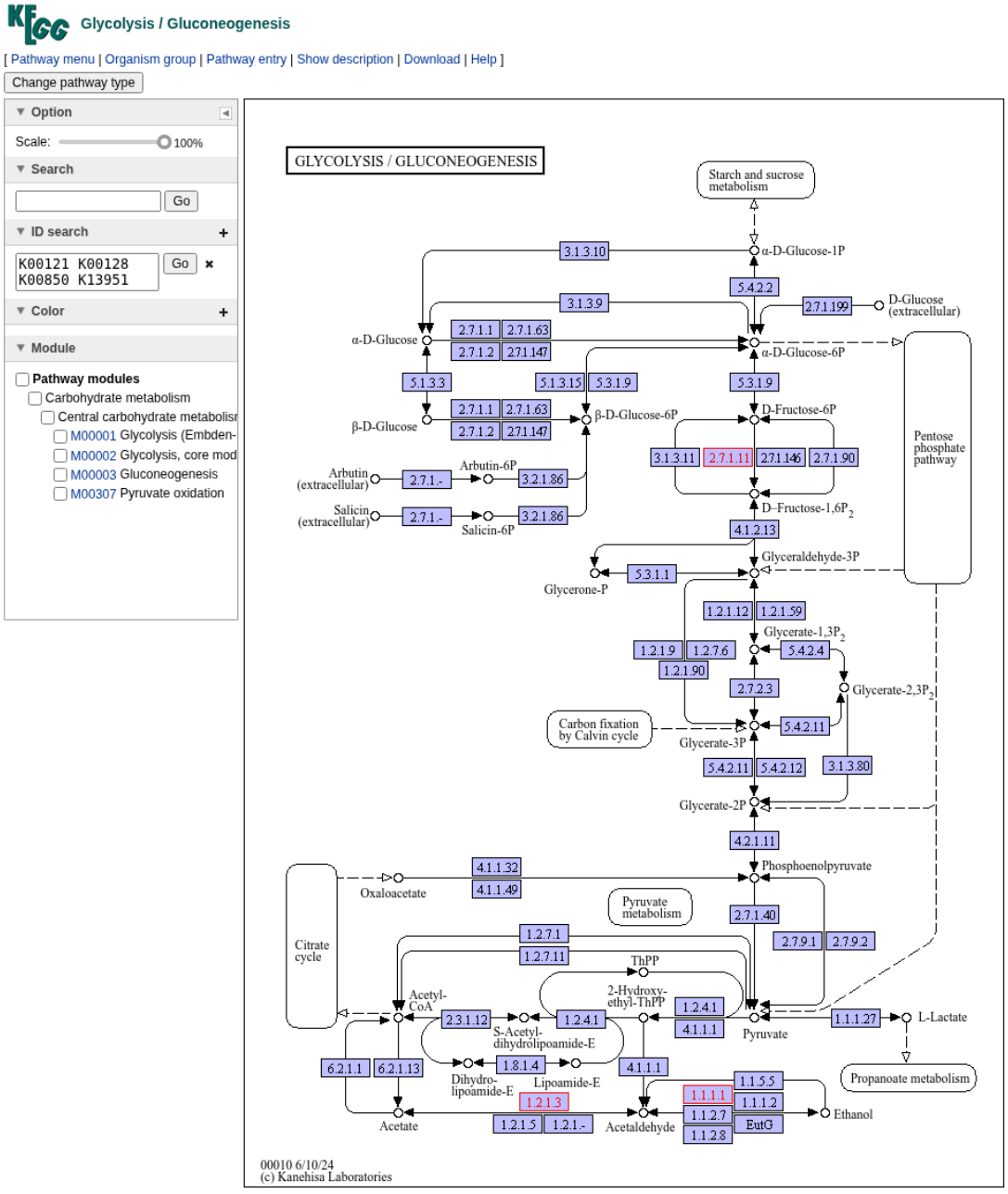

Figure 16: Example of Glycolysis for Comparison-1 DGE: Purple color entry has been assigned to at least one gene entry whereas a pink color entry corresponds to an ortholog table entry, which corresponds to a gene that is a part of a functional unit on the pathway.

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

Table 11: KEGG pathway statistics for Significant differentially expressed transcripts for Comparison-1

#Level 1	Level 2	Counts
09100 Metabolism	09101 Carbohydrate metabolism	37
09100 Metabolism	09102 Energy metabolism	18
09100 Metabolism	09103 Lipid metabolism	34
09100 Metabolism	09104 Nucleotide metabolism	3
09100 Metabolism	09105 Amino acid metabolism	32
09100 Metabolism	09106 Metabolism of other amino acids	14
09100 Metabolism	09107 Glycan biosynthesis and metabolism	20
09100 Metabolism	09108 Metabolism of cofactors and vitamins	9
09100 Metabolism	09109 Metabolism of terpenoids and polyketides	2
09100 Metabolism	09110 Biosynthesis of other secondary metabolites	2
09100 Metabolism	09111 Xenobiotics biodegradation and metabolism	25
09120 Genetic Information Processing	09121 Transcription	18
09120 Genetic Information Processing	09122 Translation	17
09120 Genetic Information Processing	09123 Folding, sorting and degradation	36
09120 Genetic Information Processing	09124 Replication and repair	12
09120 Genetic Information Processing	09125 Information processing in viruses	20
09120 Genetic Information Processing	09126 Chromosome	17
09130 Environmental Information Processing	09132 Signal transduction	375
09130 Environmental Information Processing	09133 Signaling molecules and interaction	61
09140 Cellular Processes	09141 Transport and catabolism	95
09140 Cellular Processes	09142 Cell motility	71
09140 Cellular Processes	09143 Cell growth and death	94
09140 Cellular Processes	09144 Cellular community - eukaryotes	71
09140 Cellular Processes	09145 Cellular community - prokaryotes	2
09150 Organismal Systems	09149 Aging	26
09150 Organismal Systems	09151 Immune system	200
09150 Organismal Systems	09152 Endocrine system	226
09150 Organismal Systems	09153 Circulatory system	28
09150 Organismal Systems	09154 Digestive system	56
09150 Organismal Systems	09155 Excretory system	14
09150 Organismal Systems	09156 Nervous system	82
09150 Organismal Systems	09157 Sensory system	27
09150 Organismal Systems	09158 Development and regeneration	46
09150 Organismal Systems	09159 Environmental adaptation	34

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in

09160 Human Diseases	09161 Cancer: overview	208
09160 Human Diseases	09162 Cancer: specific types	145
09160 Human Diseases	09163 Immune disease	29
09160 Human Diseases	09164 Neurodegenerative disease	174
09160 Human Diseases	09165 Substance dependence	26
09160 Human Diseases	09166 Cardiovascular disease	107
09160 Human Diseases	09167 Endocrine and metabolic disease	73
09160 Human Diseases	09171 Infectious disease: bacterial	123
09160 Human Diseases	09172 Infectious disease: viral	228
09160 Human Diseases	09174 Infectious disease: parasitic	68
09160 Human Diseases	09176 Drug resistance: antineoplastic	28
09180 Brite Hierarchies	09181 Protein families: metabolism	134
09180 Brite Hierarchies	09182 Protein families: genetic information processing	586
09180 Brite Hierarchies	09183 Protein families: signaling and cellular processes	370
09190 Not Included in Pathway or Brite	09191 Unclassified: metabolism	3
09190 Not Included in Pathway or Brite	09192 Unclassified: genetic information processing	2
09190 Not Included in Pathway or Brite	09193 Unclassified: signaling and cellular processes	22

Note:

(a) Non-specific organism pathway hits are obtained in pathway results due to lenient threshold (bit score: 60) applied in KAAS webserver.

(b) Pathway stats for only one comparison has been shown above, however stats and images for all other combinations has been provided as a part of data deliverables "*07_Significant_DGE_pathways*".

5.8.6. Over-representation (or enrichment) analysis

Over-representation (enrichment) analysis is a statistical approach used to identify biological functions, pathways, or processes that are significantly over-represented among a given list of genes compared to a reference background. This helps in interpreting large-scale transcriptomic data by linking differentially expressed genes (DEGs) to specific molecular mechanisms and biological pathways.

Functional enrichment analysis was performed using the **clusterProfiler** R package to identify over-represented Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways among the significantly differentially expressed genes (DEGs). The functions **enrichGO** and **enrichKEGG** were applied independently to both up-regulated and down-regulated gene sets.

Multiple hypothesis testing correction was performed using the **Benjamini–Hochberg (BH)** method, and GO and KEGG terms with pvalueCutoff = 0.05, and qvalueCutoff = 0.2 were considered statistically significant.

The results of the enrichment analyses were visualized using **dot plots** and **bar plots**, highlighting the most significantly enriched GO terms and KEGG pathways that may represent key biological processes and molecular mechanisms underlying the observed expression changes.

Figure 17: Dot plot of GO enrichment for Comparison1 of down-regulated genes showing the most significantly enriched Gene Ontology (GO) terms in the differentially expressed genes (DEGs).

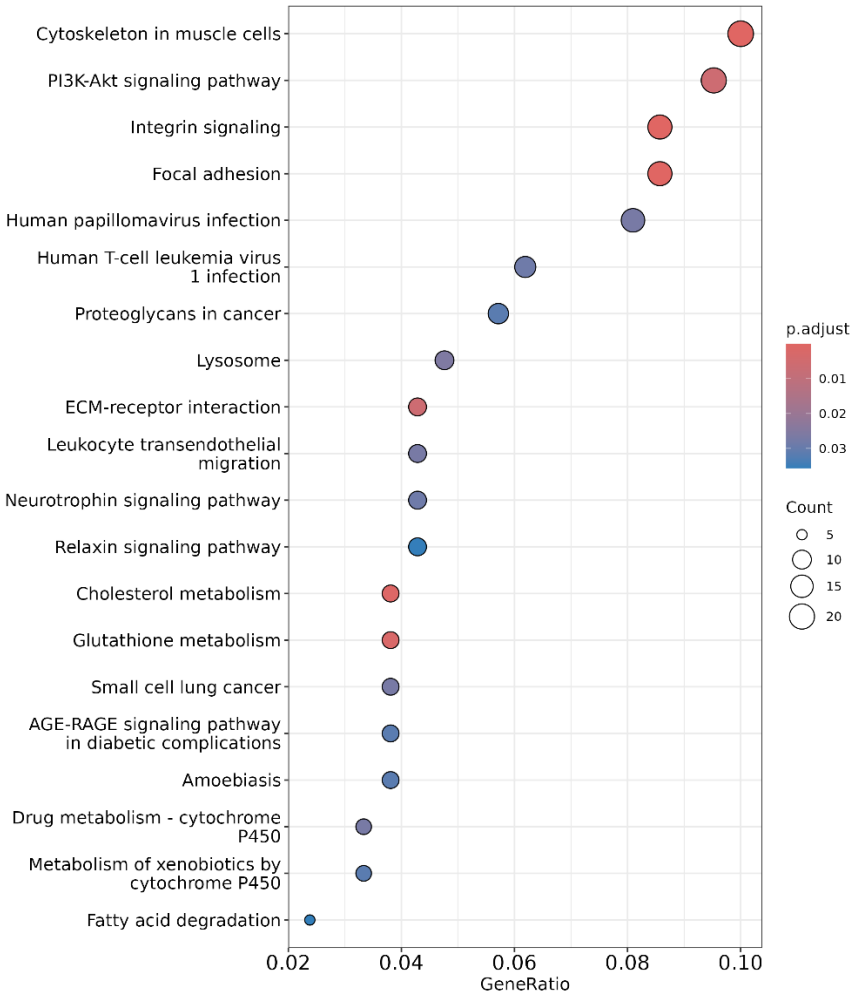

Figure 18: Dot plot of KEGG pathway enrichment for Comparison1 of down-regulated genes, highlighting the most significantly enriched KEGG pathways in the differentially expressed genes (DEGs).

Note: Dot plot for only one comparison has been shown above, however plots for upregulated genes as well as all other combinations have been provided as a part of data deliverables "*08_Significant_DGE_Enrichment/*".

Please note: No enrichment was obtained for the following comparisons: C5_up_KEGG.

6. Deliverables

```
├── 01_Raw_Data
│   ├── ICGA-BC-NT-125_S182_L006_R1_001.fastq.gz
│   ├── ICGA-BC-NT-125_S182_L006_R2_001.fastq.gz
│   ├── ICGA-BC-NT-126_S184_L006_R1_001.fastq.gz
│   ├── ICGA-BC-NT-126_S184_L006_R2_001.fastq.gz
│   ├── ICGA-BC-NT-127_S186_L006_R1_001.fastq.gz
│   ├── ICGA-BC-NT-127_S186_L006_R2_001.fastq.gz
│   ├── ICGA-BC-NT-129_S188_L006_R1_001.fastq.gz
│   ├── ICGA-BC-NT-129_S188_L006_R2_001.fastq.gz
│   ├── ICGA-BC-TT-125_S183_L006_R1_001.fastq.gz
│   ├── ICGA-BC-TT-125_S183_L006_R2_001.fastq.gz
│   ├── ICGA-BC-TT-126_S185_L006_R1_001.fastq.gz
│   ├── ICGA-BC-TT-126_S185_L006_R2_001.fastq.gz
│   ├── ICGA-BC-TT-127_S187_L006_R1_001.fastq.gz
│   ├── ICGA-BC-TT-127_S187_L006_R2_001.fastq.gz
│   ├── ICGA-BC-TT-129_S189_L006_R1_001.fastq.gz
│   ├── ICGA-BC-TT-129_S189_L006_R2_001.fastq.gz
│   ├── NGS_240592_md5sum.txt
│   └── NGS_240592_Raw_Stats.txt
├── 02_reference_genome_and_gff
│   ├── GCF_000001405.40_GRCh38.p14_genomic.fna
│   └── GCF_000001405.40_GRCh38.p14_genomic.gff
├── 03_transcript_assembly_gtf
│   ├── gffcompare_results.gtf
│   ├── merged_transcripts.gtf
│   ├── NGS_240592_ICGA-BC-NT-125Aligned_1.gtf
│   ├── NGS_240592_ICGA-BC-NT-126Aligned_1.gtf
│   ├── NGS_240592_ICGA-BC-NT-127Aligned_1.gtf
│   ├── NGS_240592_ICGA-BC-NT-129Aligned_1.gtf
│   ├── NGS_240592_ICGA-BC-TT-125Aligned_1.gtf
│   ├── NGS_240592_ICGA-BC-TT-126Aligned_1.gtf
│   ├── NGS_240592_ICGA-BC-TT-127Aligned_1.gtf
│   ├── NGS_240592_ICGA-BC-TT-129Aligned_1.gtf
│   ├── novel.isoforms.gtf
│   ├── ReadMe.txt
│   └── sed1Jy5Yk
└── 04_transcript_sequences_fasta
    ├── Merged.fasta
    ├── NGS_240592_ICGA-BC-NT-125Aligned_transcript.fasta
    ├── NGS_240592_ICGA-BC-NT-126Aligned_transcript.fasta
    ├── NGS_240592_ICGA-BC-NT-127Aligned_transcript.fasta
    ├── NGS_240592_ICGA-BC-NT-129Aligned_transcript.fasta
```

- NGS_240592_ICGA-BC-TT-125Aligned_transcript.fasta
- NGS_240592_ICGA-BC-TT-126Aligned_transcript.fasta
- NGS_240592_ICGA-BC-TT-127Aligned_transcript.fasta
- NGS_240592_ICGA-BC-TT-129Aligned_transcript.fasta

05_differential_expression_analysis
- Comparison1_DGE.xlsx
- Comparison1_Heatmap.png
- Comparison1_MA_and_Volcano.pdf
- Comparison2_DGE.xlsx
- Comparison2_Heatmap.png
- Comparison2_MA_and_Volcano.pdf
- Comparison3_DGE.xlsx
- Comparison3_Heatmap.png
- Comparison3_MA_and_Volcano.pdf
- Comparison4_DGE.xlsx
- Comparison4_Heatmap.png
- Comparison4_MA_and_Volcano.pdf
- Comparison5_DGE.xlsx
- Comparison5_Heatmap.png
- Comparison5_MA_and_Volcano.pdf

06_Significant_DGE_GO
- Comparison1_Significant_DGE_with_GO.xlsx
- Comparison2_Significant_DGE_with_GO.xlsx
- Comparison3_Significant_DGE_with_GO.xlsx
- Comparison4_Significant_DGE_with_GO.xlsx
- Comparison5_Significant_DGE_with_GO.xlsx

07_Significant_DGE_pathways
- Comparison1_Significant_DGE_Pathway_images
- Comparison1_Significant_DGE_Pathways.xlsx
- Comparison2_Significant_DGE_Pathway_images
- Comparison2_Significant_DGE_Pathways.xlsx
- Comparison3_Significant_DGE_Pathway_images
- Comparison3_Significant_DGE_Pathways.xlsx
- Comparison4_Significant_DGE_Pathway_images
- Comparison4_Significant_DGE_Pathways.xlsx
- Comparison5_Significant_DGE_Pathway_images
- Comparison5_Significant_DGE_Pathways.xlsx

08_Significant_DGE_Enrichment
- C1_down_barplot_GO_enrichment.png
- C1_down_barplot_KEGG_enrichment.png
- C1_down_dotplot_GO_enrichment.png
- C1_down_dotplot_KEGG_enrichment.png
- C1_down_GO_enrichment_results.xlsx
- C1_down_KEGG_enrichment_results.xlsx

```
├── C1_up_barplot_GO_enrichment.png
├── C1_up_barplot_KEGG_enrichment.png
├── C1_up_dotplot_GO_enrichment.png
├── C1_up_dotplot_KEGG_enrichment.png
├── C1_up_GO_enrichment_results.xlsx
├── C1_up_KEGG_enrichment_results.xlsx
├── C2_down_barplot_GO_enrichment.png
├── C2_down_barplot_KEGG_enrichment.png
├── C2_down_dotplot_GO_enrichment.png
├── C2_down_dotplot_KEGG_enrichment.png
├── C2_down_GO_enrichment_results.xlsx
├── C2_down_KEGG_enrichment_results.xlsx
├── C2_up_barplot_GO_enrichment.png
├── C2_up_barplot_KEGG_enrichment.png
├── C2_up_dotplot_GO_enrichment.png
├── C2_up_dotplot_KEGG_enrichment.png
├── C2_up_GO_enrichment_results.xlsx
├── C2_up_KEGG_enrichment_results.xlsx
├── C3_down_barplot_GO_enrichment.png
├── C3_down_barplot_KEGG_enrichment.png
├── C3_down_dotplot_GO_enrichment.png
├── C3_down_dotplot_KEGG_enrichment.png
├── C3_down_GO_enrichment_results.xlsx
├── C3_down_KEGG_enrichment_results.xlsx
├── C3_up_barplot_GO_enrichment.png
├── C3_up_barplot_KEGG_enrichment.png
├── C3_up_dotplot_GO_enrichment.png
├── C3_up_dotplot_KEGG_enrichment.png
├── C3_up_GO_enrichment_results.xlsx
├── C3_up_KEGG_enrichment_results.xlsx
├── C4_down_barplot_GO_enrichment.png
├── C4_down_barplot_KEGG_enrichment.png
├── C4_down_dotplot_GO_enrichment.png
├── C4_down_dotplot_KEGG_enrichment.png
├── C4_down_GO_enrichment_results.xlsx
├── C4_down_KEGG_enrichment_results.xlsx
├── C4_up_barplot_GO_enrichment.png
├── C4_up_barplot_KEGG_enrichment.png
├── C4_up_dotplot_GO_enrichment.png
├── C4_up_dotplot_KEGG_enrichment.png
├── C4_up_GO_enrichment_results.xlsx
├── C4_up_KEGG_enrichment_results.xlsx
├── C5_down_barplot_GO_enrichment.png
├── C5_down_barplot_KEGG_enrichment.png
```

Lab Facility: 2A,3A,3B PASL -House, Beside Sahjanand college, Opposite Kamdenu Complex, Panjarapole, Ambawadi, Ahmedabad-380015, Gujarat
Ph+91-79-49006800 | WhatsApp: 6356005900 | Email:info@unigenome.in | Website: www.unigenome.in


```
    ├── C5_down_dotplot_GO_enrichment.png
    ├── C5_down_dotplot_KEGG_enrichment.png
    ├── C5_down_GO_enrichment_results.xlsx
    ├── C5_down_KEGG_enrichment_results.xlsx
    ├── C5_up_barplot_GO_enrichment.png
    ├── C5_up_barplot_KEGG_enrichment.png
    ├── C5_up_dotplot_GO_enrichment.png
    ├── C5_up_dotplot_KEGG_enrichment.png
    ├── C5_up_GO_enrichment_results.xlsx
    └── C5_up_KEGG_enrichment_results.xlsx
├── Readme.txt
└──NGS_240592_Final_Report_RNASeq_09.12.2025.pdf
```

[END OF REPORT]