Detection of fusion genes and their expression in gastric cancer patients in Mizo population

By

Ranjan Jyoti Sarma

Registration No. and Date: MZU/M.Phil./473 of 03/05/2018

Dissertation submitted in fulfilment of the Requirements for the degree of Master of Philosophy in Biotechnology

Under the supervision of

Dr. N. Senthil Kumar

Professor, Department of Biotechnology

School of Life Sciences, Mizoram University,

Aizawl, Mizoram-796004, India

2019

ACKNOWLEDGEMENT

I am thankful to **Prof. N Senthil Kumar** for his constant support and motivation throughout the work.

I am thankful to the sample donors which made possible to conduct the study.

A special thanks to **Dr. Jeremy L. Pautu**, Regional Cancer Research Centre, Zemabawk; **Dr. John Zohmingthanga**, Civil Hospital, Aizawl and **Dr. Lalawmpuii Pachuau**, Civil Hospital, Aizawl.

Special thanks to **Dr. Arindam Maitra**, Scientist, NIBMG for providing the next-generation sequencing facility.

I am also thankful to **Dr. J. Bhattacharya (Head)** along with all the teaching , non-teaching staffs and all the research scholar of Department of Biotechnology, Mizoram University.

I am also thankful to **DBT, New Delhi** for (**DBT-Biotech Hub**) the fellowship and **DBT-BIF** for providing the computational facility.

I am thankful to my family members for lifting my confidence up with constant support.

Dated:

Place: Aizawl, Mizoram **Ranjan Jyoti Sarma**

TABLE OF CONTENTS

Content	Page Number
Abbreviations	
List of Figures	
List of Tables	
Introduction and Literature Review	1-14
Objective	15
Materials and Methods	16-26
Result	27-39
Discussion	40-44
Summary	45
Appendices	46-54
Appendix -I: The preview of the questionnaire followed for the sample collection of the study.	46-50
Appendix-II: List of Significant genes filtered from Cuffdiff data.	51-54
References	55-58
Biodata	59-61

ABBREVIATIONS

Abbreviations	Definition
RNA-Seq	RNA Seqeuencing
GC	Gastric Cancer
TNF	Tumor Necrosis Factor
RIN	RNA Integrity Number
GTF	Gene Transfer Format
DAVID	Database for Annotation, Visualization and Integrated Discovery
CAM	Cell adhesion Molecule

LIST OF FIGURES

Figure	Title	Page Number
Fig 1	(a) Electrophoresis summary of the RNA isolated from the samples. (b) RIN categories: shorter fragment sizes reflect the degradation of rRNA	18
Fig 2	Dispersion Plot: Genes counts varies among samples	28
Fig 3	Boxplot: (a) Variation Seen in across both the conditions (B) Variation seen across the condition with all replicates	29
Fig 4	M vs A plot (Avg. intensity vs log ratio) for a given pair of conditions across all FPKMs	30
Fig 5	(a)The log2FoldChange and negative log10 (Adjusted P-Value) (FDR) was plotted in x and y axis repectively. (b) The red points show differentially expressed genes with FDR less than 0.05	31
Fig 6	Heatmap constructed for the upregulated genes using CummeRbund.	32
Fig 7	Heatmap constructed for the Down-regulated genes using CummeRbund.	35
Fig 8	Plot of Benjamini score assigned to the upregulated gene involving pathway	34
Fig 9	Plot of Benjamini score assigned to the Down-regulated gene involving pathway	37

LIST OF TABLES

Table	Title	Page Number
Table 1	Details of the Tumor and Adjacent Normal sample selected for RNA Isolation.	17
Table 2	Samples selected for RNA-Seq library preparation	19
Table 3	Quality status before and after trimming of forward read (read-1) and reverse read (read-2) analysed in FastQC tool	21
Table 4	Fusion gene found in Normal sample	27
Table 5	Fusion gene found in Tumor sample	27

INTRODUCTION AND REVIEW OF LITERATURE

Cancer now is a major threat in terms of incidence and mortality and is stated as the second most mortality related cause in both males and females. Among the major cancer types, Gastric cancer (GC) or Stomach cancer is the fifth most common type (approximately 10 million diagnosed cases in 2018) and ranked third in term of mortality (approximately 7.83 Million Death) (Bray *et al.,* 2018). A recently published data by the IARC (International Agency for Research on Cancer) and WHO (World Health Organization) showed that the gastric cancer incidence is in fourth rank globally in male and seventh rank in female in all ages (Bray *et al.,* 2018).

Cancer in the populations in the Asian countries are more than western countries (Bray *et al.,* 2018).As per the data available in Cancer Today by IARC, in India, GC occurrence is also high and ranked sixth (http://gco.iarc.fr/today/home). However, regional incidence of gastric cancer varies and GC is the second most cancer-causing death among Indian male and female in the age range of 15 to 44 (Murugesan *et al.,* 2018). National Centre for Disease informatics and Research (NCDIR) reported in 2016-17 that expectancy of cancer will increase from 1.38 million (in 2015) to 1.73 Million (in 2020). As per the report of Population Based Cancer Registry (PBCR) by ICMR, the crude rate (CR) of cancer is highest in Aizawl district of Mizoram among both male and female. As per the report, over 700

people have been killed by cancer each year from 2012-14. The report also revealed that there were 4656 new cases of cancer reported between 2012 and 2014 with number of females at 2089 and males at 2567. On an average, there were 1,552 new cancer cases per year in Mizoram (PBCR Report-2012-14, ICMR).

GASTRIC CANCER AND ITS TYPES

The stomach lining is made-up of columnar epithelial cells and glands which forms the gastric mucosa. Chronic inflammation in gastric mucosa, known as atrophy gastritis, can lead to peptic ulcers, and ultimately gastric cancer (Rawla and Barsouk, 2019). Anatomically, gastric cancer is dived into two types, namely Cardia type (the proximal part of the stomach joining the oesophagus) and non-cardia type (from the mid to distal part of the stomach). It has been reported that cardia cancers (when the tumor occurs at the gastroesophageal junction) may be related to gastroesophageal reflux disease (GERD) similar to oesophageal adenocarcinoma (EAC). Studies has fund that *Helicobacter pylori* is strongly associated with gastric cancer development (Mukaisho *et al.,* 2015).

According to Lauren's classification, histologically gastric cancer is Adenocarcinoma of Intestinal and diffuse types. The morphology of both the types are distinctly different and share differences in epidemiology, etiology and pathology. The intestinal type cancer cells are seen to have association with intestinal metaplasia (Lauren, 1965). The intestinal type is easier to classify histologically than diffuse type carcinoma and is associated with vascular or

lymphatic invasion. Intestinal type of GC mostly seen to occur in male patients over 50 years of age, and carcinogenesis is associated with *H. pylori* causing gastritis, atrophy gastritis, intestinal metaplasia, dysplasia and finally to gastric carcinoma (Ma *et al.,* 2016). The diffuse type is associated with young aged patients and seen to be more in female. It is associated with atrophy gastritis (chronic inflammation in mucosa) have less association with environmental factors; however, it may also have association with *H. pylori* infection. Diffuse type is usually less well differentiated, with rare appearance of signet ring cells due to the pushing of the nucleus of the cell to aside by extra-cellular mucus. The pathogenesis of both types of carcinoma include both genetic and epigenetic modification like DNA methylation, histone protein modifications and chromosomal rearrangement (van der Woude *et al.,* 2003; Ma *et al.,* 2016).

RISK FACTORS OF GASTRIC CANCER

There are multifactorial processes involved for the stomach mucosa to attain malignancy. Environmental factors and genetic predisposition influence the occurrence of gastric cancer. *H. pylori* infection is considered as one of the major risk factors followed by obesity, smoking, smoked food, red meat, alcohol, low intake of fresh vegetables and low socioeconomic status (Zali *et al.,* 2011).

In 1994, *H. pylori*, a gram-negative bacterium, was classified as type I carcinogen by WHO and IARC. It is most closely associated with intestinal type of GC (Zali *et al.,* 2011). It is a widely distributed gastric pathogen and considered as a

strong risk factor in intestinal type. It colonises the gastric epithelium and triggers the development of gastritis in mucosa which leads to intestinal metaplasia (Diaz *et al,* 2018). It can stay in stomach for many decades and inflammation in stomach can persist along with it. However, some persons may not get the adverse effect of the presence of the pathogen in their entire life-time but they get the derived benefit from it (Wroblewski *et al.,* 2010). The *H. pylori* infection outcome depend on the type of strain or immune responses by the host which have influential role in the pathogen and host interaction (Blaser *et al.,* 2001). *H. pylori* induces carcinogenesis through increasing apoptosis (Moss *et al.,* 1996).

The gene *cagA* (cytotoxin-associated gene A) of *H. pylori* that belong to the *cag* pathogenicity island and persons harbouring *cagA*$^{+}$ strains tend to have higher proliferation of gastric epithelial compared to *cagA*$^{-}$ strains or uninfected persons (Peek *et al.,* 1997). Another polymorphic locus *vacA* of *H. pylori*, which encodes a major toxin vacuolating cytotoxin and strains belong to s1a *vacA* subtype are usually more toxigenic which leads to severe clinical outcomes. Another *iceA* gene of *H. pylori* get induced after contact with stomach epithelial cells. Similarly, *iceA* gene has two genotypes namely, *iceA1* and *iceA2*, and *iceA1* strains has association with peptic ulcer. cagA^{+} vacA s1a iceA1 strains of *H. pylori* are mostly associated with disease, however, geographic differences influence the susceptibility to the disease (Israel *et al.,* 2001; Wroblewski *et al.,* 2010). Polymorphism of key genes in the patients having *H. pylori* infection increases the risk of gastric carcinoma by altering the gene expression and gene functions. Genome Wide Association Study (GWAS)

also found that Single Nucleotide Variation (SNP)s in TNF-α gene that codes for tumor necrosis factor alpha (TNF) and interleukin gene such as IL-1 and IL-10 have been associated with *H. pylori* associated gastric cancer (Ishaq and Nunn, 2015).

Epstein-Barr virus or EBV is well known for oncogenesis by activating the cell survival and growth signalling pathway. EBV infects epithelial cells and human B lymphocytes. B cells poses major glycoprotein namely, gp350/220 which can mediate the viral attachment to CR2 (Complement Receptor Type 2) and CD21 on the surface of the B cell. EBV infects the epithelial cell less efficiently than B lymphocyte as epithelial express low CD21 and no expression at all (Lizasa *et al.,* 2012). Stusy has found that expression of tumor regulatory genes often get lowered upon EBV infection and the immune response gene especially the cytokine pathway genes undergo alteration with significant altered expression (chong *et al.,* 2007). Amplification of CD274 locus 9p24.1 accompany the gastric cancer associated with EBV. CD274 encodes for PD-L1 (Programmed Death Ligand 1) which plays role in suppression immune response for tumor through the interaction with the PD-1 (programmed death receptor-1), a co-inhibitory molecule expressed by T cells (Blank *et al.,* 2007) which subsequently inhibits the T cell proliferation, cytotoxicity survival, and cytokine release. It also induces tumor-specific T cells apoptosis and promotes the CD4+ T cells differentiation into regulatory T Cells, and supress the immunity by making it unable to attack the tumor cells by the cytotoxic T lymphocyte (Cho *et al.,* 2016).

Studies have suggested that particulate matter in polluted air contribute in gastric cancer as well as other type of cancer (Nagel *et al.,* 2018). Human exposed to *n*-nitroso compounds have role in increasing risk of gastric cancer. However, vitamin C plays a role of reducing the risk of cancer by inhibiting the formation of nitroso compounds (Shi *et al.,* 1991; Lee *et al.,* 2013). Indirect evidence suggests that Polycyclic aromatic hydrocarbons are generated from sources like air pollution tobacco smoke and smoked meat (Larsson *et al.,* 2006). Consumption of salty food, Packed food, salt-preserved meat also contribute in developing gastric cancer (Liao *et al.,* 2014).

GC have been associated with low Socio-economic status who are non-accessible to fresh food, fruits and vegetables and consumption of stored food are strongly related with the lower socioeconomic status (Karimi *et al.,* 2014). Although *H. pylori* is widely distributed all over the world, higher rate of occurrence is more in developing countries than developed countries (Diaz *et al.,* 2018).

Gastric cancer is also hereditary due to genetic predisposition and increase the risk of occurrence of the disease among the family members, also referred to as Hereditary diffuse gastric cancer (Petrovchich *et al.,* 2016). However, the major genetic causes in vast majority of populations are still unknown and yet to be studied.

CDH1 that encodes for epithelial-cadherin and its mutation in germline is one of the major causes of familial GC and it inherits as autosomal dominant and affects

cell adhesion pathway (Liu *et al.,* 2014; Petrovchich *et al.,* 2016). Signet ring cell appearance in diffuse type of gastric cancer often have alteration in MAP3K6 gene. Whole exome sequencing has also identified pathogenic variants of ATM gene and PALB2 associated with gastric cancer (Diaz *et al.,* 2018). Mutation in BRCA1, BRCA2 and TP53 causes the loss of genome integrity also have major role in GC, however mutation in BRCA1 and BRCA2 in GC is comparatively low. ARIDIA, MLL3 are associated in chromatin remodelling with microsatellite instability. Apart from CDH1, mutation in FAT4, CTNNA1 also affects cell adhesion in GC. CTNNA1 mutation also found in hereditary gastric cancer patient. Mutation on ROCK1/2, RAC1, RHOA affects the Cytoskeleton and cell motility in GC. Pathogenic variants of Wnt signalling pathway genes such as CTNNB1, APC, RNF43 have adverse role in Signal transduction. ERBB1/2/3/4, MET, FGFR2, KRAS, RASA1, PIK3R1, PTEN, AKT1/2/3, MAP3K4/6 mutaion affects the RTK (Receptor Tyrosine Kinase) pathway. Mutation related to RTK pathway has association with uncontrol proliferation and metastasis. RTK pathway gene mutation strongly related with Epstein-Barr Virus (Morishita *et al.,* 2014; Lin *et al.,* 2015). DUS4L–BCAP29 novel gene fusion was found in gastric cancer cell line (Kim *et al.,* 2014). CLDN18-ARHGAP26 fusion transcript causes loss of the epithelial integrity. Rearrangements in genome can have significant impact on function of the gene by deletion, amplification gene or disruption and contribute in fusion gene formation with altered function. (Yao *et al.,* 2015). GS (Genomic Stability) group of the TCGA data which is comprised of majority of gastric cancer sample was also found to have CLDN18–ARHGAP fusion (Katona *et al.,* 2017).

Role of Fusion Gene in Gastric Cancer

Fusion genes which are the result of Chromosomal mutation such as insertion, deletion, translocation as well as transversion and plays major role in attaining malignancy by activating different pathway oncogene and supressing tumor suppressor gene. The discovery of fusion gene, *BCR-ABL* in leukaemia gave a lead to study of other fusion genes in cancer, which led to the discovery of many fusion in cancer and subsequently some potent therapeutics. The real role of the fusion gene as a diagnostic tool detect the presence and absence of the fusion with the appearance and disappearance of the tumor tissue, respectively. Transcriptome sequencing has revolutionized genomic discovery including fusion gene (Parker *et al.,* 2013). Chromosome rearrangements that give rise to fusion gene and that appear in both germline and somatic cells, would be a type of abnormal chromosome constitution (Poot *et al.,* 2015). Fusion transcript resulting from the gene fusion are used as diagnostic and prognostic monitoring in many types of cancer. Apart from cancer cells, fusion transcripts are also found in normal cell line which depicts the normal activity of the cell. Moreover, fusion transcripts are also generated from intergenic and intragenic splicing events in mRNA (Kumar *et al.,* 2016).

Genome sequencing techniques are newly applied for fusion detection from tumor samples. Identification of fusion transcripts is revolutionized by transcriptome sequencing as it allows to detect the chimeras in transcript level (Fernandez-Cuesta *et al.,* 2015; Schroder *et al.,* 2019). Fusion genes also belong to

structural variants and recently developed algorithms allow to detect these structural variants from transcriptome data by aligning the reads to the reference genome. These algorithms try to search adjacent sequence as well as the breakpoints and assemble the breakpoints and report the fusion with supportive read evidence. They further look for type of the reads and their orientation that fused together which helps to know the potential viability of a gene fusion (Fernandez-Cuesta *et al.,* 2015; Schroder *et al.,* 2019).

CD44-SLC1A2 is a fusion found in gastric cancer tissue and silencing of that fusion results in significant reduction of anchorage-independent growth, cell proliferation and invasion (Tao *et al.,* 2011). A fusion of SLC45A3 which codes solute carrier protein to the FGFR2 which codes for fibroblast growth factor receptor 2 protein, encodes RTK and plays a major role in prostate and other types of cancers including GC (Kumar *et al.,* 2016). CLDN18-ARHGAP26 fusion protein can cause degradation of cell-ECM disrupts the cell adhesion and receptor tyrosine kinase signalling thereby plays major role in carcinogenesis. CLDN18-ARHGAP26 contributes to GC by loss of CLDN18 and gain of ARHGAP26 functions, and hampers wound-healing (Yao *et al.,* 2013; Tanaka *et al.,* 2018). DUS4L–BCAP29 novel gene fusion was found in gastric cancer cell line (Kim *et al.,* 2014). CLDN18-ARHGAP26, CTNND1-ARHGAP26, and ANXA2-MYO9A novel fusion found in GC and CLDN18-ARHGAP26, CTNND1-ARHGAP26 are recurrent and prognostic in diffuse gastric cancer (Yang *et al.,* 2018). Functional analysis on cell line proved that CLDN18-ARHGAP26 is a major cause of independent cell growth (Nakayama *et al.,* 2018).

Therapeutic treatment can also be improved with the precise detection of fusion genes. Several drugs for treating cancer with fusion already in market which includes Imatinib mesylate for treating *BCR-ABL1* and Crizotinib for *EML4-ALK* fusions. Precise diagnosis of gene fusion helps in predicting prognosis of the disease and, possibility of patient survival and further treatment steps (Heyer *et al.,* 2019).

Differential Gene Expression (DGE) and Analysis of RNA-Seq data

Transcriptome is defined as the total RNA present in the cell at a given time. DGE analysis is the way to find the distribution of gene expression in the replicates with different conditions (Froussios *et al.,* 2016). Apart from fusion detection, transcriptome sequencing technology is commonly used techniques for differential gene expression study between two or more conditions with three or more replicates. Eukaryotic transcriptome achieves more complexity when the genes undergo alternative splicing which results in variation in protein structure and function that it codes for (Ghos *et al.,* 2016). The affordability of NGS in transcriptome analysis has played a tremendous role in producing huge amount of expression data in Medicine as well as in Cancer biology. Subsequent bioinformatics analysis is playing the key role in translating the big data by developing new algorithms and tools and due to the demand of big data in biology and cancer genomics, bioinformatics tools development in is increasing day by day (Raplee *et al.,* 2019).

Differential expression is analyzed by taking the normalized fragment count data and applying different statistical models to discover biologically significant quantitative changes in expression levels between case and controls or any

experimental groups of interest. Transcriptome sequencing has advantage over microarray in the sense that it can detect novel gene and isoforms expression. High-throughput transcriptome sequencing (RNA-Seq) helps in correctly interpreting DGE between specific conditions which is the key in the understanding the mechanism lies in phenotypic variation (Petryszak *et al.,* 2014 ; Ritchie *et al.,* 2015).

Transcriptome sequencing data analysis is computationally costly as the data size increases along with the sample size. There are number of steps involved in RNA-seq data analysis for DGE experiments. The basic steps are discussed below:

Raw reads QC: QC (Quality control) for checking the statistics relating to sequence quality, per sequence GC content, overrepresented reads, PCR duplicates or contaminations. As per recommendation, outliers more than 30% can be thrown out. FastQC (Andrew *et al.,* 2010) is a widely used open source tool to perform quick analyses of the NGS data from most of the sequencing plateforms. Low quality bases are often can be seen at the 3' end and can be trimmed off using Trimmomatic along with the adapter sequence to increase the mapability (Bolger *et al.,* 2014).

Alignment to Reference Genome: The reads after trimming are aligned to the reference sequence of all the dna (genome) or reference transcriptome. The aligning of RNA-seq and DNA seq reads are way different. Its very important to find out the splice junctions to align the exon models (RNA-Seq reads). Tophat is widely

used as splice junction aligner tool to align the RNA-Seq reads to reference genome and to the reference transcriptome together. Tophat identifies the potential splice junction site by aligning the reads to the reference genome and thereafter form a database of the splice site for confirmation (Tranpell *et al.,* 2012). The PCR bias can be detected by GC content and high percentage of GC content is not allowed for the alignment process. Qualimap, RSeQC , Picard are also some tools available for quality control (Conesa *et al.,* 2016).

Transcript Assembly and Relative Abundance: Transcript assembly is crucial step in Differential Gene Expression. Fragment assembly is process of identification of the pairs of incompatible fragments followed by connecting them when they are compatible and the alignments of the transcripts overlap in the genome. Post-transcriptional processing impact can be understood by quantify the relative change in abundance of the transcripts formed due to alternative splicing of the primary transcript. Such analysis could be helpful to detect how the gene is regulating (Tranpell *et al.,* 2010).

Transcript quantification: RNA-seq is most commonly use to calculate the expression of gene and transcript. This can be easily done by using cufflinks. Cufflinks counts all fragments, including compatible or non-compatible with the reference transcript, and report the expression summery in FPKM (Fragments Per Kilobase of exon model per Million mapped reads) unit. The reference transcript annotation record in the format of GTF (Gene transfer File) originated from the

same assembly of reference genome is required to get the gene level transcript quantification. FPKM and RPKM (Read Per Kilo base transcripts per Million mapped reads) are equivalent for single end reads. However, RPKM which is based on read count normalization method can bias estimates of Differential Expression (DE) as some poor-quality reads may be discarded in the trimming process (Bullard *et al.,* 2010; Hansen *et al.,* 2010). FPKM is convertible into Transcript Per Million also referred to as TPM (Conesa *et al.,* 2016).

Differential gene expression (DGE) analysis

DGE analysis is actually done to compare the gene expression values among samples in the form of RPKM, FPKM, and TPM which accounts for the number of transcripts. However, selection of tools to construct a streamline pipeline for RNA-Seq data to reach the research goal is still challenging. The statistical models and optimal methods applied varies depending on the complexity of the organism's genome. Transcriptome analysis often become challenging when the organism's reference genome is not available. Moreover, data QC (quality control) should be at every steps of data analysis to ensure both reliability and reproducibility of the results (Conesa *et al.,* 2016).

Quantification of gene expression is major challenge in DGE experiments. Currently, reference guided process is widely used in DGE for better interpretation (Conesa *et al.,* 2016). Data analysis face major challenges when sequence reads are very short because it may align to any position in the reference genome results in

many false positive results (Zhao *et al.,* 2019). The usefulness of transcriptome sequencing lies in the fact that both novel discovery and quantitative information can be deduced (Ghos *et al.,* 2016; Conesa *et al.,* 2018).

The data can be analysed using Cufflinks to create transcripts by combining the junction and exon level data of each gene (Ghos *at el.,* 2016). Cufflinks has many other tools that do the deep analysis of the transcriptome data and compare the transcripts between two groups. Cuffmerge is a tool inside the cufflinks package with is used for Transcript assembly. Once the assembly is completed the transcript abundance is quantified by the Cuffquant and subsequently DGE is estimated by Cuffdiff in FPKM unit (Trapnell *et al*. 2012).

OBJECTIVES

To detect the fusion genes and their expression profiles associated with Mizo gastric cancer patients.

MATERIALS AND METHODS

SAMPLE COLLECTION

Total six samples of tumor along with Adjacent normal tissue were taken for the study and Institutional Ethics Committee (No.B.12018/1/13-CH(A)/IEC), Civil Hospital of Aizawl, Mizoram has approved the study. Patients with gastric neoplasm and previous history of other malignancies were excluded from the study. Information on treatment, clinical stage, dietary and life style habits, previous disease history, tobacco and alcohol use was collected using a structured questionnaire (**Appendix –I)**. Tumor tissues and adjacent normal tissues were subjected to pathological review to confirm histology and tumor cell content for identifying the stages. Pathological identification was done by (i) Microscopic examination of H & E stain tissue (ii) Histological Grading & typing and (iii) TNM staging (Tumor/ Node/ Metastasis) (Table 1). The samples were immediately transferred to RNAlater (ThermoFicher Scientific) and kept at 4^{o} C overnight and transferred to -80^{o} C prior to RNA isolation.

RNA ISOLATION

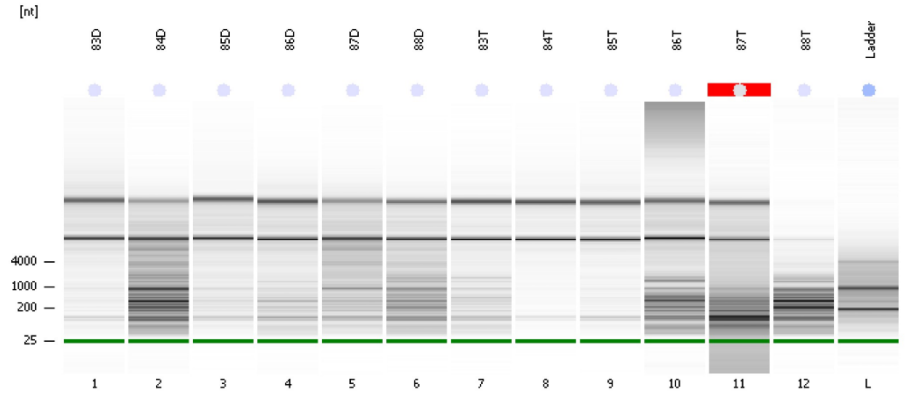
TRIzol and PureLink RNA mini kit (Ambion) method was followed for total RNA isolation. The all the spteps followed according to the manufacturer's protocol. Agilent RNA 6000 Nano chips in 2100 Bioanalyzer (Agilent) was used to check the quality of isolated total RNA and quantitation was done by Qubit using Quant-iT RNA assay kit broad range and NanoDrop spectrophotometer (Thermo Scientific).

Table 1: Details of the Tumor and Adjacent Normal sample selected for RNA Isolation.

Sample ID		**Tumor Type**	**Gender**	**Age**	**TNM**	***H. pylori***	**EBV**
Adjacent Normal	**Cancer**						
D83	T83	Moderately differentiated adenocarcinoma	Male	61	T_2 N_2	_	-
D84	T84	Moderately differentiated adenocarcinoma	Female	58	T_{4a} N_{3a}	-	-
D85	T85	Poorly differentiated adenocarcinoma; Completely resected	Male	65	T_2 N_0	-	-
D86	T86	Poorly differentiated adenocarcinoma	Female	45	T_{4a} N_{3a}	+	-
D87	T87	Moderately differentiated adenocarcinoma; Completely resected	Male	78	T_3 N_0 M_X	+	-
D88	T88	Moderately differentiated adenocarcinoma	Male	70	T_2 N_o	+	-

RNA Integrity Number Checking

(a)



(b)

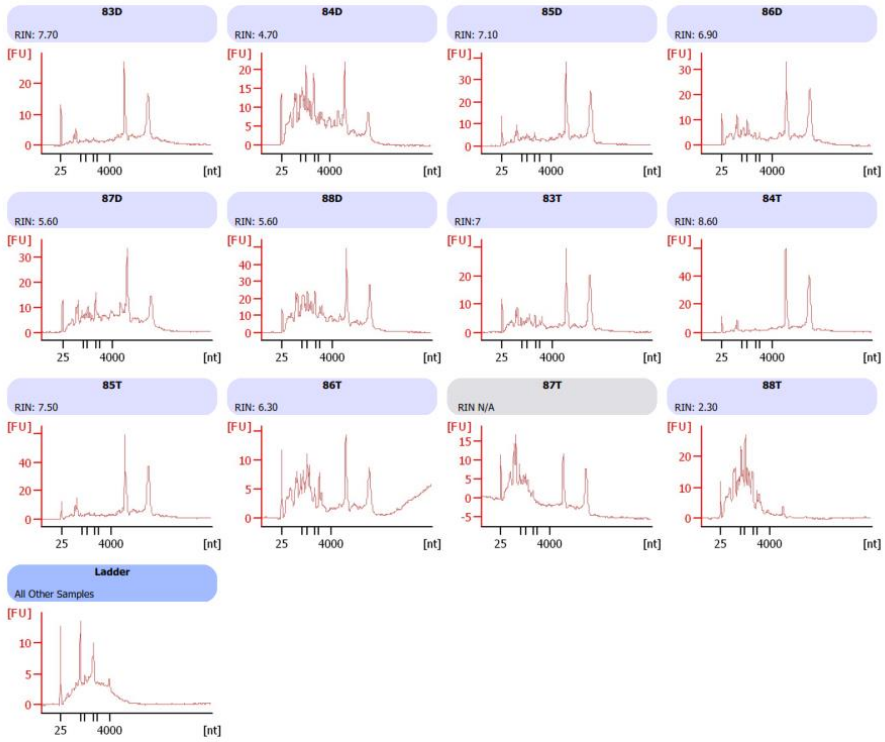


Fig 1: (a) Electrophoresis summary of the RNA isolated from the samples. (b) RIN categories: shorter fragment sizes reflect the degradation of rRNA. RIN value with 10 signify that the RNA is intact and less than 2 is degraded.

RNA-Seq Library Preparation

RIN values of all the samples were analysed and the samples bearing RIN $\geq$ 6 were selected for Illumina sequencing library preparation (Fig 1). Sequencing library was prepared using three adjacent normal and four tumor samples (Table 2).

Table 2: Samples selected for RNA-Seq library preparation:

SR. NO.	SAMPLE ID	NANODROP CONC (ng/ul)	260/280	RIN
1	T 83	1209	2.11	7
2	T 84	1247	2.09	8.6
3	T 85	724	2.1	7.5
4	T 86	1143	2.09	6.3
5	D 83	1297	2.11	7.7
6	D 85	317	2.09	7.1
7	D 86	1927	2.09	6.9

T83-86 are the tumor sample while D83,85 and D86 are the Adjacent normal.

1 µg of total RNA sample measured by qubit was used for library preparation using Illumina TruSeq Stranded Total RNA Library preparation kit. Manufacturer's instructions were followed in the kit method. The rRNAs were removed prior to fragmentation and adapter ligation. Then, first and second strand cDNA synthesis from rRNA-depleted fragmented total RNA, both ends of cDNA were repaired and adapters were ligated. Finally, libraries were enriched for limited cycle PCR. The high sensitivity D1000 ScreenTape was used to check the quality of prepared RNA-Seq libraries in Agilent 2200 TapeStation system and quantitative Real Time PCR was used to quantify the final library. Paired end 2 x 100 bp sequencing of these

libraries were performed in Illumina HiSeq-2500 (Illumina). The RNA isolation to Illumina paired-end sequencing was done in NIBMG, Kalyani, West Bengal.

DATA Quality Check (QC): The sequencing data obtained from the sequencing machine was in FASTQ format. The basic QC was done using FASTQC tool (Andrew *et al.,* 2010). Data were checked for Basic statistics, GC content, adapter contamination etc.

Quality Trimming: The low-quality bases and the adapters were removed using a java-based tool Trimmomatic version 0.38 using Trueseq3 paired end library.

```
java -jar /Trimmomatic-0.38/trimmomatic-0.38.jar PE -threads 6 -phred33 T86_R1.fastq T86_R2.fastq paired_r1.fq.gz unpaired_r1.fq.gz paired_r2.fq.gz unpaired_r2.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:70
```

- Remove adapters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10)
- Remove leading low quality or N bases (below quality 3) (LEADING:3)
- Remove trailing low quality or N bases (below quality 3) (TRAILING:3)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15)
- Drop reads below the 70 bases long (MINLEN:70)

Post Trimming Quality Check: The trimmed fastq files were again checked for basic statistics, GC content and adapter contamination. Adapter sequences were removed in the trimmed file (Table 3).

Table 3: Quality status before and after trimming of forward read (read-1) and reverse read (read-2) analysed in FastQC tool.

SampleID_ read-ID	Before Trimming			After Trimming		
	Total Sequences	GC Content (%)	Adapter	Total Sequences	GC Content	Adapter
D83_read-1	19469977	46	Yes	17176911	45	No
D83_read-2						
D85-read-1	17385748	47		15614115	47	
D85-read-2						
D86-read-1	19289000	47		17269912	46	
D86-read-2						
T83-read-1	19561401	45		17266844	46	
T83-read-2						
T84-read-1	18595426	45		16913754	45	
T84-read-2						
T85-read-1	15656665	45		14454534	44	
T85-read-2						
T86-read-1	16685764	44		15474623	43	
T86-read-2						

Reference data preparation: The reference genome (in FASTA) and transcript annotation file in Gene Transfer Format (GTF) were downloaded from Illumina's

iGenome project. The reference assembly was of hg19 version of UCSC where TSS-ID and P_ID were added separated and prepared for the RNA-Seq community by Illumina's iGenome project (http://sapac.support.illumina.com/sequencing/sequencing_software/igenome.html).

Reference Genome Index and Transcriptome index:

The reference genome indexing was done using Bowtie2, fast gapped aligner for RNA-seq data. Bowtie2 is a high speedy, sensitive and accurate aligner and it uses dynamic programming algorithms. Tophat (Tranpell *et al.,* 2012) which use bowtie2, was used for transcriptome indexing. It first creates the transcriptome from the reference genome using the transcript annotation file and index the transcriptome.

Genome Indexing:

```
Bowtie2-index –f genome.fa  genome
```

Transcriptome Indexing:

```
tophat -G gene.gtf --transcriptome-index ./transcriptome_index/ genome.fa
```

Fusion Gene Analysis: GeneFuse tool was used in fusion analysis. Being a faster tool to detect fusion from directly fastq file, genefuse uses a curated list of fusion from COSMIC as reference and extract the fusion.

```
genefuse -t 16 -r ./Ref/genome.fa  -f ./genefuse/genes/druggable.hg19.csv  -1
Normal/D83/d83_r1.fq\  Normal/D83/d83_r2.fq -h Fusion-result/normal/d83/d83_report.html >
Fusion-result/normal/d83/d83_result
```

- The reference genome fasta file, specified by -r or –ref
- The fusion setting file, specified by -f or --fusion=
- The fastq file(s), specified by -1 or --read1= for single-end data. If dealing with pair-end data, specify the read2 file by -2 or --read2=
- Option -h or --html= to specify the file name of HTML report
- The plain text result is directly printed to STDOUT, you can pipe it to a file using a >

Differential gene Expression:

Tophat software was used for the alignment of reads to the reference genome (hg19). The command used for alignment was:

```
tophat -p 14 -o ~/Desktop/d86-aligned --transcriptome-index hg19/transcriptome_index/genes hg19/Bowtie2Index/genome d86_read1.fq d86_read2.fq
```

The Aligned files in BAM files (Binary form of Sequence Alignment File) were used for the reference guided transcripts assembly using Cufflinks 2.0 against the Boost version of 10 in ubuntu 16.04. the command used for transcript assembly was:

```
cufflinks -p 16 -o ./d83__cufflinks/ -g ~/Desktop/cufflinks_ref/genes.gtf d83.bam
```

Likewise, the assembly was done for all the samples including normal and tumor. After the assembly process, Cufflinks produces assembled transcripts file in GTF format. The final assembly was done by using Cuffmerge. Cuffemrge, another tool available in Cufflinks. Cuffemerge merge all the assembles transcript into one GTF file. The mega merge was done for tumor and Adjacent normal separately.

```
cuffmerge -o cuffmerge_out -p 16 --ref-gtf /cufflinks_ref/genes.gtf --ref-sequence /cufflinks_ref/genome.fa transcripts-normal_and_tumor.txt
```

The “transcripts-normal_and_tumor.txt" file is composed of the list of all the assembled transcript GTF file as below:

/normal/d83__cufflinks/transcripts.gtf

/normal/d85__cufflinks/transcripts.gtf

/normal/d86__cufflinks/transcripts.gtf

/tumor/t83__cufflinks/transcripts.gtf

/tumor/t84__cufflinks/transcripts.gtf

/tumor/t85__cufflinks/transcripts.gtf

/tumor/t86__cufflinks/transcripts.gtf

The Cuffmerge tool assembled all the transcripts and create a file “merged.gtf”. The pre-calculation of expression level of genes was done by using Cuffquant, another tool available in cufflinks. This process was done for the adjacent normal and the tumor separately.

For Adjacent Normal:

$cuffquant -p 14 -o ./cuffquant/normal/d83 -u -b ~/Desktop/cufflinks_ref/genome.fa ./assembled_transcript/cuffmerge_out/merged.gtf ./Aligned-Bam/normal/d83.bam

$cuffquant -p 14 -o ./cuffquant/normal/d85 -u -b ~/Desktop/cufflinks_ref/genome.fa ./assembled_transcript/cuffmerge_out/merged.gtf ./Aligned-Bam/normal/d85.bam

$cuffquant -p 14 -o ./cuffquant/normal/d86 -u -b ~/Desktop/cufflinks_ref/genome.fa ./assembled_transcript/cuffmerge_out/merged.gtf ./Aligned-Bam/normal/d86.bam

For Tumor:

$cuffquant -p 14 -o ./cuffquant/tumor/t83 -u -b ~/Desktop/cufflinks_ref/genome.fa ./assembled_transcript/cuffmerge_out/merged.gtf ./Aligned-Bam/tumor/t83.bam

$cuffquant -p 14 -o ./cuffquant/tumor/t84 -u -b ~/Desktop/cufflinks_ref/genome.fa ./assembled_transcript/cuffmerge_out/merged.gtf ./Aligned-Bam/tumor/t84.bam

$cuffquant -p 14 -o ./cuffquant/tumor/t85 -u -b ~/Desktop/cufflinks_ref/genome.fa ./assembled_transcript/cuffmerge_out/merged.gtf ./Aligned-Bam/tumor/t85.bam

$cuffquant -p 14 -o ./cuffquant/tumor/t86 -u -b ~/Desktop/cufflinks_ref/genome.fa ./assembled_transcript/cuffmerge_out/merged.gtf ./Aligned-Bam/tumor/t86.bam

After the transcript abundance estimation by Cuffquant, Cuffdiff was used to estimate the differential expression level between Adjacent normal and tumor samples.

$cuffdiff -p 16 -o dge-cuffdiff/ final_assembly/merged.gtf -u -b genome.fa -L Normal, Cancer cuffquant/normal/d83/abundances.cxb,normal/d85/abundances.cxb,normal/d86/abundances.cxb tumor/t83/abundances.cxb,tumor/t84/abundances.cxb,/tumor/t85/abundances.cxb,/tumor/t86/abundances.cxb

Data Visualization: CummeRbund, a Bioconductor package for R veersion 3.5 was used for visualization of differential gene expression data.

Pathway analysis using DAVID 6.8

The upregulated and downregulated genes were filtered from the significant expression gene list. Lunux grep and awk commands were used for filtering the differentially expressed genes from the Cuffdiff data. The criteria for searching the

significant gene expression gene was qvalue <=0.05 and abs fold change > 2 or log2 fold change > 1.

Finding out significant expression gene list:

```
cat gene_exp.diff |grep 'yes' |awk '{if (($10 >= 1)||($10<=1)||($8<=0.05)) print }' > Significant_gene_exp.txt
```

Finding out upregulated gene list:

The upregulated genes were filtered from the significant genes by setting the log2 fold change greater than one.

```
cat Significant_gene_exp.txt | awk '{if ($10 > =1)}' | awk '{print $3}'> gene_Upregulated.txt2fold
```

Finding out down regulated gene list:

The upregulated genes were filtered from the significant genes by setting the log2fold change greater than one.

```
cat Significant_gene_exp.txt | awk '{if ($10 <1)}' | awk '{print $3}'> gene_Downregulated.txt
```

The upregulated genes and Down-regulated genes were analysed to in DAVID 6.8 to enriched the genes in KEGG pathway (Huang *et al.,* 2007). Based on the Benjamini score the pathway were selected for further interpretation. The upregulated genes were searched in cBioPortal (https://www.cbioportal.org/) against the TCGA (The Cancer Genome Atlas) the stomach adeno carcinoma datasets to find out mutational relation to their expression level (Gao et al., 2013). And also in The Human Protein Atlas (https://www.proteinatlas.org/) to check their expression in stomach cancer in other patients.

RESULT

Fusion genes

After the analysis with the Genefuse, One fusion gene were detected one Adjacent normal in one sample (D83) and three fusion genes were detected in Tumor samples in two tumor samples (T83 and T85).

Table 4: Fusion gene found in Normal sample:

Adjacent Normal	Status (Number)	Fusion
D83	Yes (1)	**ETV6**_ENST00000396373.4:intron:1|-chr12:11835082__**NTRK3**_ENST00000394480.2:intron:14|+chr15:88532314
D85	No	---
D86	No	---

Table 5: Fusion gene found in Tumor samples.

Tumor	Status (Number)	Fusion
T83	Yes (2)	**ABL1**_ENST00000318560.5:+chr9:133688513__**ALK**_ENST00000389048.3:intron:5|-chr2:29601685 **ALK**_ENST00000389048.3:intron:13|+chr2:29461163__**STRN**_ENST00000263918.4:intron:9|+chr2:37106315
T84	No	---
T85	Yes (1)	**CLTC**_ENST00000269122.3:intron:1|-chr17:57710715___**TPR**_ENST00000367478.4:intron:27|+chr1:186311260
T86	No	---

Differential gene Expression

The cuffdiff output files were analysed using CummeRbund, a bioconductor package in R3.5. The dispersion plot, sample-wise dendrogram, Box-plots, Heatmaps for upregulated and downregulated genes were constructed.

Dispersion Plot

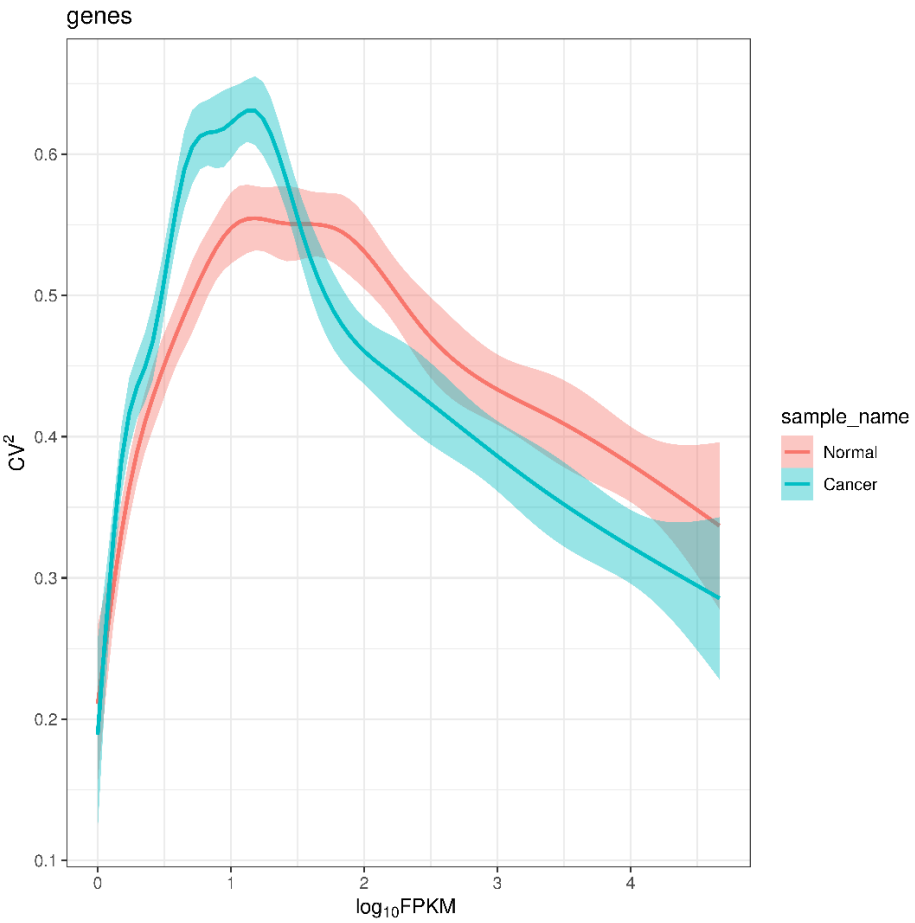
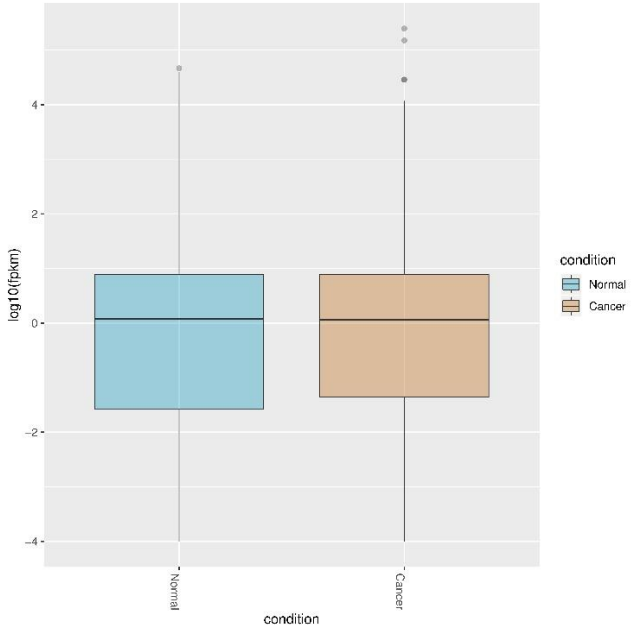


Fig 2: Dispersion Plot: Genes counts varies among samples: Overlapping regions implies less coefficient of vitiation (CV). The plot hast Log_{10}FPKM values in X axis and square of Coefficient of Variation (CV^2) in Y-axis.

Box plot

(a)



(b)

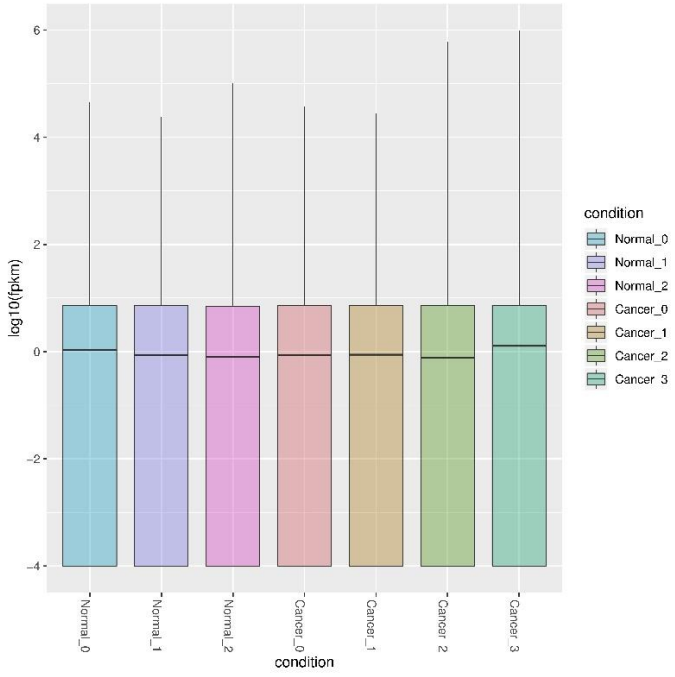


Fig 3: Boxplot: (a) Variation Seen in across both the conditions (B) Variation seen across the condition with all replicates: Normal_0, Normal_2, Normal_2 are D83, D85, D86 and Cancer_0 to Cancer_3 are T83, T84, T85 and T86 respectively.

MA Plot

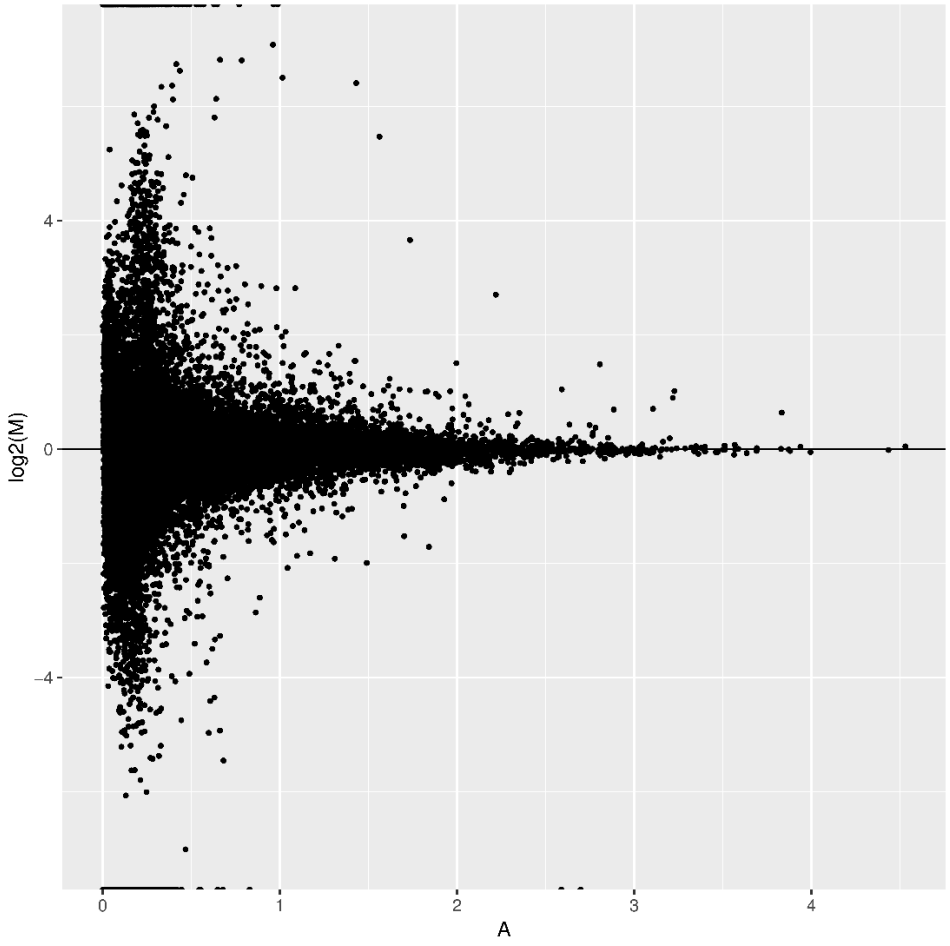
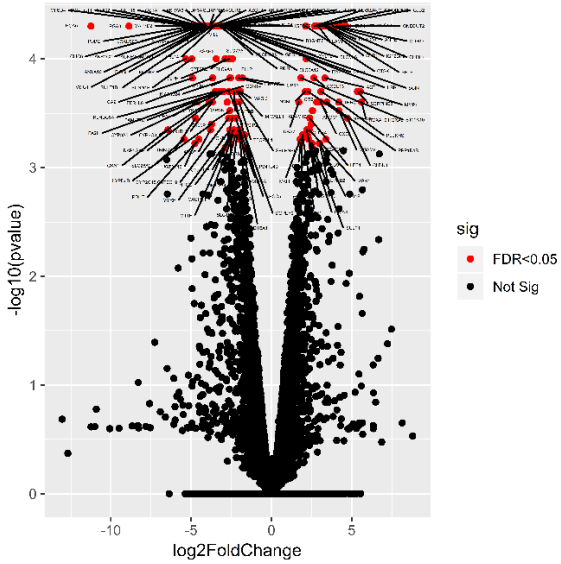


Fig 4: M vs A plot (Avg. intensity vs log ratio) for a given pair of conditions across all FPKMs: MA Plot of two condition taking log transformation of fold-change (M). Genes with similar expression values in both normal and tumor samples are clustered i.e., genes expressed with no significant differences in between the two conditions (M=0).

Volcano Plot

(a)



(b)

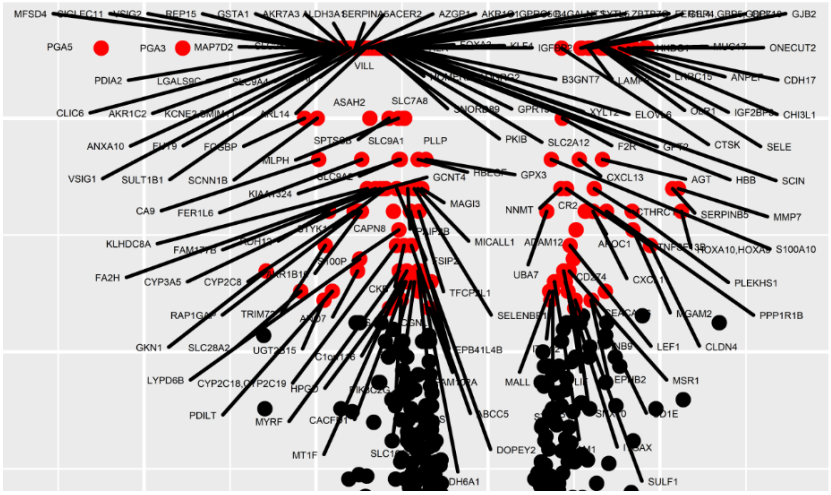


Fig: 5 (a)The log2FoldChange and negative log10 (Adjusted P-Value) (FDR) was plotted in x and y axis repectively. (b) The red points show differentially expressed genes with FDR less than 0.05. The Volcano plot was constructed using R ggplot2 package (Wickham, 2016).

Heatmap of upregulated genes

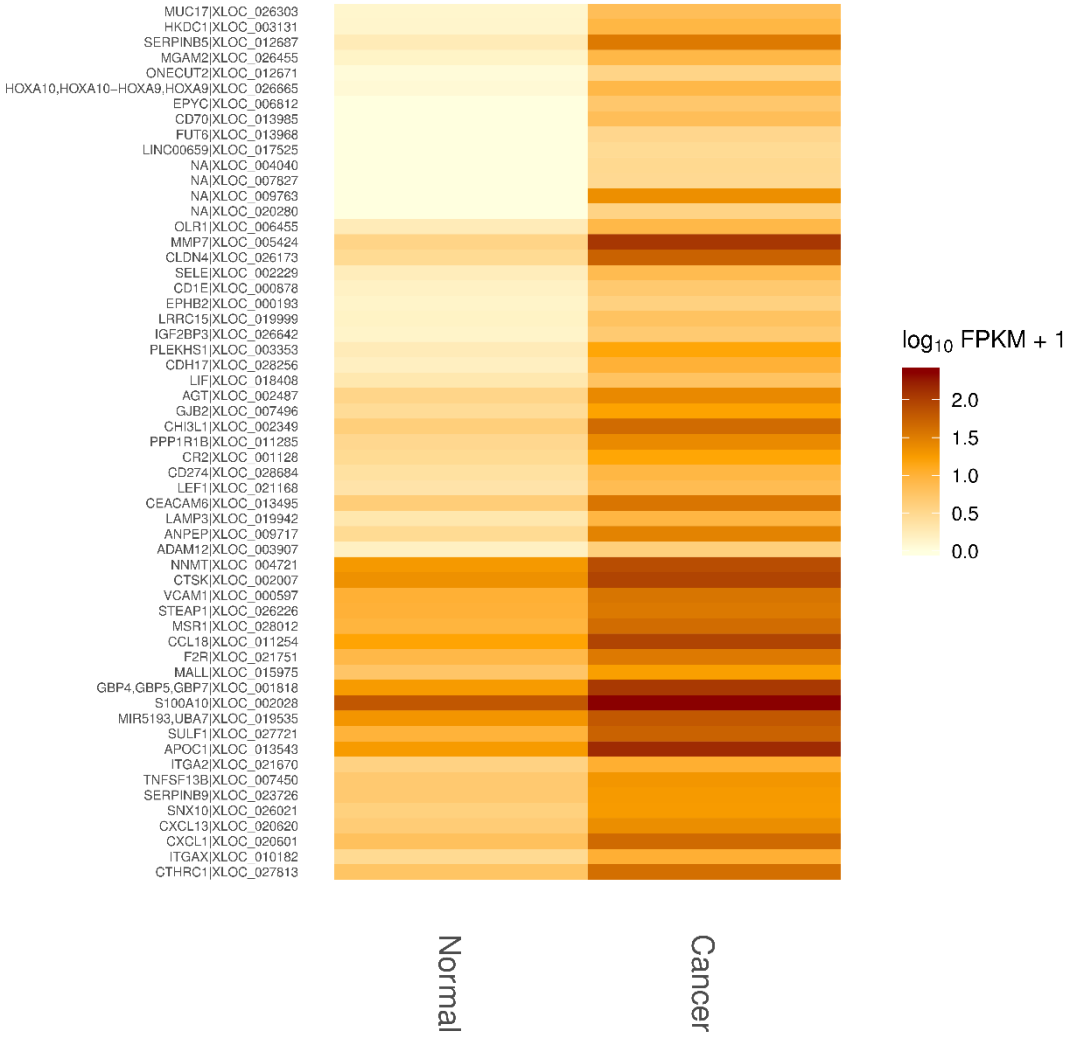


Fig 6: Heatmap constructed for the upregulated genes using CummeRbund. Dark red colore implies the most upregulated genes.

MMP7, S100A10, APOC1 showed high level of expression in Cancer tissue. *SERPINB5, CDLN4, CTSK, CCL18, GBp4-7, MIR5193, UBA7, SULF1, CXCL1, CTHRc1* fall within the range of medium to high expression. TP53 is mostly altered gene in gastric cancer (Fenoglio-Preiser *et al.,* 2003) but no significant expression was found in this study.

Heatmap of the Down-regulated gene

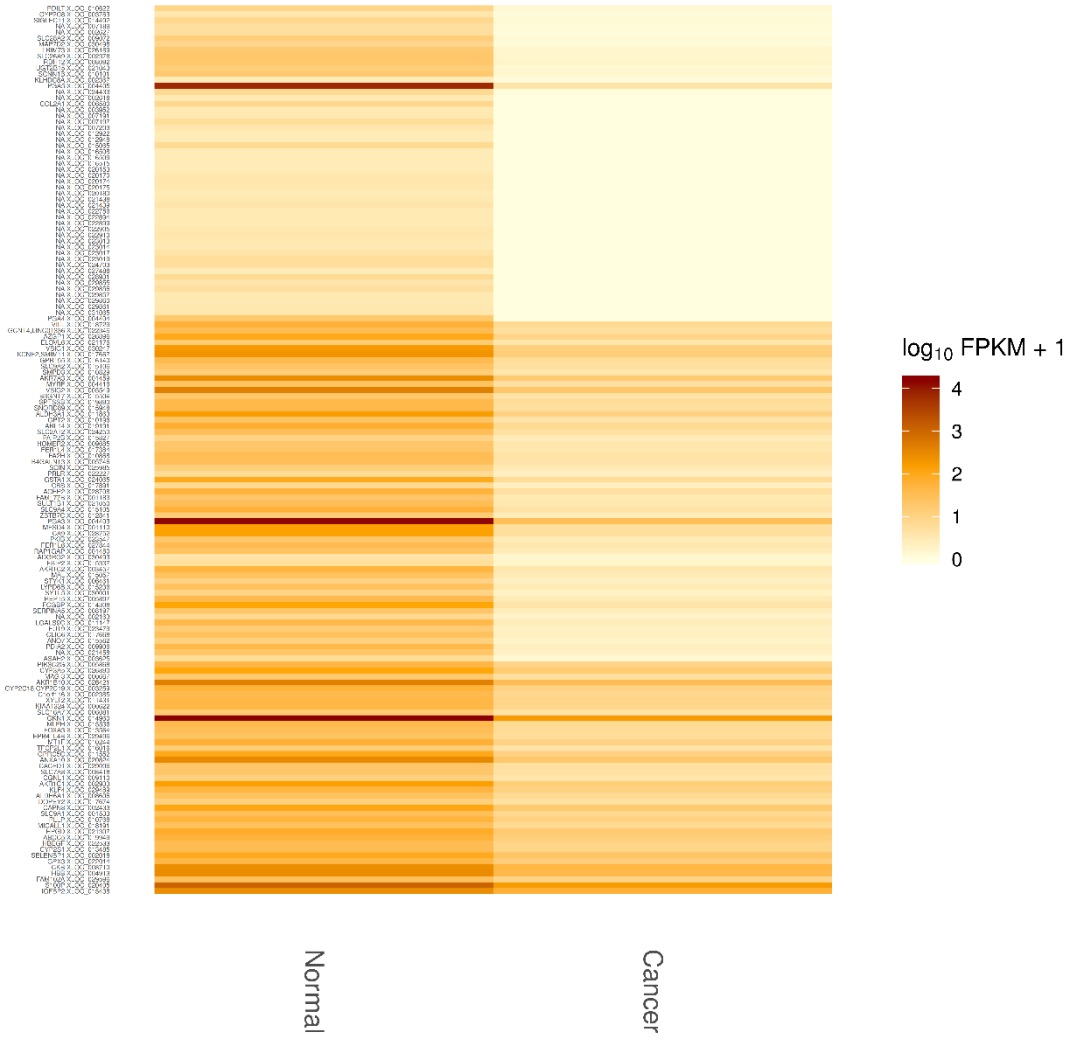


Fig 7: Heatmap constructed for the Down-regulated genes using CummeRbund. Dark red colour implies the most upregulated genes in Normal and corresponding light red band implies Down-regulated genes in Cancer.

PGA5, PGA3 and *GKN1* are three genes mostly downregulated in Cancer. PGA5 (Pepsinogen 5, Group I) encodes a protein precursor of the digestive enzyme pepsin. Diseases associated with PGA5 include Atrophic Gastritis and Gastritis (*Genecardss*). *PGA3* (Pepsinogen 3, Group I (Pepsinogen A) is a Protein Coding gene. Diseases associated with PGA3 include Atrophic Gastritis and Gastritis (*Genecardss*).

Pathway analysis with DAVID 6.8

The Gen set enrichment analysis was done using DAVID 6.8. The analysis was done for both upregulated genes and downregulated genes to investigate the which pathways are affecting due to the differential expression of the genes. The Upregulated gene list was uploaded in DAVID database and KEGG pathways were seen and analysed.

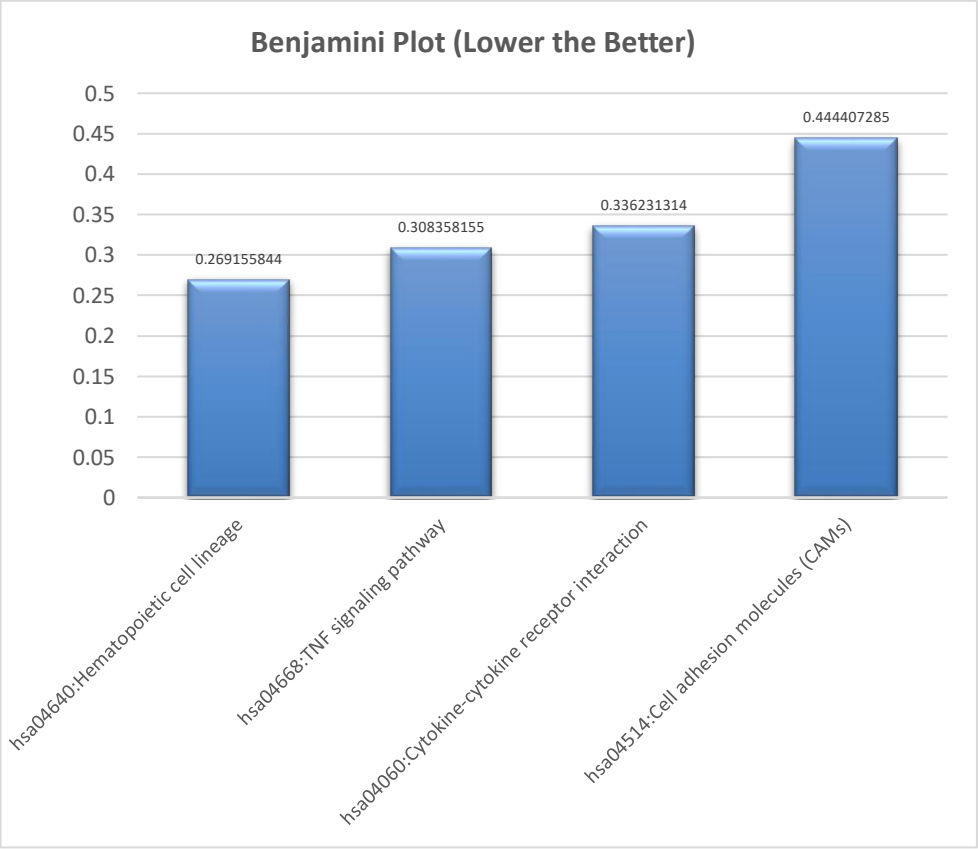


Fig 8: Plot of Benjamini score assigned to the upregulated gene involving pathway

Upon analysis, 4 genes were enriched in the Hematopoietic cell lineage, 4 Genes were enriched in the TNF Signalling Pathway, 6 Genes were enriched in the Cytokine-Cytokine receptor interaction pathway, 4 Genes were enriched in Cell

Adhesion Molecule pathway. The function of each Genes were extracted from the *Genecards* database (https://www.*Genecards*s.org/).

The gene enriched in hematopoietic cell lineage pathway are *CD1E, ANPEP, CR2, ITGA2*. CD1e molecule (CD1E) proteins helps in self of microbial lipid and glycolipid antigens presentation to the T cells. The Alanyl aminopeptidase, membrane (*ANPEP*) gene plays a role in the peptide digestion and also involved in the peptide processing including peptide hormones. The Complement C3d receptor 2(*CR2*) acts as receptor of human B lymphocytes and often highly expressed when the person et infected with EBV virus. The Integrin subunit alpha 2(*ITGA2*) gene codes for transmembrane receptor (Alpha sub unit) for collagens. Integrin are belonged to cell adhesion molecules and cell surface receptors, they aid in cell to cell or cell to ECM (Extracellular matrix) interactions. (*Genecards*; Böger *et al.,* 2015).

The function of genes involved in TNF signalling pathways were retrieved from *Genecards* database. The gene C-X-C motif chemokine ligand 1 (*CXCL1*) is an antimicrobial gene encodes for chemokines that paly major role in inflammation. Abnormal expression of this protein is associated with tumor growth progression. LIF Interleukin 6 Family Cytokine (*LIF*) is involved in the induces hematopoietic differentiation. Selectin E (*SELE*) has role in accumulation of blood leukocytes in the inflammation site. Vascular cell adhesion molecule 1 (*VCAM1*) aids in adhesion of leukocyte-endothelial cell and play major role in signal transduction *(Genecardss).*

6 genes were enriched in Cytokine-cytokine receptor path. These include C-C Motif chemokine ligand 18 (*CCL18*) is involved in Immunoregulatory and inflammatory processes and C-X-C Motif chemokine ligand 1 (*CXCL1*) acts as B lymphocyte chemo-attractant and is an antimicrobial peptide and CXC chemokine are expressed in spleen and lymph nodes. C-X-C Motif chemokine ligand 13 (*CXCL13*) also acts as chemo-attractant to B lymphocyte. CD70 Molecule (*CD70*) encodes cytokine that belongs to the tumor necrosis factor (*TNF*) ligand family. It induces proliferation of T cells, aid in increasing the cytolytic T cells, and major role in T cell activation. This cytokine is also reported to play a role in regulating B-cell activation, cytotoxic function of natural killer cells, and immunoglobulin synthesis. Tumor necrosis factor superfamily member 13b (*TNFSF13B*) belongs to the tumor necrosis factor (TNF) ligand family. This cytokine is expressed in B cell lineage cells, and acts as a potent B cell activator. LIF Interleukin 6 Family Cytokine (LIF) was also enriched in the cytokine-cytokine receptor pathway (*Genecardss*).

The genes enriched in Cell adhesion molecule (CAM) CD274 molecule (*CD274*) encodes an immune inhibitory receptor ligand that is mostly seen to express in T cell, B cells and also in various types of tumor cells. In tumor microenvironments, this interaction provides an immune escape for tumor cells through cytotoxic T-cell inactivation. Claudin 4 (*CLDN4*) encodes for integral membrane proteins high expression of gene is related affects the solute movement among the cells. Similarly, Selectin E(*SELE*), Vascular cell adhesion molecule 1 (*VCAM1*) were also enriched in CAM pathway.

The upregulated genes were searched in cBioPorat (Gao *et al.,* 2013) against the TCGA (The Cancer Genome Atlas) the stomach adeno carcinoma datasets namely, Stomach Adenocarcinoma (Pfizer and UHK, Nat Genet 2014), Stomach Adenocarcinoma (TCGA, Nature 2014) and Stomach Adenocarcinoma (U Tokyo, Nat Genet 2014). Only the datasets Stomach Adenocarcinoma (TCGA, Nature 2014) has the expression summary in cBioPortal. The *genes MMP7, CLDN4, CTSK, CCL18, GBP4, GBP5, SULF1* were frequently mutated in the dataset and showed medium to high level expression.

The Gene set enrichment analysis was also done for the Down regulated genes. Three pathways with lower Benjamini score was selected for the analysis.

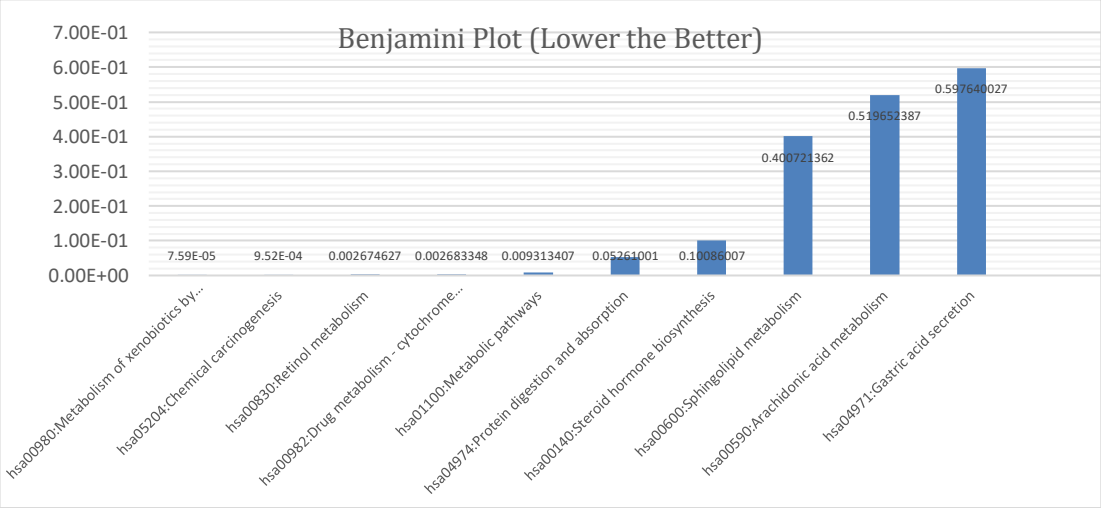


Fig 9: Plot of Benjamini score assigned to the Down-regulated gene involving pathway

The pathway selected for the down-regulated genes were Xenobiotics metabolism, Chemical carcinogenesis and retinol metabolism pathways. The information of the genes with their major function were retrieved from Genecards database.

Genes involved in Metabolism of Xenobiotics Pathway includes UDP glucuronosyltransferase family 2 member B15 (*UGT2B15*) gene that encodes a glycosyltransferase that involves in toxic compounds elimination from the body. Aldehyde dehydrogenase 3 family member A1 (*ALDH3A1*) encodes for Aldehyde dehydrogenase enzyme which is a novel gastric cancer marker with potential prognostic values (Wu *et al*., 2016). Aldo-keto reductase family 1 member C1 (AKR1C1) codes for enzyme that catalyses the reaction of progesterone to the inactive form. Aldo-keto reductase family 1 member C2 (*AKR1C2*) gene encodes a member of the Aldo/keto reductase superfamily, many more known enzymes and proteins. Mutation in Aldo-keto reductase family 7 member A3 (*AKR7A3*) causes disease that include carcinogenesis in pancreas. Cytochrome P450 play major role in protective mechanism against hepato-carcinogens. Cytochrome P450 family 2 subfamily S member 1 (*CYP2S1*) and Cytochrome P450 family 3 subfamily A member 5 (*CYP3A5*) codes for the cytochrome P450 proteins which catalyse many reactions involved in drug metabolism and cholesterol and lipid synthesis. Glutathione S-transferase alpha 1 (*GSTA1*) is responsible for degrading xenobiotics.

The genes involved in chemical carcinogenesis pathway are UDP glucuronosyltransferase family 2 member B15(UGT2B15), Aldehyde dehydrogenase

3 family member A1(*ALDH3A1*), Cytochrome P450 family 2 subfamily C member 18(*CYP2C18*), Cytochrome P450 family 2 subfamily C member 19(*CYP2C19*), Cytochrome P450 family 2 subfamily C member 8(*CYP2C8*), Cytochrome P450 family 3 subfamily A member 5(*CYP3A5*), Glutathione S-transferase alpha 1(*GSTA1*).

Genes involved in retinol metabolism pathway are UDP glucuronosyltransferase family 2 member B15 (*UGT2B15*), Cytochrome P450 family 2 subfamily C member 18 (*CYP2C18*), Cytochrome P450 family 2 subfamily C member 8(CYP2C8), Cytochrome P450 family 2 subfamily S member 1 (*CYP2S1*), Cytochrome P450 family 3 subfamily A member 5 (*CYP3A5*) and *RDH12*. The cytochrome P450 proteins involves in drug metabolism and synthesis of cholesterol, steroid. *RDH12* encodes a retinal reductase (NADPH-dependent) who has highest activity towards all-trans-retinol (*Genecards*).

DISCUSSION

The fusion genes play a major role in carcinogenesis and a many fusion genes have been detected in different cancer types. ETV6-NTRK3 fusion has not been reported in Stomach cancer, but EN Fusion leads to the activation of two of the Ras-MAPK mitogenic pathway and the PI3K (phosphatidyl inositol-3-kinase) pathway and AKT cell survival factors are subsequently get activated (Lannon and Sorensen, 2005). AKT (Protein Kinase B) enhances transcription of anti-apoptotic genes by inhibiting inhibits transcription factors helps in cell death genes expression (Song *et al*., 2007). Dysregulation of AKT/PI3K pathway is frequent several cancers including gastric cancer (Singh *et al.,* 2015).

ABL1 (9q34.12) protein functions as a kinase and ABL1 kinase should be activated normally to perform as tumor suppressor (Dasgupta *et al*., 2016). The ALK gene is located on chromosome 2 and encodes a transmembrane tyrosine kinase. Mutation in ALK including fusion has been reported in neuroblastoma, Lung cancer and anaplastic large cell lymphoma. ABL1- ALK Fusion has not been reported in any type of cancer. No record was found in TCGA Fusion database and COSMIC Fusion Record. ALK gene arrangement is very rare in gastrointestinal cancers (Alese *et al.,* 2015). Further validation is required.

STRN (2p22.2) involves in Neurophysiological process, Glutamate regulation of Dopamine D1A receptor signaling and Plasma membrane estrogen receptor signalling pathways (*Genecards*). It is also a prognostic marker in renal cancer (The

Human Protein Atlas). STRN-ALK Fusion is a potential therapeutic target in Thyroid Cancer (Kelly *et al.*, 2014) and in colorectal Cancer (Yakirevich *et al.,* 2014). COSMIC Fusion database is reported in only one STRN fusion i.e. STRN-ALK based on different breakpoint in 5' partner and 3' partner from tissue Lung, Peritoneum and Thyroid. STRN-ALK fusion is declared as therapeutic target in gastric cancer as it was tested by an ALK inhibitor in STRN-ALK positive patient (Li *et al.,* 2017; Patent: WO2017153932A1).

However, ALK-STRN is different from STRN-ALK due to its sequence arrangement (Fusion Breakpoint) and is not reported in Gastric cancer and further validation is required. *CLTC* (Clathrin Heavy chain-17q23.1) which codes for protein is present in the Cytoplasmic face of the vesicles. TPR (1q31.1) gene encodes a large coiled-coil protein that forms intranuclear filaments attached to the inner surface of nuclear pore complexes (NPCs) (Genetic Home reference, NIH). CLTC-TPR fusion is not reported in any other cancer type. CLTC-TFF3 has been found in RCC (Takeuchi *et al.,* 2019), CLCT-ALK in Human Diffuse Large B Cell Lymphomas (Cerchietti *et al.,* 2011).

Differential Gene Expression between Normal and Tumor always gives the idea about the genes that are affecting the phenotype. Over-expression of **MMP7** is associated with cell invasion and metastasis and is up-regulated by catecholamines

in gastric cancer (Shi *et al.,* 2010). Previous Study suggests very poor prognostic effect of ***MMP7*** association with aggressive type of Gastric cancer. *MMP7: 181A→G* (rs11568818) polymorphism in the *MMP7* promoter modulates gene expression and possibly affects cancer progression (Soleyman-Jahi *et al.,* 2015). Similarly, S100A10 is also reported to have strong expression in Gastric cancer and is associated with poor differentiation and metastasis. APOC1 gene is also a commonly upregulated gene in gastric cancer (Oue *et al.,* 2004).

A systematic meta- and bioinformatics analysis through multiple online databases up to Feb 10, 2017 reported down regulation of SERPINB5 in Gastric cancer (Zheng *et al.,* 2017). Thus, Upregulation of SERPINB5 in the present study conflicts with the previous findings. High expression of CLDN4 is associated with increasing tumor size, and lymph node metastasis in patients with GC (Chen *et al.,* 2016). CCL18 is reported as a prognostic indicator in Gastric cancer: Expressed in tumor-associated macrophages (Leung *et al.,* 2004). CCL18 (binds with CCR8) and CCL18 is reported as Independent Prognostic Factor and CCR8 correspondingly expressed highly in TCGA RNA-Seq data (Islam *et al.,* 2013). Dysregulation of CCL18/CCR8 could predict the poor prognosis in patients with GC and provide a potential antitumor target for the treatment of GC. No significant expression CCR8 found in our data. SERPINB5 gene encodes Maspin is a mammary serine protease inhibitor and inhibits invasion and metastasis of cancer cells as a tumor suppressor. No reported information about MIR5193 (Micro RNA 5193) association with gastric exist. The Human Protein Atlas showed the high to medium level expression of

GBP5 in gastric Adenocarcinoma (https://www.proteinatlas.org/ENSG00000154451-GBP5/pathology). High to medium level expression of GBP5 in gastric Adenocarcinoma (https://www.proteinatlas.org/ENSG00000213512-GBP7/pathology). Overexpression to medium expression of SULF1 in gastric cancer is reported. (Junnila *et al.,* 2010; https://www.proteinatlas.org/ENSG00000137573-SULF1/pathology). PGA5 was found to be downregulated in four gastric cancer (GSE545) patients (Oue *et al.,* 2004). There was no report of PGA5 and PGA3 expression summary in The Human Protein Atlas. Loss of GKN1 expression is frequently detected in gastric Cancer and Gastric Mucosa with *Helicobacter pylori* (Yoon *et al.,* 2014).

Pathway analysis helps in identifying the pathways that are getting affected by downregulation and upregulation of the genes. The Human Protein Atlas and our DGE result implies medium expression of CD1e and ANPEP. The Human Protein Atlas reports no expression of CR2, but upregulated expression was detected in our study. *ITGA2* showed high expression in Gastric cancer. Expression *CD1e* and *ANPEP* might causes aggressive tumour type. Generally, cell adhesion proteins (e.g. cadherins with mucins) are involved in epithelial-to-mesenchymal transitions through oncogenic pathways. The Human Protein Atlas reports no expression of *SELE, VCAM* (TNF Signaling pathway Gene gene) very low expression of *CXCL13* (Cytokine-Cytokine Receptor pathway gene), *CD274* (Cell Adhesion Pathway gene) .

Likewise, UGT2B15 (Xenobiotic Metabolism Pathway gene) and CYP2C19 (Chemical carcinogenesis Pathway gene) has no expression record in The Human Protein Atlas.

Fusion ETV6-NTRK (Translocation) in adjacent normal and ABL-ALK (Translocation), ALK-STRN (Deletion), CLCT-TPR (Translocation) in gastric cancer are found in this study which have not been reported earlier for gastric cancer. Even though this fusion gene has to be verified with other techniques, the present detection was done through GeneFuse software which was proven to have high accuracy in the detection (Chen *et al.,* 2018). The individual expression of gene candidate partner in fusions were insignificant which also implies that Cuffquant might not measure the transcript abundance for those gene due to chimeric events. Comparison with TCGA expression data (TCGA nature 2014) in cBioPortal (Gao *et al.,* 2013) and our up-regulated expression data suggest further mutational studies are required in *MMP7, CLDN4, CTSK, CCL18, GBP4, GBP5, SULF1* genes. High expression of MMP7, S100A10, APOC1 and low Expression of *PGA5, PGA3* and *GKN1* suggest that these can be predictive Biomarker in Gastric cancer among Mizo population. Pathway analysis implies that the upregulated genes are mostly involving in Immune response. Downregulation of the genes involves in Xenobiotic metabolism pathway also indicates the progression of the disease. Downregulation of *GSTA1, CYP1A1, CYP1/23* involved in chemical carcinogenic pathway may have role in gastric cancer progression by forming DNA Adducts. This can be correlated with the previous findings in *GST* gene polymorphism with low expression in gastric cancer in Mizo population (Ghatak *et al.,* 2016).

SUMMARY

The salient findings of this research can be summarised below:

- ETV6-NTRK3 fusion was detected in adjacent normal of one patent sample (T83) and it has not been reported in Stomach cancer or adjacent normal before in any other population.
- ABL1- ALK Fusion was found in Stomach cancer and it has not been reported in any type of cancer in any other population.
- Likewise, the fusions ALK-STRN and CLCT-TPR were not reported in any type of cancer in any other population.
- Breakpoint analysis implied that ETV6-NTRK, ABL-ALK, CLCT-TPR fusions arises due to Translocation and ALK-STRN fusion due to deletion.
- TCGA expression data and our up-regulated expression data suggests that further mutational studies are required in MMP7, CLDN4, CTSK, CCL18, GBP4, GBP5, SULF1 genes in Mizo population.
- Very low expression of *PGA5, PGA3* and *GKN1* were observed in gastric cancer tissue which are involved in gastric acid secretion.
- High expression of MMP7, S100A10, APOC1 and low Expression of PGA5, PGA3 and GKN1 suggests that these can be predictive biomarker in Gastric cancer among Mizo population.

APPENDICES

Appendix -1: The preview of the questionnaire followed for the sample collection of the study.

Questionnaire for Epidemiological Study of Gastric Cancer

Referring Dr:_____ MSCI/Civil Hospital No._____/_____

Referring Unit:_____ Reg Date:_____

PROFORMA

MZU,MSCI,CIVIL Hospital & NIBMG

PERSONAL HISTORY

Hming (*Name)*: Mipa/Hmeichhia (*Male/Female*):

Kum (*Age):*

Tawng hman (*Language*): Nupui/pasal nei/neilo (*Marital status):*

Pian ni*(Date of birth):* *Nupui/pasal neiha kum zat* (*Age at the time of marriage*):

Rihzawng (Weight): San zawng (*Height*):

Lehkha zir chen(*Education*): Eizawnna (*Occupation)*:

Unau engzat nge in nih*? (No. of Siblings):* [ ] Mipa (*Male*) [ ] Hmeichhia (*Female*) [ ]

Fa I nei em? (*Do you have children?):* Aw/*Yes* [ ] Aih/*No* [ ]

I neih chuan, fa engzat nge I neih? (*If yes, how many children do you have?):* [ ]

Mipa/Hmeichhia engzat nge? (*Gender of the children):* Mipa(*Male)* [ ] Hmeichhia(*Female)* [ ]
(Thi sa a piang chhiar tel tur, chhiat erawh chhiar tel loh tur) (*Please include stillbirths; it is not necessary to include miscarriages)*

Addres:_____
_____Pin Code_____

Tel No._____Mob.No._____

E mail:_____

Cancer Diaognosis/Treatment_____

Engtik kumah nge cancer I vei tih hmuhchhuah a nih? *(Year of cancer detected?):*_____

Tumor	Site	Age	Histopathology	Surgery Date	Chemotherapy Date	Radiation
1st Primary						
2nd Primary						
3rd Primary						

Environmental/Lifestyle Factors

What has been your main occupation?_____

Hengah tc hian hna I thawk cm? I hnathawhnaah hetiang te hi I in chiahpiah tir em? *(Do you have Occupational exposure to?)*		No. of years	Age (From/to)	Nature of use	Name of company/brand
Radiation(eg. In a factory,laboratory/medical setting)	Yes No Don't Know				
Plastic	Yes No Don't Know				
Agriculture/Rubber plant (If yes C4A)	Yes No Don't Know				
Pesticides/Pest control/ Mosquito Repellant	Yes No Don't Know				
Chemical/Dyes	Yes No Don't Know				
Any other exposure (Asbestors,Chromium or Lead)	Yes No Don't Know				

i) Was your mother an agriculture worker around the time of your birth? Yes/No

ii) Has DDT ever been used in or around your household? Yes/No

iii) What is your water supply source? River [ ] Tube well[ ] Govt./municipal [ ]

iv) Other_____

I hna a hahthlak viau em, zan lam ah hna I thawk em(night duty)? (*Is your job stressful or do you perform shift work (night duty)?)*: Aw/*Yes* [ ] Aih/*No* [ ]

In in bulah cell phone tower a awm em?*(Is there a cell phone tower near your house?)*:

Aw/Yes[ ] Aih/No [ ]

TASTE PREFERENCES:

Do you consume (I ei ngai em)	0(Never)	1(Little) 1 days in a week	2(Average) 2-4 days in a week	3(Heavy) 5-7 days in a week
Spicy food				
Western food (Pizza,burgers,fries)				
Burmies product				
Sour test (tamarind ,lime juice etc)				
Bawngsa (*Beef*				
Vawksa (*Pork)*				
Kelsa (*Mutton)*				
Arsa (*Chicken)*				
Artui (*Egg)*				
Sangha (*Fish)*				
fermented fish				
Bekang/fermented pulse				
Sa-Um				
Extra salt with food				
Pickles/chutneys				
Smoked vegetables				
Smoked meat				
Fat intakc				
Boiled food				
Fried food				
Smoked food				
Fibers food/fruits (Banana				

I nghei tawh anih chuan, engtik atangin? (*If quit already, since when?):*

If consumption has changed during life record highest consumption.

Beverage	Yes/No	From age	To age	Units/Day	Days/Week

Mei I zu em? (*Do you smoke?):* Aw/*Yes* [ ] Aih/*No* [ ]

Aw (*If yes*): In reng (*Regularly*) [ ] A chang chang in (*Occasionally*) [ ]

Nikhat ah engzat nge I zuk thin tlangpui *(Average Number of smoke per day)*:

Engtik atangin nge I zuk tan? (*When did you start smoking?)*:

Eng nge I zuk? (*Type of Smoke)*:Zozial*(Local)* [ ]; Biri [ ];Cigarette(Eng siam?)*(Which brand?):*

I nghei tawh anih chuan, engtik atangin? *(If quit already, since when?)*:

Has there ever been a time when you smoked at least one cigarette per day for three months or longer?

[ ] Yes [ ] No [ ] Don't know

If yes list consumption (excluding times when the subject did not smoke)

Product	Yes/No	Used From/To	Frequency	Av. Quantity per day
Cigarette				
Biri				
Zozial				

Vaihlo a siam thil dang tih I nei em? (*Do you consume other tobacco products?):*Aw/Yes [ ] Aih/No [ ]

Aw *(If yes):* Ti reng *(Regularly)* [ ] A chang chang in(O*ccasionally)* [ ]

Have you ever chewed pan or tobacco regularly?(At least once a week for six months or more)Yes [ ] No [ ] Don't Know [ ]

Type	Yes/No	From age	To age	No. per day
Chewing with tobacco and lime(khaini) Pan+tabacco+betelnut+lime+catechu(mewa)				
Gutka				
Sahdah(*Oral snuff*)				

impression_____

_____

Colonoscopy/Endoscopy: Regions_____Date_____

Impression_____

_____

Natna/Damlohna dang I nci cm? (*Do you have any other diseases?):* Aw/*Yes* [ ] Aih/*No* [ ]

I neih chuan, eng natna nge? (*If yes, what type of disease?):* _____

H. pylori [ ] Diabetes [ ] obesity [ ] HIV [ ] HbsAg[ ] HCV[ ]
EBV [ ] Gastric atrophy [ ]

Surgery: Site/Procedure_____

Pathological Staging-pTNM_____Date_____

Histopathological Report: Specmen_____Path No._____

Date_____Impression_____

_____

_____

IHC: Hormone receptor status

Tumor details: Specimen_____Path No._____

Report Date_____Grade_____Size of the tumor_____

cm. Tumor emboli_____Lymphovascular Invasion_____

_____

_____

Appendix-II: List of Significant genes filtered from Cuffdiff data.

Gene	p_value	q_value	significant
EPHB2	0.0006	0.044208	yes
VCAM1	0.0005	0.039361	yes
KIAA1324	0.00015	0.017392	yes
MAGI3	0.0002	0.020489	yes
CD1E	0.0006	0.044208	yes
MFSD4	5.00E-05	0.006925	yes
CR2	0.00025	0.02337	yes
FAM177B	0.0002	0.020489	yes
AKR7A3	5.00E-05	0.006925	yes
RAP1GAP	0.00025	0.02337	yes
SLC9A1	0.00015	0.017392	yes
GBP4,GBP5,GBP7	5.00E-05	0.006925	yes
CTSK	5.00E-05	0.006925	yes
SELENBP1	0.0002	0.020489	yes
S100A10	0.00015	0.017392	yes
SELE	5.00E-05	0.006925	yes
CHI3L1	5.00E-05	0.006925	yes
KLHDC8A	0.0002	0.020489	yes
SLC26A9	5.00E-05	0.006925	yes
C1orf116	0.0005	0.039361	yes
CAPN8	0.00025	0.02337	yes
AGT	0.00015	0.017392	yes
AKR1C1	5.00E-05	0.006925	yes
HKDC1	5.00E-05	0.006925	yes
CYP2C18, CYP2C19	0.00035	0.030794	yes
PLEKHS1	0.00025	0.02337	yes
AKR1C2	5.00E-05	0.006925	yes
ASAH2	0.0001	0.012569	yes
CYP2C8	0.00025	0.02337	yes
ADAM12	0.0003	0.027528	yes
PGA3	5.00E-05	0.006925	yes
PGA4	5.00E-05	0.006925	yes
PGA5	5.00E-05	0.006925	yes
MYRF	0.00045	0.037186	yes
NNMT	0.0002	0.020489	yes
HBB	5.00E-05	0.006925	yes
MMP7	0.0002	0.020489	yes
VSIG2	5.00E-05	0.006925	yes

B4GALNT3	5.00E-05	0.006925	yes
PIK3C2G	0.0006	0.044208	yes
REP15	5.00E-05	0.006925	yes
SLC16A7	0.0005	0.039361	yes
OLR1	5.00E-05	0.006925	yes
STYK1	0.0002	0.020489	yes
COL2A1	5.00E-05	0.006925	yes
EPYC	5.00E-05	0.006925	yes
TNFSF13B	0.0002	0.020489	yes
GJB2	5.00E-05	0.006925	yes
RDH12	0.00025	0.02337	yes
SERPINA5	5.00E-05	0.006925	yes
SLC7A8	0.0001	0.012569	yes
ALDH6A1	0.00045	0.037186	yes
CKB	0.00035	0.030794	yes
SLC28A2	0.00055	0.041758	yes
CGNL1	0.00055	0.041758	yes
HOMER2	5.00E-05	0.006925	yes
ANPEP	5.00E-05	0.006925	yes
PDIA2	5.00E-05	0.006925	yes
SCNN1B	0.0001	0.012569	yes
ITGAX	0.0005	0.039361	yes
GPT2	5.00E-05	0.006925	yes
MT1F	0.00065	0.047194	yes
PDILT	0.00055	0.041758	yes
PLLP	0.0001	0.012569	yes
SMPD3	0.0005	0.039361	yes
FA2H	0.0002	0.020489	yes
LGALS9C	5.00E-05	0.006925	yes
CCL18	5.00E-05	0.006925	yes
PPP1R1B	0.00025	0.02337	yes
XYLT2	5.00E-05	0.006925	yes
GPRC5C	5.00E-05	0.006925	yes
ALDH3A1	5.00E-05	0.006925	yes
ONECUT2	5.00E-05	0.006925	yes
SERPINB5	0.0002	0.020489	yes
ZBTB7C	5.00E-05	0.006925	yes
CYP2S1	0.00045	0.037186	yes
CEACAM6	0.00055	0.041758	yes
APOC1	0.00025	0.02337	yes
FOXA3	5.00E-05	0.006925	yes
FUT6	5.00E-05	0.006925	yes

CD70	5.00E-05	0.006925	yes
FCGBP	0.0001	0.012569	yes
SIGLEC11	5.00E-05	0.006925	yes
GKN1	0.00045	0.037186	yes
MAL	5.00E-05	0.006925	yes
SLC9A4	5.00E-05	0.006925	yes
SLC9A2	0.00015	0.017392	yes
LYPD6B	0.0004	0.034583	yes
FSIP2	0.0002	0.020489	yes
IGFBP2	5.00E-05	0.006925	yes
B3GNT7	5.00E-05	0.006925	yes
MLPH	0.0001	0.012569	yes
ANO7	0.00045	0.037186	yes
PAIP2B	0.0002	0.020489	yes
SNORD89	5.00E-05	0.006925	yes
MALL	0.00055	0.041758	yes
TFCP2L1	0.00025	0.02337	yes
GPR155	5.00E-05	0.006925	yes
FER1L4	5.00E-05	0.006925	yes
LINC00659	5.00E-05	0.006925	yes
KCNE2, SMIM11	5.00E-05	0.006925	yes
CLIC6	5.00E-05	0.006925	yes
DOPEY2	0.0005	0.039361	yes
CBS	0.00045	0.037186	yes
MICALL1	0.0002	0.020489	yes
LIF	0.00055	0.041758	yes
VILL	5.00E-05	0.006925	yes
ARL14	5.00E-05	0.006925	yes
UBA7	0.00025	0.02337	yes
SPTSSB	0.0001	0.012569	yes
LAMP3	5.00E-05	0.006925	yes
ABCC5	0.00035	0.030794	yes
LRRC15	5.00E-05	0.006925	yes
S100P	0.0003	0.027528	yes
CXCL1	0.00025	0.02337	yes
CXCL13	0.00015	0.017392	yes
ANXA10	5.00E-05	0.006925	yes
UGT2B15	0.0006	0.044208	yes
SULT1B1	5.00E-05	0.006925	yes
LEF1	0.00035	0.030794	yes
ELOVL6	5.00E-05	0.006925	yes
HPGD	0.0005	0.039361	yes

ITGA2	0.0006	0.044208	yes
F2R	0.0001	0.012569	yes
GPX3	0.00015	0.017392	yes
PRLR	5.00E-05	0.006925	yes
GCNT4	0.0002	0.020489	yes
HBEGF	0.00015	0.017392	yes
FUT9	5.00E-05	0.006925	yes
PKIB	5.00E-05	0.006925	yes
SERPINB9	0.00045	0.037186	yes
GSTA1	5.00E-05	0.006925	yes
SLC2A12	5.00E-05	0.006925	yes
SCIN	5.00E-05	0.006925	yes
SNX10	0.00065	0.047194	yes
CLDN4	0.00035	0.030794	yes
TRIM73	0.00035	0.030794	yes
STEAP1	0.00055	0.041758	yes
MUC17	5.00E-05	0.006925	yes
AKR1B10	0.00025	0.02337	yes
MGAM2	0.0003	0.027528	yes
IGF2BP3	5.00E-05	0.006925	yes
HOXA10, HOXA9	0.00025	0.02337	yes
CYP3A5	0.0002	0.020489	yes
AZGP1	5.00E-05	0.006925	yes
SULF1	0.0006	0.044208	yes
CTHRC1	0.0002	0.020489	yes
FER1L6	0.00015	0.017392	yes
MSR1	0.00045	0.037186	yes
CDH17	5.00E-05	0.006925	yes
CD274	0.0004	0.034583	yes
ACER2	5.00E-05	0.006925	yes
CA9	0.00015	0.017392	yes
CACFD1	0.0005	0.039361	yes
KLF4	5.00E-05	0.006925	yes
EPB41L4B	0.00025	0.02337	yes
FAM102A	0.00065	0.047194	yes
SYTL5	5.00E-05	0.006925	yes
VSIG1	5.00E-05	0.006925	yes
ADGRG2	5.00E-05	0.006925	yes
MAP7D2	5.00E-05	0.006925	yes

REFERENCES

Anders S and Huber W. (2010) Differential expression analysis for sequence count data. Genome Biol. 11:R106, doi.org/10.1186/gb-2010-11-10-r106.

Andrews S. (2010) FastQC: a quality control tool for high throughput sequence data. Ann Intern Med. 1463, 4054.

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, and Jemal A. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 68(6):394-424.

Chen S, Liu M, Huang T, Liao W, Xu M, and Gu J. (2018) GeneFuse: detection and visualization of target gene fusions from DNA sequencing data. Int J Biol Sci. 14(8):843-848.

Conesa A, Madriga P, Tarazona S, Gomez-Cabrero D et al. (2016) A survey of best practices for RNA-seq data analysis. Genome Biology. 17:13. https://doi.org/10.1186/s13059-016-0881-8

Crew KD, and Neugut AI. (2006) Epidemiology of gastric cancer. World J Gastroenterol, 12: 354-362.

Eftang LL, Esbensen Y, Tannæs TM, Blom GP, Bukholm IR and Bukholm G. (2013). p-regulation of CLDN1 in gastric cancer is correlated with reduced survival. BMC Cancer. 13:586. doi: 10.1186/1471-2407-13-586

Fernandez-Cuesta L et al. (2015) Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. Genome Biol. 16(1):7

Froussios K, Schurch NJ, Mackinnon K, Gierlinski M, Duc C, Simpson GG, and Barton GJ. (2016) How well do RNA-Seq differential gene expression tools perform in higher eukaryotes? bioRXiv (preprint)

Ghatak S, Yadav RP, Lalrohlui F, Chakraborty P, Ghosh S, Das M, Pautu JL, Zohmingthanga J, and Senthil KN. (2016) Xenobiotic Pathway Gene Polymorphisms Associated with Gastric Cancer in High Risk Mizo-Mongoloid Population, Northeast India. Helicobacter. 1(6):523-535.

Ghosh S, and Chan CKK. (2016) Analysis of RNA-Seq Data Using TopHat and Cufflinks. In: Edwards D. (eds) Plant Bioinformatics. Methods in Molecular Biology. Vol: 1374. Humana Press, New York, NY

Hansen KD, Brenner SE, and Dudoit S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 38(12):e131.

Heyer EE, Deveson IW, Wooi D, and Selinger C. (2019) Diagnosis of fusion genes using targeted RNA sequencing.Nat Commun. 10(1):1388.

Hu KW, Chen FH, Ge JF, Cao LY, and Li H. (2012) Retinoid receptors in gastric cancer: expression and influence on prognosis. Asian Pac J Cancer Prev. 13(5):1809-17.

Ishaq S, and Nunn L. (2015) Helicobacter pylori and gastric cancer: a state of the art review. Gastroenterol Hepatol Bed Bench. 8: S6–S14.

Israel DA, Salama N, Arnold CN, et al.(2001) Helicobacter pylori strain-specific differences in genetic content, identified by microarray, influence host inflammatory responses. J Clin Invest. 107(5):611–620.

Karimi P, Islami F, Anandasabapathy S, Freedman ND, and Kamangar F. (2014) Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention. Cancer Epidemiol Biomarkers Prev. 23(5):700–713.

Lauren P. (1965) The two histologically main types of gastric carcinoma: Diffuse and so-called intestinal-type carcinoma: An attempt at a histo-clinical classification. Acta Pathol Microbiol Scand. 64:31-49.

Li WX, He K, Tang L, et al., (2017) Comprehensive tissue-specific gene set enrichment analysis and transcription factor analysis of breast cancer by integrating 14 gene expression datasets. Oncotarget. 8(4):6775-6786.

Lin Y, Wu Z, and Guo W. (2015) Gene mutations in gastric cancer: a review of recent next-generation sequencing studies. Tumor Biol. 36: 7385. https://doi.org/10.1007/s13277-015-4002-1

Masugi Y, Nishihara R, Yang J, et al. (2016) Tumour CD274 (PD-L1) expression and T cells in colorectal cancer. Gut-bmj,66(8):1463-1473.

Morishita A, Gong J, and Masaki T. (2014) Targeting receptor tyrosine kinases in gastric cancer. World J Gastroenterol. 20(16):4536-45.

Moss SF, Calam J, Agarwal B, Wang S, and Holt PR. (1996) Induction of gastric epithelial apoptosis by Helicobacter pylori. Gut. 38(4):498-501.

Mukaisho K, Nakayama T, Hagiwara T, Hattori T, and Sugihara H. (2015) Two distinct etiologies of gastric cardia adenocarcinoma: interactions among pH, Helicobacter pylori, and bile acids. Front Microbiol. 6:412

Oue N, Hamai Y, Mitani Y, Matsumura S, Oshimo Y, Aung PP, Kuraoka K, Nakayama H and Yasui W.(2004) Gene expression profile of gastric carcinoma: identification of genes and tags potentially involved in invasion, metastasis, and carcinogenesis by serial analysis of gene expression. Cancer Res. 64(7):2397-405

Petrovchich I, and Ford JM. (2016) Genetic predisposition to gastric cancer. Seminars in Oncology. 43(5):554-559

Raplee ID, Evsikov AV, and Marín de Evsikova C. (2019) Aligning the Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression Quantification Tools for Clinical Breast Cancer Research. J Pers Med. 9(2). pii: E18. doi: 10.3390/jpm9020018

Rawla P, and Barsouk A. (2019) Epidemiology of gastric cancer: global trends, risk factors and prevention. Prz Gastroenterol. 14(1):26-38.

Servarayan Murugesan C , Manickavasagam K, Chandramohan A Jebaraj A, Jameel ARA, Jain MS, and Venkataraman J.(2018) Gastric cancer in India: epidemiology and standard of treatment. Updates in Surgery. 70: 233

Shen S, Jiang J, and Yuan Y.(2017) Pepsinogen C expression, regulation and its relationship with cancer. Cancer Cell International. 17:57. doi: 10.1186/s12935-017-0426-6.

Siewert JR, and Stein HJ. (1998) Classification of adenocarcinoma of the oesophagogastric junction. Br J Surg. 85:1457-9

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, and Pachter L.(2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7(3):562-78.

Trapnell C, Williams BA, Pertea G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28(5):511–515.

van der Woude CJ, Kleibeuker JH, Tiebosch AT, Homan M, Beuving A, Jansen PL, and Moshage H. (2003) Diffuse and intestinal type gastric carcinomas differ in their expression of apoptosis related proteins. J Clin Pathol. 56(9):699-702.

Wickham H. (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. ISBN 978-3-319-24277-4.

Wroblewski LE, Peek RM, and Wilson KT. (2010) Helicobacter pylori and gastric cancer: factors that modulate disease risk. Clin Microbiol Rev. 23(4):713-39.

Yoon JH, Choi WS, Kim O, and Park WS. (2014) The role of gastrokine 1 in gastric cancer. J Gastric Cancer, 14(3):147-55.

Zali H, Rezaei-Tavirani M, and Azodi M. (2011) Gastric cancer: prevention, risk factors and treatment. Gastroenterol Hepatol Bed Bench. 4(4):175-85.

Bio-data

Name: Ranjan Jyoti Sarma

M.Phil. Research Scholar

Address:

Department of Biotechnology

Mizoram University

Aizawl -796 004, Mizoram, India

Email: ranjanjsarma@gmail.com

Mobile: 9774440639.

Educational Qualification:

Qualification	Board/University	Year	Division	Percentage
M.Phil Biotechnology	Mizoram University	2019	Pursuing	Not declared
M.Sc Bioinformatics	Pondicherry University	2017	I	75%
B.Sc Biotechnology	Northeastern Hill University	2015	I	63%
Class XII	AHSEC, Assam	2011	I	66%
Class X	SEBA, Assam	2009	I	75%

Workshop Attended:

- **Analysis of Genome Scale Data from Bulk and Single-cell Sequencing**. 19th - 23rd Nov 2018, Conducted by EMBL-EBI & NIBMG at NIBMG, Kalyani.
- **3rd Advanced Research Training Workshop on Understanding Human Disease and Improving Human Health Using Genomics-Driven Approaches**. 23rd -31st July 2018, Conducted by NIBMG, Kalyani.
- **The Concept and Application of Genomics in Clinical Medicine**. 11th August 2018, Conducted by CSIR- Institute of Genomics and Integrative Biology(CSIR-IGIB) , New Delhi.
- **Hands-on Training Workshop on Cancer Genomics**.19th -23rd March 2018, Organized by DBT- NER Biotechnology/Bioinformatics Centre, Advanced

Centre for Treatment, Research & Education in Cancer, Khargar, Navi Mumbai.

- **Recent Advances in Cancer Research-2018 (RACR-2018).**5th -7th March 2018, organized by Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati and sponsored by Department of Biotechnology, Government of India.

- **A Brief Introduction to Bioinformatics and Systems Biology**.13th -14th December 2018, organized by Bioinformatics Infrastructure Facility (BIF), department of Biotechnology, Mizoram University sponsored by Department of Biotechnology (DBT), New Delhi.

- **Statistical Methods in Biological Research**.3rd – 5th november 2017, organized by Bioinformatics Infrastructure Facility (BIF), Department of Biotechnology, Mizoram University sponsored by Department of Biotechnology (DBT), New Delhi.

- **Application of NGS in Microbial Ecology**.30th – 2nd November 2017, organized by Bioinformatics Infrastructure Facility (BIF), Department of Biotechnology, Mizoram University sponsored by Department of Biotechnology (DBT), New Delhi.

- **Research Training Workshop on Understanding Human Disease and Improving Human Health Using Genomics-Driven Approaches.**19th – 24th November 2017, Conducted by NIBMG AT NIBMG, Kalyani.

Conference attended:

- The 12th Annual Convention Of association of Biotechnology and Pharmacy (ABAP) & International Conference on Biodiversity, Environment and Human Health: Innovation and Emerging Trends (BEHIET 2018).12th – 14th November 2018, organized at the School of Life Sciences, Mizoram University, Aizawl , Mizoram- 796004

Poster presentation:

- Poster presentation on **"Differential Gene expression profiling of Gastric cancer RNA-Seq data"** at the 12th Annual Convention Of association of Biotechnology and Pharmacy (ABAP) & International Conference on Biodiversity, Environment and Human Health: Innovation and Emerging Trends (BEHIET 2018).12th – 14th November 2018, organized at the School of Life Sciences, Mizoram University, Aizawl , Mizoram- 796004.

Paper Communicated:

- Subbarayan Sarathbabu, **Ranjan Jyoti Sarma**, H. Lalhruaitluanga, Devadasan Velmurugan, Selvi Subramanian, Nachimuthu Senthil Kumar. (2019). In vitro DNA binding activity and molecular docking reveals pierisin-5 as an anti-proliferative agent against Gastric cancer. Journal of Biomolecular Structure & Dynamics (Communicated and under Review) IF : 3.107

Skills Developed:

- Linux Shell Scripting, MS Windows, Biolinux8.
- Coding in Perl.
- NGS Data Analysis (RNA-Seq Data QC, Alignment, Assembly, Gene Expression, Fusion Gene Detection, Analysis using R, Whole Exome Sequencing Data Analysis).
- R/ R-Studio.
- MySQL, PHP, HTML/CSS-Basic.
- Cheminformatics Software Applications.
- MS Word, Excel, PowerPoint, LibreOffice (Linux)

Award:

- Anundaram Barua Award, Govt. of Assam, 2009