IDENTIFICATION OF RECURRENT GENOMIC ALTERATIONS IN GASTRIC ADENOCARCINOMA IN MIZO POPULATION

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

PAYEL CHAKRABORTY

MZU REGN NO.: 1600803

PH.D REGN NO.: MZU/PH.D./1045 OF 26.05.2017

DEPARTMENT OF BIOTECHNOLOGY

SCHOOL OF LIFE SCIENCES

FEBRUARY 2021

IDENTIFICATION OF RECURRENT GENOMIC ALTERATIONS IN GASTRIC ADENOCARCINOMA IN MIZO POPULATION

BY

PAYEL CHAKRABORTY

Department of Biotechnology

Name of the Supervisor: Prof. N Senthil Kumar

Submitted

In partial fulfillment of the requirement of the Degree of Doctor of Philosophy in Biotechnology of Mizoram University, Aizawl

CERTIFICATE

This is to certify that the thesis entitled "**Identification of recurrent genomic alterations in Gastric adenocarcinoma in Mizo population**" submitted to the Mizoram University; in partial fulfillment for the degree of Doctor of Philosophy in Biotechnology is a record of research work carried out by **Ms. Payel Chakraborty** under our personal supervision and guidance.

No part of this thesis has been reproduced elsewhere for any degree.

Dated: 26/02/2021

Dr. N. Senthil Kumar

DECLARATION
Mizoram University
February 2021

I, **Payel Chakraborty**, hereby declare that the subject matter of this thesis entitled "**Identification of recurrent genomic alterations in Gastric adenocarcinoma in Mizo population**" is the record of work done by me, that the contents of this thesis did not form basis of the award of any previous degree to me or to the best of my knowledge to anybody else, and that the thesis has not been submitted by me for any research degree in any other University/Institute.

This is being submitted to the Mizoram University for the degree of Doctor of Philosophy in Biotechnology.

Date: 26/02/2021

(Candidate)

(Dr. N. Senthil Kumar)
(Supervisor)

ACKNOWLEDGEMENT

First and foremost, I am grateful to **God**, the Creator and **my parents**, **my family** and to whom I owe my very existence, and for providing me the opportunity to step in the excellent world of science. I am thankful to him for giving me the grace and privilege to pursue this program and successfully complete it in spite of many challenges faced. The journey has been quite remarkable, and it is a unique stepping stone to many exploits ahead. I have also been supported and supervised by many people to whom I would like to express my deepest gratitude.

I would like to acknowledge my indebtedness and render my warmest gratitude to my supervisor **Prof. N. Senthil Kumar,** Department of Biotechnology, Mizoram University, who is captivating, honest, and the true embodiment of a mentor. His hard working nature, unconventional approach, openness and flexibility at accepting new ideas and creativity provided me with a perfect environment in which I was able to improve myself both at the personal and academic levels to achieve the very best and to explore new frontiers with devotion and confidence. His sacrifice of time and patience in listening to my ideas and problems, untiring suggestions and thoughtful criticisms are things that I would always cherish in the days to come. Learning from him was enormous, his intuition has made him as a constant oasis of ideas and passions in science to provide me unflinching encouragement to transform me a budding researcher and I believe his wishes are immensely valuable for my career ahead. I would also like to pay my respects to all the faculty member**,** Department of Biotechnology, Mizoram University for extending all required facilities to complete my research work. All the faculty members from Department of Biotechnology, Mizoram University were kind enough for their help and encouragement at various phases of research, whenever I approached them, and I do hereby acknowledge all of them. I am thankful to the office and technical staff of the Department for their constant help and support.

I owe my gratitude to **Dr. Jeremy L. Pautu**, Oncologist, Mizoram State Cancer Institute, Zemabawk, Aizawl, Mizoram; **Dr. John Zohmingthanga**, Head pathologist, Civil Hospital, Aizawl, Mizoram for providing me samples and scientific suggestions**.** I would like to express my deepest thank to **Dr. Arindam Maitra,** Associate Professor, **Sillarine Kurkalang and Arindam Palodi**, National Institute of Biomedical Genomics, Kalyani, West-Bengal for his inspiration, constant support and for giving new ideas during research. I owe special mention to my friends, **Souvik Ghata**, **Subhajit Mukherjee, Surajit De Mandal, Ravi Prakash Yadav, S. Sarath Babu, Brinda Senthil Kumar, Freda Lalrohlui, Doris Zodinpuii, Christine Sailo, Dr. Lalawmpuii Pachuau, Andrew Vanlallawma, Zothan Zami, Ranjan Jyoti Sarma, Lallianmawii Pachuau, Vabeiryureilai Mathipi, Puja Pandey, Michael Vanlalchhuana, Basanti Mukherjee, Litika Varmani, Dipika Malakar, David K. Zorinsanga and Pu Rina** who worked in lab and helped me throughout the completion of my research as well as for sharing the joyous, sad feelings and giving me the golden moments during this period.

Specially, I also thankful to **Department of Biotechnology (DBT),** Government of India, New Delhi for providing the fellowship to complete my research works **under Department of Biotechnology (DBT), New Delhi sponsored projects on Gastric Cancer (BT/551/NE/U-Excel/2014) and (DBT-NER/health/46/2015) to Mizoram University.**

At last but not the least, I acknowledge all those who provide their samples and consent for collecting samples and those who knowingly and unknowingly contributed in making my work easier and a real success.

Date: 26/02/2021 Payel Chakraborty

List of Tables

List of figures

Introduction and Review of Literature

Worldwide stomach cancer occupies the fifth position and is recorded as the third lethal cancer as per the mortality rate (Bray et al. 2018). The incidence rate of gastric cancer (GC) always differs greatly across geography, ethnicity and gender. Stomach cancer rates are double in males in comparison with females (Ferlay et al. 2012). The occurrence rate is higher in Eastern Asia, remarkably in Korea (holding the highest position). Southeastern Asia and Southcentral Asia also has high incidence rate (Bray et al. 2018). Numerous countries of Western Asia have reported that GC is the main reason for death among male patients. However, the occurrence rate is declining steadily, but still, it is the third most deadly malignancy in the world (Rawla et al. 2019). National Cancer Registry Programme in India has reported stomach cancer as the third most prevalent cancer among males, according to incidence and it is the fourth most prevalent cancer in the North Eastern region of India (Mathur et al. 2020). Mizoram, a northeastern state of India, recorded the highest occurrence rate of Gastric Cancer in India (Ibrahim et al. 2017) and globally holds a fifth position (Phukan et al. 2004).

Gastric cancer is mainly the cancer of the stomach, which has three distinct parts: fundus, body or corpus and pyloric antrum. The stomach has mucosa lining which is comprised of three glands: i) cardiac glands containing foveolar cells which produce mucus, ii) oxyntic glands containing parietal cells, which produces hydrochloric acid and chief cells which produces pepsinogen, and iii) gastrin secreting endocrine G cells (Tan et al. 2015). GC can be histologically classified into: i) intestinal and ii) diffuse (Lauren et al. 1965) (Figure 1). Intestinal type is identifiable by glands that sort from well-differentiated to moderately differentiated tumors and occasionally found with poorly differentiated tumors. Intestinal mucins are present in higher amounts in comparison with diffuse-type GC and are frequently found on a background of intestinal metaplasia. Diffuse type GC comprises poorly cohesive cells, diffusely infiltrating the gastric wall (Machlowska et al. 2019). These tumors have similarities with signet-ring cell tumors. In Diffuse types of tumors, desmoplasia is more common and inflammation is less obvious in comparison with intestinal-type GC. There is a third subtype which contains both intestinal and diffuse type of cells and is called mixed gastric carcinoma.

Besides Lauren classification, WHO classified GC into four main types: i) Tubular adenocarcinomas, ii) Papillary adenocarcinomas, iii) Mucinous adenocarcinomas and iv) Signet-ring cell carcinomas (Lauwers et al. 2010). Tubular adenocarcinoma is a very common subtype of GC and histologically the cells are irregularly separated, fused and branching tubules vary in size (Hu et al. 2012). It contains intraluminal mucus, nuclear and inflammatory debris as well. Papillary adenocarcinoma develops in the proximal part of the stomach and is also related to liver metastasis and a higher amount of lymph node involvement (Hu et al. 2012). This is well-differentiated exophytic carcinoma, histologically characterized as cylindrical or cuboidal cells, an elongated finger-like structure that is maintained by fibrovascular connective tissue cores. Histologically, Mucinous adenocarcinoma comprises extracellular mucinous pools occupying more than 50% volume of tumor and is used to form glandular structure and irregular cell clusters. In Signet-ring cell carcinoma, an isolated and small group of malignant cells containing intra-cytoplasmic mucin occupies more than 50% volume of the tumor. Morphologically, Signet-ring cell carcinoma can appear in five forms: i) clear signet ring-like when nuclei are in the reverse of the cell membrane due to expanded globoid clear cytoplasm, ii) another type where cells look like histiocytes with central nuclei and without or with little mitotic activity, iii) tinny eosinophilic cells with very clear, but small mucin containing cytoplasmic granules, iv) tinny cells with very less or no mucin, and v) anaplastic cells with very less or without mucin (Lauwers et al. 2010) (Figure 1). These forms are used to blend with each other which in turn produces a different proportion of tumor.

The main 4 types of GC are prime subtypes mentioned by WHO and other rare subtypes are squamous carcinoma, adenosquamous carcinoma, hepatoid adenocarcinoma, choriocarcinoma, carcinoma with lymphoid stroma, parietal cell carcinoma, malignant rhabdoid tumor, panteth cell carcinoma, mucoepidermoid carcinoma, undifferentiated carcinoma, mixed adeno-neuroendocrine carcinoma, embryonal carcinoma, endodermal sinus tumor, oncocytic adenocarcinoma and pure gastric yolk sac tumor (Hu et al. 2012).

Clinically, Gastric cancer can be classified as early and advanced stages to determine the proper interventions or therapy. American joint committee on Cancer (AJCC), 8th edition on cancer staging classified Pathological tumor, node and metastasis information (pTNM) in four stages: I, II, III and IV, wherein stage I is divided into IA & IB, stage II subdivided into IIA & IIB, Stage III subdivided into IIIA, IIIB and IIIC. AJCC also classified grading as: well-differentiated, moderately differentiated and poorly differentiated types (Lauwers et al. 2010).

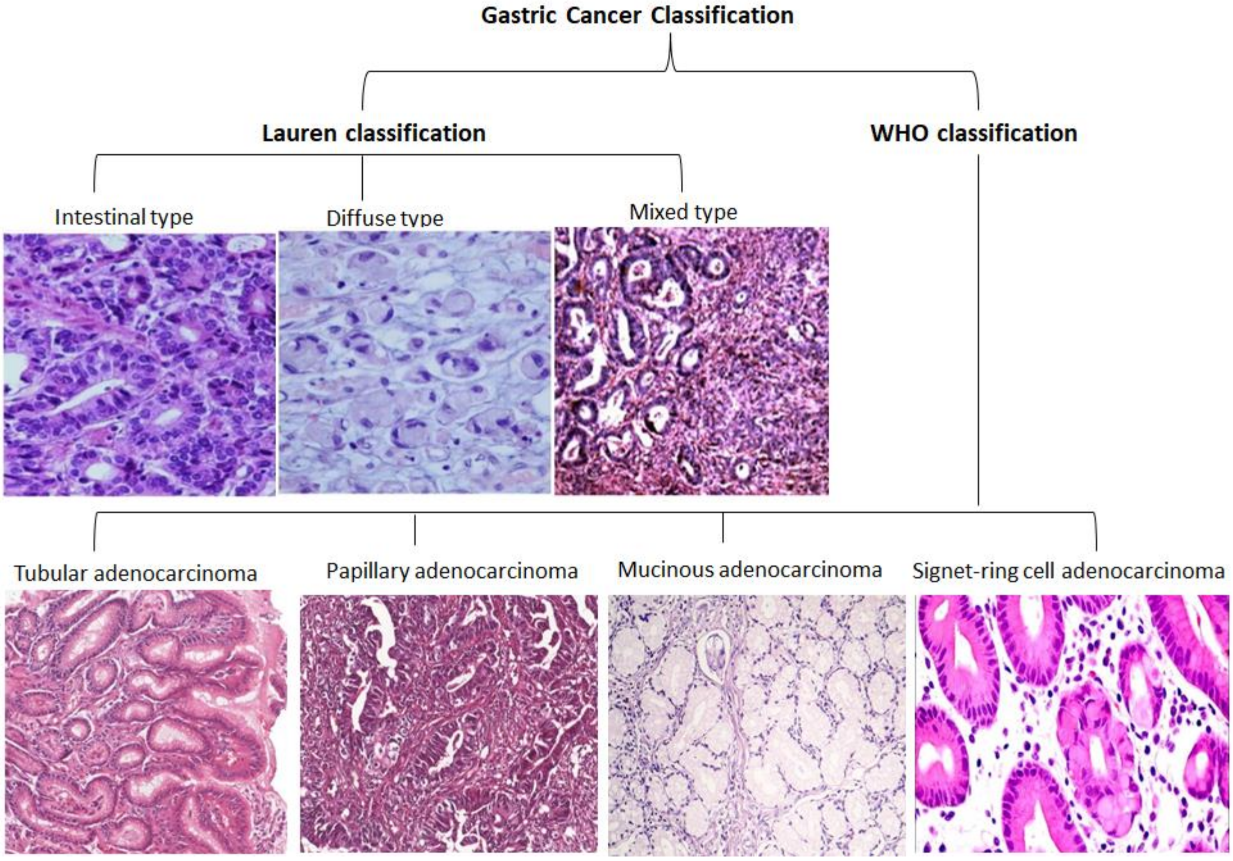


Figure 1: Classification of Gastric Cancer

(Source: https://www.slideshare.net/PritikaNehra1/gastric-cancer-pathology-seminar; Zhu et al. 2014; Ghita et al. 2011; https://www.dreamstime.com/mucinous-carcinoma-stomach-mucinous-carcinoma-stomach light-micrograph-photo-under-microscope-image126790777;http://currinfo.blogspot.com/2018/01/signet-ring cell-carcinoma.html)

Gastric cancer has a pre-malignant stage and role of inflammation is an important factor for developing intestinal-type Gastric Cancer. The stages are chronic gastritis, atrophic gastritis, intestinal metaplasia (IM), dysplasia and adenocarcinoma. Chronic, atrophic and IM are the prime stages for the development of Gastric pathogenesis. Chronic gastritis is the starting of inflammation in normal mucosa. Atrophic mucosa is a stage where replacement of glandular cells occurred with fibrosis cells or metaplastic cells (Tan et al. 2015). IM also is considered as a pre-neoplastic stage, where the transformation of gastric mucosa occurred into an intestinal-like structure full of goblet cells and intestinal mucin. Studies have reported that overexpression of *CDX2*, a homeobox transcription factor is one of the responsible factors for intestinal metaplasia (Almeida et al. 2003). This stage has a high risk to develop Gastric cancer. Another phenomenon of GC development is spasmolytic peptide expressing metaplasia, it is a metaplastic change in gastric mucosa that used to occur due to injury in mucosa induced after *H. pylori* infection and chronic gastritis (Kusters et al. 2006). Chronic inflammation for a longer period of time can develop into GC by activating the NF-kB transcription factor, an important factor in the progression of the tumor (Karin et al. 2005). Due to chronic gastritis, leukocytes and macrophages produce reactive oxygen species and nitrosamines, respectively to increase the oxidative stress which can alter the proliferation of cells. Additionally, during this inflammation, chemokines and cytokines are also produced which may promote the risk of cancer along with leukocyte migration (Tan et al. 2015).

Etiologically, Gastric cancer is heterogeneous and develops due to a multitude of risk factors like environmental factors, *H. pylori* infection, diet, smoking, alcohol drinking and genomic as well as epigenetic alterations (Tan et al. 2015). Epstein Bar Virus (EBV) is associated with GC in 5-10% of cases (Shinozaki-Ushiku et al. 2015). Salt and salted food especially salted fish, cured meat, pickled vegetables, and salt-preserved foods are always the risk factors for atrophic gastritis (Tsugane et al. 2007). The mode of action of salt to develop GC may include: i) create suitable conditions for colonization of virulent form of *H. pylori* in the stomach, ii) can disturb the viscosity of stomach protecting mucus layer and as a result, the stomach gets exposed to carcinogens like N-nitroso compounds and iii) salt can cause inflammation in stomach epithelial cell, as a response it can increase proliferation of epithelial cells and some endogenous mutation can also arise (Ang et al. 2014, Tsugane et al. 2007, Wang et al. 2009). A diet containing less fruits and fresh vegetables is also a risk factor for Gastric cancer (Nemati et al. 2012) while consuming a sufficient amount of fruits and fresh vegetables can reduce the risk of gastric cancer (Denova-Gutierrez et al. 2014; Fang et al. 2015). Consuming smoked food in excess amounts is also a risk factor for GC (Wu et al. 2013; Hamidi et al. 2016). Tobacco consumption or smoking is also a primer factor for GC and the risk is more in smoker males than smoker females and non-smokers are at low risk (Ladeiras et al. 2008; Nishino et al. 2006). Studies have reported that former smokers are at low risk compared to occasional smokers and smokers with higher consumption of cigarettes (Nishino et al. 2006). In some studies, alcohol consumption is also a factor for developing GC (Phukan et al. 2005; Steevens et al. 2010; Verma et al. 2012). In India, some studies have reported that higher intake of rice, salt, preserved food, pickled food, spicy food, consumption of extra chilly, warmed foods, smoked meat, salted fish and using soda as a food ingredient are positively associated with the increased risk for GC (Mathew et al. 2000; Rao et al. 2002; Wang et al. 2009).

Mizo people have their unique food (smoked and fermented) and smokeless tobacco (tuibur) habits which can enhance the risk for Gastric Cancer development. Saum is fermented pork fat and was reported as a habitat for pathogens that may affect human health (De Mandal et al. 2018). Tuibur is tobacco-infused water which is alkaline in nature and used to contain polyaromatic hydrocarbons and carbonyl compounds (Lalruatfela et al. 2017; Madathi et al. 2018). Tuibur is a risk factor for Gastric Cancer in many studies (Phukan et al. 2005; Mukherjee et al. 2020). People belonging to the lower socio-economic group are more prone to GC (Sitraz et al. 2018).

Helicobacter pylori (H. pylori), a class I carcinogen has been considered as an environmental risk factor for GC (IARC 1994; Lu and Li, 2014). Studies have proved extra salt consumption along with *H. pylori* make the condition favorable for GC (Loh et al. 2007; Zhang et al. 2017). *H. pylori* is a common flora as about 50% of the population are infected, but only 1- 2% develop Gastric Cancer (Peleteiro et al. 2014). *H. pylori* have two virulence factors: cytotoxin-associated gene A (CagA) and vacuolating cytotoxin A (VacA). It has been experimentally proved that virulence factor cytotoxin-associated gene A within the bacterial cag pathogenicity island (cagPAI) is associated to develop Gastric Cancer (Odenbreit et al. 2000; Wroblewski et al. 2013; Peleteiro et al. 2014). Asian *H. pylori* strains have genetic alteration in the CagA gene which is also found associated with chronic gastritis and Gastric adenocarcinoma in humans (Higashi et al. 2002). This CagA gene in the epithelial cell goes through tyrosine phosphorylation to activate SHP-2 (Src homology 2-containing tyrosine phosphatase) by Src kinase and which is an inducer of the Ras-ERK pathway (Higashi et al. 2002). CagA can also interact with Met tyrosine kinase and E-cadherin which can interrupt the binding between E-cadherin and β-catenin which inturn activated β-catenin-dependent transcription (Murata-Kamiya et al. 2007). Many studies have reported that eradication of *H. pylori* can reduce the risk of GC (Wong et al. 2004; Fuccio et al. 2009).

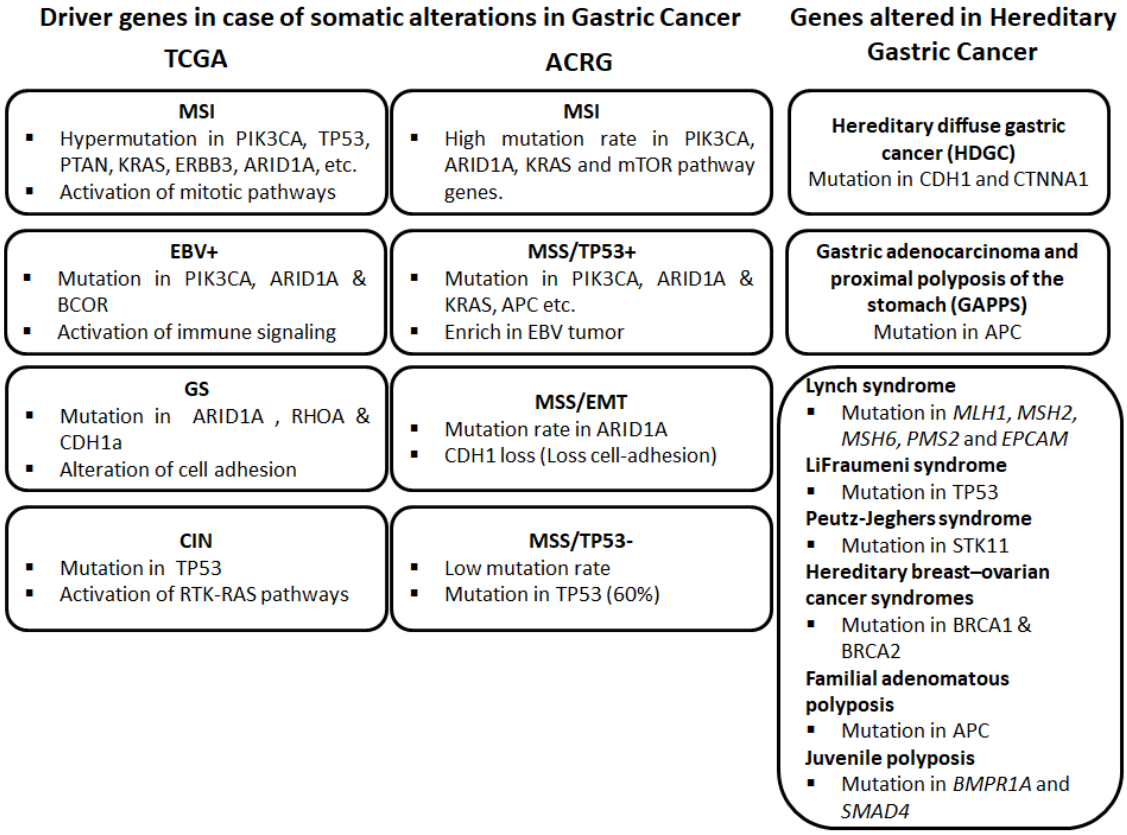
Epstein Bar Virus is an important risk factor for Cancers (Shannon-Lowe et al. 2019), especially with the diffuse type of GC (10%) (Carrasco-Avino et al. 2017). Epstein Bar Virus (EBV) infection is very common worldwide (90%), but very rarely infects epithelial cells (Smatti et al. 2018). The statistics of EBV-associated gastric cancer incidence based on the geographical area are as follows: Germany (17%) followed by Poland (12.5%), Russia (8.1%), France (7.7%) and China (4%) (Sitarz et al. 2018). According to TCGA based on gene expression, EBV positive cases are a subclass of Gastric cancer along with Microsatellite Instability MSI-H, genomically stable and Chromosomal instability (CIN) group (Cancer Genome Atlas Research Network 2014). EBV-associated GC is more prevalent in males in the age group of 50-68 years (Lee et al. 2009; Truong et al. 2009; Murphy et al. 2009). EBV-infected cases were frequently found in the cardia and the body of the stomach (Murphy et al. 2009). EBV infected GCs, methylation of CpG Island in cancer-associated genes is a prime feature of this subgroup due to the expression of viral protein LMP2A (Kaneda et al. 2012). EBV encoded microRNAs are mostly found in the BHRF1 and BART regions of the EBV genome. miRNA can suppress the expression of some cellular protein and miR-BART4-5p found to play a vital role in developing GC by regulating the apoptosis process (Pfeffer et al. 2004; Shinozaki-Ushiku et al. 2015).

TCGA and Asian Cancer Research Group (ACRG) groups have categorized MSI as a molecular subtype of GC (The Cancer Genome Atlas Research Network, 2014; Cristescu et al. 2015). MSI-H cases have a major association with an overall survival rate of GC (Zhu et al. 2015). MSI-associated GC cases showed better prognosis and were used to respond to treatments (The Cancer Genome Atlas Research Network, 2014; Cristescu et al. 2015). MSI-associated GC was reported as a different category, which was significantly associated with aged patients, female patients, distal location of stomach and intestinal-type GC cases (Smyth et al. 2017; Kim et al. 2010; Kim et al. 2013; Polom et al. 2018). Studies have reported an 18.5% prevalence of MSI in Chinese GC patients and were higher in advanced stage GC like in Western countries (Hamada et al. 2019). It has been reported that smoking has a significant association with MSI-H cases in colorectal cancer (Carr et al. 2018).

Now a day's, next-generation sequencing (NGS) technology has become a revolutionary tool for cancer research, to detect disease-associated variants and to understand the pathways mechanism for developing the disease. Presently, NGS technology has become less expensive and useful compared to Sanger sequencing. Whole exome sequencing (WES), and deep targeted re-sequencing methods can explore recurrent, unique, driver and passenger variants at high-depth for a large number of samples. Several studies have reported several driver genes and alterations related to Gastric Cancer development by high throughput next-generation sequencing (NGS) approach (Yamamoto et al. 2014). NGS is a systematic novel approach to identify disease-related gene alterations in GC which can help the researchers to reveal and understanding the pathogenesis of GC and identifying new genes for therapeutic targets. We can develop useful biomarkers to detect Gastric Cancer in an early stage by this promising approach of NGS technology. The other NGS approach like whole-genome sequencing and whole-transcriptome sequencing are also able to find out new driver genes and alterations related to the disease. These new high throughput advanced technologies pave way for molecular research in relation to GC for taking new measures in diagnostic and therapeutic fields by transforming the genomic data to clinical fields.

Hereditary GC can be related to three main syndromes: hereditary diffuse gastric cancer (HDGC), gastric adenocarcinoma and proximal polyposis of the stomach (GAPPS), and familial intestinal gastric cancer (FIGC) (Lauren et al. 1965). Out of these three syndromes, germline *CDH1* mutations are related with HDGC and *CTNNA1 is* also reported as significant for this type. Germline mutation of the *APC* gene has an association with GAPPS and molecular characterization of FIGC syndromes is not yet understood properly. Other cancer-associated syndromes which are associated with GC are as follows: Lynch (genes related to this syndrome are *MLH1*, *MS2*, *MSH6*, *PMS2* and *EPCAM*), LiFraumeni (TP53), Peutz-Jeghers (*STK11*), hereditary breast-ovarian cancer syndromes (*BRCA1* and *BRCA2*), familial adenomatous polyposis (*APC*), and juvenile polyposis (*BMPR1A* and *SMAD4*) (Lauren et al. 1965; Katona et al. 2017) (Figure 2).

Till date, only about 10-15% of familial gastric cancer cases have been reported compared to sporadic cases in different studies and first degree relatives with a history of Gastric cancer have two-timed higher risk of GC (Tramacere et al. 2012; Tsugane et al. 2007; Caldas et al. 1999). 40% of the familial gastric cancer (FGC) and 23-30% HDGC were accounted pathogenic mutation in E-cadherin gene, *CDH1*and the effect of this gene on disease progression is well studied (Brooks-Wilson et al. 2004; Figueiredo et al. 2013; Blair et al. 2020). Some studies have reported that germline mutation of mitogen-activated protein kinase kinase kinase 6 (*MAP3K6*) is associated with FGC cases, as there was no coding mutation recorded in *CDH1* (Gaston et al. 2014). Studies have reported germline mutation of *MLL3* associated with Lynch Syndrome and increased risk of colorectal cancer (Villacis et al. 2016). Nowadays, studies have reported that *BRCA1/2* mutations are also gaining importance in stomach Cancer, with a 6 fold higher risk with first-degree relatives of *BRCA1/2* mutation carriers (Cavanagh et al. 2015).



Driver genes in case of somatic alterations in Gastric Cancer

TCGA

MSI
- Hypermutation in PIK3CA, TP53, PTAN, KRAS, ERBB3, ARID1A, etc.
- Activation of mitotic pathways

EBV+
- Mutation in PIK3CA, ARID1A & BCOR
- Activation of immune signaling

GS
- Mutation in ARID1A , RHOA & CDH1a
- Alteration of cell adhesion

CIN
- Mutation in TP53
- Activation of RTK-RAS pathways

ACRG

MSI
- High mutation rate in PIK3CA, ARID1A, KRAS and mTOR pathway genes.

MSS/TP53+
- Mutation in PIK3CA, ARID1A & KRAS, APC etc.
- Enrich in EBV tumor

MSS/EMT
- Mutation rate in ARID1A
- CDH1 loss (Loss cell-adhesion)

MSS/TP53-
- Low mutation rate
- Mutation in TP53 (60%)

Genes altered in Hereditary Gastric Cancer

Hereditary diffuse gastric cancer (HDGC)
Mutation in CDH1 and CTNNA1

Gastric adenocarcinoma and proximal polyposis of the stomach (GAPPS)
Mutation in APC

Lynch syndrome
- Mutation in *MLH1, MSH2, MSH6, PMS2* and *EPCAM*

LiFraumeni syndrome
- Mutation in TP53

Peutz-Jeghers syndrome
- Mutation in STK11

Hereditary breast–ovarian cancer syndromes
- Mutation in BRCA1 & BRCA2

Familial adenomatous polyposis
- Mutation in APC

Juvenile polyposis
- Mutation in *BMPR1A* and *SMAD4*

Figure 2: List of driver genes in different groups of Gastric Cancer

TCGA and ACRG identified distinct molecular subtypes of Gastric Cancer by NGS technology. TCGA study has described four distinct molecular classifications associated with GC. In the case of the MSI subgroup: *PIK3CA, TP53, PTAN, KRAS, ERBB3* and *ARID1A* were the hyper-mutated genes. EBV (+) subgroup had significant mutations in *PIK3CA, ARID1A* and *BCOR*. The hypermutation was observed in *ARID1A, RHOA* and *CDH1*a genes in genomically stable (GS) cases. TP53 gene was mutated in case of chromosomal instability (CIN) cases (The Cancer Genome Atlas Research Network, 2014). ACRG group again reported four subgroups associated with GC. In the case of the MSI subgroup, the hypermutated genes were common like the TCGA study (*PIK3CA, KRAS* and *ARID1A*). *PIK3CA, KRAS, ARID1A* and *APC* were highly mutated in MSS/TP53^{+} subgroup. *ARID1A* gene was mutated in MSS/EMT subgroup and *TP53* was mutated in MSS/TP53^{-} subgroup (Cristescu et al. 2015) (Figure 2).

Several studies have reported that *TP53* and *CDH1* are the two driver genes for developing GC in most of the population (Bellini et al. 2012; Vander et al. 2015). Studies have confirmed that the mutations which were responsible for the loss of *TP53* function are the most pathogenic alterations and are associated with cancer development in GI Track including the stomach (Bellini et al. 2012; Oki et al. 2009). *TP53* was the top frequently mutated genes in both TCGA (50% in non-hyper-mutated cases and the case of CIN it was 71%) and ACRG (33%) studies of GC (The Cancer Genome Atlas Research Network, 2014; Cristescu et al. 2015). *CDH1* mutations are significantly associated with hereditary Gastric Cancer, mainly with diffuse-type (Vander et al. 2015; Boland et al. 2017) In the TCGA study, the CDH1 gene was mutated in 11% of GC cases and 9% of cases in the ACRG group (The Cancer Genome Atlas Research Network, 2014; Cristescu et al. 2015). Other driver genes like *BRCA2* mutations were also found in 9% of GC cases in the Chinese population (Chen et al. 2015) and *CTNNA1* was also found to be mutated in diffuse-type Gastric Cancer cases (Majewski et al. 2013). Some whole-exome sequencing (WES) studies have reported recurrent somatic mutations in the new driver gene-*ARID1A* (chromatin remodeling gene) and *FAT4* (cell adhesion gene). TCGA and other studies have reported mortality and cytoskeleton-related gene (*ROHA*) as a new driver gene with hot spot mutation in Gastric Cancer cases (The Cancer Genome Atlas Research Network, 2014; Kakiuchi et al. 2014). Studies have reported mortality and cytoskeleton-related gene (*MACF1*) mutations associated with GC cases by whole-exome sequencing studies (The Cancer Genome Atlas Research Network, 2014). Several studies have reported the Wnt signaling pathway for developing different types of Cancer along with GC. The genes of the Wnt signaling pathway like *APC, CTNNB1* and *RNF43* were mutated in the case of GC in TCGA and ACRG study (The Cancer Genome Atlas Research Network, 2014; Koushyar et al. 2020; Oliveira et al. 2015). Next-generation sequencing has reported mutations in the RTK pathway in relation to GC. *PIK3CA, KRAS, ERBB2, ERBB3*, *ERBB4* and *EGFR* were the most mutated gene of the RTK pathway in relation to GC (The Cancer Genome Atlas Research Network, 2014; Cristescu et al. 2015). TCGA study has reported that *PIK3CA* has mutated in 80% of the EBV (+) cases and 42% of the MSI cases. Studies have reported that *KRAS* was mutated in 9% of the GC cases

(Deng et al. 2012). The genes of the ERBB family (*EGFR, ERBB2, ERBB3,* and *ERBB4*) were mutated in less frequency compared to *PIK3CA* and *KRAS* genes. In the case of copy number variation, the ERBB2 gene showed a significant result with GC in different studies (Pan et al. 2018). TCGA study has reported some genes which were mutated in less than 20% of cases and they are as follows: *FBXW7, VPS13A, BCORL1, CIC, ERBB3, ZBTB20*, and *HDAC4* (The Cancer Genome Atlas Research Network, 2014). Another two genes: *SMAD4* and *MUC6* were also mutated in GC cases in TCGA and ACRG studies.

Recently, NGS studies have reported novel 7 genes associated with GC (*FBXW7*, *XIRP2, NBEA, COL14A1, CNBD1, AKAP6*, and *ITGAV*) (Li et al. 2016). Mutations in chromatin remodeling genes (*ARID1A, MLL3* and *MLL)* have been found in 47% of GCs (Watson et al. 2013). Studies have reported TP53 as the most mutated gene along with *ERBB2, FBXW7* and *MLL3* followed by *MTOR, NOTCH1, PIK3CA, KRAS, ERBB4* and *EGFR* (Pan et al. 2018). Multiple pathways are involved in developing GC. Studies have reported that alteration in the function of several genes and pathways playing an important role in developing gastric adenocarcinoma. GC is used to develop for abnormalities in developmental pathways like Wnt/β-catenin signaling, Hedgehog signaling, Hippo pathway, Notch signaling, along with nuclear factor-kB, and epidermal growth factor receptor (Molaei et al. 2017).

In the case of genetic testing of GC, The University of Chicago has customized a panel for Hereditary Gastric Cancer detection and registered for lab testing (GTR Lab ID: 1238) in NCBI (Guilford et al. 1998; Chompret et al. 2004; Fitzgerald et al. 2010; Veiga et al. 2010; Kluijt et al. 2012; Chun et al. 2012). This panel contains 19 genes related to hereditary Gastric Cancer syndrome and other syndromes which are associated with Gastric Cancer development (Table 1). To date, there is no somatic gene panel customized for Gastric cancer study.

APC	**BMPR1A**	**CDH1**
CTNNA1	**EPCAM**	**MSH2**
MSH6	**NF1**	**PDGFRA**
PMS2	**SDHA**	**SDHB**
SDHC	**SDHD**	**SMAD4**
STK11	**TP53**	

Table 1: Gene Panel of Hereditary Gastric Cancer detection

Immunohistochemistry (IHC) is a staining method of Formalin-fixed Paraffin-embedded (FFPE) tissues extensively used in Pathology lab for obtaining histological information from cancer tissue. IHC helps to get into deep in the area of tumor classification, pathology, multi-lineage expression, pathogenic infection status and disease progression. Further for developing biomarkers, IHC is a commonly used technique by which the behavior and progression of the tumor can be understood easily, which in turn can provide information on the biological behavior and prognosis of a tumor. The biomarkers developed by IHC are commonly used in routine screening of cancer tissues for several institutional review board protocols, for getting accurate expression information to find out the druggable target for better therapeutic treatment. Therefore, IHC is the bridge of information between surgical and molecular pathology and the mediator to transform our basic knowledge of science into the clinical field for the development of new drugs (Machado et al. 1996).

Till date, studies have reported a wide range of immunohistochemical markers for GC in relation to disease prognosis (*HER2, VGEF, hERG1, KLF5, CA IX, PKP3, MMP2, HDAC*, *BCL-6*, E-cadherin, *COX-2, TSP-1* and *BAX*), therapeutic response (*HER2* and *VGEF*), histological subtypes of GC (*HER2, VGEF* and E-cadherin), tumor progression (*VEGF)*, stage (*KLF5, PKP3, SATB1* and TGF *β*), grade (*KLF5* and E-cadherin), lymph node metastasis (*HER2, KLF5, CA IX, Ki67*, *BCL-2, SATB1* and *c-myc2)* and invasion (E-cadherin) (Garcia et al. 2003; Tanner 2005; Cutsem et al. 2009; Yuan et al. 2009; Zhou et al. 2010; Mutze et al. 2010;

Cheng et al.2010; Kato et al. 2010; Schimanski et al. 2011; Demirag et al. 2011; Gou et al. 2011; Ananiev et al. 2011; Nakao et al.2011; Liu et al. 2011; Li et al. 2012; Xu et al. 2012).

Studies have reported *HER2/ERBB2* as a biomarker for detection of Gastric cancer along with *VEGF* and *HERG1* and *EGFR* expression was higher in intestinal type of GC (Lastraioli et al. 2012; Birkman et al. 2018). Studies have reported *HER2* (the oncogene) as a single predictive biomarker in the case of target therapy of Gastric Cancer patients (Brickman et al. 2017). Studies have reported that *HER2* is a prime regulator for the development of tumors in Breast cancer (Holbro et al. 2003). In the case of GC, *HER2* plays a role of oncogene and the overexpression of this gene is linked to poor prognosis, more aggressive tumors and also responsible for poor survival (Allagyer et al. 2000; Giuffre et al. 2001; Park et al. 2006; Zhang et al. 2009; Gravalos et al. 2008). The anti-human epidermal growth receptor 2 (*HER2*) is an important gene of the RTK pathway which we can targetable for personalized therapy and trastuzumab therapy became successful against *HER2* (+) Gastric Cancer cases (Bang et al. 2010). *TP53* is a tumor suppressor gene and is used to regulate the cell cycle and imitate carcinogenesis. Studies have reported that *TP53* is a prognostic marker of GC and has significantly high expression in intestinal type of GC. In the case of mucinous and poorly differentiated GC cases, *TP53* expression was higher (Lazar et al. 2010). Alterations of the *TP53* gene are very common in all types of cancer including GC but it may play a role to develop the intestinal type of GC (Fukunaga et al. 2016). Several studies have reported that deficiency in the expression of *ERCC1*, a DNA repair protein is a significant marker to predict clinical response and disease prognosis in a different type of cancer including GC (Qlaussen et al. 2006; Kim et al. 2008; Wang et al. 2014; Kwon et al. 2007). It has been reported that *H. pylori* can cause a deficiency in the expression of the *ERCC1* gene. Studies have reported that patients having a low expression of the *ERCC1* gene can be treated with platinum-based adjuvant chemotherapy (De Dosso et al. 2013). *BAX* is the apoptosis regulatory protein of the *BCL-2* family which is used to regulate the sensitivity of a cell towards apoptotic stimulation. Studies have reported that *BAX* expression is used to correlate significantly with histological type, anatomical site and lymph node metastasis of GC cases (Liu et al. 2011). Low expression of *BAX* can predict independently the poor survival of patients (Wang et al. 2019).

In this present study, we hypothesized that a high incidence of Gastric Cancer in the Mizo population might be due to the effect of environmental risk exposure including unique dietary habits and lifestyle factors along with pathogen (*H. pylori* and EBV) infection. Furthermore, as the Mizo tribal population is homogeneous, a unique set of driver genes with pathogenic alterations may play a role to initiate the progression of Gastric cancer. Very few studies have been reported from India to understand Gastric Cancer genomics and from Mizoram as well. In this study, samples from the Mizo ethnic group were collected and screened for Pathogen detection as well as for microsatellite instability. The epidemiological information to analyze the major risk factors for GC was also collected. Targeted re-sequencing was performed with a panel of driver genes to identify the frequently mutated genes and alterations that might play an important role to develop GC. Whole Exome Sequencing was performed to identify the novel driver genes related to GC in this population. This study is important to find out the significant etiological factors and prediction of driver gene alterations related to GC in this population to find out whether "the population is genetically predisposed with pathogenic mutation related to GC?". The results obtained from this study can be translated to the clinical field for therapeutic improvement for this high-risk Gastric Cancer population.

Objectives

- To find out the association of *Helicobacter pylori* and Epstein – Barr virus genotypes with Microsatellite Instability status in gastric cancer cases.
- To identify the driver gene mutations for gastric cancer and their association with demographic and clinicopathological features.
- To predict the prognostic Immunohistochemical markers for detection of oncogenes associated with Gastric Cancer.

Materials and Methods

Sample Description

The ethical committees of Civil Hospital, Aizawl (B.12018/1/13-CH(A)/IEC dtd. 18/04/2014), and Human Ethical Committee, Mizoram University (MZU/IHEC/2015/008 dtd. 14/12/15) approved the study. Samples from 80 patients were collected from four different hospitals: Civil Hospital Aizawl, Ebenezer Hospital, Aizawl Hospital and Green Wood Hospital, Aizawl from September 2016 to January 2019. Fresh gastric tumor and adjacent normal tissue for each patient were collected in the PBS buffer solution and stored in -80ºC freezer. Peripheral blood samples for each patient and healthy controls were collected in 3 ml in EDTA Vial and stored in -80ºC freezer. Formalin-fixed paraffin-embedded tissue blocks (Tumor and Adjacent Normal) were collected for each patient and stored at room temperature.

This study was conducted to find out significant risk exposure for GC patients in Aizawl, Mizoram, Northeast India by comparing demographic and epidemiological data between GC patients and healthy controls. Controls and cases were randomly selected at a 2:1 ratio in respect of their age and gender. Eighty patients (53 males and 27 females) were included in this study after conforming histologically as a case of stomach adenocarcinoma by pathologists. The samples were confirmed histologically. Pathological identification was done by (i) Microscopic examination of H&E stain tissue (ii) Histological Grading & typing (if applicable) and (iii) TNM staging (Tumor/ Node/ Metastasis) were done if a suitable specimen is available for examination. Tissue with 80% tumor cell was included in the present study after confirmed histological review. The age range of patients was from 31 to 86 (60.11 ± 11.40) years. A total of 160 controls (79 males and 81 females) were randomly selected from the same ethnic group from where the patients were selected and belong with an almost similar age range from 31 to 85 (57.96 ± 11.48) years.

The inclusion and exclusion criteria for selecting patient samples to conduct the study were:

Subject inclusion criteria

- Patients with Gastric Cancer and without any pre-treatment for cancer were included.
- Cases clinically diagnosed by an oncologist and confirmed by a pathologist.
- Samples were collected only from Mizo ethnic tribe.

Subject exclusion criteria

- Gastric cancer patients with other chronic diseases were excluded.
- Patients who were pre-treated for any other type of cancer were excluded.

Data collection

A well-designed and informative questionnaire was collected from each participant with a duly informed consent form. The patient group and healthy controls were interviewed by a telephonic interview for the follow-up study. In the questionnaire, the Lifestyle habits were categorized as follows: a) smoking, categorized as smokers (who used to smoke at least once a week for three months or more) and non-smokers (if the person never smoked before or left smoking for more than 5 years); b) chewing tobacco in smokeless form, categorized as a consumer (who used to take at least once a week for six months or more) and non-consumer (if the person never consumes before or left more than 5 years before); c) tuibur or tobacco infused water, categorized as drinkers (if the person used to drink at least once in a week) and non-drinkers (if the person never drinks); and d) alcohol, categorized as drinkers (if the person used to drink at least one day in a week) and non-drinkers (if the person never drink). The questionnaire also had detailed information on food habits such as a) extra salt intake, categorized as consumers (if the person takes extra salt at least once in their meal in a week) and as non-consumers (if the person never takes extra salt with their daily food for once); b) smoked food, categorized as consumers (if the person ate at least for one day in a week) and as non-consumers (if the person did not eat even for a single day in a week); and c) sa-um or fermented pork fat, categorized as consumers (if the person ate at least for once in a week) and as non-consumers (if the person did not eat even for once in a week). The excess body weight [body mass index (BMI) ≥25] was categorized as obese.

DNA isolation from Tumor Tissue and Blood samples

Genomic DNA was extracted from the tissue using a commercially available QIAamp® DNA Tissue Kit and DNA was extracted from blood samples using a commercially available QIAamp® Blood DNA mini kit. Genomic DNA from tissue and blood was also isolated by a conventional method using the phenol-chloroform method according to Ghatak et al. (2013). The conventional protocol was the same for extraction of DNA from blood and tissue, except in the case of blood sample the RBC was lysed in the first step. The hypotonic RBC lysis buffer (ammonium bicarbonate and ammonium chloride, Hi-media) was used to lyse RBC and to separate the WBC from the whole blood. The extraction buffer (10 mM Tris-HCL, 10 mM EDTA, 50 mM NaCl and 2% SDS) was used to extract nucleic acid. Separation of nucleic acids was followed as per the modified protocol of phenol-chloroform extraction (Ghatak et al. 2013). Precipitation was done by chilled isopropanol and sodium acetate and elution was done by MilliQ water. The DNA was stored in a freezer for long-term storage.

DNA visualization was done in electrophoresis by using 0.8% agarose gel, 1X TAE and 10 mg/ml ethidium bromide and documentation was done on ChemiDoc (XRS$^+$) (BIORAD, USA). Quantification was done by using Picogreen dye in Qubit 2.0 Fluorimeter (Invitrogen). Invitrogen, Qubit dsDNA BR Assay Kit (Q32850) was used in this study. For Qubit quantification, 200 µl solution for each sample and two standard DNA (supplied by manufacturer by adding 199 µl of dsDNA Buffer and 1 µl of dye) were prepared. After that two µl sample DNA was mixed with 198 µl of the solution and 10 µl for standard DNA was mixed with 190 µl solution by gentle vortex followed by a spin. The reading was measured on a Qubit instrument.

Pathogen Genotyping

Detection of *Helicobacter pylori* in GC patients was by PCR amplification of specific 16SrRNA region and *UraC* gene. Genotyping of *H. pylori* was by PCR amplification of CagA and VacA genes. The detection and genotyping of *Epstein Barr Virus* (*EBV*) type1/ type 2 infections was determined by using a standard PCR assay of EBNA3C - Epstein–Barr virus nuclear antigen 3C gene using distinct primer sets according to Fassone et al. (2000). The PCR reaction volume of 10 µl contained: 1x PCR buffer with, 1 unit of Taq DNA Polymerase, 0.2 mM dNTPs (All from the Thermo Scientific, USA), and 0.2 picomol primer (Active Oligo-ILS, Bangalore, India). The reaction mixture (10 µl) was PCR amplified for initial denaturation at 95ºC for 5 min, followed by 35 cycles at 95°C for 1 min., n°C (depending on primer) for 40 s, 72°C for 40 sec/1 min followed by extension at 72°C for 5 min. (Table 2). *H. pylori* and EBV positive and negative control samples were used for confirmation in PCR assays.

Gene	Primer (5'to 3')	Product Size (bp)	Annealing Temperature	Annealing Time	Reference
16SrRNA	F- CTGGAGAGACTAAGCCCTCC R- ATTACTGACGCTGATTGTGC	109	60°C	40 sec	Ren et al. 2012
UraC	F- AAGCTTTTAGGGGTGTTAGGGGTTT R- AAGCTTACTTTCTAACACTAACGC	294	54°C		Ho et al. 2004
CagA	F – AATACACCAACGCCTCCAAG R- TTGTTGCCGCTTTTGCTCTC	340	55°C		
VacA	F- GAGCGAGCTATGGTTATGAC R- ACTCCAGCATTCATATAGA	500	53°C		Chisholm et al. 2001
EBNA 3C	F- AGAAGGGGAGCGTGTGTTGT R- GGCTCGTTTTTGACGTCGGC	Type I- 153 Type II- 246	59°C		Fassone et al. 2000

Table 2: Primer sequences for *H. pylori* and *EBV* genotyping

PCR amplification of microsatellite loci

The determination of MSI/MSS associated GC cases were carried out by allele comparison of the mononucleotide repeat markers BAT-25, BAT-26, and dinucleotide repeat markers D2S123, D17S250, D16S752, D16S265, D16S398, D16S496, D18S58, and D16S3057 (Suraweera et al. 2002; Sarrio et al. 2003; Losso et al. 2012; Pećina-Šlaus et al. 2017; Forster et al. 2018) in tumor and corresponding blood samples and also in healthy control blood samples (Table 3). The forward primers for the markers were labeled with fluorescent dye 6-FAM, VIC, NED, and PET. The PCR reaction volume of 10 µl contained: 1x PCR buffer, 1 unit of Taq DNA Polymerase, 0.2 mM dNTPs, and 0.15 Picomol primers (Thermo Scientific). PCR (Master cycler Eppendorf, nexus GX2) protocol included initial denaturation (95^{0}C for 10 min), followed by 35 cycles (94^{0}C for 1 min; 55^{0}C for 40 sec; 72^{0}C for 40 sec) and a final extension step (72^{0}C for 7 min) (Table 3).

Fragment Analysis

Fragment analysis was performed using the Automated ABI sequencer model 3500 Genetic Analyzer (Applied Biosystems, Singapore) to analyze the amplified loci. In brief, 8.7 µl deionized formamide was combined with 0.3 µl GeneScan^{Tm}-600 size standards (Applied Biosystems, V-2.0) and 1 µl PCR product in a Genetic Analyzer sample plate. The plate was uniformly sealed by septa, and mild vortexing was followed to mix well. The denaturation step was carried out at 90°C for 2 min, followed by incubation on ice, and mini-plate centrifugation for 1 min. The loci were predicted as MSI when there was an allele shift or (and) novel peaks; MSI or MMR deficient, if at least two or more than two markers were having instability and the instability was found only in BAT-25 /BAT-26 Maker. If instability was not present in any of the markers, then the sample was classified as MSS or MMR proficient (Warneke et al. 2013).

Microsatellite Marker Name	Primer Sequences	Dye 5’ labeled in Forward primer	Marker Size (bp)	Repeat	Annealing Temperature	Gene Name & Chromosome numbers
BAT25	F: 5’ -TCGCCTCCAAGAATGTAAGT - 3’ R: 5’ - TCTGCATTTTAACTATGGCTC - 3’	PET	110 - 133	(T)25	56℃	KIT proto-oncogene receptor tyrosine kinase (KIT) , chromosome 4
BAT26	F: 5’ - TGACTACTTTTGACTTCAGCC -3’ R: 5’ - AACCATTCAACATTTTTAACC C -3’	NED	95 - 120	(A)26	56℃	mutS homolog 2 (MSH2), chromosome 2
D2S123	F: 5’- AAACAGGATGCCTGCCTTTA - 3’ R: 5’ - GGACTTTCCACCTATGGGAC -3’	NED	194 - 230	(CA)29	59℃	DNA segment containing (CA) repeat, chromosome 2
D17S250	F: 5’ - GGAAGAATCAAATAGACAAT - 3’ R: 5’ - GCTGGCCATATATATATTTAAACC - 3’	VIC	140 - 170	(CA)19	52℃	DNA segment containing (CA) repeat, chromosome 17
D16S752	F: 5'-AATTGACGGTATATCTATCTGTCTG-3' R: 5'-GATTGGAGGAG GGTGATTCT-3'	6-FAM	92-126	(CTAT)11	57℃	CDH1, chromosome 16
D16S265	F: 5'-CCAGACATGGCAGTCTCTA-3' R: 5'-AGTCCTCTGTGCAC TTTGT-3'	VIC	95 - 115	(CA)21	58℃	CDH1, chromosome 16
D16S398	F: 5'-CTTGCTCTTTCTAAACTCCA-3' R: 5'-GAAACCAAGTGGGT TAGGTC-3'	PET	175 - 195	(CA)23	55.5℃	CDH1, chromosome 16
D16S496	F: 5’- GAAAGGCTACTTCATAGATGGCAAT-3’ R: 5’- ATAAGCCACTGCGCCCAT-3’	VIC	200 - 230	(T)13 and (CA)21	61℃	CDH1, chromosome 16
D18S58	F: 5’-GCTCCCGGCTGGTTTT-3’ R: 5’- GCAGGAAATCGCAGGAACTT -3’	6-FAM	140 - 155	(CA)18	60℃	DNA segment containing (CA) repeat, chromosome 18
D16S3057	F: 5’-CCTGTGTGTATAACTATGTCAAAAT-3’ R: 5’-GCCCTTGAAACTAGGCAATA-3’	6-FAM	190 - 207	(CG)19	57℃	DNA segment containing (CG) repeat, chromosome 17

Table 3: List of Microsatellite Markers used in the study

Targeted re-sequencing approach to finding out driver gene alterations

Forty-eight patients were selected for targeted re-sequencing based on the pathogen and MSI status. Among them, 42% (20) and 65% (31) of patients were found to be infected with *H. pylori* and EBV, respectively and 42% (20) of patients were Microsatellite Instable. Paired tumor and blood samples were used for sequencing (Figure 3).

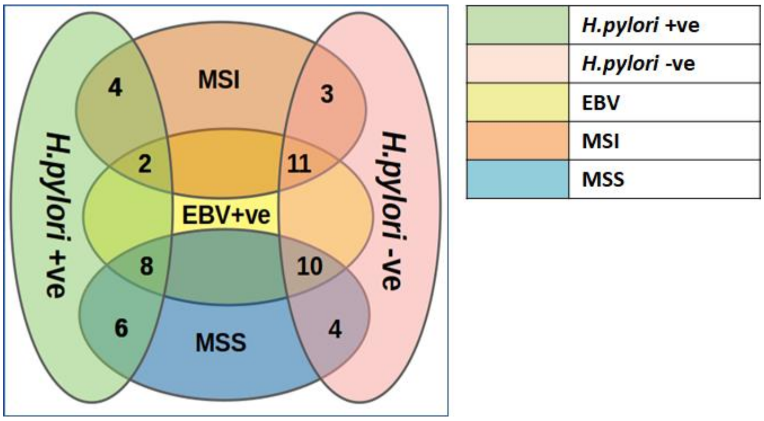
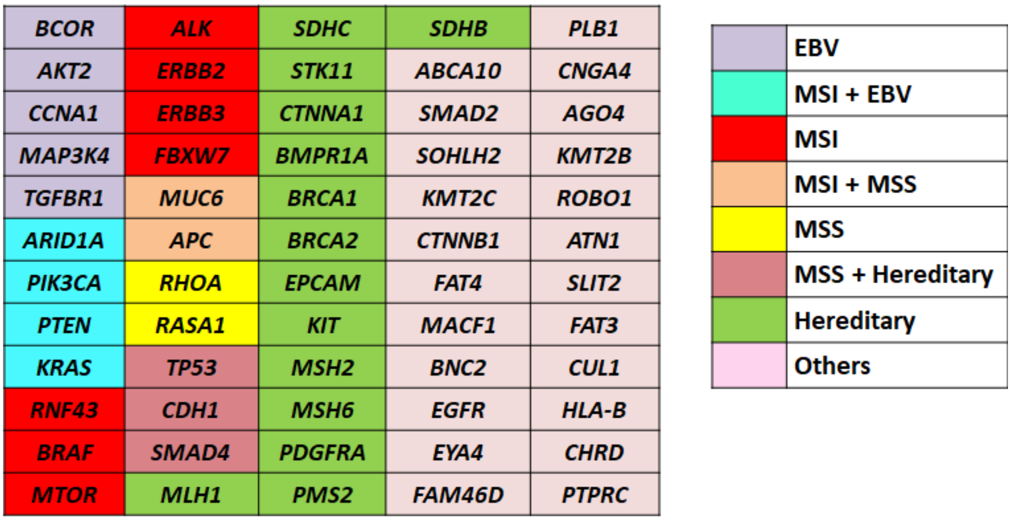


Figure 3: Characteristics of samples used for targeted re-sequencing.

A panel of 60 genes which used to play a driving role for developing GC in different populations and previously implicated in Gastric Cancer studies of TCGA, ACRG, and other large studies by reviewing journals (The Cancer Genome Atlas Research Network, 2014; Cristescu et al. 2015; Liang et al. 2014; Hereditary Gastric Cancer Panel, University of Chicago). A panel of 60 genes of 284.262 kb region size was designed with 401.060 kb probes size and 100% converge by Agilent SureSelect to cover the interesting region of panel genes (Figure 4).



BCOR	*ALK*	*SDHC*	*SDHB*	*PLB1*
AKT2	*ERBB2*	*STK11*	*ABCA10*	*CNGA4*
CCNA1	*ERBB3*	*CTNNA1*	*SMAD2*	*AGO4*
MAP3K4	*FBXW7*	*BMPR1A*	*SOHLH2*	*KMT2B*
TGFBR1	*MUC6*	*BRCA1*	*KMT2C*	*ROBO1*
ARID1A	*APC*	*BRCA2*	*CTNNB1*	*ATN1*
PIK3CA	*RHOA*	*EPCAM*	*FAT4*	*SLIT2*
PTEN	*RASA1*	*KIT*	*MACF1*	*FAT3*
KRAS	*TP53*	*MSH2*	*BNC2*	*CUL1*
RNF43	*CDH1*	*MSH6*	*EGFR*	*HLA-B*
BRAF	*SMAD4*	*PDGFRA*	*EYA4*	*CHRD*
MTOR	*MLH1*	*PMS2*	*FAM46D*	*PTPRC*

🟪	EBV
🟦	MSI + EBV
🟥	MSI
🟧	MSI + MSS
🟨	MSS
🟫	MSS + Hereditary
🟩	Hereditary
⬜	Others

Figure 4: Custom gene panel of 60 genes for a Mizo population study

Wet lab method of NGS sequencing

The method employed by the Agilent Sure Select™ Target Enrichment System extracts target regions from genomic libraries by hybridization to in-solution biotinylated cRNA probes, or "baits". This hybrid capture-based library preparation helps in the elimination of amplification and sequencing artifacts that limit the sensitivity of sequencing. The experiment was started with 10 to 200 ng gDNA diluted with 1X Low TE Buffer to make a final volume of 50 µl. The 50 µl gDNA samples were fragmented using the Covaris E220 instrument (Condition: Peak power – 175, Duty factor – 10, cycle/brust – 100, Duration – 200 second and bath temperature was 8^{0} C). End repair "A" tailing was done by sheared DNA end the mix of end repair buffer and end repair enzyme by following the program: 20°C for 15 minutes, 72°C for 15 minutes and 4°C hold in a PCR machine. Ligation with P5 index step was done by adding 25 µl of prepared Ligation master Mix (T4 DNA ligase), 5 µl of appropriate index P5 and 70 µl of end repair product by following the program: at 20°C for 30 minutes and followed by 4°C hold. Cleanup was done by AMPure beads and ethanol. Amplification of the adapter ligate library was followed by adding 13.5 µl of PCR master mix to 34.5 µl ligate library and 2 µl of appropriate P7 index in a touchdown PCR and the condition was 98°C for 2 minutes, (98°C for 30 seconds, 60°C for 30 minutes and 72°C for 1 minute) for 8 cycles, 72°C for 5 minutes and hold at 4ºC. Again purification was done by AMPure cleanup protocol. QA/QC was done by Tape station chip and the fragment size was 300 - 400 bp.

After the whole genome library was prepared, hybridization was done for capturing the interesting region. Hybridization methods required 500-1000 ng of prepared whole-genome DNA library in a volume of 12 µl (the maximum amount should be I ug). All the reagents, Blocker solution, RNase solution and capture library were thawed for preparing the hybridization mix. 05 µl Blocker solution was added first to the library and incubated at 95ºC for 5 min., 65ºC for 10 min and 65ºC for 1 minute (pause). During this pause, 13 µl of hybridization mix was added to each well of plate and incubation was followed by 65ºC for 1 minute, 37ºC for 3 seconds and hold at 65ºC for 30 cycles. After hybridization, capturing of the hybridized DNA was done by streptavidin magnetic beads (Dynabeads My One Streptavidin TI magnetic beads). Washing was followed by two wash buffers, to wash off the unbound beads. After washing, the remaining 25 µl library was again amplified with 25 µl Amplified master mix (Post captured master mix) for 12 cycles at 98ºC for 2 minutes, 98ºC for 30 seconds, 60ºC for 30 seconds, 72ºC for 1 minute followed by a final extension at 72ºC for 5 minutes and hold at 4ºC. Post captured PCR product was cleaned by AMPure cleanup method and 25 µl of nuclease-free water (Milli Q) was used to elute the final library. Final QA/QC was done by Tape station High Sensitivity chip and dilution was done for each sample to get a final product of 4 - 15 nM. Capture hybrids of this panel of genes and paired tumor and blood DNA samples from each patient were amplified, pooled and sequenced in HiSeq-2500 (Illumina). A mean coverage depth of 1000X was achieved for GC tumor DNA and 600X for matched normal blood cells. Data were analyzed for finding both somatic and germline variants.

Bioinformatics pipeline for analyzing somatic variants

The sequence reads obtained were mapped to hg19 reference sequence with BWA MEM aligner. Variant calling was done by 2 variant callers, VarScan2 (Koboldt et al. 2012) and Base by Base (BBB) in the house (NIBMG) developed pipelines (India Project Team of the International Cancer Genome Consortium. 2013). Both the vcf files were annotated by the CRAVAT annotation tool (Douville et al. 2013). Then, the union of coding variants of BbB and Verscan 2 was considered and three filters were applied: i) removal of somatic variants with VAFs ≤ 0.05 (Tumor) or ≥ 0.02 (Blood). ii) Selection of variants =< 0.01 allele frequency in 1000 genome database, and iii) exclusion of synonymous variants, respectively to get the discovery set (Figure 5).

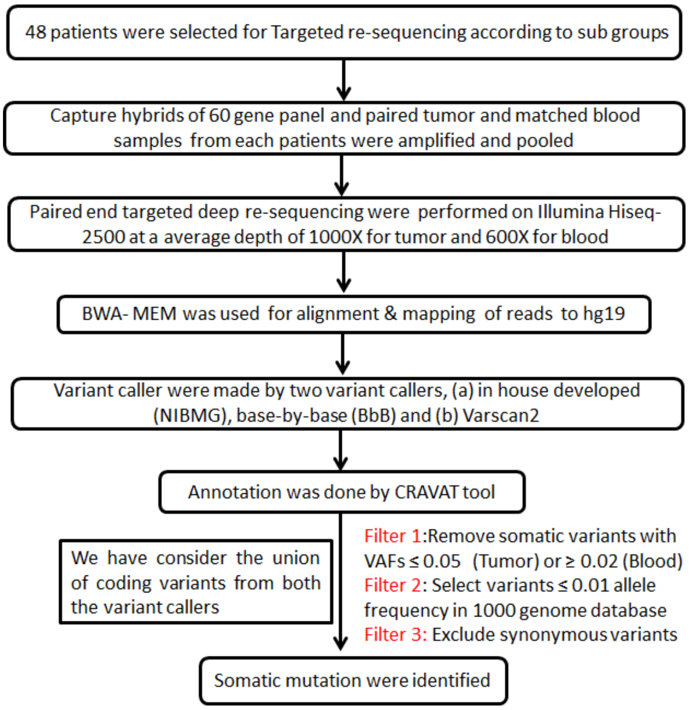

48 patients were selected for Targeted re-sequencing according to sub groups
Capture hybrids of 60 gene panel and paired tumor and matched blood samples from each patients were amplified and pooled
Paired end targeted deep re-sequencing were performed on Illumina Hiseq-2500 at a average depth of 1000X for tumor and 600X for blood
BWA- MEM was used for alignment & mapping of reads to hg19
Variant caller were made by two variant callers, (a) in house developed (NIBMG), base-by-base (BbB) and (b) Varscan2
Annotation was done by CRAVAT tool
We have consider the union of coding variants from both the variant callers
Filter 1:Remove somatic variants with VAFs ≤ 0.05 (Tumor) or ≥ 0.02 (Blood)
Filter 2: Select variants ≤ 0.01 allele frequency in 1000 genome database
Filter 3: Exclude synonymous variants
Somatic mutation were identified

Figure 5: Bioinformatics pipeline for identifying somatic variants

Bioinformatics pipeline for analyzing germline variants

The sequence reads obtained were mapped to hg19/GRCH37 reference genome using BWA-MEM. Sequence and variant calls were identified using GATK v3.8.0 suite’s Haplotype Caller and annotation was done by ANNOVAR database (Wang et al. 2010). After annotation, five filters were applied to get unique variants for the study population. First, only the exonic variants were selected and the Second filter was to discard off-target genes beyond the gene panel. Then, the third filter was to find out the unique variants (by excluding the common variants in other populations) by selecting variants with ≤ 0.01 allele frequency in 1000 Genomes. Then we have excluded the synonymous variants (Figure 6). Lastly, the variants which were present (mutation) in all the patients were excluded, as it is a germline analysis (Suzuki et al. 2020).

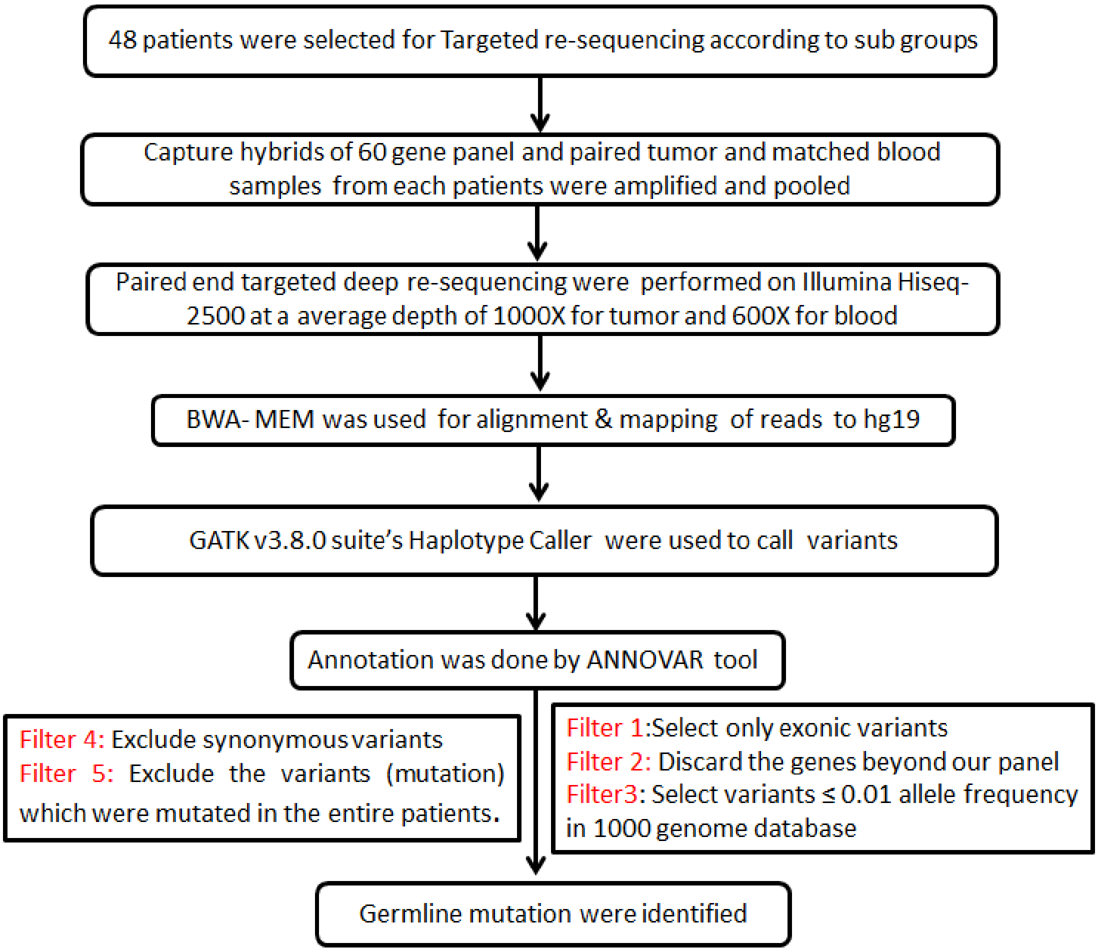

48 patients were selected for Targeted re-sequencing according to sub groups
Capture hybrids of 60 gene panel and paired tumor and matched blood samples from each patients were amplified and pooled
Paired end targeted deep re-sequencing were performed on Illumina Hiseq-2500 at a average depth of 1000X for tumor and 600X for blood
BWA- MEM was used for alignment & mapping of reads to hg19
GATK v3.8.0 suite's Haplotype Caller were used to call variants
Annotation was done by ANNOVAR tool
Filter 4: Exclude synonymous variants
Filter 5: Exclude the variants (mutation) which were mutated in the entire patients.
Filter 1:Select only exonic variants
Filter 2: Discard the genes beyond our panel
Filter3: Select variants ≤ 0.01 allele frequency in 1000 genome database
Germline mutation were identified

Figure 6: Bioinformatics pipeline for identifying germline variants

Whole Exome Sequencing (WES)

Whole exome sequencing was done for 37 patient samples and 4 healthy controls. Seventeen samples were taken from the previous batch of targeted re-sequencing and 20 new samples were selected for this analysis. Paired-end sequencing was performed for matched blood and tumor samples on Illumina Hiseq-2500 at an average depth of 90 X. BWA-MEM was used for alignment and mapping of reads with hg19 reference genome. GATK v3.8.0 suite's Haplotype Caller was used for variant calling (Poplin et al. 2017). The variants were annotated by the ANNOVAR tool (Wang et al. 2010). After annotation, five filters were applied to get unique variants for the study population. Only the exonic variants were selected by applying the first filter. The second filter was to exclude the common variants in other populations by selecting variants with ≤ 0.01 allele frequencies in 1000 Genomes to find out the unique variants of the population. The synonymous variants and the variants which were present in healthy controls were excluded (Figure 7). Finally, the variants which were present (mutation) in all the patients were excluded as it is a germline analysis (Suzuki et al. 2020).

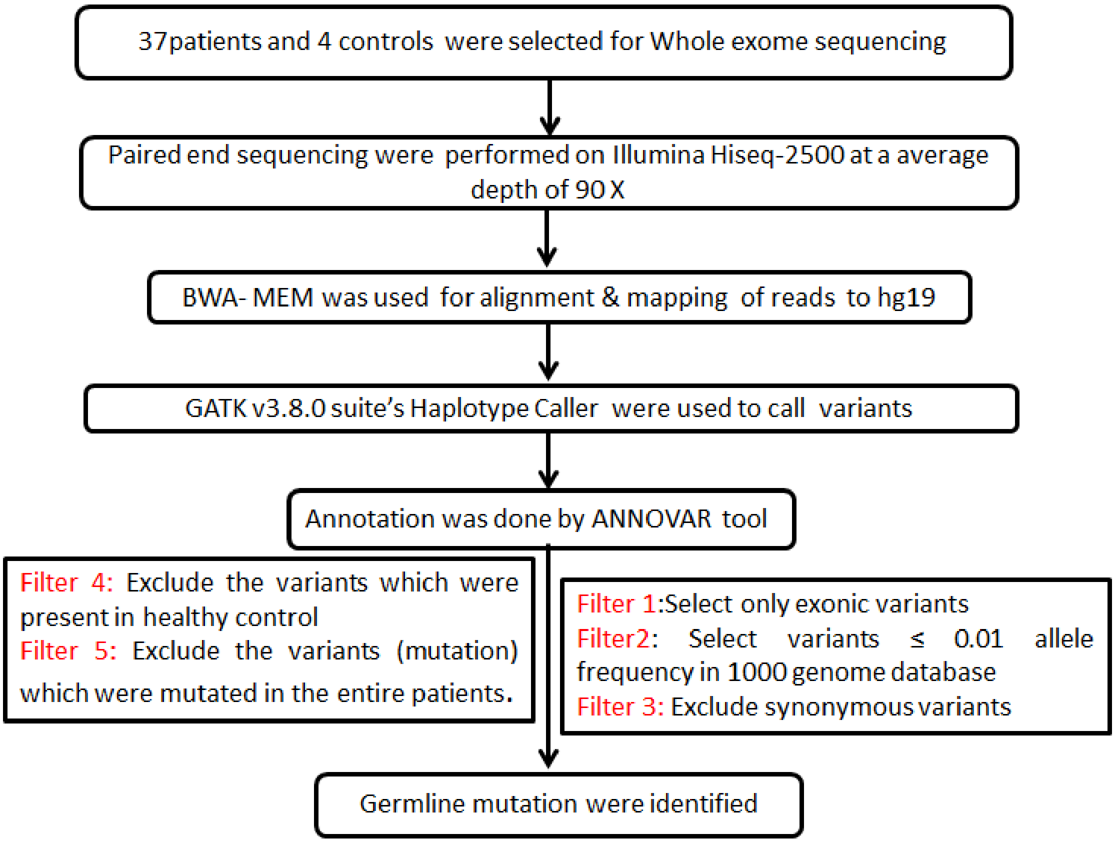

37patients and 4 controls were selected for Whole exome sequencing
Paired end sequencing were performed on Illumina Hiseq-2500 at a average depth of 90 X
BWA- MEM was used for alignment & mapping of reads to hg19
GATK v3.8.0 suite's Haplotype Caller were used to call variants
Annotation was done by ANNOVAR tool
Filter 4: Exclude the variants which were present in healthy control
Filter 5: Exclude the variants (mutation) which were mutated in the entire patients.
Filter 1:Select only exonic variants
Filter2: Select variants ≤ 0.01 allele frequency in 1000 genome database
Filter 3: Exclude synonymous variants
Germline mutation were identified

Figure 7: Bioinformatics pipeline for identifying germline variants from WES

Pathogenicity prediction

Prediction of pathogenicity of known variants was done by ClinVar (Landrum et al. 2014) and COSMIC database (Forbes et al. 2008). Prediction of novel missense variants was done by Mutation taster (Schwarz et al. 2014), Polyphen

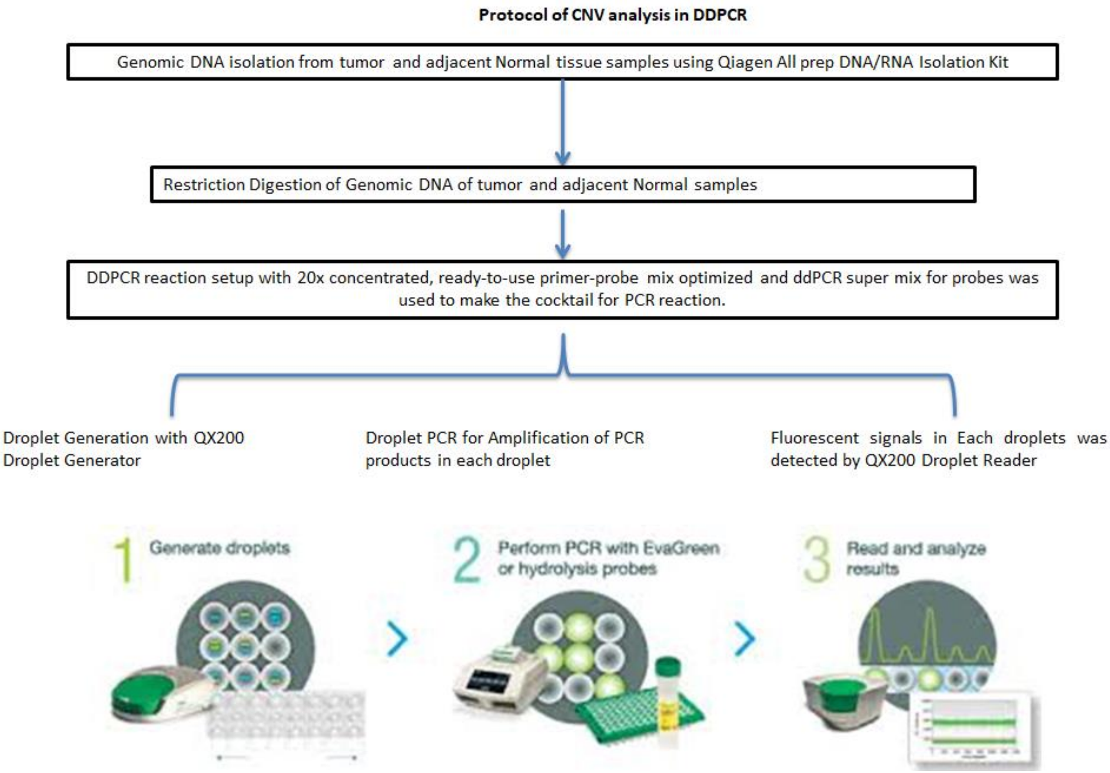
2 (Adzhubei et al. 2010), PROVEAN (Choi et al. 2012), and PANTHER (Thomas et al. 2003). Variants were classified as i) pathogenic and ii) benign.

Copy Number Variation Analysis

The copy number variation (**CNV**) is defined as the variation in the number of copies of a particular gene from one individual to the other. As every gene has two copies, there will be a change in copy number if there is a duplication or deletion. Seventeen (17) samples were selected for this analysis on the basis of mutation data derived from targeted re-sequencing using Droplet Digital PCR (ddPCR). Two genes were targeted, ERBB2 (Oncogene) and TP53 (Tumor suppressor gene), to compare the copy number status with mutated patient samples and EFTUD2 was used as a reference gene (Table 4). Genomic DNA from tumor and adjacent normal tissue samples were taken and restriction digestion was done by HindIII and BAMH1. ddPCR reaction setup was done with sample DNA (75 ng), primers (target and reference genes), ready-to-use primer-probe mix (20X) and super mix for probes were used to make the cocktail for PCR reaction (Figure 8). For each sample, the reaction was performed in triplet form along with triplet of NTC. Droplet Generation was done with QX200 Droplet Generator. Droplet PCR (C1000 Touch Thermal Cycler, BIORAD) was used for amplification of PCR products in each droplet and fluorescent signals in each droplet were detected with QX200 Droplet Reader.

Gene	**Primers**	**Length**	**GC content (%)**	**Melting Temp. (ºC)**
TP53	**Forward**- 5'- CACCAGCAGCTCCTACAC -3'	18	61.1	61.3
	Reverse - 5'- AAGAAGCCCAGACGGAAAC -3'	19	52.6	61.8
	Probe - 5'- CCCTGTCATCTTCTGTCCCTTCCC -3'	24	58.3	66.9
EFTUD2	**Forward**- 5'- CAGATGATGGAGTCCAGTTTCA -3'	22	45.5	61.7
	Reverse- 5'- GGTGCATATCTGGGAGTCTTC -3'	21	52.4	61.5
	Probe- 5'- TCATCCTCCAGGGTGTAGTTCTCCC -3'	25	56	67.6
HER2	**Forward**- 5'- CAACCAAGTGAGGCAGGTC -3'	19	57.9	62.2
	Reverse- 5'- AGCGGGTCTCCATTGTCTA -3'	19	52.6	62.1
	Probe- 5'- TAGTTGTCCTCAAAGAGCTGGGTGC -3'	25	52	67.4

Table 4: List of Primers used in ddPCR assay



Protocol of CNV analysis in DDPCR

Genomic DNA isolation from tumor and adjacent Normal tissue samples using Qiagen All prep DNA/RNA Isolation Kit

Restriction Digestion of Genomic DNA of tumor and adjacent Normal samples

DDPCR reaction setup with 20x concentrated, ready-to-use primer-probe mix optimized and ddPCR super mix for probes was used to make the cocktail for PCR reaction.

Droplet Generation with QX200 Droplet Generator

Droplet PCR for Amplification of PCR products in each droplet

Fluorescent signals in Each droplets was detected by QX200 Droplet Reader

1 Generate droplets
2 Perform PCR with EvaGreen or hydrolysis probes
3 Read and analyze results

Figure 8: Flowchart of ddPCR protocol

Protein Expression study using Immunohistochemistry (IHC)

This application was applied to see the expression of apoptotic and cell proliferating gene BAX (ab32503), TP53 (ab80645), ERBB2 (D8F12) XP – 4290T and ERCC1 (D6G6) XP – 12345T in tumor and adjacent normal tissues. TP53 was raised in mice, while BAX ERBB2 and ERCC1 were rabbit monoclonal antibodies. Two types of secondary antibody [Anti mouse, HRP linked Secondary antibody-7076P2 and HRP Rabbit (8114S), Cell Signaling] was used in this study Twenty four (24) patient samples were selected on the basis of clinical and mutation data and 3 µm sections were done by microtome (LEICA RM2125 RTS) and placed on the coated slide. Baking was done at 60ºC temperature for 45 min. – 1 hr. De-paraffinization was done by exposing the slides in three changes of xylene (5 min. for each change). Hydration was followed by de-centering order of graded alcohol (100% for three changes, 95% and 75% 1 min for each change) and three times washed in distilled water for 5 min, in each change. Antigen retrieval was done by citrate buffer (10 mM sodium citrate buffer, pH 6.0) at 98ºC for 25 minutes. Washing was done by TBST (Tris-Buffered Saline, 0.1% Tween 20 in 1 liter) for I min. Blocking was done by blocking solution (5% BSA in TBST) for 25 minutes. Incubation with primary antibody was followed overnight at 4ºC. The dilution used for primary antibody was as follows; BAX (1:250), TP53 (1:400), HER2/ERBB2 (1:200) and ERCC1 (1:125). Washing was done in TBST 3 times (10 min. each). Incubation with secondary antibody was done at room temperature for 30 minutes followed by washing with TBST for 3 times (10 min. each). After that, DAB chromogen solution was applied for 25-30 minutes followed by washing with TBST for 4 times (5 min. each). Counter-staining was done by hematoxylin followed by washing with distilled water for 1 minute. The dehydration process was done by dipping the slides in ascending order of graded alcohol (in 80% and 95% for 2 changes, in 100% for 3 changes) for 1 minute for each change. Again washing was performed by dipping the slides in 3 changes of xylene 1 minute each. At last, mounting was done by DPX with applying coverslip and observed under a microscope (Figure 9).

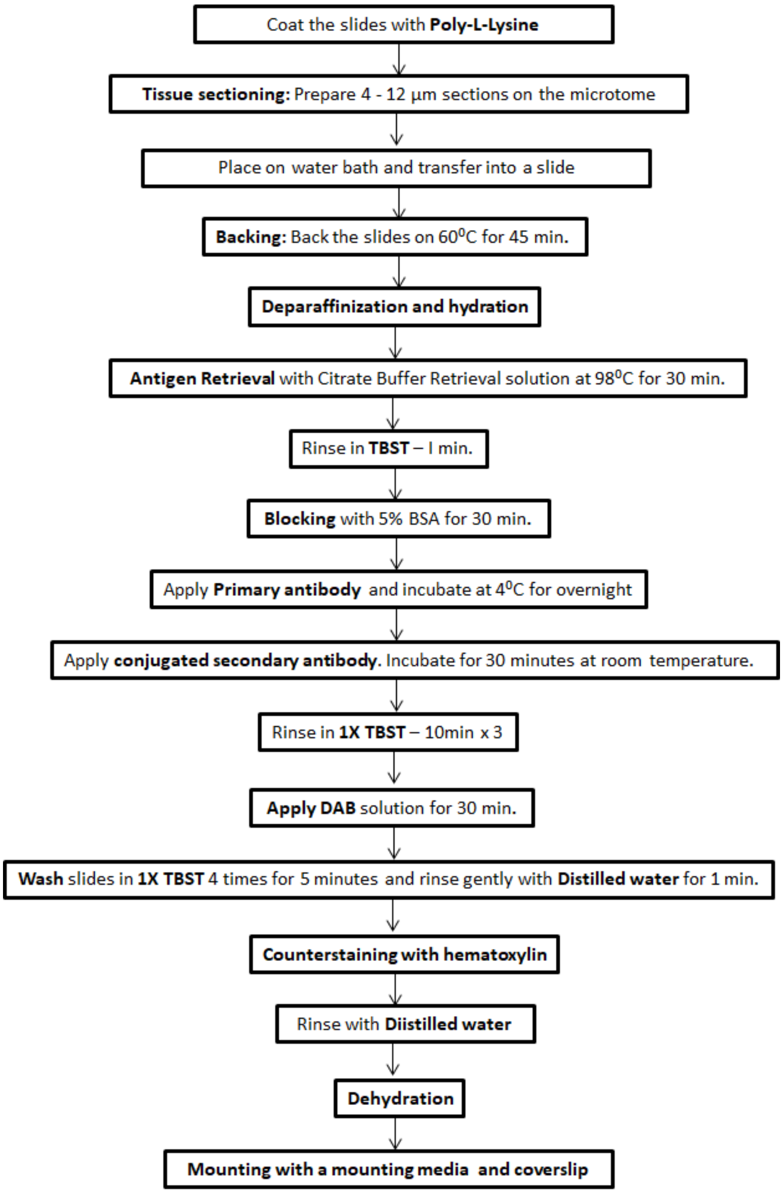

Coat the slides with Poly-L-Lysine
Tissue sectioning: Prepare 4 - 12 µm sections on the microtome
Place on water bath and transfer into a slide
Backing: Back the slides on 60⁰C for 45 min.
Deparaffinization and hydration
Antigen Retrieval with Citrate Buffer Retrieval solution at 98⁰C for 30 min.
Rinse in TBST – I min.
Blocking with 5% BSA for 30 min.
Apply Primary antibody and incubate at 4⁰C for overnight
Apply conjugated secondary antibody. Incubate for 30 minutes at room temperature.
Rinse in 1X TBST – 10min x 3
Apply DAB solution for 30 min.
Wash slides in 1X TBST 4 times for 5 minutes and rinse gently with Distilled water for 1 min.
Counterstaining with hematoxylin
Rinse with Diistilled water
Dehydration
Mounting with a mounting media and coverslip

Figure 9: Flowchart of IHC protocol

The interpretation of staining was done as positive expression and negative expression. Positive expression cases were further divided into low, moderated and high expression.

Statistical Analysis

The association of demographic factors among case-control subjects was tested for Hardy–Weinberg equilibrium by a chi-square test with one degree of freedom (df) (Gunathilake et. al. 2018). Non-parametric T-test was also performed. The odds ratio (OR) and 95% confidence intervals (CIs) were estimated for determining association in each group of factors among case-control subjects and among each subgroup [*H. pylori*, *EBV* infection and MMR deficient (MSI)/MMR proficient (MSS)] and factors by binary logistic regression (Univariate and Multivariate analysis) (Denis et al. 2018). The likelihood test was utilized to choose whether to hold each covariate in the model. Variables between groups of interest were compared using Pearson's $\chi 2$-test or Fisher's exact test for categorical variables, and Mann-Whitney U test or t-test for continuous variables. Then, the independent impact of hazard components was explored in a multivariate model (presenting all factors and terms of connections) keeping only those statistically significant or demonstrating a confounding effect on the contemplated elements.

Overall survival was calculated using the Cox proportional-hazards regression model (using three years cut-off). The log-rank test, Kaplan-Meier survival analysis was used to analyze the impact of the variables on survival rate (Moghimi-Dehkordi et al. 2009). To evaluate the association of different gene mutations and pathogen interaction with gastric cancer tissue with OS and DFS, univariate and multivariate Cox proportional hazards regression model were applied, and hazard ratios (HRs) together with 95% confidence intervals (CI) were calculated to determine the risk of death or cancer recurrence. The multivariate model was adjusted for established prognostic factors such as age, sex, tumor-node-metastasis (TNM) stage. All the patients with incomplete or missing cores were excluded from the analysis. The receiver-operating characteristic (ROC) curves were used to calculate the area under the curves (AUC) to determine the predicting ability of the final model compared to models with only one factor or the basic model. Statistical analyses were performed using SPSS version 23.0 (SPSS, IBM, Armonk, NY, USA) and GraphPad Prism version 8.0a (GraphPad Software, Inc., San Diego, CA, USA) and R packages. A two-sided P-value <0.05 was considered statistically significant.

Results

Epidemiological factors, Pathogen and Microsatellite status

The characteristics of GC patients of this cohort are represented in Table 5. The age range from 40-69 years exhibited the highest number of GC cases (75%), and male patients (66.25%) were more prone to GC than females. The family history (for any cancer type) in the first-degree relative was found in 32.5% of patients and the distal part of the stomach was reported to have the highest tumor cases (73.75%) (Table 5). Out of the total 80 GC patients, 50% of the cases were found in stage III, well-differentiated cases were found in 8.75% of the patient, 46.25% were moderately differentiated and poorly differentiated cases were found in 32.5% of the patient. In this study, most of the patients were in the advanced stage at the time of diagnosis.

Factors	**N (Total =80)**	**%**
Median Age ± SD	59.5 ± 11.40	
Age (years)		
<40	1	1.25
40-69	**60**	**75**
>69	19	23.75
Sex		
Male	**53**	**66.25**
Female	27	33.75
Family History of Cancers		
Yes	28	35
1st-degree relative	26	32.5
2nd-degree relative	2	2.5
No	52	65
Anatomy		
Distal	**59**	**73.75**
Proximal	11	13.75
Data Not available	10	12.5
Stage		
I	20	25
II	14	17.5
III	**40**	**50**
IV	2	2.5
Data Not Available	4	5
Differentiation		
Well Differentiated	7	8.75
Moderately Differentiated	**37**	**46.25**
Poorly Differentiated	26	32.5
Data Not available	4	5

Table 5: Characteristics of Gastric Cancer Patient samples

The distribution of demographic and lifestyle habits among GC patients and controls is presented in Table 6. Extra salt consumption was the highest significant risk factor (*p*-value < 0.0001) followed by smoked food consumption (*p*-value = 0.01), smoking (*p*-value < 0.0001) and alcohol drinking (*p*-value < 0.0001) and they are the high risk factors for developing GC. Other factors like Sa-um, Paan with betel nut, tobacco chewing and tuibur were not significantly associated with GC cases. Circos plot representing the frequency and association of demographic factors and lifestyle habits between GC patients and healthy control (HC) is given in Figure 10.

Factors	**[a]HC (n = 160)**	**[b]GC (n = 80)**	**[c]ORs (95% CI)[d]**	***p* value**
Age (Years ± SD)	57 ± 11.48	59.5 ± 11.40	-	-
Gender				
Male	79 (49.37%)	53 (66.25%)	-	-
Female	81 (50.62%)	27 (33.75%)		
Extra salt				
Consumers	150 (93.75%)	56 (70%)	**0.15 (0.07 – 0.34)**	**<0.0001**
Non-consumers	10 (6.25%)	24 (30%)		
Sa-um				
Consumers	132 (82.5%)	66 (82.5%)	1.00 (0.49 – 2.02)	1.00
Non- consumers	28 (17.5%)	14 (17.5%)		
Smoked food				
Consumers	126 (70%)	51 (63.75%)	**0.47 (0.26 – 0.85)**	**0.01**
Non-consumers	34 (30%)	29 (36.25%)		
Paan with betel nut				
Consumers	97 (60.62%)	50 (62.5%)	1.08 (0.62 – 1.88)	0.77
Non-consumers	63 (39.37%)	30 (37.5%)		
Chewed tobacco				
Consumers	63 (39.37%)	41 (51.25%)	1.61 (0.94 – 2.78)	0.08
Non- consumers	97 (60.62%)	39 (48.75%)		
Tuibur				
Consumers	27 (16.87%)	21 (26.25%)	1.45 (0.91 – 3.35)	0.08
Non- consumers	133 (83.12%)	59 (73.75%)		
Smoking				
Smokers	34 (21.25%)	52 (65%)	**6.88 (3.79 – 2.48)**	**<0.0001**
Non-smokers	126 (78.75%)	28 (35%)		
Alcohol drinking				
Drinkers	4 (2.5%)	29 (36.25%)	**22.17 (7.44- 66.10)**	**<0.0001**
Non-drinkers	156 (97.5%)	51 (63.75%)		

Table 6: Distribution of demographic and lifestyle habit factors among GC patient and Healthy controls (HC)

OR – Odd Ratio; p-value

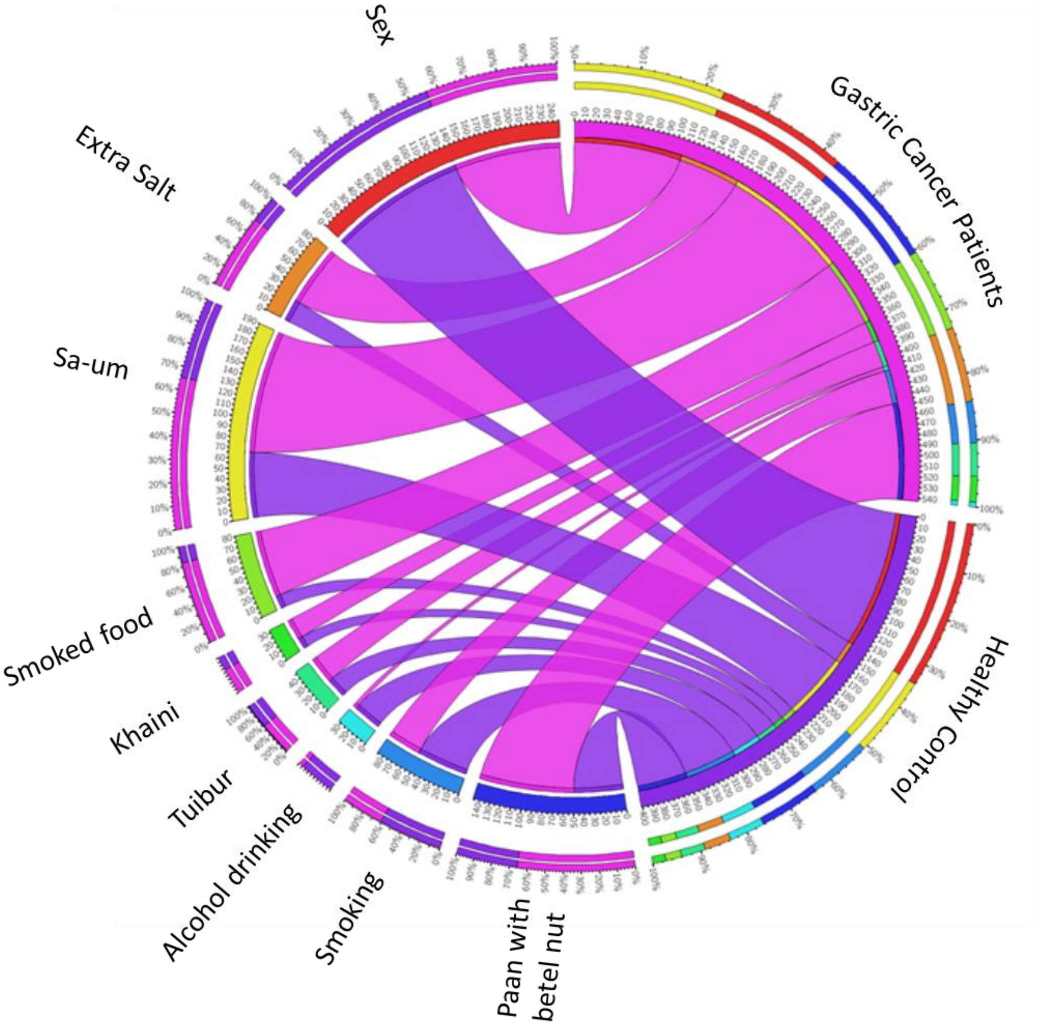


Figure 10: Frequency distributions of each demographic factors in the gastric cancer patients (pink ribbon) and healthy control (blue ribbon) groups in the study cohort.

The data were visualized via. Circos software. The frequency of different demographic factors associated with gastric cancer and healthy control groups is depicted in the outer ring. The inner ring of the circos plot depicts the subject number exposed with different demographic risk factors. Each factor has been assigned a specific color. The arc originates from gastric cancer and healthy control groups and terminates at different demographical factors to compare the association between the origin and terminating factors. The area of each colored ribbon depicts the frequency of the samples.

The univariate binary logistic regression analysis was performed for sex, BMI, dietary and lifestyle habits. BMI information was not available for 7 patients and 7 healthy controls (HC), so the analysis was done for 73 patients and 153 HC. Sex (*p*-value = 0.019) and BMI (*p*-value = 0.0001) were significant factors for the gastric cancer patients (Table 7).

Factors	ODDS ratio (95% CI)	*p* value
	Univariate analysis	
Sex	0.50 (0.28 – 0.89)	0.019
Age	1.01 (0.99 – 1.04)	0.07
BMI	**0.63 (0.56 – 0.72)**	**0.0001**
Extra Salt	**0.59 (0.41 – 0.86)**	**0.007**
Sa-um	0.75 (0.50 – 1.13)	0.180
Smoked Food	**0.49 (0.34 – 0.70)**	**0.0001**
Tuibur	**1.48 (1.09 – 2.00)**	**0.011**
Alcohol drinking	**3.11 (1.96 – 4.92)**	**0.0001**
Smoking	**7.50 (4.03 – 13.94)**	**0.0001**
Paan with betel nut	0.99 (0.56 – 1.76)	0.984
Multivariate analysis (logistic model)		
Sex	0.58 (0.24 – 1.40)	0.230
BMI	**0.69 (0.60 – 0.79)**	**0.0001**
Extra Salt	**0.68 (0.41 – 1.14)**	**0.042**
Smoked Food	**0.64 (0.40 – 1.04)**	**0.001**
Tuibur	1.30 (0.80 – 2.12)	0.285
Alcohol drinking	**1.83 (1.03 – 3.26)**	**0.001**
Smoking	**4.41 (1.86 – 10.43)**	**0.0007**

Table 7: Univariate and multivariate analysis of the risk factors compared between Gastric Cancer patients and Healthy Controls.

Among the dietary factors, extra salt consumption (*p*-value = 0.007), smoked food consumption (*p*-value = 0.0001), Smokeless tobacco (tuibur) intake (*p*-value =

0.011), smoking (*p*-value = 0.0001) and alcohol consumption (*p*-value = 0.0001) are the major significant risk factors for the GC (Table 7).

Further, multivariate analysis was performed with the seven significant factors for finding out the major risk factors and confounding factors that are associated with GC risk. Five factors were predicted as significantly associated with GC risk with high OR and 95% CI in multivariate analysis. BMI (*p*-value = 0.0001), Extra salt consumers (*p*-value = 0.042), smoked food consumers (*p*-value = 0.001), smokers (*p*-value = 0.0007) and alcohol drinkers (*p*-value = 0.001) were the high-risk groups associated with GC development (Table 7).

A risk score was estimated with the five factors using a logistic model and validated in the GC clinical cohort (Stage I, N = 20; Stage II, N = 14; Stage III, N = 44; Stage IV, N = 2) with the healthy controls (Figure 11A). The exposer of five-panel epidemiological factors might be successful in predicting the GC risk with different early symptoms (area under the curve – AUC = 0.91; *p*-value < 0.0001) (Figure 11B). This five-panel epidemiological factor achieved a high-risk score with significant-high positive probability values for GC patients with high sensitivity (79.45%) and specificity (91.72%) (Figure 11C).

For predicting GC at the early-stage, a risk score was estimated with the same 5 factors using a logistic model and was validated in the early stage (Stage I, N = 20 and II, N = 14) GC clinical cohort with the healthy control (Figure 11D). The exposer of five-panel epidemiological factors might be successful in predicting the GC risk during the premalignant stage with different early symptoms with a higher AUC value (0.946; *p*-value < 0.0001) (Figure 11E). This 5-panel epidemiological factor achieved a high-risk core with significant-high positive probability values for GC patients with high sensitivity (96.67%) and specificity (80.89%) (Figure 11F). The estimated significant factors (BMI, extra salt consumption, smoked food, alcohol drinking, and smoking) were the major risk factors associated with GC development. This significant panel of epidemiological factors can be used to detect GC patients at an early stage by counseling and proper public health practices.

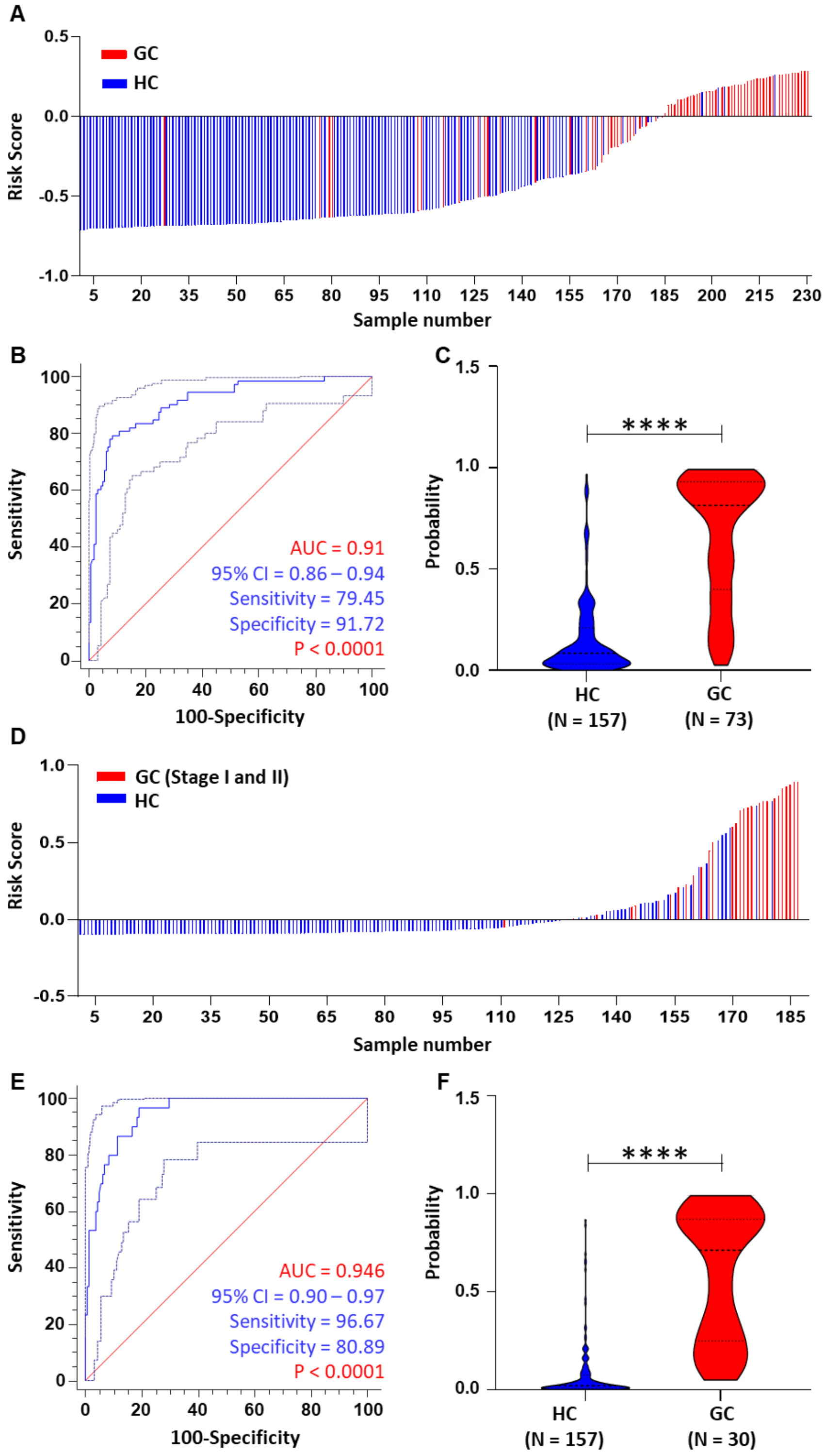


Figure 11: Estimation of accuracy value of the significant epidemiological factors based on the logistic model between gastric cancer and healthy control samples.

(A) Waterfall plot and risk score estimation for stage-I, II, III and IV samples, (B) Receiver operating curve (ROC) and accuracy estimation of epidemiological factors panel (BMI, extra salt consumptions, smoked food consumptions, alcohol drinking and smoking) (C) Significant association of the estimated probability values of the epidemiological factors panel between gastric cancer (n = 73) and healthy controls (n = 157), (D) Waterfall plot and risk score estimation for stage-I and II samples, (E) Receiver operating curve (ROC) and accuracy estimation of epidemiological factors panel. (F) Significant association of the estimated probability values of the epidemiological factors panel between stage-I and II gastric cancer (n = 30) and healthy controls (n = 157).

Genomic DNA was isolated from tissue of all the patients and blood DNA was isolated from all the patients and 25 healthy control (Figures 12 and 13). Screening and Genotyping of *H. pylori* and EBV were done for 80 patients. Out of 80 samples, 71 (88.75%) cases were positive for the pathogens and 9 (11.25%) of them were negative for pathogens. EBV positive cases were 32 (40%), 50 (63%) were detected positive for *H. pylori* and, 11 (13.75%) were positive for both the pathogen. Out of 50 *H. pylori*-positive cases (Figure 14), 46 cases were CagA, 17 were VacA, and 13 were both positive for both the genotypes (Figure 15 and 16). Out of 32 EBV cases, 29 were Type I, 7 were Type II positive and, 4 of them were having both genotypes (Figure 17).

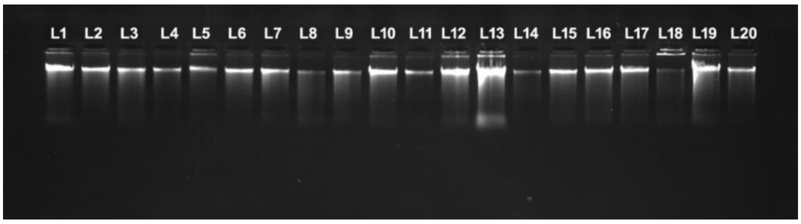


Figure 12: Gel picture of extracted DNA from the blood of representative patient samples (L1-L20)

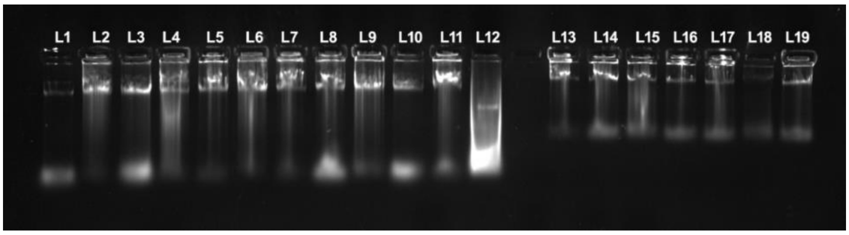


Figure 13: Gel picture of extracted DNA from tissue of representative patient samples (L1-L19)

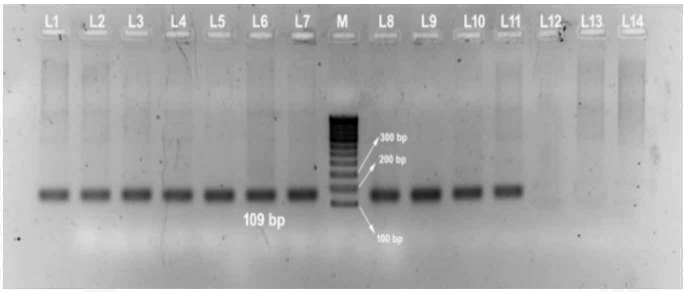


Figure 14: Gel picture of amplified16srRNA gene product of representative patient samples (L1- L14)

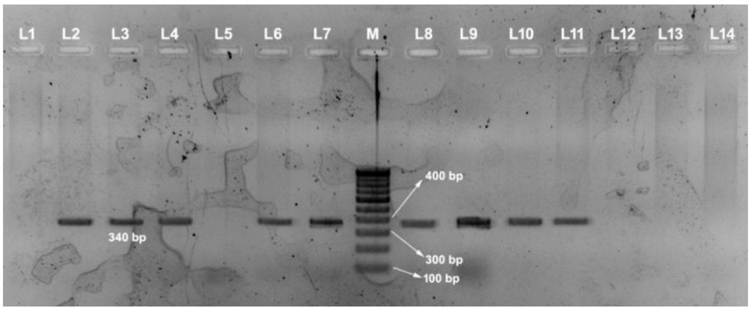


Figure 15: Agarose gel picture of amplified CagA gene product of representative samples (Pateint1-14).

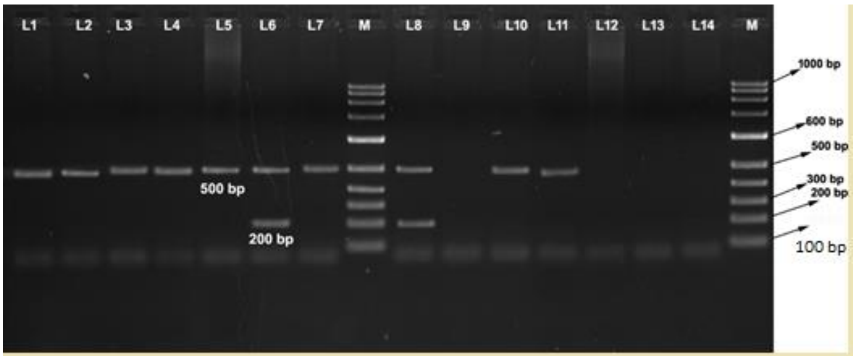


Figure 16: Agarose gel picture of amplified VacA gene product of representative patient samples (L1 - L14)

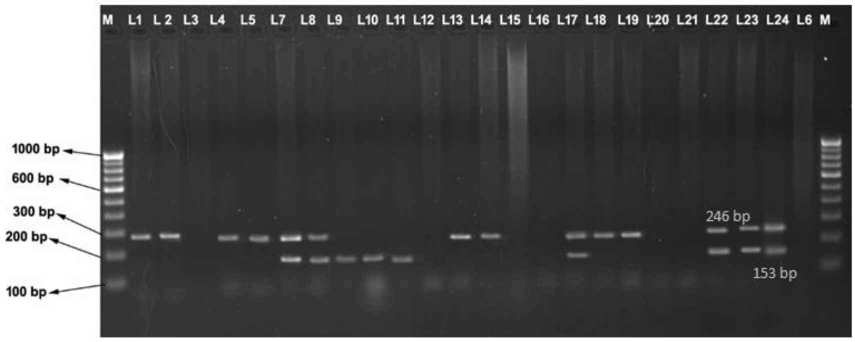


Figure 17: Agarose gel picture of an amplified EBNA3C gene product of representative patient samples (L1 - L5, L7 - L21 & L6)

The distribution of pathogen genotypes in GC patients was analyzed (Figure 18). Some patients are affected by one genotype of any one of the pathogen, whereas few patients were found to be affected by both the genotypes of *H. pylori* or EBV. Some patients were affected by one genotype of either *H. pylori* (CaGA or VacA) or one genotype of EBV (Type I or Type2). Some patients were found with all the genotypes from both the pathogens.

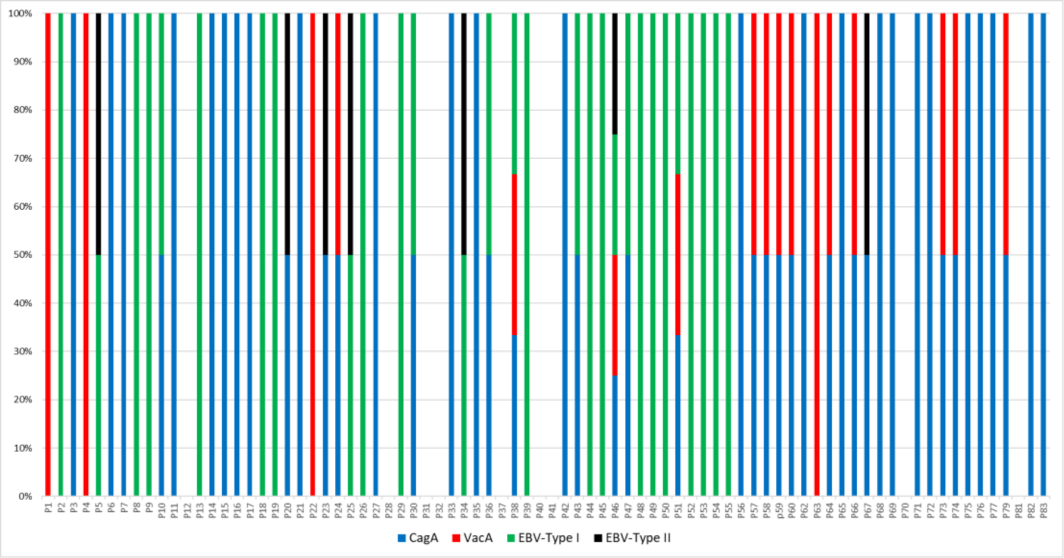
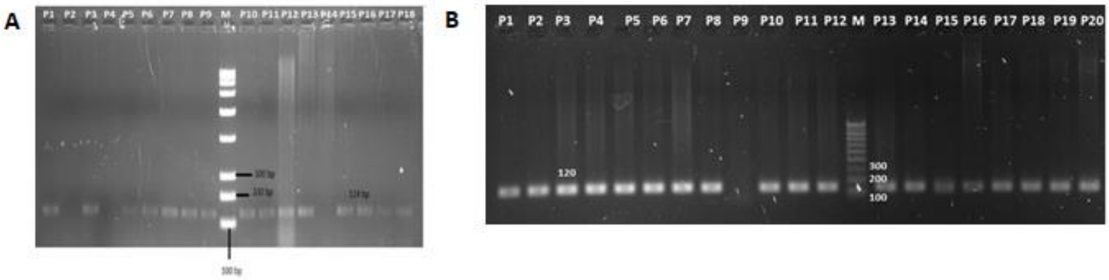


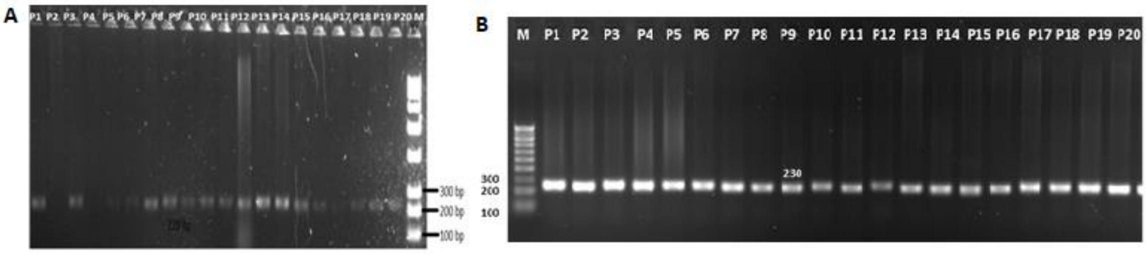
Figure 18: Distribution analysis of the Pathogen genotypes in GC patients (P1 – P 83).

In the case of MSI analysis, PCR amplification was done for each marker and the representative gel images are given (Figure 19-22). After screening 80 patients for MSI detection, 32 of them were detected as MSI-H, 30 of them were detected as MSI-L, and 18 cases were found to be Microsatellite stable. But, for statistical analysis, MSI-L was considered as MSS. Hence, 40% of cases were reported as Microsatellite instable and 60% of cases were reported as Microsatellite stable. If instability was found in two or more than two markers out of ten then it was considered as MSI cases (Figure 23A) and if there was no instability found in any marker then it was considered as an MSS case (Figure 23B).



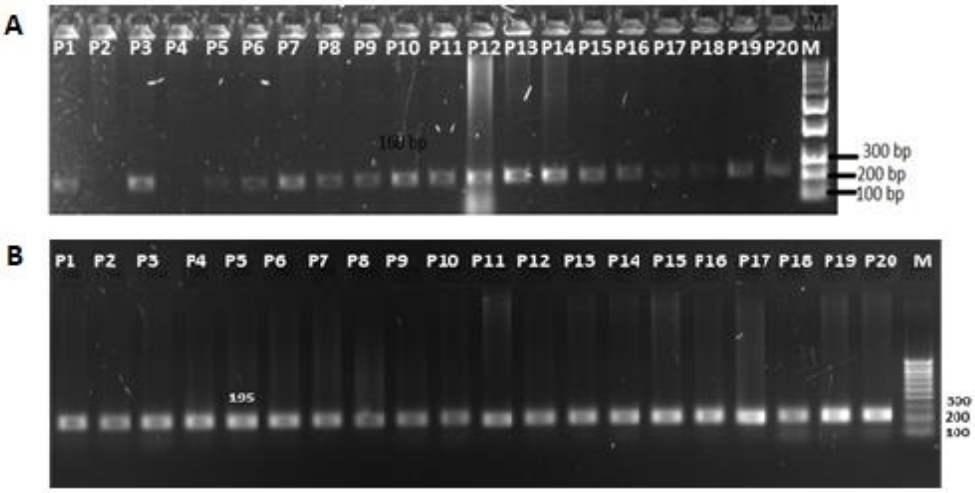
A
P1 P2 P3 P4 P5 P6 P7 P8 P9 M P10 P11 P12 P13 P14 P15 P16 P17 P18
100 bp
B
P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 M P13 P14 P15 P16 P17 P18 P19 P20
120
300
200
100

Figure 19: Agarose gel image of (A) BAT25 and (B) BAT 26 Marker representative samples (P1 – P18) and (P1 – P20) respectively



A
P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15 P16 P17 P18 P19 P20 M
300 bp
200 bp
100 bp
B
M P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15 P16 P17 P18 P19 P20
230
300
200
100

Figure 20: Agarose gel image of (A) D16S496 and (B) D2S123 Marker representative samples (P1 - P20).



A
P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15 P16 P17 P18 P19 P20 M
160 bp
300 bp
200 bp
100 bp
B
P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15 P16 P17 P18 P19 P20 M
195
300
200
100

Figure 21: Agarose gel image of (A) D16S3057 and (B) D16S398Marker representative samples (P1- P20)

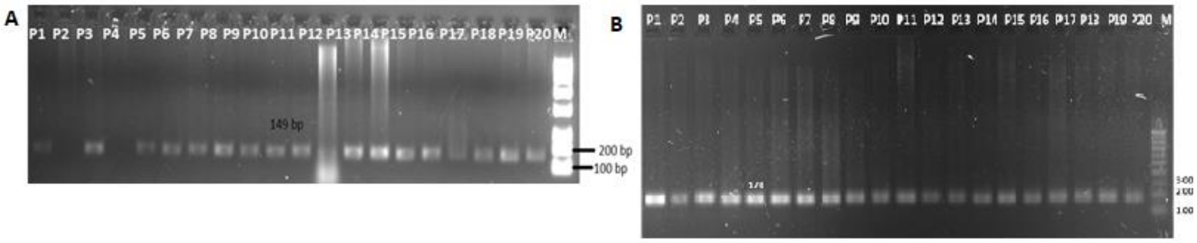

A
P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15 P16 P17 P18 P19 P20 M
149 bp
200 bp
100 bp
B
P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15 P16 P17 P18 P19 P20 M
174
300
200
100

Figure 22: Agarose gel image of (A) D18S58 and (B) D17S250Marker representative samples (P1 - P20)

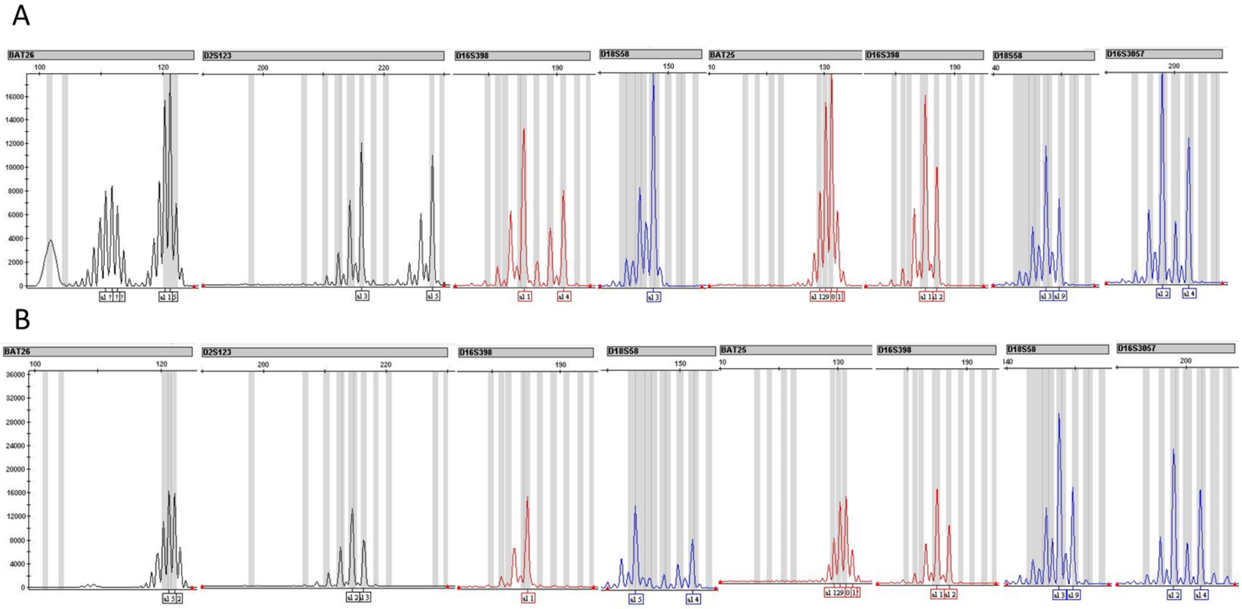


Figure23: (A) Representative MSI case and (B) Representative MSS case. Here, a comparison was done between the Tumor and blood sample of the studied patient.

In 32 MSI cases, 18 (56%) were positive for *H. pylori* infection and, 13 (41%) were EBV positive and one of them was negative for both pathogens. In the case of *H. pylori*-infected cases, 17 (53%) were CagA positive cases, 7 (23%) were VacA positive cases, six were positive for both the genotype. In EBV-infected cases, all 13 were positive for the Type I genotype (41%) and 13% cases were Type II positive. Four (04) of them were positive for both Type I and Type II genotypes (Table 8). EBV Type II cases were found more in the MSI subgroup and other all the genotypes were evenly distributed in both the subgroup. There was no direct relationship found between pathogen genotype and MSI-associated cases. EBV type II associated cases are showing a correlation with the MSI subgroup.

	MSI **(n = 32)**	MSS **(n=48)**	Total **(n = 80)**
H. pylori	56%	67%	63%
CagA	53%	60%	58%
VacA	23%	21%	21%
EBV	41%	40%	40%
Type_I	41%	40%	36%
Type_II	13%	6%	9%

Table 8: Distribution of pathogen-associated cases among the MSI and MSS subgroup

A proportion analysis among *H. pylori* and EBV-associated cases with MSI status was performed (Figure 24). EBV (+) cases were found more in the MSS subgroup, though it was not significant (Figure 24A). The proportion of *H. pylori* (+) cases was more in the MSS subgroup than *H. pylori* (-) cases (Figure 24B). CagA genotype-associated cases were higher in the MSS subgroup, but interestingly CagA (-) cases were not found in the MSI subgroup (Figure 24C). The proportion of VacA(+) cases was similar in both the subgroup, but VacA(-) cases were found more in the MSS subgroup (Figure 24D). The proportion of EBV type I (+) cases was more in the MSS subgroup (Figure 24E) and EBV type II (+) cases were found in a similar proportion in both the subgroups (Figure 24F).

Further, the GC samples were classified as *H. pylori* (+), *H. pylori* (-), *EBV* (+), *EBV* (-), MMR deficient and MMR proficient and a comparison was made between all the subgroups with clinical, demographic, and lifestyle habit data to find out significant factors with each subgroup of GC patients. The frequency distribution of clinical factors among the subgroups of GC patients is presented in Table 9. The tumor of the MMR deficient (87.5%) and *H. pylori*-positive (70%) patients group

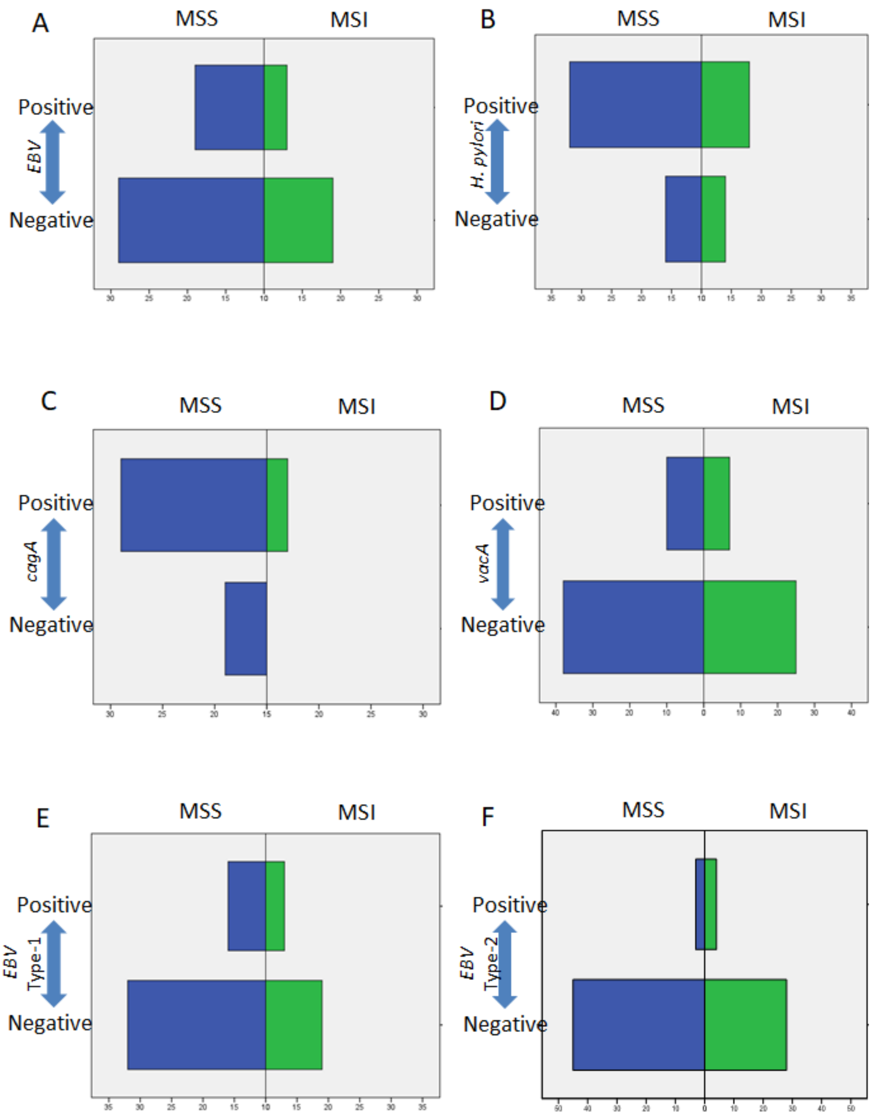


Figure 24: Proportion analysis of different genotypes in MSI and MSS subgroups

was located at high frequency in the distal portion of the stomach, whereas the tumor of *EBV* positive (65.62%) patient group was observed at less frequency. The poorly differentiated adenocarcinoma cases were observed at high frequency in MMR deficient, EBV positive and *H. pylori*-positive group whereas MMR proficient, EBV negative, and the moderately differentiated adenocarcinoma cases were observed at high frequency in *H. pylori*-negative group.

Factors	***H. pylori* (+) cases (n = 50)**	***H. pylori* (-) cases (n = 30)**	***EBV* (+) cases (n = 32)**	***EBV* (-) cases (n = 48)**	**MMR gene deficient (n = 32)**	**MMR gene proficient (n = 48)**
Anatomy						
Proximal	8 (16%)	3 (10%)	4 (12.5%)	7 (14.58%)	3 (9.37%)	8 (16.66%)
Distal	35 (70%)	24 (80%)	21 (65.62%)	38 (79.16%)	28 (87.5%)	31 (64.58%)
Data not available	7 (14%)	3 (10%)	7 (21.87%)	3 (6.25%)	1 (3.12%)	9 (18.75%)
TNM Stage						
I	11 (22%)	9 (30%)	8 (25%)	12(25%)	9(28.12%)	11(22.91%)
II	9 (18%)	5 (16.66%)	5 (15.62%)	9 (18.75%)	5 (15.62%)	9 (18.75%)
III	24 (48%)	16 (53.33%)	17 (53.12%)	23 (47.91%)	17 53.12%)	23 (47.91%)
IV	2 (4%)	0	1 (3.12%)	1 (2%)	0	2(4.16%)
Data Not available	4 (8%)	0	1 (3.12%)	3(6.25%)	1(3.12%)	3(6.25%)
Grade						
[a]WD	4 (8%)	3 (10%)	2 (6.25%)	5(10.41%)	2(6.25%)	5(10.41%)
[b]MD	23 (46%)	15 (50%)	12 (37.5%)	26 (54.16%)	12 (37.5%)	26 (54.16%)
[c]PD	20 (40%)	11 (36.66%)	16 (50%)	15 (31.25%)	17 (53.12%)	14 (29.16%)
Data Not available	3 (6%)	1 (3.3%)	2 (6.25%)	2(4.16%)	1(3.12%)	3(6.25%)
Family history of Cancer						
Yes	13 (26%)	14 (46.66%)	12 (37.5%)	15 (31.25%)	13 (40.62%)	14 (29.16%)
No	37 (74%)	16 (53.33%)	20 (62.5%)	33 (68.75%)	19 (59.37%)	34 (70.83%)

Table 9: Distribution of clinical factors among the various sub-groups in the gastric cancer patients' cohort (n = 80).

[a]WD - Well Differentiated, [b]MD - Moderately Differentiated, [c]PD - Poorly Differentiated.

The chi-square distribution test was performed to find out significant risk factors with each subgroup. Among all the risk factors, only smoked food consumption was significantly associated with the *H. pylori*-positive patient group (*p*-value= 0.006) and EBV-infected patient group (p-value = 0.002) (Table 10). Smoked food is the prime risk factor for developing pathogen-associated GC. Smokeless tobacco (tuibur) consumers (p-value = 0.06) were at low risk for developing *EBV* associated GC. Two lifestyle factors, tobacco chewing and alcohol drinking were found as a significant risk factor with high OR, 95% CI (p-value = 0.04) and (p-value= 0.03), respectively for MMR deficient patients group (Table 10).

For further verification, binary logistic regression was performed for determining the odd ratio and 95% CI. A significant association was found between *H. pylori*-infected GC patients with consumption of smoked food (*p*-value = 0.007) (Table 11, Figure 25A). Smoked food consumption (*p*-value=0.003) and tuibur intake (*p*-value = 0.05) were significant factors for *EBV* infected GC patients and tuibur consumption (Table 11, Figure 25C). Significant association was observed with chewing tobacco (p-value = 0.04) and alcohol drinking (p-value = 0.03) for the MMR deficient (MSI) patient group (Table 11, Figure 25E). Factors such as smoked food and tuibur consumption are found to be the major risk for pathogen infection in GC patients and chewing tobacco, alcohol drinking as lifestyle factors were the risk factors for MMR deficient GC patients.

Factors	*H. pylori* (+) cases (n = 50)	*H. pylori* (-) cases (n = 30)	EBV (+) cases (n = 32)	EBV (-) cases (n = 48)	MMR gene deficient (n = 32)	MMR gene proficient (n =48)
Age (mean)	59.5 ± 12.37	59.5 ± 9.76	59.5 ± 9.94	59.5 ± 12.36	56.5 ± 12.31	60 ± 10.60
Sex						
Male	34 (68%)	19 (63.33%)	20 (62.5%)	33 (68.75%)	12 (37.5%)	31 (64.58%)
Female	16 (32%)	11 (36.66%)	12 (37.5%)	15 (31.25%)	20 (62.5%)	17 (35.41%)
Extra salt						
Consumers	36 (72%)	20 (66.66%)	20 (62.5%)	36 (75%)	22 (68.74%)	34 (70.83%)
Non-consumers	14 (28%)	10 (33.33%)	12 (37.5%)	12 (25%)	10 (31.25%)	14 (29.16%)
ORs (95% CI), p value	*1.32 (0.49 – 3.51); 0.57*		*0.55 (0.21 – 1.46); 0.23*		*0.90 (0.34 – 2-39); 0.84*	
Sa-um						
Consumers	42 (84%)	24 (80%)	25 (78.12%)	41 (85.41%)	29 (90.62%)	37 (77.08%)
Non- consumers	8 (16%)	6 (20%)	7 (21.87%)	7 (14.58%)	3 (9.37%)	11 (22.91%)
ORs (95% CI), p value	*1.31 (0.40 – 4.23); 0.64*		*0.60 (0.19 – 1.94); 0.40*		*2.87 (0.73 – 11.26); 0.12*	
Smoked food						
Consumers	26 (52%)	25 (83.33%)	27 (84.37%)	24 (50%)	22 (68.74%)	29 (60.41%)
Non-consumers	24 (48%)	5 (16.66%)	5 (15.62%)	24 (50%)	10 (31.25%)	19 (39.58%)
ORs (95% CI), p value	***0.21 (0.07 – 0.65); 0.006***		***5.40 (1.78 – 16.37); 0.002***		*1.44 (0.56 – 3.70); 0.44*	
Paan with betel nut						
Consumers	30 (60%)	20 (66.66%)	21 (65.62%)	29 (60.41%)	23 (71.87%)	27 (56.25%)
Non-consumers	20 (40%)	10 (33.33%)	11 (34.37%)	19 (39.58%)	9 (28.12%)	21 (43.75%)
ORs (95% CI), p value	*0.75 (0.29 – 1.93); 0.55*		*1.25 (0.49 – 3.17); 0.63*		*1.98 (0.76 – 5.18); 0.16*	

Chewed tobacco						
Consumers	26 (52%)	15 (50%)	15 (46.87%)	26 (54.16%)	12 (37.5%)	29 (60.41%)
Non- consumers	24 (48%)	15 (50%)	17 (53.12%)	22 (52.08%)	20 (62.5%)	19 (39.58%)
ORs (95% CI), p value	*1.08 (0.43 – 2.67); 0.86*		*0.74 (0.30 – 1.83); 0.52*		***0.39 (0.15 – 0.98); 0.04***	
Tuibur						
Consumers	13 (26%)	8 (26.66%)	12 (37.5%)	9 (18.75%)	8 (25%)	13 (27.08%)
Non- consumers	37 (74%)	22 (73.33%)	20 (62.5%)	39 (81.25%)	24 (75%)	35 (72.91%)
ORs (95% CI), p value	*0.96 (0.34 – 2.69); 0.94*		*2.60 (0.93 – 7.20); 0.06*		*0.89 (0.32 – 2.49); 0.83*	
Smoking						
Smokers	35 (70%)	17 (56.66%)	19 (59.37%)	33 (68.75%)	21 (65.62%)	31 (64.58%)
Non-smokers	15 (30%)	13 (43.33%)	13 (40.62%)	15 (31.25%)	11 (34.37%)	17 (35.41%)
ORs (95% CI), p value	*1.78 (0.69 – 4.57); 0.22*		*0.66 (0.26 – 1.68); 0.39*		*1.04 (0.40 – 2.67); 0.92*	
Alcohol drinking						
Drinkers	17 (43%)	12 (40%)	10 (31.25%)	19 (39.58%)	16 (50%)	13 (27.08%)
Non-drinkers	33 (66%)	18 (60%)	22 (68.75%)	29 (60.41)	16 (50%)	35 (72.91%)
ORs (95% CI), p value	*0.77 (0.30 – 1.97); 0.58*		*0.69 (0.26 – 1.78); 0.44*		***2.69 (1.05 – 6.89); 0.03***	

Table 10: Distribution of demographic factors and lifestyle habits among the various sub-groups in the gastric cancer patients' cohort (n = 80), ORs - ODDS Ratios

Factors	OD (95% CI)	*p*-value
H. pylori		
Age	0.70 (0.26 – 1.91)	0.49
Sex	0.81 (0.31 – 2.10)	0.66
Extra salt consumption	1.28 (0.48 – 3.42)	0.61
Smoked food consumption	**0.21 (0.07 – 0.65)**	**0.007**
Sa-um consumption	1.31 (0.40 – 4.23)	0.64
Paan with betel nut consumption	0.75 (0.29 – 1.93)	0.55
Tuibur intake	0.96 (0.34 – 2.69)	0.94
Chewing tobacco	1.08 (0.43 – 2.67)	0.86
Smoking	1.78 (0.69 – 4.57)	0.22
Alcohol intake	0.77 (0.30 – 1.97)	0.58
EBV		
Age	2.14 (0.77 – 5.95)	0.14
Sex	1.32 (0.51 – 3.38)	0.56
Extra salt consumption	0.55 (0.21 – 1.46)	0.23
Smoked food consumption	**5.40 (1.78 – 16.37)**	**0.003**
Sa-um consumption	0.61 (0.91 – 1.94)	0.40
Paan with betel nut consumption	1.25 (0.49 – 3.17)	0.63
Tuibur intake	**2.60 (0.90 – 7.20)**	**0.05**
Chewing tobacco	0.74 (0.30 – 1.83)	0.52
Smoking	0.66 (0.26 – 1.68)	0.39
Alcohol drinking	0.69 (0.27 – 1.78)	0.44
MMR genes		
Age	0.48 (0.18 – 1.27)	0.14
Sex	0.82 (0.32 – 2.15)	0.70
Extra salt consumption	0.90 (0.34 – 2.39)	0.84
Smoked food consumption	1.44 (0.56 – 3.70)	0.44
Sa-um consumption	2.87 (0.73 – 11.26)	0.13
Paan with betel nut consumption	1.98 (0.76 – 5.18)	0.16
Tuibur intake	0.89 (0.32 – 2.49)	0.83
Chewing tobacco	**0.39 (0.15 – 0.98)**	**0.04**
Smoking	1.04 (0.40 – 2.67)	0.92
Alcohol drinking	**2.69 (1.05 – 6.89)**	**0.03**

Table 11: Univariate analysis of the association of demographic factors with pathogens and MMR genes status in Gastric Cancer patients’ cohort

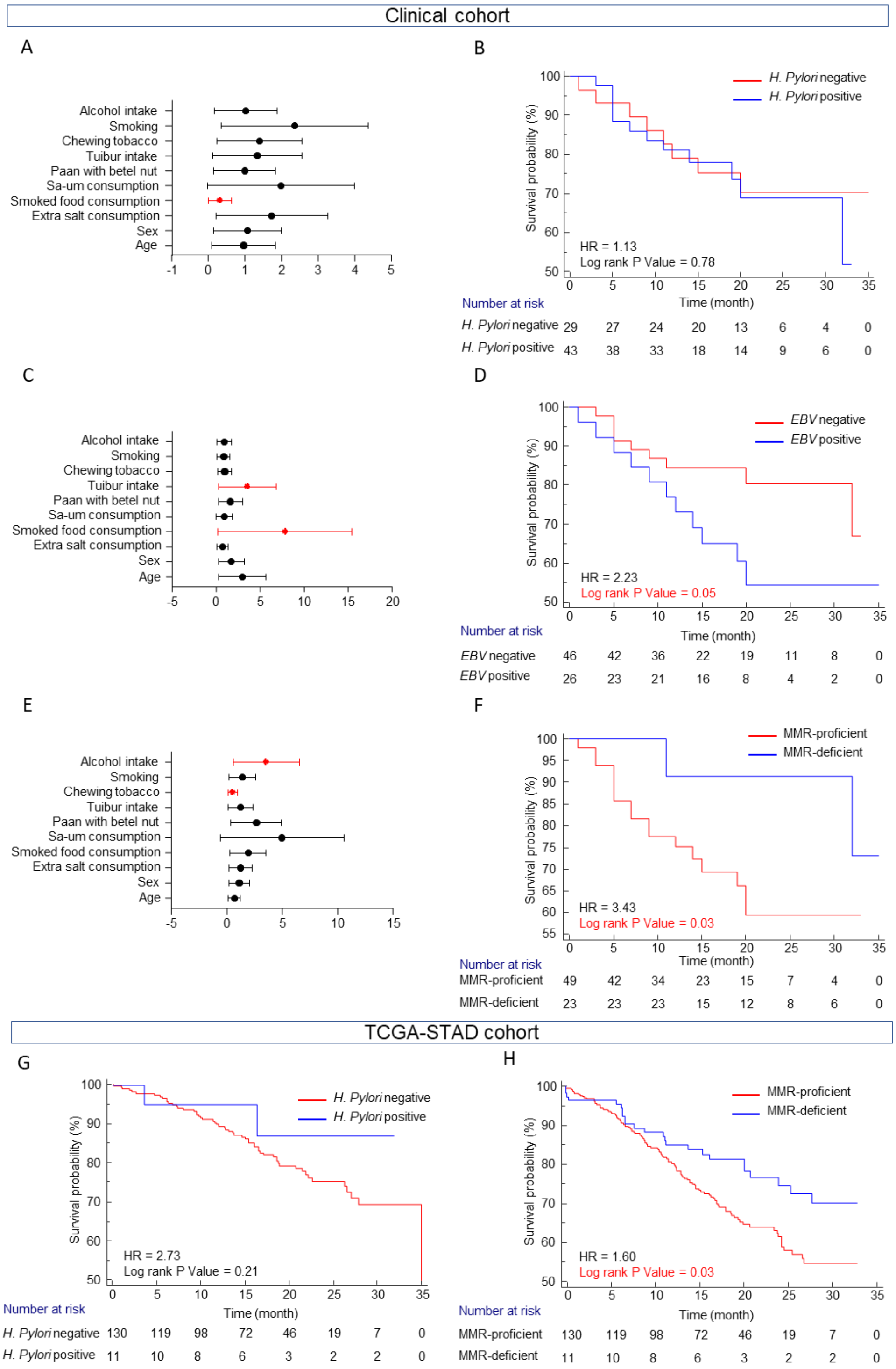


Figure 25: Association of overall survival probability and demographic factors with the pathogen and MMR gene status in gastric cancer patients.

Odds ratios and 95% confidence interval of the demographic factors presented for the *H. pylori* (A), *EBV* (C) and MMR gene status (E). Association between overall survival and the *H. Pylori* (G) and MMR gene status (H) in TCGA-STAD cohort. *EBV* status could not be analyzed due to less sample size in the TCGA-STAD dataset.

In this cohort, the follow-up data for 3 years was used to study the overall survival (OS) rate of patients with the subgroup [*H. pylori* (+), *H. pylori* (-), *EBV* (+), *EBV* (-), MMR deficient, and MMR proficient] by unadjusted analysis using the Kaplan Meier curve to find out prognostic factors. A univariate Cox proportional hazards model demonstrated that *H. pylori* infection does not have significant relation for GC patient's prognosis with stage I, II, and III (HR: 1.13, 95% CI: 0.86 - 1.73; *p*-value = 0.13; Figure 25B). *EBV* infection and MSI were independent prognostic predictors for GC patients with stages I, II, and III (Figures 25D and 25F). The GC patients group with *EBV* infection showed poor prognosis (HR: 2.22, 95% CI: 0.92 - 2.97; *p*-value = 0.05) with stages I, II, and III and were observed as a high-risk group. The comparison between MMR deficiency and proficiency exhibited a significant prognostic predictor for stages I, II, and III GC patient groups (HR: 3.43; 95% CI: 0.95 - 4.08; *p*-value = 0.03). In this cohort, MSI/MMR deficient cases showed a good prognosis for GC patients, whereas MSS/MMR proficient cases exhibited a poor prognosis for GC patients (Figure 25F). Further, we performed a comparison study by retrieving the data of gastric cancer patients with the *H. pylori*, *EBV,* and MMR gene status as independent prognostic factors for stages I, II, and III gastric cancer patients group from the TCGA-STAD cohort. In this approach, the Cox proportional-hazards regression model showed that *H. pylori* status has no significant log-rank value and *p*-value (Figure 25G), whereas MMR gene status exhibited as an independent prognostic factor in the TCGA-STAD cohort (HR: 1.60; 95% CI: 1.04 – 1.91; *p*-value = 0.03) (Figure 25H).

Targeted re-sequencing

In the case of somatic data analysis of the targeted re-sequencing (48 paired blood and tumor samples), two variant callers BbB and VarScan2 were used. BbB variant caller detected 1105 somatic variants in 58 genes. Among them, 297 were in the coding region and, 808 were in the non-coding region. Varscan 2 detected 2195 somatic variants in 58 genes, of which 369 were in the coding region and 1826 in the non-coding region. Now, a total of 666 coding variants from both the callers were filtered. But among them, 55 variants were common in both the variant callers. So, a total of 611somatic coding variants from both the caller were used to apply the first filter to identify 501 variants. Further, 271 variants were filtered out by applying the second filter and finally 183 somatic non-synonymous discovery variants were identified using the third filter (Figure 26). 183 non-silent somatic mutations were detected in 45 genes (in 32 patients), out of which 24 (13.11%) were indels and the remaining 159 (86.88%) were single-nucleotide substitutions. Among the single-nucleotide substitutions, 9 stopped gain, 10 were splice-site mutations and 140 were missense mutations (Figure 27A). The mutation signature of C>T transition was found in more than 75% of cases (Figure 27B)

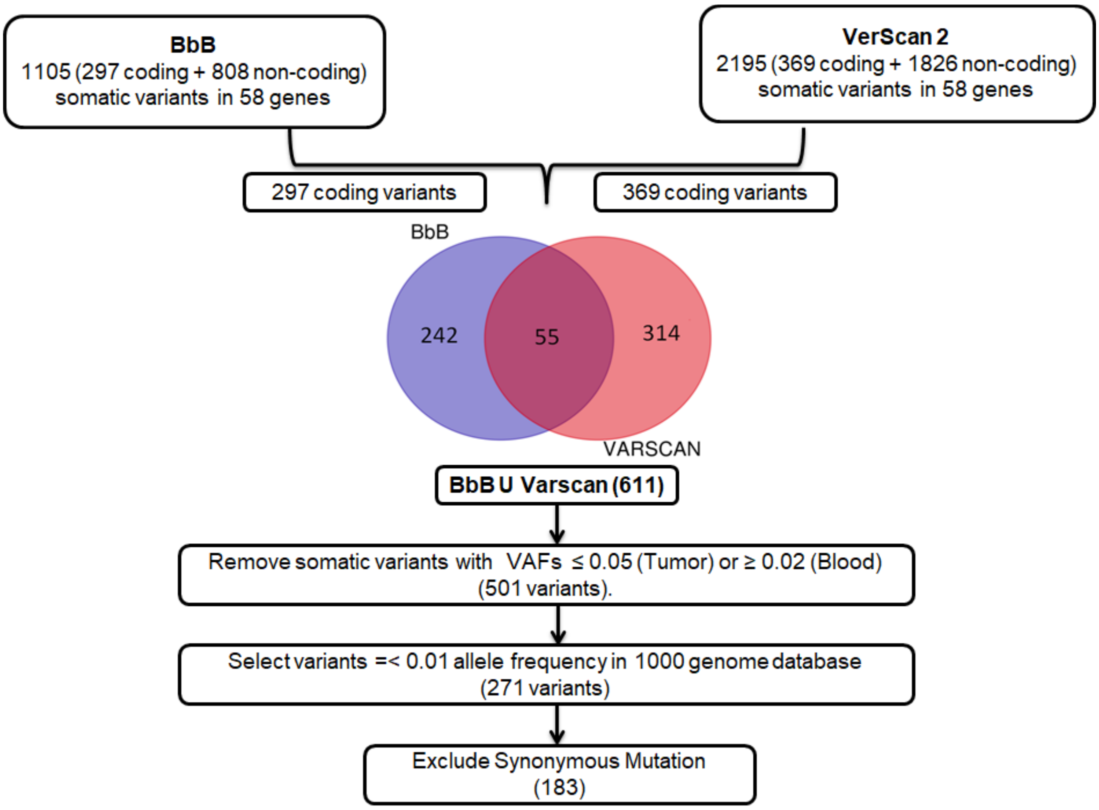

BbB
1105 (297 coding + 808 non-coding) somatic variants in 58 genes
VerScan 2
2195 (369 coding + 1826 non-coding) somatic variants in 58 genes
297 coding variants
369 coding variants
BbB
242
55
314
VARSCAN
BbB U Varscan (611)
Remove somatic variants with VAFs ≤ 0.05 (Tumor) or ≥ 0.02 (Blood) (501 variants).
Select variants =< 0.01 allele frequency in 1000 genome database (271 variants)
Exclude Synonymous Mutation (183)

Figure 26: Pipeline for applying a filter to get discovery set of variants

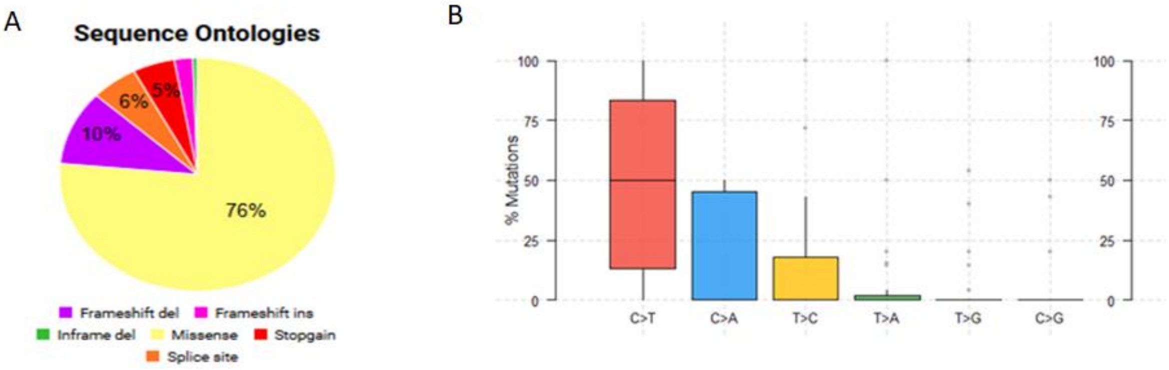


Figure 27: (A) Sequence ontology and (B) Mutation signature graph of somatic data

The top ten somatic mutated genes were *TP53* (47%), followed by *MUC6, FAT4*, *RNF43*, *BCOR*, *PTPRC, ERBB2*, *CTNNB1*, *SOHLH2,* and *FBXW7* (Figure 28A). The data was compared with the TCGA and Asian Cancer Research Group (ACRG) study of GC (Figure 28B). *TP53* and *FAT4* were found to be mutated in all the studies. *MUC6* and *APC* were found to be mutated in the Mizo population study and the ACRG study. The similarity between the top ten mutated genes of the ACRG group and our study was more than the TCGA group.

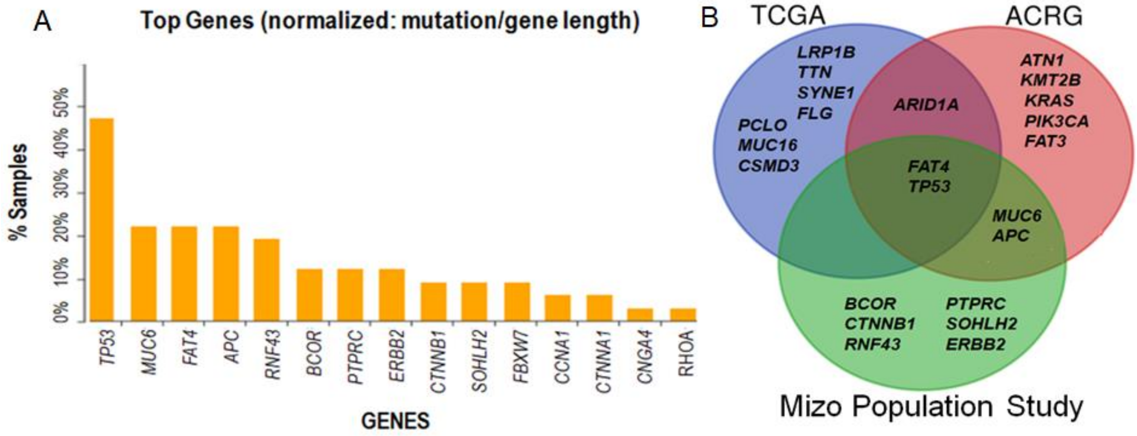


Figure 28: (A) Top mutated genes in gastric cancer in the Mizo population. (B) Comparison between top 10 mutated genes in TCGA, ACRG and Mizo population study.

Gene	***EBV* (+) cases (n = 22)**	***EBV* (-) cases (n = 10)**	**Gene**	**MSI cases (n = 13)**	**MSS cases (n = 19)**
APC	7 (31.81%)*	0 (0%)	BNC2	2 (15.38%)	0 (0%)
RNF43	6 (27.27%)	0 (0%)	CTNNB1	2 (15.38%)	2 (10.52%)
ARID1A	4 (18.18%)	0 (0%)	FAT3	4 (30.76%)	4 (21.05%)
ERBB2	4 (18.18%)	0 (0%)	BCOR	0 (0%)	4 (21.05%)
TP53	8 (36.36%)	7 (70%)	PTPRC	0 (0%)	4 (21.05%)
		Subtypes	***EBV* (+) (Frequency)**	**MSI (Frequency)**	**MSS (Frequency)**
TCGA Study		**Gene**	PIK3CA (80%)	ARID1A (73%)	CDH1 (30%)
			ARID1A (57%)	KMT2D (69%)	FAT4 (16%)
			CTNNA1 (20%)	FAT4 (58%)	ROHA (12%)
			BCOR (20%)	LRP1B (56%)	TP53 (12%)
			RELN (13%)	ACVR2A (53%)	ARID1A (12%)
ACRG Study		**Gene**	ARID1A (47%)	ARID1A (47%)	TP53 (45%)
			PIK3CA (40%)	PIK3CA (40%)	ARID1A (11%)
				TP53 (26%)	APC (11%)
				KRAS (23%)	PIK3CA (8%)
				APC (16%)	KRAS (6%)
Mizo population Study		**Gene**	TP53 (36.36%)	TP53 (46%)	TP53 (47%)
			APC (31.81%)	MUC6 (31%)	PTPRC (21%)
			RNF43 (27.27%)	FAT3 (31%)	BCOR (21%)
			ARID1A (18.18%)	FAT4 (23%)	APC (21%)
			ERBB2 (18.18%)	APC (23%)	FAT4 (21%)

Table 12: Frequency of mutated genes in gastric cancer subtypes in Mizo population and comparison between mutated genes in gastric cancer subtypes in TCGA, ACRG and Mizo population study. [Z-test P-value for equality of proportions: * $p < 0.05$]

The frequently mutated genes were analyzed according to the subgroups and the *APC* gene (32%) was significantly mutated with EBV (+) gastric cases (Table 12). Enrichment of RNF43, ARID1A and ERBB2 mutations were found in EBV (+) subtypes and mutation of these genes were absent in EBV (-) cases. The low frequency of *TP53* mutation was found in EBV (+) cases compared to EBV (-) gastric cancer subtypes. *BNC2* was found to be mutated only in MSI compared to MSS gastric cancer subtypes. *BCOR* and *PTPRC* were found to be mutated in MSS cases, but mutation of these genes was absent in MSI gastric cancer subtypes (Table 12).

Further, the subgroup-specific mutated genes data were compared with TCGA and ACRG datasets (Table 12). In the case of the EBV (+) subgroup, AID1A was the only gene that was common frequently in all the studies. *PIK3CA* was mutated in TCGA and ACRG studies, but not in the present study. In the case of the MSI subgroup, *FAT4* was mutated in the TCGA study and our study. This was the only gene that was similar to the TCGA study in the MSI subgroup. In TCGA and ACRG groups, *ARID1A* was the top mutated gene in the MSI subgroup. TP53 and APC were the two mutated genes that were commonly mutated in ACRG and our study. In the case of the MSS subgroup, *CDH1* was the top mutated gene in the TCGA study while *TP53* was the top mutated gene in ACRG and this study. *TP53* and *FAT4* were mutated commonly in TCGA and this study. *TP53* and *APC* were commonly mutated in ACRG and this study. *TP53* was mutated in all the studies for the MSS subgroup (Table 12). After observing the entire top mutated gene in each group we found that there is a similarity in mutation pattern between ACRG and in this Mizo population study.

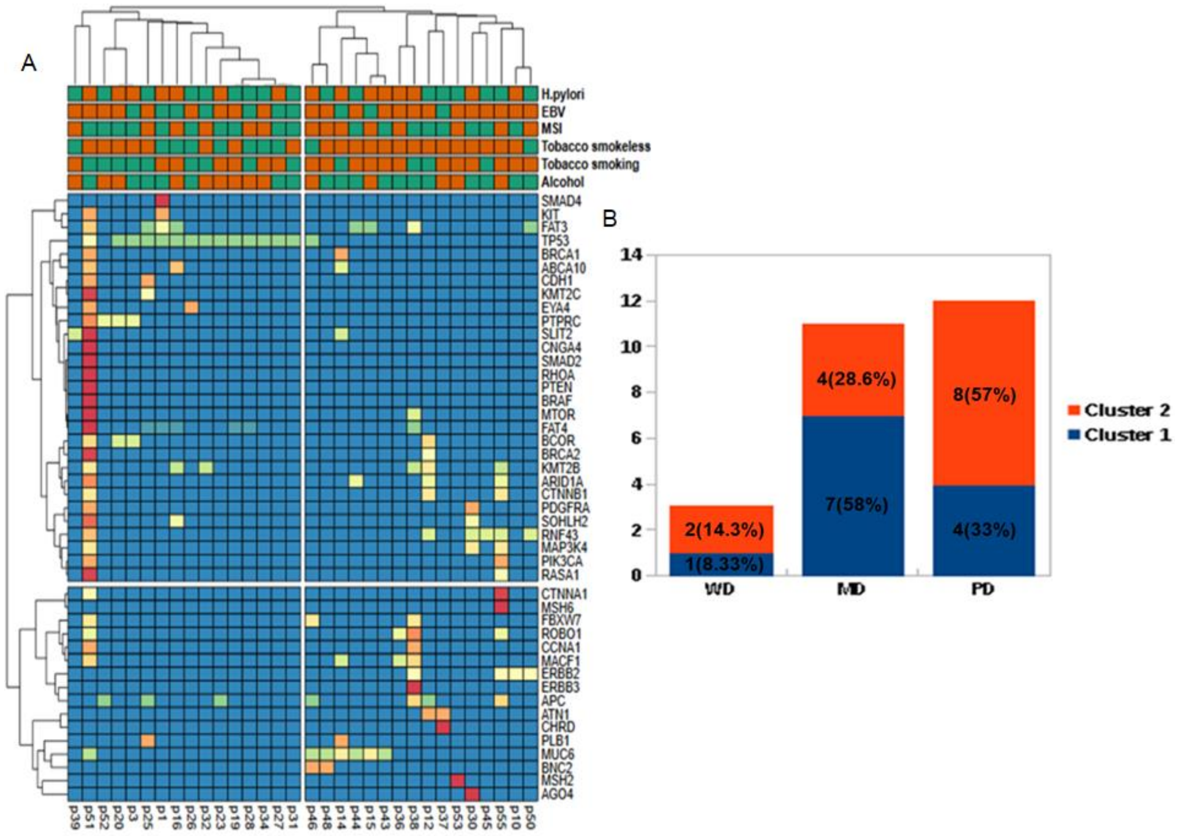


Figure 29: (A) Mutational landscape on the basis of subtypes and risk exposure in the Mizo population. (B) Frequency of cancer grade in each cluster.

[WD: Well-differentiated; MD: Moderately differentiated; PD: Poorly differentiate]

	Cluster 1 (n = 16)		Cluster 2 (n = 16)	
Subtypes	**Case Frequency**		**Case Frequency**	
TP53 Mutation	14 (88%)***		1 (6.3%)	
EBV	9 (56%)		13 (81.3%)	
H,Pylori	7 (44%)		8 (50%)	
MSI	6 (38%)		8 (50%)	
MSS	10 (63%)		8 (50%)	
	Gene	**Mutation Frequency**	**Gene**	**Mutation Frequency**
Frequency of mutated genes in both the cluster	PTPRC	4 (25%)	ERBB2	4 (25%)
	KIT	2 (12.5%)	BNC2	2 (12.5%)
	CDH1	2 (12.5%)	ATN1	2 (12.5%)
	KMT2C	2 (12.5%)	CHRD	1 (6.25%)
	EYA4	2 (12.5%)	AGO4	1 (6.25%)
	BRAF	1 (6.25%)	MSH2	1 (6.25%)
	SMAD2	1 (6.25%)	MSH6	1 (6.25%)
	RHOA	1 (6.25%)	ERBB3	1 (6.25%)
	PTEN	1 (6.25%)		
	CNGA4	1 (6.25%)		
	SMAD4	1 (6.25%)		

Table 13: Molecular subtypes in cluster 1 and 2 and Frequency of somatic mutations in each cluster

[*Z*-test *P*-value for equality of proportions*:* * * * $p < 0.001$]

The data is presented as a heat map and two prominent patient clusters were obtained (Figure 29A). TP53 was significantly mutated with the cluster 1 group compared to cluster 2. The EBV (+) group was dominant in cluster 2, while only one sample exhibited TP53 mutation in this cluster. There were no significant differences in *H. pylori*, MSI, or MSS subgroups between the two clusters. We have compared the mutated genes in both the cluster on the basis of their frequency. There was an enrichment for *PTPRC* (25%) gene in cluster 1 while enrichment of *ERBB2* (25%)

was observed in cluster 2 (Table 13). *KIT*, *CDH1*, *KMT2C*, *EYA4* genes exhibited two mutations each and *BRAF*, *SMAD2*, *RHOA*, *PTEN*, *CNGA4*, and *SMAD4* exhibited one mutation each only in cluster 1. *BNC2*, *ATN1* exhibited two mutations each and *CHRD*, *AGO4 MSH2*, *MSH6*, and ERBB3 genes were mutated with 6.25% frequency only in cluster 2 (Table 13).

The risk factors and clinical data were compared with mutations on the heat map. In the case of risk factors, there was no significant relationship between the two clusters. In the case of clinical data, 58% of moderately differentiated cases were found in the cluster 1 group. 57% of poorly differentiated cases were found in cluster 2, showing that patient samples with aggressive tumors were found in high EBV infected groups (Figure 29B). One hyper-mutated patient sample (stage IV) was found in the study with a mutation in most of the genes and two mutations in the *TP53* gene were also identified.

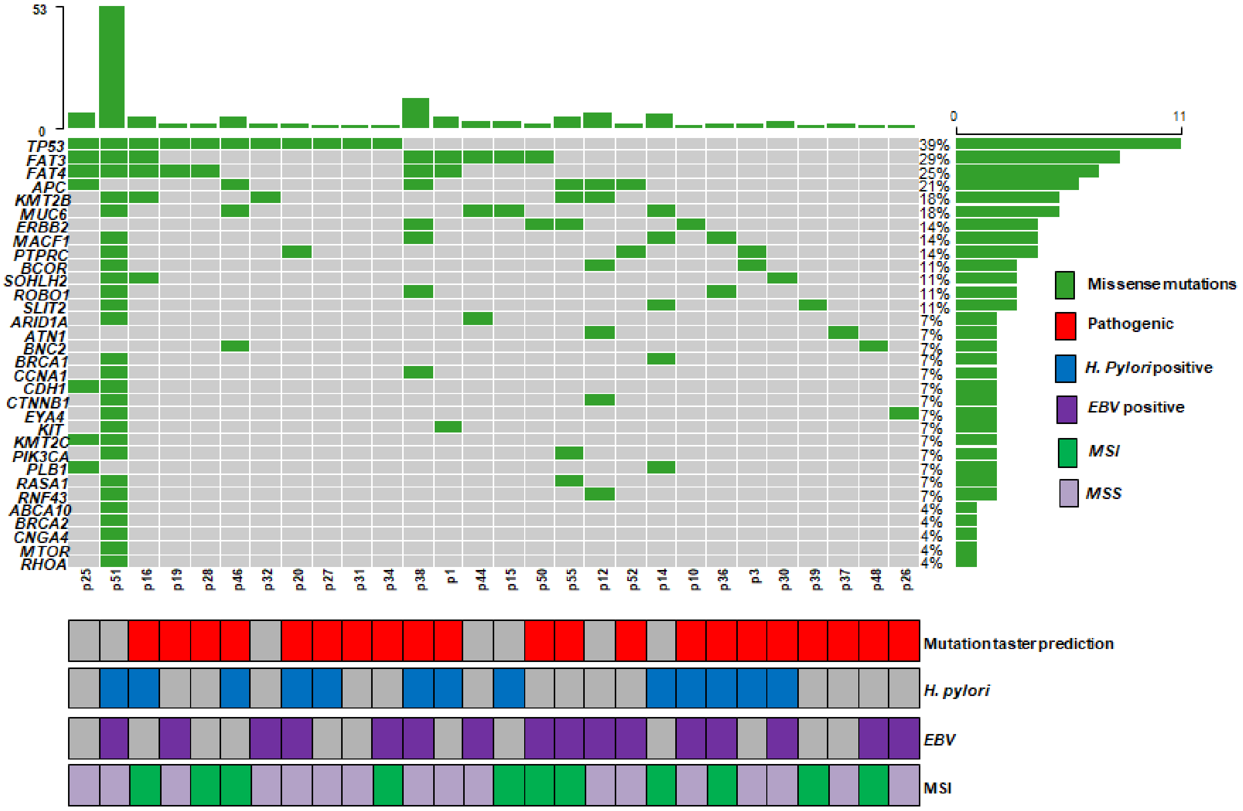


Figure 30: Oncoplot of somatic mutations data in GC samples

The X-axis represents patient samples and Y-axis represents the mutated genes

In this oncoplot, the top mutated gene is *TP53* which can develop cancer by altering the TP53 pathway (Figure 30). Another set of important genes are *FAT3* and *FAT4* which can develop cancer by altering the Hippo pathway. *APC* gene was also mutated frequently in this study which can develop cancer by WNT signaling pathway and *ERBB2*, a gene of RTK-RAS pathway is another mutated gene of this study. These pathways might play a role in the development of Gastric cancer in the Mizo population.

In this study, 183 variants were obtained, out of the 11 variants were predicted as pathogenic in the CLINVAR database. In the case of these 11 variants, 8 (R306*, G245S & R175H of *TP53*, D769Y and V842I of *ERBB2*, E545K and H1047R of *PIK3CA* and R876* of *APC* gene) were reported as pathogenic stomach cancer mutations in other populations. One pathogenic stop-loss (_352_) of *TP53* of liver cancer and two pathogenic stop gain variants (R1450* & R332*) of APC gene of large intestinal cancer were found in this study (Table 14). All the variants have occurred with a 2% frequency.

Gene	Mutation type	Protein alteration	COSMIC	dbSNP	Frequency (%)
TP53	Stop loss	_332_			2
TP53	Stop Gain	R306*	COSM10663	rs121913344	2
TP53	Missense	G245S	COSM6932	rs28934575	2
TP53	Missense	R175H	COSM10648	rs28934578	2
ERBB2	Missense	D769Y	COSM1251412		2
ERBB2	Missense	V842I	COSM14065		2
PIK3CA	Missense	E545K	COSM763	rs104886003	2
PIK3CA	Missense	H1047R	COSM775	rs121913279	2
APC	Stop Gain	R1450*	COSM13127	rs121913332	2
APC	Stop Gain	R332*	COSM19239	rs775126020	2
APC	Stop Gain	R876*	COSM18852	rs121913333	2

Table 14: List of pathogenic variants reported in CLINVAR

Gene	Mutation type	Protein alteration	COSMIC	dbSNP	Frequency (%)
TP53	MS	R273C	COSM10659	rs121913343	4
	MS	G266V	COSM10958		2
	MS	P250L	COSM10771		2
	MS	R175H	COSM10648		2
	MS	S215N	COSM44093		2
	MS	L194R	COSM44571		2
	MS	L137Q	COSM44745		2
	MS	E358V	COSM44081	rs773553186	2
MTOR	MS	A1792V	COSM1215724		2
BNC2	MS	S575R	COSM1184805		2
KMT2B	MS	S1747L	COSM3198517		2
ERBB2	MS	S310F	COSM48358		2
ERBB2	MS	Y781C	COSM85895		2
RHOA	MS	R5Q	COSM190569	rs11552758	2
RNF43	MS	H86R	COSM4755838		4
FAT3	MS	Y4395C	COSM5473266		2
CTNNA1	SG	R98*	COSM1241051		2
KMT2C	MS	G1517R	COSM28365	rs776685589	2
SLIT2	MS	R352C	COSM3132436	rs368061718	2
FAT3	MS	R3784H	COSM1357795	rs202061798	2
APC	MS	R332Q	COSM3428822	rs377665107	2
FAT4	MS	I3602L	COSM5008355		2

Table 15: List of pathogenic variants reported in COSMIC

Twenty-one missense variants and one-stop gain were reported as pathogenic stomach cancer variants in the COSMIC database. The most frequently mutated gene was *TP53* with 8 variants (R273C, G266V, P250L. R175H, S215N, L194R, L137Q & E358V) followed by *ERBB2* with 2 variants (S310F &Y781C) and *FAT3* (Y4395C & R3784H). A1792V of *MTOR*, S575R of *BNC2*, S1747L of *KMT2B*, G1517R of *KMT2C*, R5Q of *RHOA*, H86R of *RNF43*, R98* of *CTNNA1*, R352C of

SLIT2, R332Q of *APC* and I3602L of FAT4 gene were also found in this study. All the variants occurred with a 2% frequency only one variant (R273C) of *TP53* occurred with a 4% occurrence frequency (Table 15).

In this study, 99 novel variants were obtained as they were not reported in dbSNP/1000 Genomes/genomAD /COSMIC/EXAC/ExPasy database and hence are expected to be driver mutations. Out of 99 variants, 78 variants were predicted as pathogenic by Mutation taster (Table 16).

Gene	**Mutation type**	**Protein alteration**	**Frequency (%)**
MUC6	FD	P767PRPSSPAASPPRTSLGQPVPPHARCWPPVLPACPPSVSLAVSAPRASTRMPTGSVCPPRSAHVSSRGSPTLEELSSTLTAGPAPAQGGGGPVSRAPTAHPPAPSTGRATSSPSTASASYSTATASTSWPRTSVVSTTHSPPSRS*	4
CNGA4	MS	Y187C	2
CNGA4	MS	Y386C	2
TP53	SS	_332_	2
TP53	SS	_126_	2
MTOR	SS	_2389_	2
MTOR	MS	H1366R	2
MTOR	FI	L439RTFCGCEV*	2
SLIT2	MS	C32R	2
SLIT2	MS	G1235S	2
SLIT2	MS	M1376L	2
SLIT2	MS	Q1382R	2
ARID1A	MS	Y148S	2
ARID1A	MS	G1483S	2
BRCA2	MS	T2783A	2
KMT2B	MS	A1667T	2
KMT2B	MS	P2251L	2
KMT2B	MS	R2409Q	2
KMT2B	MS	V2472A	2
AGO4	MS	G730D	2
SOHLH2	MS	L99S	2
SOHLH2	SS	_6_	2
CCNA1	MS	D316V	2
CCNA1	MS	T361A	2
MACF1	MS	Q2473H	2
MACF1	MS	S4262P	2

MACF1	MS	T5516A	2
BCOR	MS	P1746S	2
BCOR	MS	T1730M	2
BCOR	SG	S542*	2
BCOR	MS	Q430P	2
BRCA1	MS	E1060G	2
CTNNB1	MS	D144A	2
CTNNB1	SG	R474*	2
MSH2	SS	_315_	2
SMAD4	MS	P356T	2
RHOA	MS	D78G	2
PDGFRA	SS	_1041_	2
KIT	MS	N99D	2
RNF43	MS	H556R	2
RNF43	SS	_126_	2
ABCA10	MS	N961D	2
ROBO1	MS	D1207E	2
ROBO1	MS	Y307C	2
ROBO1	MS	E64D	2
RASA1	MS	R707H	2
RASA1	SS	_1021_	2
PTEN	SS	_212_	2
FAT3	MS	V527A	2
FAT3	MS	T921I	2
FAT3	MS	D1645H	2
FAT3	MS	E2046G	2
FAT3	MS	V2622G	2
FAT3	MS	V3677F	2
APC	MS	D1266E	2
APC	FD	P1634PGMICHGCIVLKGHL*	2
APC	MS	R2521I	2
FAT4	MS	F49L	2
FAT4	SG	Q397*	2
FAT4	MS	V570A	2
FAT4	MS	V659A	2
FAT4	MS	L1062P	2
FAT4	MS	D1412N	2
FAT4	MS	I3250L	2
FAT4	MS	T3315S	2
FAT4	MS	V3423A	2
FAT4	MS	F3988L	2
FAT4	MS	V4094A	2
EYA4	MS	Q437H	2

Gene	Mutation type	Protein alteration	Frequency (%)
FBXW7	FI	M429IERQHHH*	2
MAP3K4	MS	Y1443H	2
PTPRC	MS	G13R	2
PTPRC	MS	A745S	2
FAT4	MS	L200M	2
FAT4	MS	K4381T	2
CTNNA1	MS	V830A	2
FBXW7	MS	M1I	2

Table 16: List of pathogenic novel variants predicted by Mutation Taster

Germline data was analyzed by the GATK tool and 1319 variants were obtained. A total of 78 variants in 32 genes (out of 60 gene panels) were mutated in entire 48 patients. Out of the 78 variants, 69 were non-synonymous variants and 9 were indels. Out of 9 indels, 6 were in-frame deletions, two were in-frame insertions and one was frameshift insertion.

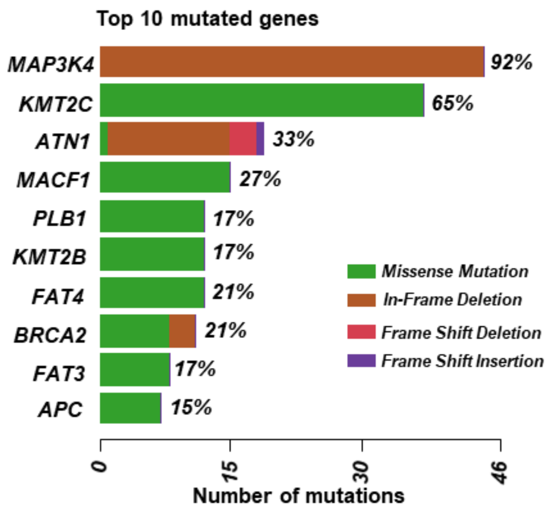


Figure 31: Top ten frequently mutated genes in germline analysis in GC samples

Out of 32 genes, *MAP3K4* (92%) is the top mutated gene in this germline study followed by *KMT2C* (65%), *ATN1* (33%), *MACF1* (27%), *BRCA2* & *FAT4* with 21%, *FAT3, KMT2B* & *PLB1* with 17% and *APC* with 15% frequency (Figure 31).

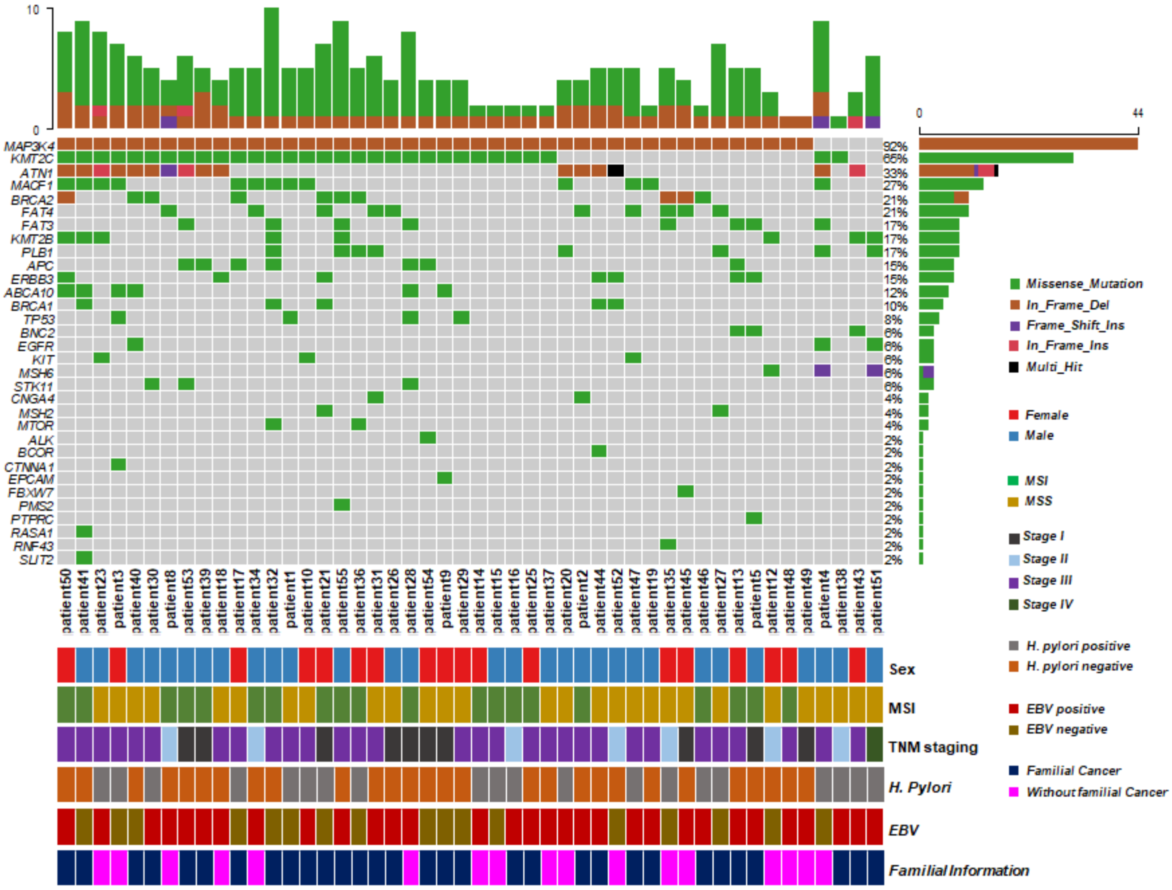


Figure 32: Oncoplot of germline mutations in GC samples

In this oncoplot, it is clearly shown that in the case of germline mutations, the *KMT2C* gene is mutated in only one set of patients (Figure 32). *MAP3K4* gene exhibited only one homozygous in-frame deletion (A1199del) with 92% occurrence frequency. There are some patient samples showing only *KMT2C* and *MAP3K4* mutations. Here, *KMT2C* (*MLL3*) gene is an important driver gene for developing GC at the germline level in this population. There was no significant relationship found between factors and genes, only in the case of male patients, *ATN1* and *KMT2C* genes were highly mutated compared to females. *BRCA2* was highly mutated in females compared to males. Surprisingly in this study, non-synonymous *CDH1* mutation was not present.

Further, binary logistic regression analysis was done to identify the significantly mutated genes or gene family with clinical factors. The most frequently mutated gene and gene families like *FAT3* and *FAT4* under FAT family, *EGFR* and

ERBB3 under EGFR family, *BRCA1* and *BRCA2* under DNA repair gene family and *MACF1* & *ATN1* independently were selected. The genes of the FAT family, *FAT3/4* were strongly significant (*p*-value = 0.003) with well and moderately differentiated cases (Table 17). Genes of the FAT family can be targeted for early diagnosis and therapeutic development as they are significant with early well and moderately differentiated cases. *MACF1* gene was significantly (*p*-value = 0.02) mutated with advanced stage and with poor survival status (*p*-value = 0.03) (Table 26). *MACF1* gene was showing a more aggressive tumor with a poor prognosis. *BRCA1/2* were showing a good prognosis (*p*-value = 0.03) (Table 17) which can be targeted for therapeutic response. *BRCA1/2* mutations are also increasing the risk of GC after Breast cancer. There was no significant association found between mutated genes and age, sex and familial information with cancer. *FAT3/4, MACF1* and *BRCA1/2* are the important genes for developing GC in this population at the germline level.

Factors	*Well and moderately differentiated cases* (n = 25)	*Poorly differentiated cases* (n = 21)	Early stage (I&II) (n = 17)	Advance stage (III&IV) (n = 30)	Family history with cancer (n = 17)	Family history without cancer (n =31)	Survival status (Alive) (n=30)	Survival status (deceased) (n=18)
FAT family (FAT4/3)								
Mutated cases	14 (56%)	2 (9.52%)	9 (52.94%)	8 (26.66%)	6 (35.29%)	11 (35.48%)	11(36.66%)	6(33.33%)
Not mutated cases	11 (44%)	19 (90.47%)	8 (47.05%)	22 (73.33%)	11 (64.70%)	20 (64.51%)	19(63.33%)	12(66.66%)
$^{\text{Non parametric T test}}$ *p* value	**0.001**		0.07		0.99		0.81	
^{LR}ORs (95% CI), *p* value	**13.41 (2.38 – 75.52); 0.003**		3.54 (0.88 – 14.20); 0.07		1.05 (0.29 – 3.79); 0.93		0.90 (0.22 – 3.63); 0.89	
EGFR family (EGFR/ERBB3)								
Mutated cases	6 (24%)	3 (14.28%)	3 (17.64%)	7 (23.33%)	3 (17.64%)	7 (22.58%)	7 (23.33%)	3 (16.66%)
Not mutated cases	19 (76%)	18 (85.71%)	14 (82.35%)	23 (76.66%)	14 (82.35%)	24 (77.41%)	23 (76.66%)	15 (83.33%)
$^{\text{Non parametric T test}}$ *p* value	0.68		0.65		0.69		0.58	
^{LR}ORs (95% CI), *p* value	1.97 (0.27 – 14.22); 0.50		0.65 (0.11 – 3.77); 0.63		0.79 (0.15 – 4.17); 0.78		1.04 (0.17 – 6.23); 0.95	
MACF1								
Mutated cases	8 (32%)	4 (19.04%)	1 (5.88%)	11 (36.66%)	5 (29.41%)	8 (25.80%)	5 (16.66%)	8 (44.44%)
Not mutated cases	17 (68%)	17 (80.95%)	16 (94.11%)	19 (63.33%)	12 (70.58%)	23 (74.19%)	25 (83.33%)	10 (55.55%)
$^{\text{Non parametric T test}}$ *p* value	0.32		**0.02**		0.79		**0.03**	
^{LR}ORs (95% CI), *p* value	2.99 (0.59 – 15.05); 0.18		**0.09 (0.01 – 0.87); 0.03**		1.16 (0.29 – 4.53); 0.83		**5.09 (1.11 – 23.39); 0.03**	
DNA Repair gene (BRCA1/2)								
Mutated cases	6 (24%)	7 (33.33%)	4 (23.52%)	10 (33.33%)	3 (17.64%)	11 (35.48%)	12 (40%)	2 (11.11%)
Not mutated cases	19 (76%)	14 (66.66%)	13 (76.47%)	20 (66.66%)	14 (82.35%)	20 (64.51%)	18 (60%)	16 (88.88%)
$^{\text{Non parametric T test}}$ *p* value	0.48		0.48		0.19		**0.03**	
^{LR}ORs (95% CI), *p* value	0.49 (0.09 – 2.59); 0.40		0.64 (0.13 – 3.05); 0.57		0.37 (0.08 – 1.68); 0.19		**0.13 (0.02 – 0.85); 0.03**	
ATN1								
Mutated cases	7 (28%)	7 (33.33%)	4 (23.527)	10 (33.33%)	6 (35.29%)	9 (29.03%)	9 (30%)	6 (33.33%)
Not mutated cases	18 (72%)	14 (66.66%)	13 (76.47%)	20 (66.66%)	11 (64.70%)	22 (70.96%)	21 (70%)	12 (66.66%)
$^{\text{Non parametric T test}}$ *p* value	0.69		0.48		0.65		0.81	
^{LR}ORs (95% CI), *p* value	0.85 (0.16 – 4.43); 0.84		0.80 (0.16 – 3.92); 0.78		1.64 (0.41 – 6.55); 0.48		1.37 (0.31 – 6.10); 0.67	

Table 17: Comparison of clinical factors with frequently mutated gene and genes family

The analysis with only familial cases of Gastric Cancer resulted in 18 mutated genes: *MAP3K4, ABCA10, BRCA1, ATN1, FAT3, APC, STK11, TP53, CTNNA1, KMT2C, FAT4, PLB1, CNGA4, MSH6, PMS2, BRCA2* and *KMT2B*. In univariate analysis out of those 18 genes, *CTNNA1, PMS2* and *KMT2C* were significantly mutated with familial cases of GC (Figure 33A). Further, multivariate analysis with these three genes (*CTNNA1, PMS2* and *KMT2C*) identified significant mutations in familial GC cases (Figure 33B). These three genes are the significant genes that might develop familial GC, besides *CDH1* in the Mizo population.

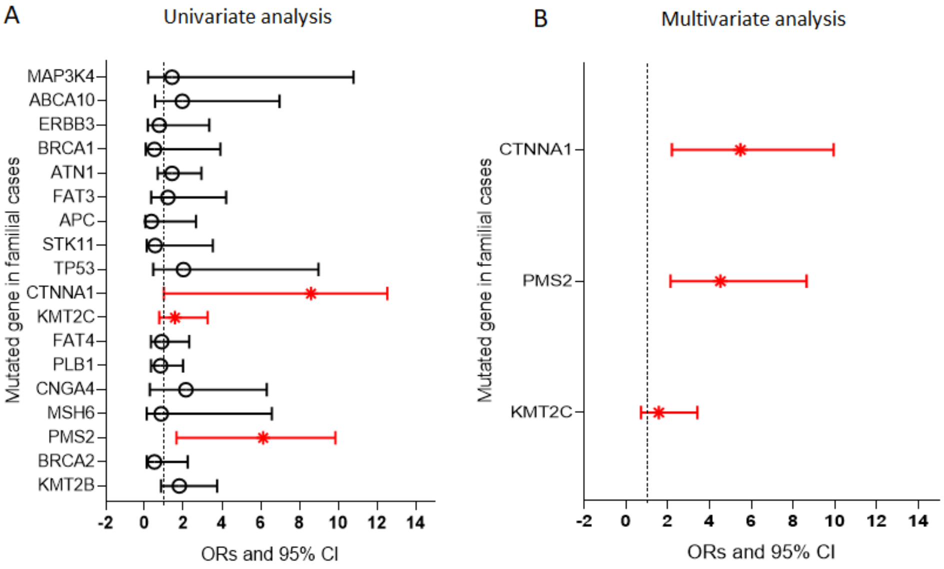


Figure 33: (A) Univariate and (B) Multivariate analysis of mutated genes in Familial GC cases

Survival analysis was done by selecting the familial patients having mutations in *CTNNA1*, *PMS2* and *KMT2C* to find out prognostic risk factors by unadjusted analysis of follow-up data using the Kaplan Meier curve. A univariate Cox proportional hazards model demonstrated that the *KMT2C* gene was an independent prognostic predictor for familial GC patients as it was showing poor prognosis (HR: 1.57, 95% CI: 0.76 - 3.26;

p-value = 0.02; Figure 34A). *PMS2* and *CTNNA1* were also showing poor prognosis with a high odd ratio and significant *p*-value (HR: 6.13, 95% CI: 1.65 – 9.85; *p*-value = 0.004; Figure 34B and HR: 8.59, 95% CI: 1.01 – 12.53; *p*-value = 0.05; Figure 34C, respectively) but we cannot consider as an independent predictor due to fewer patients with those gene mutations (it may be a case of data overfeeding).

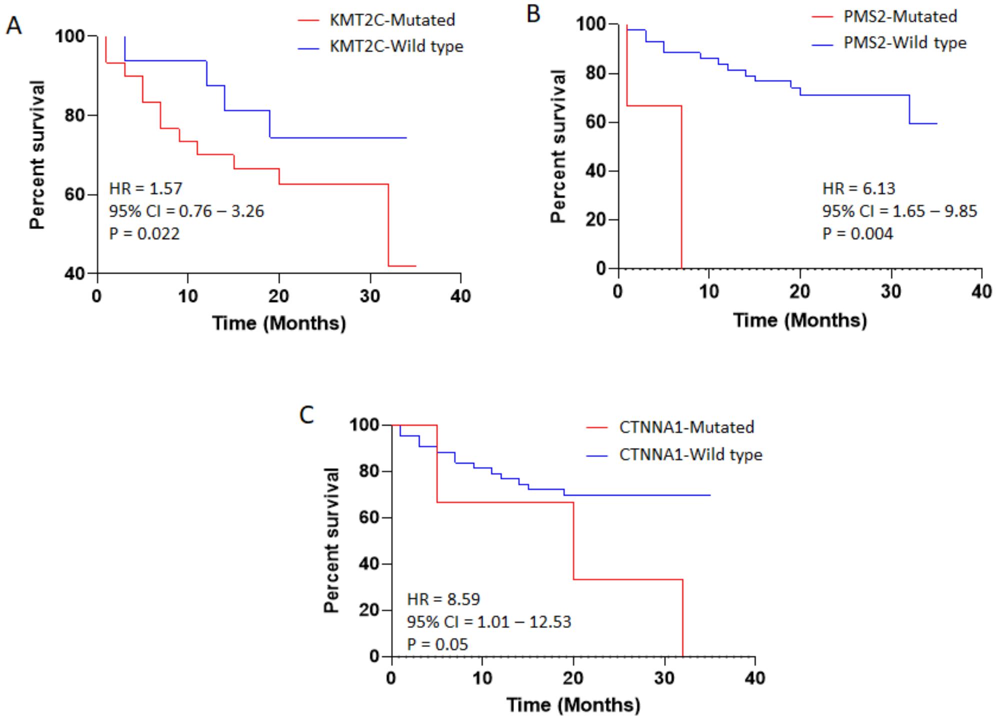


Figure 34: Survival analysis of (A) KMT2C, (B) PMS2 and (C) CTNNA1 mutated patients

Further, survival analysis was done by combining those samples with KMT2C, PMS2 and *CTNNA1* for getting a panel of genes to predict Familial GC cases. The panel of three genes was a strong prognostic predictor with a significant *p*-value (HR: 1.82, 95% CI: 0.68 - 4.85; *p*-value = 0.04; Figure 35A) as it was showing poor prognosis. A risk score was estimated with the same panel of three genes using a logistic model (Figure 35B). The panel of three genes might be successful in predicting the familial GC risk in this population with a higher AUC value (0.68; *p*-value = 0.03) (Figure 35B). The estimated significant gene mutations were the major risk factors associated with Familial GC development. This significant panel of mutated genes can be used to detect Familial GC patients in this population.

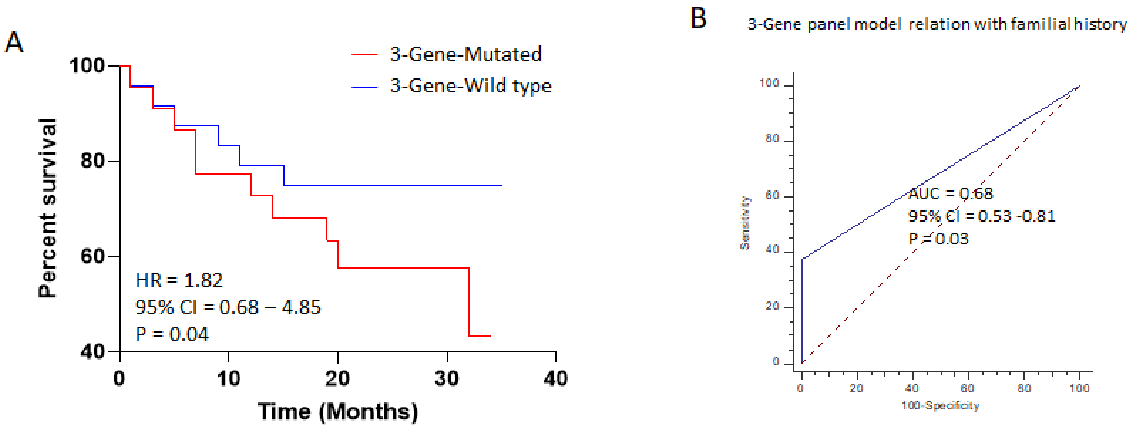


Figure 35: (A) Survival analysis of a panel of three genes (B) Accuracy score of panel gene to predict familial GC

The pathway analysis was done to find out their significant role in GC development in this population. The RTK-RAS pathway is an important pathway related to the development of GC. *ERBB3, EGFR, KIT, ALK* and *RASA1* were the frequently mutated gene of the RTK-RAS pathway in this study (Figure 36). Another important pathway is the Hippo-signalling pathway for developing cancer. *FAT3* and *FAT4* genes of the hippo signaling pathway were frequently mutated in this study. Wnt signaling was altered in this study due to alteration found in *APC* and *RNF43.* PI3K pathways were also got altered in this study due to alteration in *STK11* and MTOR gene (Figure 36). Another important cancer-related pathway TP53 pathway was altered due to germline mutation of the TP53 gene. Notch signaling was also altered in this study due to mutation in BRCA 2 gene. So RTK-RAS, Hippo signaling, Wnt, PI3K, TP53 and NOTCH signaling pathway alterations might be responsible for developing Gastric cancer in this population (Figure 36).

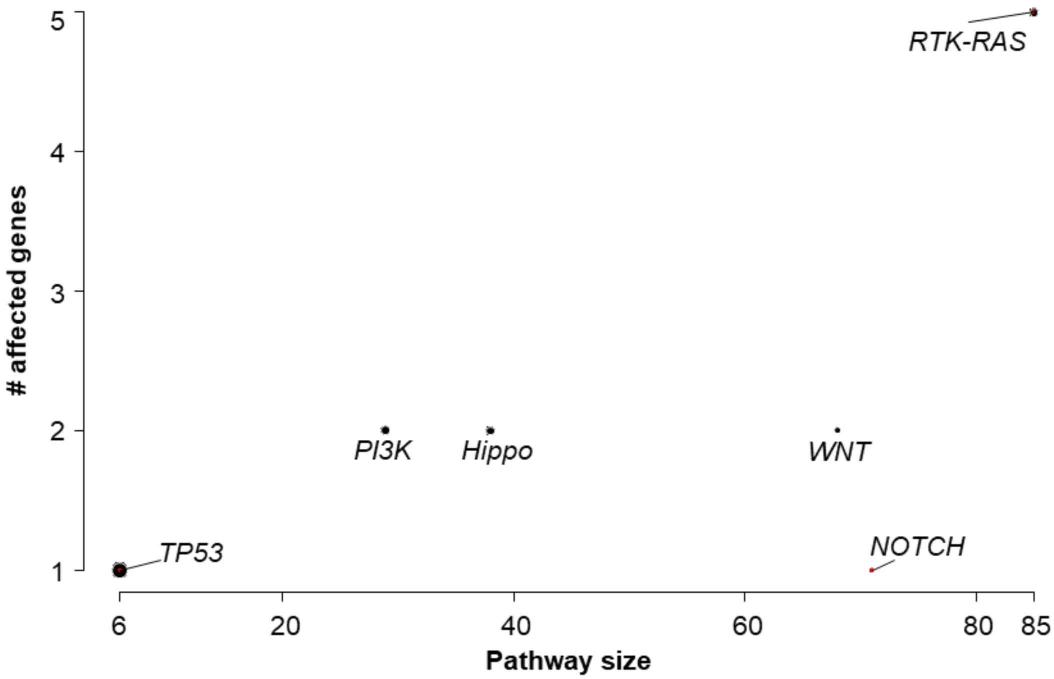


Figure 36: Significant pathways associated with GC development in this study

In this study, out of 78 variants, 23 were novel variants. Pathogenicity prediction was done for all the non-synonymous variants by four prediction tools (SIFT, PROVEAN, Polyphen2 and Mutation Taster) (Table 18). We predicted the variants as pathogenic if the variants were found to be predicted as damaging or deleterious in all the tools. Among all the missense or non-synonymous variants, only 12 variants were predicted as pathogenic in all the tools and they are as follows: C3121Y, P4952L and R5357Q (*MACF1* with 8.33%, 4.17% and 2.08% frequency, respectively), P922R and W4352G (*KMT2C* with 4.17% and 2.08% frequency, respectively), A2066G (*FAT3* with 2.08% frequency), Y856H (*BRCA1* with 8.33% frequency), P587R (KMT2B with 10.42% frequency), A667T (*MSH2* with 2.08% frequency), Q965L (*ABCA10* with 2.08% frequency) G2608A (*FAT4* with 2.08% frequency) and L114F (*PMS2* with 2.08% frequency) (Table 18). Out of 12 pathogenic variants, 3 were novel variants (P4952 L - MACF1, Q965L - ABCA10, G2608A - FAT4) (Table 18).

Gene	Mutation type	Protein Change	Variant information	Frequency of this mutation (%)	SIFT	Polyphen2	Mutation Taster	PROVEAN
MACF1	**nonsynonymous**	**R5357Q**	**Reported**	**2.08**	**D**	**D**	**D**	**D**
KMT2C	nonsynonymous	R909K	Reported	60.42	T	D	D	N
KMT2C	nonsynonymous	P309S	Reported	12.50	T	D	D	D
TP53	nonsynonymous	E319V	Reported	6.25	T	B	D	N
FAT4	nonsynonymous	I3602L	Reported	16.67	D	B	D	N
CNGA4	nonsynonymous	L174V	Reported	4.17	D	B	D	N
MACF1	nonsynonymous	N2198Y	Reported	10.42	D	P	N	D
CTNNA1	nonsynonymous	N283S	Reported	2.08	D	P	D	N
ABCA10	nonsynonymous	L663S	Reported	10.42	D	P	D	D
MACF1	**nonsynonymous**	**P4952L**	**Novel**	**4.17**	**D**	**D**	**D**	**D**
PLB1	nonsynonymous	Y525C	Reported	2.08	D	D	N	D
EGFR	nonsynonymous	K253R	Reported	2.08	T	B	N	N
KMT2C	**nonsynonymous**	**P922R**	**Reported**	**4.17**	**D**	**D**	**D**	**D**
FAT3	nonsynonymous	R2606T	Novel	2.08	T	D	D	N
PTPRC	nonsynonymous	N199S	Reported	2.08	.	B	N	.
FAT3	nonsynonymous	Q3375R	Reported	2.08	T	B	N	N
ERBB3	nonsynonymous	R967K	Reported	2.08	T	D	D	N
EPCAM	nonsynonymous	R153T	Reported	2.08	T	P	D	N
MACF1	nonsynonymous	K853T	Novel	4.17	D	B	D	D
KIT	nonsynonymous	I438V	Reported	6.25	T	B	D	N
MSH6	nonsynonymous	E1163V	Reported	2.08	D	P	D	D
KMT2B	nonsynonymous	P2351L	Reported	4.17	T	B	N	N
APC	nonsynonymous	T1261I	Novel	8.33	D	P	D	D
FAT3	**nonsynonymous**	**A2066G**	**Reported**	**2.08**	**D**	**D**	**D**	**D**
ERBB3	nonsynonymous	K498I	Reported	10.42	T	B	D	D
BRCA2	nonsynonymous	P389Q	Reported	10.42	T	B	N	N
MACF1	**nonsynonymous**	**C3121Y**	**Reported**	**8.33**	**D**	**D**	**D**	**D**
PLB1	nonsynonymous	S1284T	Reported	14.58	T	P	N	N

Table is continued [D=Deleterious, T=Tolerated, P-Possibly damaging, Damaging, N=Neutral and B = Benign]

Gene	Mutation type	Protein Change	Variant information	Frequency of this mutation (%)	SIFT	Polyphen2	Mutation Taster	PROVEAN
MSH2	nonsynonymous	T8M	Reported	2.08	T	P	D	N
BRCA1	**nonsynonymous**	**Y856H**	**Reported**	**8.33**	**D**	**D**	**D**	**D**
MACF1	nonsynonymous	N2544S	Novel	2.08	T	D	D	D
KMT2B	nonsynonymous	P435A	Reported	6.25	D	B	N	N
KMT2B	**nonsynonymous**	**P587R**	**Reported**	**10.42**	**D**	**D**	**D**	**D**
KMT2B	nonsynonymous	V2174A	Novel	2.08	D	P	N	N
PLB1	nonsynonymous	R935W	Reported	6.25	D	D	N	D
PLB1	nonsynonymous	G1265R	Reported	2.08	T	D	N	D
MSH2	**nonsynonymous**	**A667T**	**Reported**	**2.08**	**D**	**D**	**D**	**D**
FAT4	nonsynonymous	R1169Q	Reported	2.08	T	B	N	N
FAT3	nonsynonymous	V2622I	Novel	2.08	T	B	D	N
TP53	nonsynonymous	C341G	Reported	2.08	D	B	D	N
ABCA10	**nonsynonymous**	**Q965L**	**Novel**	**2.08**	**D**	**D**	**D**	**D**
STK11	nonsynonymous	D359N	Novel	4.17	T	B	D	N
MTOR	nonsynonymous	M1590V	Reported	2.08	T	P	D	N
APC	nonsynonymous	S2498C	Novel	4.17	T	P	N	N
FAT3	nonsynonymous	T770M	Reported	8.33	T	B	N	N
KMT2B	nonsynonymous	A2010S	Reported	2.08	T	B	N	N
RNF43	nonsynonymous	A365T	Reported	2.08	T	B	N	N
MTOR	nonsynonymous	L413F	Novel	2.08	D	B	D	D
BRCA2	nonsynonymous	M1149V	Reported	2.08	T	B	N	N
EGFR	nonsynonymous	G614S	Reported	4.17	T	B	N	N
SLIT2	nonsynonymous	S582P	Novel	2.08	D	P	D	D
RASA1	nonsynonymous	E161D	Novel	2.08	T	B	D	N
KMT2C	**nonsynonymous**	**W4352G**	**Reported**	**2.08**	**D**	**D**	**D**	**D**
ERBB3	nonsynonymous	R453C	Reported	2.08	T	D	D	D

Table is continued [D=Deleterious, T=Tolerated, P-Possibly damaging, Damaging, N=Neutral and B = Benign]

Gene	Mutation type	Protein Change	Variant information	Frequency of this mutation (%)	SIFT	Polyphen2	Mutation Taster	PROVEAN
BCOR	nonsynonymous	V926E	Novel	2.08	D	D	D	N
FAT4	**nonsynonymous**	**G2606A**	**Novel**	**2.08**	**D**	**D**	**D**	**D**
FBXW7	Startloss	C2_M36del	Novel	2.08	T	B	D	N
FAT4	nonsynonymous	A339T	Novel	2.08	D	D	D	N
FAT4	nonsynonymous	A339V	Novel	2.08	T	D	D	N
MACF1	nonsynonymous	K853T	Novel	2.08	D	B	D	D
ATN1	nonsynonymous	V906M	Reported	2.08	T	P	D	N
BRCA1	nonsynonymous	I1129V	Novel	2.08	T	B	N	N
STK11	nonsynonymous	T363I	Reported	2.08	T	B	D	N
ALK	nonsynonymous	E862Q	Novel	2.08	T	D	D	N
APC	nonsynonymous	E1216D	Novel	2.08	T	B	N	N
PMS2	**nonsynonymous**	**L114F**	**Reported**	**2.08**	**D**	**D**	**D**	**D**
BRCA2	nonsynonymous	I1846V	Reported	2.08	D	B	N	N
BRCA2	nonsynonymous	R2341H	Novel	2.08	D	D	N	N

Table 18: Germline variant list obtained from Targeted re-sequencing. (Bold rows are showing the pathogenic mutations.)

[D=Deleterious, T=Tolerated, P-Possibly damaging, Damaging, N=Neutral and B = Benign]

Gene	Mutation Type	Protein Change	Variant information	Frequency of this mutation (%)
ATN1	frameshift insertion	Q486Hfs*53	Reported	2.08
ATN1	Inframe deletion	Q502del	Reported	6.25
ATN1	Inframe insertion	Q502_H503insQ	Reported	2.08
ATN1	Inframe insertion	Q502_H503insQQQQ	Reported	4.17
ATN1	Inframe deletion	Q500_Q502del	Reported	4.17
ATN1	Inframe deletion	Q498_Q502del	Reported	16.67
ATN1	Inframe deletion	Q496_Q502del	Novel	2.08
BRCA2	Inframe deletion	L1740_S1741del	Reported	6.25
MAP3K4	**Inframe deletion**	**A652del**	**Reported**	**91.67**

Table 19: Germline indels list found in this study (Bold row is showing variants occurred with higher frequency in this study)

In this study, 9 indels were found (Table 19). Among them, *ATNI* exhibited one novel In-frame deletion (L1740_S1741del) with a 6.25% frequency. *MAP3K4* gene exhibited one in-frame deletion (A652del) in 92% frequency which was the highest occurrence frequency of variants in this study.

Whole exome sequencing

Whole exome sequencing for 37 samples was performed and 50875 variants were obtained from the variant caller. After applying all the filters, 12144 variants were obtained as the discovery set. Most of the variants were missense type and C>T transition substitution type was more prevalent compared to other signatures.

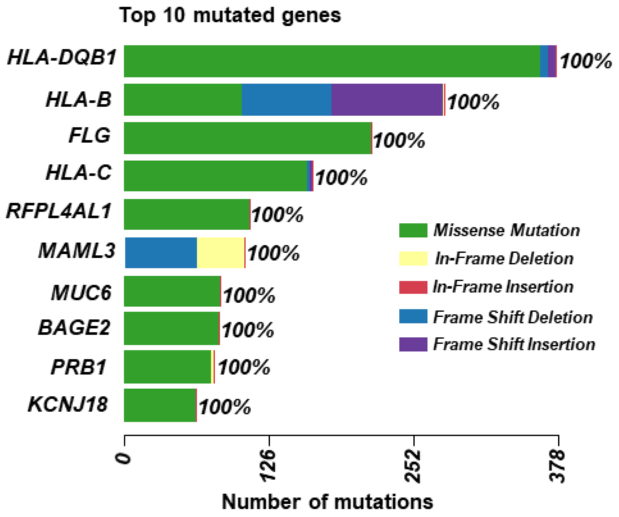


Figure 37: List of top ten genes mutated in germline whole-exome data

The top ten mutated genes were *HLA-DRB1, HLAB, FLG, HLAC, RFPL4AL1, MAML3, MUC6, BAGE5, PRB1* and *KCNJ12* (Figure 37). Out of the top ten genes, *HLA-DRB1, HLAB* and *HLAC* play a key role in the immune system. This result is indicating that the immune-related genes were mutated frequently in this population.

In whole-exome sequencing data, 34 genes were mutated frequently in more than 90% of cases (Figure 38). In the case of the germline, analysis variants were considered as polymorphism, if they occur in higher frequency. Most of them were polymorphism but some of them were pathogenic variants and those were not present in healthy controls samples.

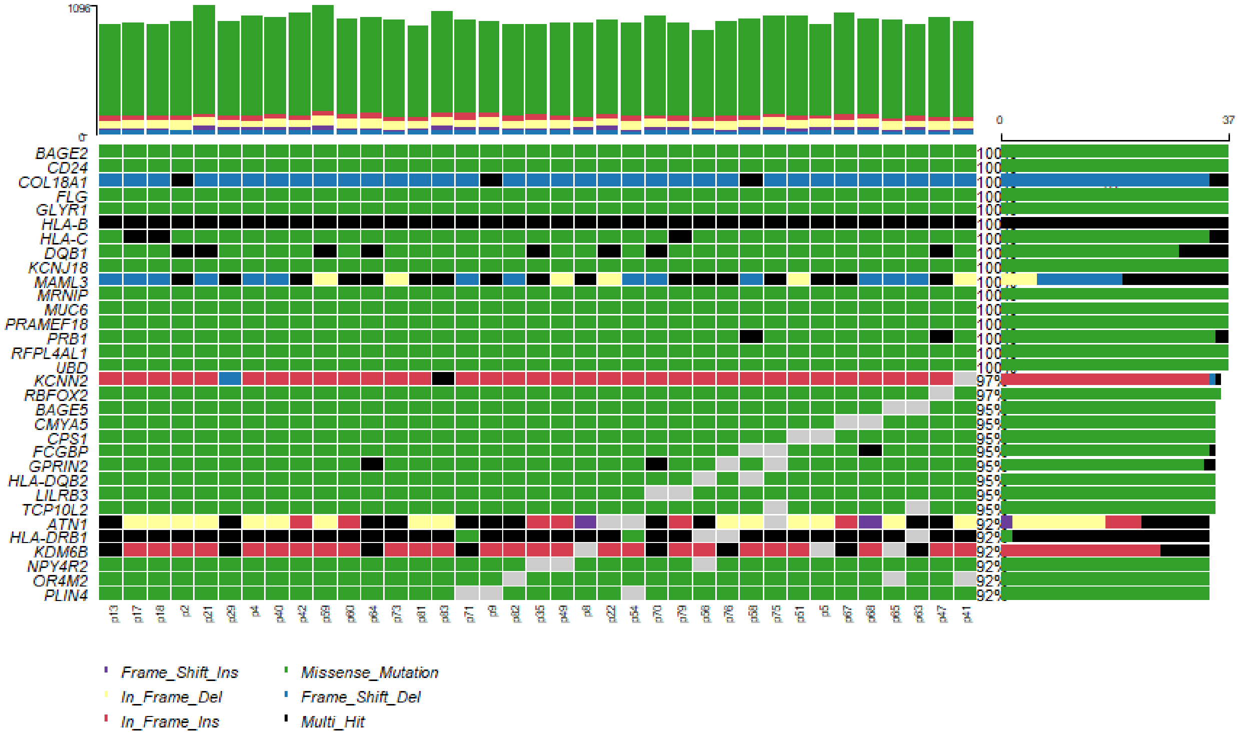


Figure 38: List of frequently mutated genes in more than 90% of samples

Among the 34 genes, six genes *COL18A1, KCNJ18, CMYA5, FCGBP, HLA-DRB1* and *OR4M2* exhibited 13 pathogenic mutations with high frequency (Table 20). Pathogenicity prediction was done by SIFT, Polyphen 2 and Mutation Taster. The variant are considered as pathogenic if it was predicted as pathogenic or damaging in all the tools. G1072R (*COL18A1* with 2.7% frequency), E430G (*KCNJ18* with 100% frequency), Y3957H, T3515N & F3628S (*CMYA5* with 2.7% frequency in each case), G3871R & C3904F (*FCGBP* with 5.4% and 2.7% frequency, respectively), T80R, D70N, Y152C, V188M & G197A (*HLA-DRB1* with 2.7%, 8.10%, 16.21%, 2,7% and 2.7% frequency, respectively) and S202C (*OR4M2* with 18.91%

Patient_ID	Ref	Alt	Genes	cDNA position	AA change	SIFT	Polyphen 2	Mutation Taster	1000 genome	NCBi-rsID
p58	G	A	COL18A1	c.G3214A	p.G1072R	Damaging	Damaging	Damaging	Reported	rs576719084
All samples	A	G	KCNJ18	c.A1289G	p.E430G	Damaging	Damaging	Damaging	Not-reported	rs5021699
p22	T	C	CMYA5	c.T11869C	p.Y3957H	Damaging	Damaging	Damaging	Reported	rs117835440
p40	C	A	CMYA5	c.C10544A	p.T3515N	Damaging	Damaging	Damaging	Not-reported	rs372352332
p60	**T**	**C**	**CMYA5**	**c.T10883C**	**p.F3628S**	**Damaging**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**
p83, p21	C	T	FCGBP	c.G11611A	p.G3871R	Damaging	Damaging	Damaging	Not-reported	rs4802062
p67	**C**	**A**	**FCGBP**	**c.G11711T**	**p.C3904F**	**Damaging**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**
p21	G	C	HLA-DRB1	c.C239G	p.T80R	Damaging	Damaging	Damaging	Not-reported	rs1059582
p2, p35, p59	C	T	HLA-DRB1	c.G208A	p.D70N	Damaging	Damaging	Damaging	Not-reported	rs56158521
p2, p21, p35, p59, p70, p83	T	C	HLA-DRB1	c.A455G	p.Y152C	Damaging	Damaging	Damaging	Not-reported	rs112796209
p59	C	T	HLA-DRB1	c.G562A	p.V188M	Damaging	Damaging	Damaging	Not-reported	rs112116022
p59	C	G	HLA-DRB1	c.G590C	p.G197A	Damaging	Damaging	Damaging	Not-reported	rs2308775
p13, p22, p56, p59, p68, p75, p76	A	T	OR4M2	c.A604T	p.S202C	Damaging	Damaging	Damaging	Not-reported	rs79101657

Table 20: List of missense pathogenic mutations found in 90% of the samples (Bold rows are showing novel variants)

Frequency) were found as pathogenic variants in this study (Table 20). E430G variant is very important for this population as it was present in all the patients. Among them, two variants, E3628S and C3904F were novel mutations.

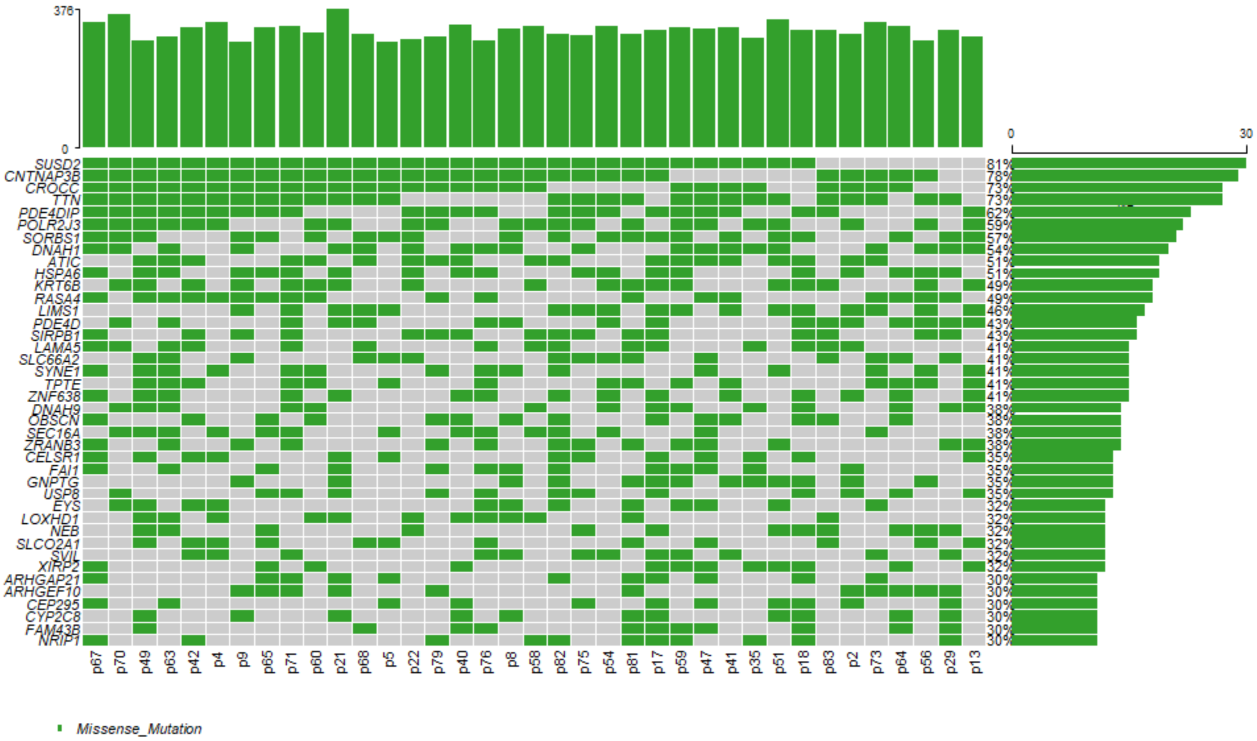


Figure 39: List of novel mutated genes found associated with GC in this study

In this study, 40 novel mutated genes for association with Gastric Cancer were identified (Figure 39). This is the first report of new germline mutated genes in association with GC. These genes might be responsible for developing Gastric Cancer in this population. The genes are as follows: *SUSD2, CNTNAP38, TTN, PDE4DIP, POLR2J3, SORBS1, DNAH1, ATIC, HSPA6, KRT6B, RASA4, LIMS1, PDE4D, SIRPB1, LAMA5, SLC66A2, SYNE1, TPTE, ZNF638, DNAH9, OBSCN, SEC16A, ZRANB3, CELSR1, FAI1, GNPTG, USP8, EYS, LOXHD1, NEB, SLCO2A1, SVIL, XIRP2, ARHGAP21, ARHGEF10, CEP295, CYP2C8, FAM43B* and *NRIP1* (Figure 39)

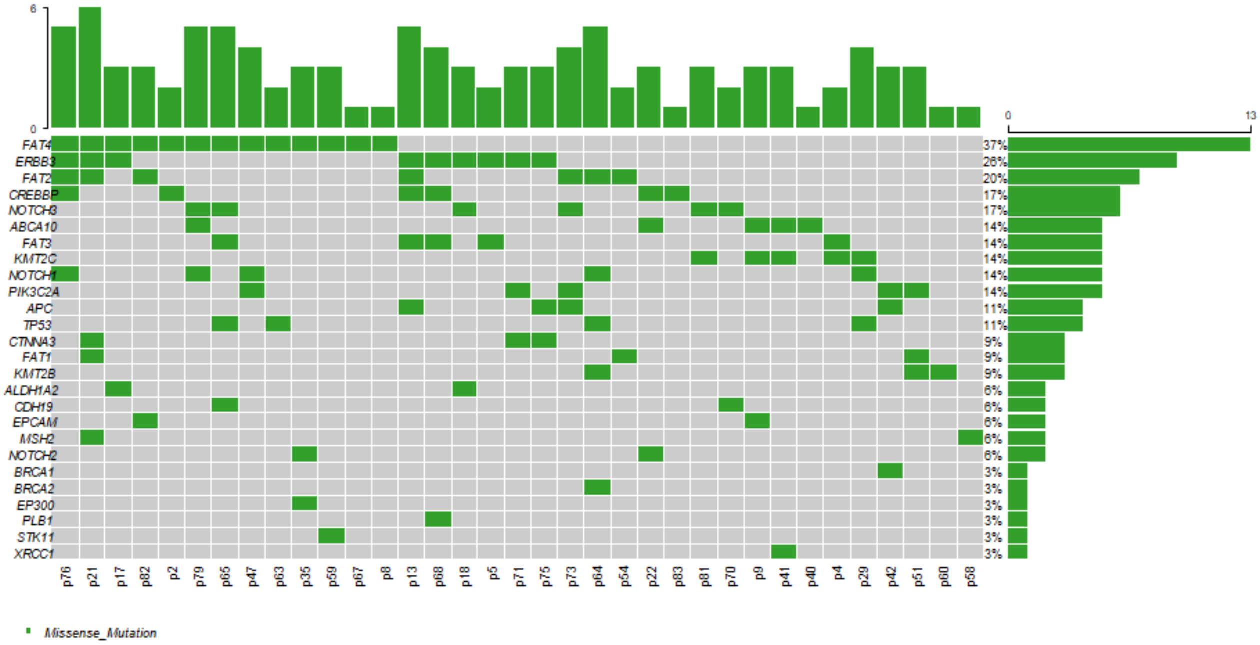


Figure 40: List of cancer-related genes reported for all type of cancer including GC

About 26 commonly mutated cancer-related genes were derived in this study. Most of these genes were also found to be mutated in targeted germline data. The genes are as follows: *FAT4, ERBB3, FAT2, CREBBP, NOTCH3, ABCA10, FAT3, KMT2C, NOTCH1, PIK3C2A, APC, TP53, CTNNA3, FAT1, KMT2B, ALDH1A2, CDH19, EPCAM, MSH2, NOTCH2, BRCA1, BRCA2, EP300, PLB1, STK11* and *XRCC1* (Figure 40). The majority of the variants were reported for Hereditary predisposing syndrome, Hereditary ovarian and Breast Cancer syndrome and LiFraumeni syndrome and Familial adenomatous polyposis. RTK-RAS, Hippo, Wnt, PI3K, TP53 and NOTCH pathway gene alterations were obtained both in targeted resequencing data as well as whole-exome sequence data.

Patient_ID	Ref	Alt	Genes	Mutation Type	cDNA position	AA change	SIFT	Polyphen2	Mutation Taster	1000 genome	NCBI-rsID
p35	C	T	EP300	missense	c.C2461T	p.P821S	Tolerate	Damaging	Damaging	Reported	rs764698803
p83	G	A	CREBBP	missense	c.C6991T	p.P2331S	Tolerate	Damaging	Damaging	Reported	rs745770513
p2, P22, P68, P76	G	T	CREBBP	missense	c.C1537A	p.L513I	Tolerate	Probably damaging	Damaging	Reported	rs61753381
p51	**C**	**G**	**FAT1**	**missense**	**c.G9158C**	**p.R3053P**	**Tolerate**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**
p54	G	A	FAT1	missense	c.C9935T	p.T3312M	Tolerate	Damaging	Damaging	Reported	rs368431115
p21	C	T	FAT1	missense	c.G3874A	p.E1292K	Tolerate	Damaging	Damaging	Reported	rs184443677
p65	**A**	**G**	**FAT3**	**missense**	**c.A4850G**	**p.E1617G**	**Tolerate**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**
p13	C	G	FAT3	missense	c.C6197G	p.A2066G	Tolerate	Damaging	Damaging	Reported	rs763919301
p4	**G**	**C**	**FAT3**	**missense**	**c.G7817C**	**p.R2606T**	**Tolerate**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**
p64	T	C	FAT2	missense	c.A13042G	p.M4348V	Tolerate	Damaging	Damaging	Reported	rs768452355
p82	**G**	**T**	**FAT2**	**missense**	**c.C12581A**	**p.P4194H**	**Damaging**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**
p13, P21, P54, P73, P76	C	T	FAT2	missense	c.G4960A	p.D1654N	Tolerate	Probably damaging	Damaging	Reported	rs150831986
p79	C	T	NOTCH1	missense	c.G1747A	p.G583S	Tolerate	Probably damaging	Damaging	Reported	rs757066417
p76	**C**	**A**	**NOTCH1**	**missense**	**c.G5168T**	**p.S1723I**	**Damaging**	**Probably damaging**	**Damaging**	**Not-reported**	**Not-reported**
p47	**T**	**C**	**NOTCH1**	**missense**	**c.A5960G**	**p.D1987G**	**Damaging**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**
p22	C	T	NOTCH2	missense	c.G710A	p.R237Q	Tolerate	Probably damaging	Damaging	Reported	rs146498360
p47	**T**	**C**	**NOTCH1**	**missense**	**c.A5960G**	**p.D1987G**	**Damaging**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**

.

Patient_ID	Ref	Alt	Genes	Mutation Type	cDNA position	AA change	SIFT	Polyphen2	Mutation Taster	1000 genome	NCBI-rsID
p22	C	T	NOTCH2	missense	c.G710A	p.R237Q	Tolerate	Probably damaging	Damaging	Reported	rs146498360
p35	T	A	NOTCH2	missense	c.A5065T	p.I1689F	Tolerate	Probably damaging	Damaging	Reported	rs60854092
p47	**C**	**A**	**NOTCH3**	**stopgain**	**c.G3961T**	**p.G1321X**	**Damaging**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**
p70	G	T	NOTCH3	missense	c.C4552A	p.L1518M	Tolerate	Damaging	Damaging	Reported	rs141320511
p79, P73	G	T	NOTCH3	missense	c.C4552A	p.L1518M	Tolerate	Damaging	Damaging	Reported	rs141320511
p65	C	T	NOTCH3	missense	c.G5510A	p.R1837H	Damaging	Damaging	Damaging	Reported	rs138265894
p81	G	A	NOTCH3	missense	c.C6097T	p.P2033S	Damaging	Probably damaging	Damaging	Reported	rs375213868
p47, p51, P73	T	C	PIK3C2A	missense	c.A3866G	p.N1289S	Tolerate	Probably damaging	Damaging	Reported	rs139012235
p71	T	C	PIK3C2A	missense	c.A1010G	p.Q337R	Tolerate	Probably damaging	Damaging	Reported	rs143829156
p13, 73, 75	C	T	APC	missense	c.C3728T	p.T1243I	.	Probably damaging	Damaging	Reported	rs1064794636
p70	**G**	**T**	**CDH1**	**missense**	**c.C1876A**	**p.L626I**	**Tolerate**	**Probably damaging**	**Damaging**	**Not-reported**	**Not-reported**
p4	**G**	**C**	**KMT2C**	**missense**	**c.C2765G**	**p.P922R**	**Tolerate**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**
p9, p29, p81	C	A	KMT2C	missense	c.G5053T	p.A1685S	Tolerate	Damaging	Damaging	Reported	rs145848316
p41	A	C	KMT2C	missense	c.T13054G	p.W4352G	Damaging	Damaging	Damaging	Reported	rs777612235
p21	C	T	MSH2	missense	c.C23T	p.T8M	Tolerate	Damaging	Damaging	Reported	rs17217716

Patient_ID	Ref	Alt	Genes	Mutation Type	cDNA position	AA change	SIFT	Polyphen2	Mutation Taster	1000 genome	NCBI-rsID
p58	G	A	MSH2	missense	c.G2197A	p.A733T	Damaging	Damaging	Damaging	Reported	rs772662439
p17	**C**	**T**	**ALDH1A2**	**missense**	**c.G196A**	**p.V66M**	**Damaging**	**Probably damaging**	**Damaging**	**Not-reported**	**Not-reported**
p41	C	T	XRCC1	missense	c.G361A	p.A121T	Tolerate	Damaging	Damaging	Reported	rs138284081
p51, p60, p64	C	G	KMT2B	missense	c.C1760G	p.P587R	Tolerate	Damaging	Damaging	Reported	rs2242519
p68	C	T	PLB1	missense	c.C878T	p.P293L	Damaging	Damaging	Damaging	Reported	rs779907565
p75, p21	C	A	CTNNA3	missense	c.G1549T	p.D517Y	Damaging	Damaging	Damaging	Reported	rs373151978
p5, p68, p71, p75	G	A	ERBB3	missense	c.G2900A	p.R967K	Tolerate	Damaging	Damaging	Reported	rs561787077
p21	C	T	MSH2	missense	c.C23T	p.T8M	Tolerate	Damaging	Damaging	Reported	rs17217716
p58	G	A	MSH2	missense	c.G2197A	p.A733T	Damaging	Damaging	Damaging	Reported	rs772662439
p47	**C**	**T**	**FAT4**	**missense**	**c.C1016T**	**p.A339V**	**Tolerate**	**Damaging**	**Damaging**	**Not-reported**	**Not-reported**
p47, p63, p67, p76	G	A	FAT4	missense	c.G1015A	p.A339T	Damaging	Damaging	Damaging	Not-reported	Not-reported
p17	G	T	FAT4	missense	c.G850T	p.D284Y	Damaging	Damaging	Damaging	Not-reported	Not-reported
p59	G	A	FAT4	missense	c.G13540A	p.A4514T	Tolerate	Damaging	Damaging	Reported	rs369929089
p65. p79	C	T	FAT4	missense	c.C11693	p.A3898V	Tolerate	Damaging	Damaging	Reported	rs138275098

					T						
p59	G	A	FAT4	missense	c.G9290A	p.S3097N	Tolerate	Probably damaging	Damaging	Reported	rs766108453
p9, p22, p40, p41, p79	A	G	ABCA10	missense	c.T1988C	p.L663S	Damaging	Probably damaging	Damaging	Reported	rs138284687
p82	A	C	EPCAM	missense	c.A331C	p.N111H	Damaging	Damaging	Damaging	Not-reported	Not-reported
p9	G	C	EPCAM	missense	c.G458C	p.R153T	Tolerate	Probably damaging	Damaging	Reported	rs189732445
p42	A	G	BRCA1	missense	c.T4948C	p.C1650R	Damaging	Damaging	Damaging	Reported	rs80356993
p64	G	A	BRCA2	missense	c.G8972A	p.R2991H	Damaging	Damaging	Damaging	Reported	rs80359150

Table 21: List of Germline pathogenic variants found in whole exome study (Bold rows denote novel variants)

Fifty three (53) pathogenic germline heterogeneous variants were identified in WES analysis and out of them 14 were novel mutations. R3053P (2.7%) of *FAT1*, E1617G (2.7%) & R2606T (2.7%) of *FAT3*, P4194H (2.7%) of *FAT2*, S1723I (2.7%) & D1987G (2.7%) of *NOTCH1*, G3961T (2.7%) of *NOTCH3*, C1876A (2.7%) of *CDH1*, C2765G (2.7%) of *KMT2C*, V66M (2.7%) of *ALDH1A2*, A339V (2.7%), A339T (10.81%) & D284Y (2.7%) of *FAT4*, and N111H (2.7%) of *EPCAM* were the novel variants (Table 21).

Gene	Variants
MACF1	N2198Y
TP53	E319V
EGFR	K253R
MACF1	N2544S
KMT2B	P587R
APC	T1261I

Table 22: List of variants found in both targeted resequencing and WES

Six germline variants obtained from targeted data were also present in Whole exome data. These are the gold standard germline variants that were present in both types of sequencing data. These variants (*N2198Y & N2544S of MACF1, E319V of TP53, K253R of EGFR, P587R of KMT2B* and T1261I of *APC*) might play important roles in developing GC in this population (Table 22).

Copy number variation analysis

The copy number analysis of *TP53*, *HER2* mutated samples and one sample without mutation in both the genes was compared by using adjacent normal and tumor tissue. CNV analysis was performed in 17 patients and 35.29% of samples had a variation for the *HER2* gene. There was a gain for the *HER2* copy number in five samples and in one sample it was a loss in copy number. *TP53* copy number was altered in 23.52% of cases, among them there was a gain in copy number in 3 samples and one sample exhibited copy number loss (Table 23). The mutation was not significantly associated with copy number alterations, but only the missense (Y781C) mutation was responsible for *HER2* copy number gain in this study (Table 23).

Patient ID	Her2 (AN)	Her2 (Tumor)	Her2 (Tumor) / Her2 (AN)	CNV Type	TP53 (AN)	TP53 (Tumor)	TP53 (AN)/ TP53 (Tumor)	CNV Type	Mutation
P1	2.17	1.97	0.91	Normal	1.69	1.88	1.11	Normal	ID/PHHERC177_
P10	2.09	2.03	0.97	Normal	1.84	1.93	1.05	Normal	MS/S310F
P16	2.09	2.13	1.02	Normal	1.85	1.67	0.90	Normal	MS/L194R
P19	1.96	3.03	1.55	Gain	1.93	1.76	0.91	Normal	MS/G266V
P20	2.06	1.98	0.96	Normal	1.84	1.95	1.06	Normal	MS/G245S
P23	2.14	1.95	0.91	Normal	1.79	1.549	0.87	Normal	SG/R306*
P25	2.03	2.01	0.99	Normal	1.95	1.99	1.02	Normal	MS/E358V
P27	1.97	2.14	1.09	Normal	1.88	1.92	1.02	Normal	MS/S215N
P28	2.03	2.38	1.17	Gain	1.99	1.97	0.99	Normal	MS/R175H
P31	1.87	2.47	1.32	Gain	1.95	1.83	0.94	Normal	MS/R273C
P38	1.98	2.01	1.02	Normal	0.621	1.9	3.06	Gain	MS/D769Y
P46	2.24	2.55	1.14	Normal	1.84	1.64	0.89	Normal	MS/L137Q
P50	2.32	2.19	0.94	Normal	0.93	1.35	1.45	Gain	MS/V842I
P51	0.86	1.203	1.40	Gain	1.91	1.95	1.02	Normal	FD/K382NSCSR QKGLTQT
P52	2.22	1.87	0.84	Loss	0.91	1.63	1.79	Gain	
P55	2.98	3.89	1.31	Gain	1.18	0.79	0.67	Loss	**MS/Y781C**

Table 23: Copy number analysis of GC samples.

(The red-colored sample was ERBB2 mutated cases, black colored samples were TP53 mutated patient and the bold patient no. was cases of without having mutations in both the genes)

Protein expression studies using IHC

In the case of the sample without any mutation, a negative relation between ERBB2 and TP53 copy number change was observed. There was a gain in *ERBB2* and a loss for *TP53* CNV. This negative correlation was found in 12% cases and 24% cases there was a gain in copy number for the *ERBB2* gene and in the case of TP53, there was no change in copy number (Table 23). Again, in 12% of cases, there was a gain in TP53 copy number while *ERBB2* copy number remained unchanged (Table 23). There were no cases found where both the genes having a gain in copy number or a loss in copy number at a time. In this study, a negative correlation with oncogene *ERBB2* and tumor suppressor gene TP53 was observed.

Immunohistochemistry staining for 24 samples with BAX antibody was done and the expression of BAX was low or moderate in adjacent normal samples (Figure 41A). In the case of the Tumor, there was a strong expression for BAX protein. The antibody expression was scored as negative, low, moderated and strong in this study (Figure 41B).

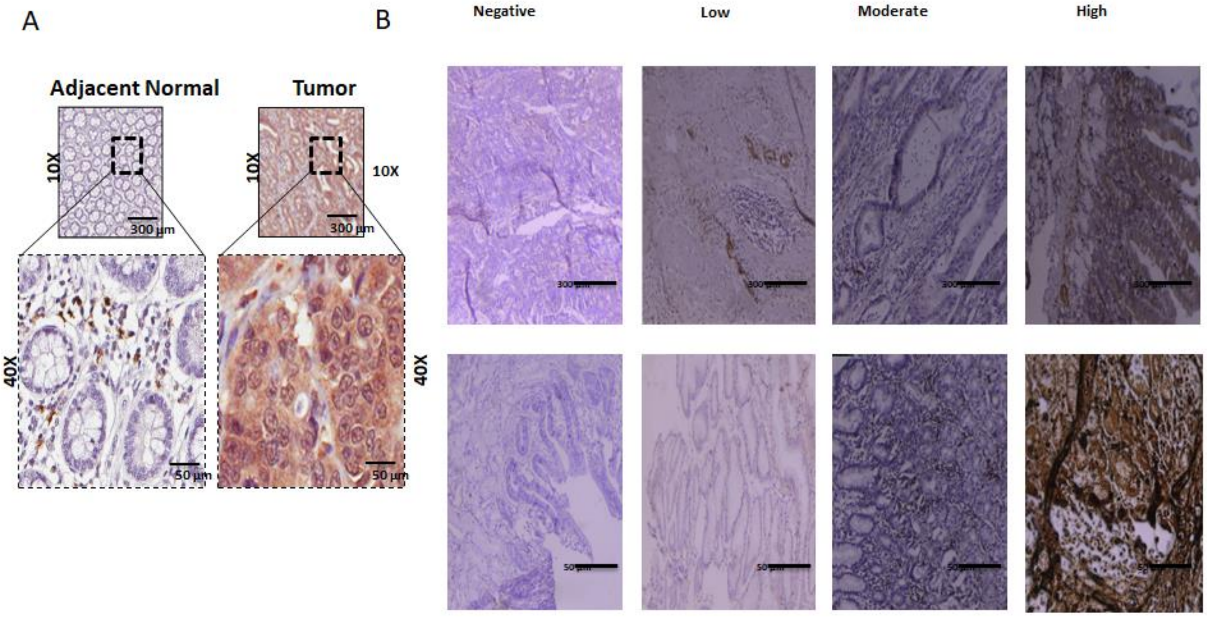


Figure 41: (A) IHC staining picture of adjacent normal and tumor case. (B) Staining picture of negative, low, moderate and high expression of BAX in 10X and 40X.

There was a higher expression of BAX protein in tumor cases compared to adjacent normal tissues (Figure 42A). Statistical comparison between expression data and clinical factors (Stage and Pathogen genotypes) was performed and in the case of the stage, the expression of BAX was higher in Stages I, II and III but unexpectedly the expression of BAX was low for Stage IV samples (Figure 42B). Interestingly BAX expression was significantly (*p*-value = 0.05) associated with EBV (+) GC cases (Figure 42C). BAX expression was not associated with *H. pylori*-infected GC cases (Figure 42D). EBV infections might be affecting apoptotic pathways of the patient as BAX is a protein having an important role in the apoptosis pathway. In survival analysis, the patient group with high BAX expression was at-risk group (HR: 1.36; 95%CI: 0.26-6.87; p-value = 0.37) compared to low expression cases, though it was not significant (Figure 42E).

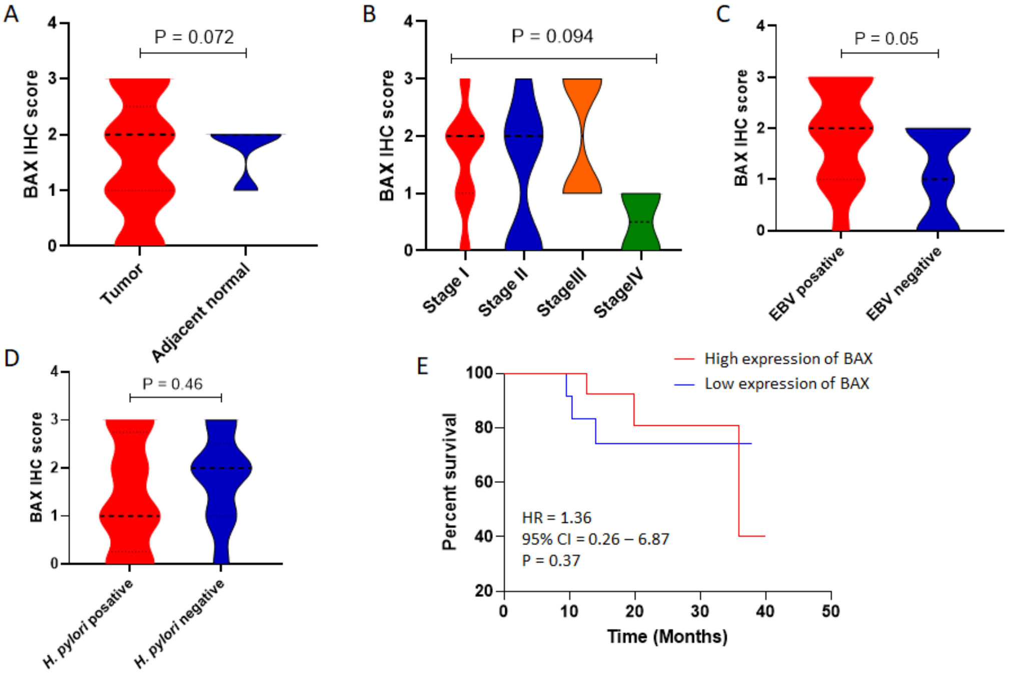


Figure 42: Comparison of BAX expression between (A) Tumor and adjacent normal, (B) Stage I, II, III and IV, (C) EBV (+) and (-) cases, (D) *H. pylori* (+) and (-) cases, (E)

survival analysis between High BAX expression cases and low BAX expression patient group.

In this study out of 13 cases, 38.46% cases were showing positive expression for TP53 46.15% cases were showing positive expression of HER2 and ERCC1 protein. Among six positive cases of HER2 showing 5 were negative expressions for TP53, only in one case both were positively expressed. This data was supporting the copy number variation result of this study. In 30.76% there was a positive correlation between TP53 and ERCC1 protein expression. Figure 43 represents the low, moderate and high expression of proteins.

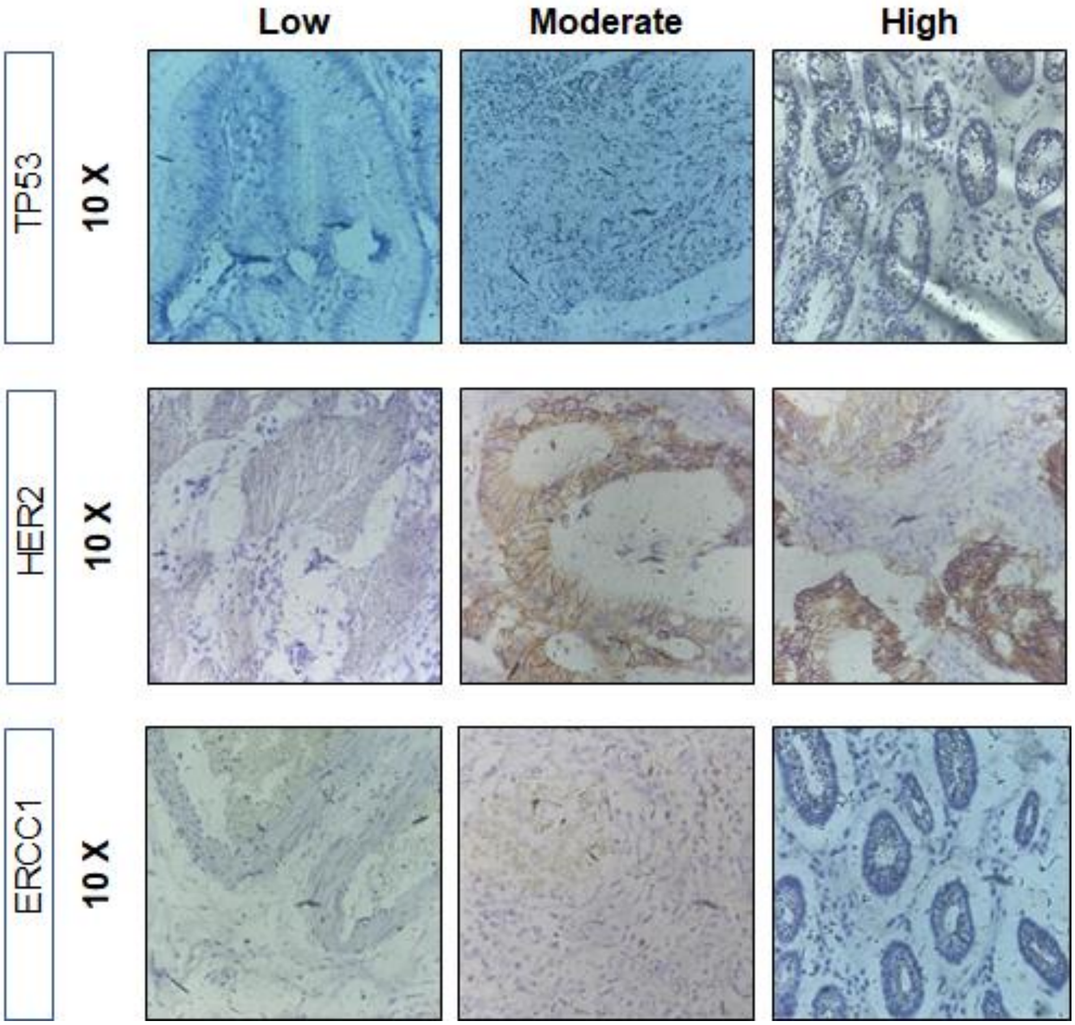


Figure 43: Staining picture of low, moderate and high expression of TP53, HER2 and ERCC1.

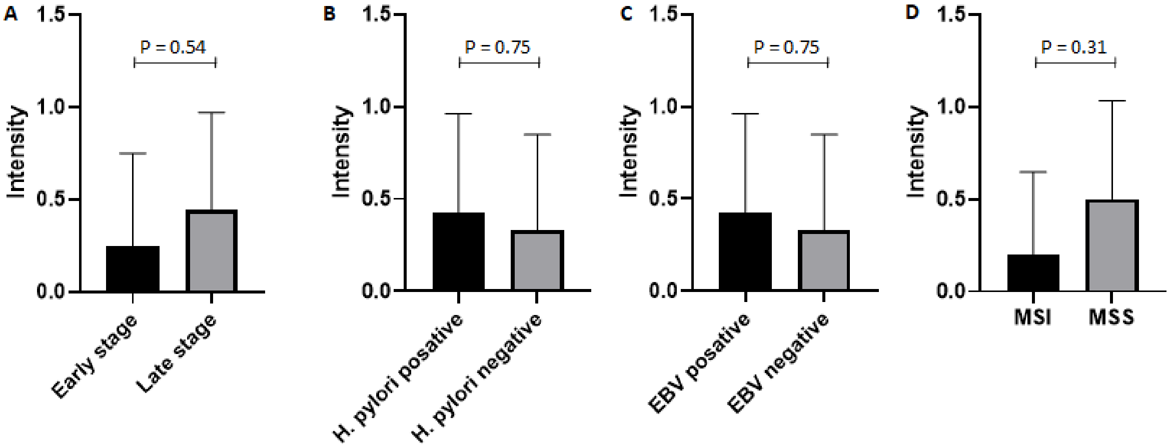


Figure 44: Comparison of TP53 expression between (A)Early and Late-stage, (B) *H.pylori* (+) and (-) cases (C) EBV (+) and (-) cases, (D), MSI and MSS cases.

TP53 expression was higher in late-stage compared to early-stage GC cases (Figure 44A). TP53 expression was similar in the case of both the pathogen (Figure 44B and 44C). In genomically stable cases, TP53 expression was higher compared to MSI-associated GC cases (Figure 44D).

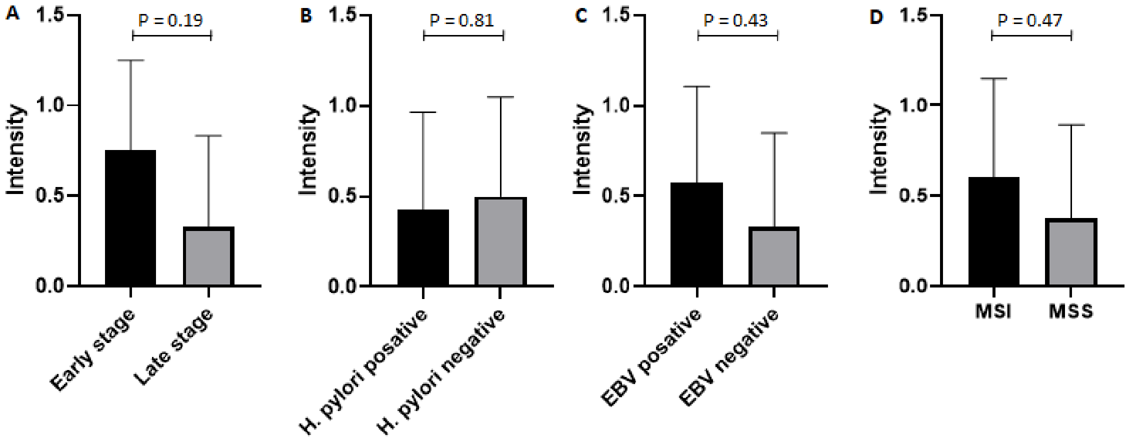


Figure 45: Comparison of HER2 expression between (A)Early and Late-stage, (B) *H.pylori* (+) and (-) cases (C) EBV (+) and (-) cases, (D), MSI and MSS cases.

HER2 expression was higher in early-stage compared to late-stage GC cases (Figure 45A). HER2 expression does not differ in *H. pylori* (+) and (-) cases (Figure 45B), while HER2 expression was higher in EBV (+) GC cases (Figure 45C). In MSI cases HER2 expression was higher compared to MSS-associated GC cases (Figure 45D).

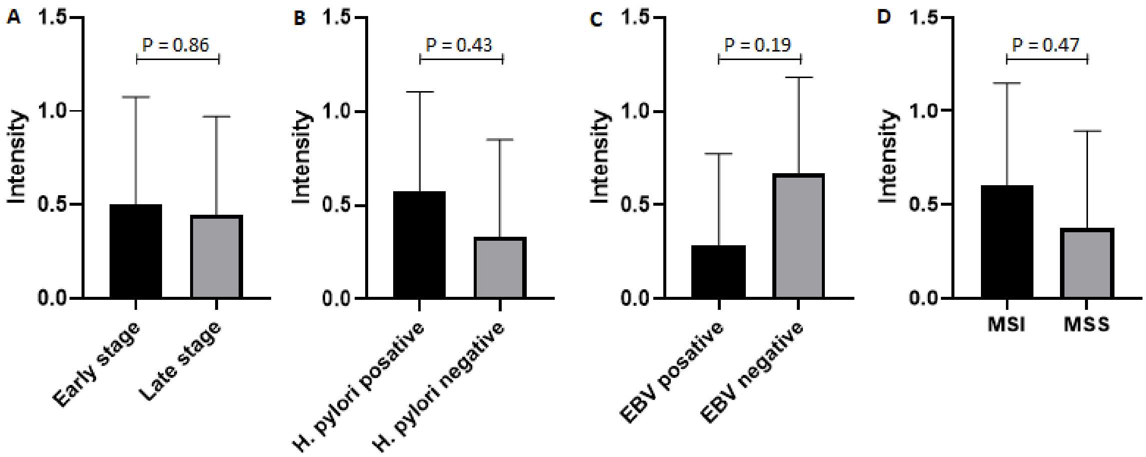


Figure 46: Comparison of ERCC1 expression between (A) Early and Late stage, (B) *H.pylori* (+) and (-) cases (C) EBV (+) and (-) cases, (D), MSI and MSS cases.

ERCC1 expression was similar in early and advanced stage GC (Figure 46A). ERCC1 expression positive cases were higher in *H. pylori* (+) cases (Figure 46B), while ERCC1 expression positive cases were in EBV (+) GC cases (Figure 46C). ERCC1 expression positive cases were higher in MSI cases compare to MSS-associated GC cases (Figure 46D).

Biomarker genes in other cancer and chronic diseases

Finally, a literature survey was done to find out possible biomarkers of GC development. The genes which were mutated in 90% of cases and among novel genes 26 were identified as predicting, a diagnostic, prognostic and therapeutic biomarker in a different type of cancer (Table 24). This might be used as a biomarker for gastric cancer also.

Gene	Marker information	Reference
HLADRB1	Marker for detecting lapatinib-induced liver injury (induced liver toxicity during treatment of early-stage breast cancer patients with lapatinib in combination with trastuzumab)	Spraggs et al. 2018
HLA-C	Biomarker in hepatocellular carcinoma (HCC)	Wang et al. 2019
HLA-B	A diagnostic marker for Lung cancer	Liu et al. 2019
MAML3	Therapeutic marker of Gastric Cancer	Li et al. 2017
MUC6	Gastric marker	Barau et al. 2006
CMAY5	Novel oncogene in Breast cancer	Colaprico et al. 2020
FCGBP	Predicting marker of HPV (+) Head and neck squamous cell carcinoma (HNSCC)	Wang et al. 2017
SUSD2	Tumor suppressor gene associated with Renal cell carcinoma and Lung cancer	Cheng et al. 2016
TTN	Therapeutic marker for immune checkpoint blockade immune therapy	Jia et al. 2019
SORBS1	The therapeutic marker which can suppress metastasis in cancer	Song et al. 2017
ATIC	Biomarker in hepatocellular carcinoma	Heo et al. 2020
HSPA6	Diagnostic biomarkers in cancers	Taha et al. 2019
KRT6B	Potential biomarkers in non-small cell lung cancer.	Xiao et al. 2017
LIMS1	The therapeutic target for cancer treatment in pancreatic cancer	Huang et al. 2019
PDE4D	Prognostic biomarker in pancreatic ductal adenocarcinoma	Liu et al. 2019
SIRPB1	Biomarker in Prostate cancer	Song et al. 2020
LAMA5	Overexpression in colorectal cancer cells.	Gordon et al. 2019
ZNF638	Diagnostic biomarker in colorectal cancer	O'Reilly et al. 2015
OBSCN	Frequently mutated in Breast and colorectal cancer	Yang et al. 2020
CELSR1	Biomarker in Breast cancer	Terkelsen et al. 2020
USP8	Predictive biomarker in Cervical Squamous Cell Carcinoma	Yan et al. 2018
SLCO2A1	molecular markers to determine malignant follicular thyroid cancer (FTC) by comparing with benign follicular thyroid adenoma (FTA)	Zhu et al. 2015
XIRP2	Prognostic marker in colon cancer	Zhou et al. 2019
ARHGEF10	Its expression can stimulate gastric tumorigenesis	Wang et al. 2020
CYP2C8	Prognostic biomarkers in hepatocellular carcinoma	Wang et al. 2019

Table 24: List of biomarker genes in other cancer and chronic diseases

Discussion

This study to the best of our knowledge is the first case-control study, designed to assess the detailed epidemiological risk factors along with the potential role of *EBV / H. pylori* infections, MMR gene status and Genomics and the prognosis of GC patients in Northeast India. The GC patients from the Mizo population exhibited higher pathogen-associated GC cases. Smoked food, extra salt consumption, smoking and alcohol are the major risk factors for developing GC and obese persons are at risk for developing GC. EBV infection was significantly associated with the unique risk factor (tuibur). *EBV* infection is a strong risk factor for GC and poor prognosis in this Indian high-risk population. TP53 mutations were also a significant factor for GC risk. This study has found that this population might be genetically predisposed with *MAP3K4, HLA-DRB1, HLAB, HLAC* and *KCNJ12* pathogenic mutations and novel genes are also found associated with GC which may develop GC by following a combination of pathways. The panel of *KMT2C, PMS2* and *CTNNA1* genes may be useful in predicting familial GC in the Mizo population.

About 75% of the GC patients were found between the age ranges of 40-69 years in this study, indicating that most of the patients were aged persons. Studies have reported that Gastric Cancer is an old age disease generally occurring after 40 years and old age is a significant risk factor for developing GC (Zali et al. 2011; Yusefi et al. 2018; Machlowska et al. 2020). Older aged people have more exposure to toxins and unhealthy food habits and some undesirable exposure like sunlight over time. Precancerous cells can develop at any time during the lifespan, but elderly people have weak immunity so it may not protect against the development of cancer cells. In this study, 66.25% of the GC patients were males and studies have reported that males have a twice higher risk of developing GC than females (GBD 2017 Stomach Cancer Collaborators 2019). In this population, gender has significant impact (6.25%; OR = 0.50; 95% CIs = 0.28 – 0.89; *p*-value = 0.019) for GC development. Male gastric cancer patients were found at significant risk group for developing GC than females in our study. *H. pylori* cause severe inflammation which can lead to Gastric cancer, male persons are found to be affected more with *H. pylori* infections (de Martel et al. 2006). The estrogen hormone can prevent the infection in women, studies have reported that increased level of estrogen is responsible for the decreasing risk of gastric cancer in females (Camargo et al. 2012)

In my findings, 73.75% of the tumor developed in the distal side of the stomach. Studies have reported that most of the GC cases found to be in the distal part and *H. pylori* present in gastric mucosa can develop a severe tissue injury in the distal stomach which may lead to Gastric Cancer (Hu et al. 2012; Piazuelo et al .2010). One study has reported that EBV (+) GC cases can be located in cardia with 58% frequency as well in non-cardia part with 42% frequency (Murphy et al. 2011). About 50% of the patients were in stage III indicating that most of the patients were diagnosed at advanced stage only and 32.5% of cases were found to be familial cases of cancer, with any type of cancer in the first degree relatives. One study has reported that 14.6% of persons have moderate risk and 7.7% have a strong risk of developing hereditary cancer in one population (Scheuner et al. 2010), which is lesser than this present study. Till date, the Mizo population practices endogamy and this might be a cause of high risk for Gastric Cancer in this population.

Excess body weight (BMI ⩾ 25) was found to be associated with an increased risk of GC (OR = 0.63; 95% CIs = 0.56 – 0.72; *p*-value = 0.0001) in this study. In multivariate analysis, obese persons with excess BMI were found to be associated with an increased risk of GC development (OR = 0.69, 95% CI = 0.60 – 0.79; p-value = 0.0001). Studies have reported that obese people are a risk for developing Gastric Cancer and a meta-analysis showed that excess BMI is a significant risk factor with gastric cancer development in the Asian population (Hirabayashia et al. 2019; Bae et al. 2020).

Consumption of extra salt, a dietary habit was found as a risk factor for developing GC in this study. Extra salt provides the possible condition for colonization of *H. pylori* by increasing the mucin level of the surface mucus in the stomach and studies have reported that *H. pylori* are a significant risk factor of stomach cancer (Fox et al. 1999, Kato et al. 2006). Expression of the carcinogenic A (CagA) gene in *H. pylori* can be significantly induced by extra salt and results in the alteration of epithelial cells to induce hypergastrinemia in GC patients (Wroblewski et al. 2010). Cell proliferation and endogenous mutation in epithelial cells can develop due to inflammatory response induced by extra salt intake (Wang et al. 2009) and moreover, extra salt intake can increase the susceptibility of epithelial cells to the carcinogenic effects of N-nitroso compounds and results in cell death (Tatematsu et al. 1975). After a wide range of literature surveys and based on the result of this study, it can be hypothesized that salt can promote gastric adenocarcinoma and the condition can be induced in combination with *H. pylori* infection. An optimum quantity of salt consumption is necessary to avoid gastric adenocarcinoma.

Another dietary factor, smoked food was found as a significant risk factor associated with GC in this population, as it was a common food habit in more than 60% of patients. Smoking in an oven or by burning of wood or charcoal and grilling method is used to cook smoked food (McDonald et al. 2015), and during this process antioxidants and antimicrobial properties along with carcinogenic chemicals like Polycyclic Aromatic Hydrocarbons (PAH) are produced (Varlet et al. 2006). Benzo[a]pyrene (BaP), a group I carcinogen, is a member of the PAH family found in smoked food and plays an important role in GC disease progression along with other cancers. Metabolic activation of cellular membrane cytochrome P450 can accumulate BaP in our body, which in turn can produce toxic byproducts and that will create DNA adducts by binding with DNA, leading to gene mutation (Rubin et al. 2001) which can alter the functional proteins through AhR/CYP450 pathway (Bersten et al. 2013). Studies have reported that in GC cell lines BaP can create proliferation by upregulation of MMP9 and c-myc expression (Wei et al. 2016). Several studies have reported that GC development is strongly associated with smoked-dried or processed food which is supporting our results (Ghatak et al. 2016; Phukan et al. 2005). This present study has observed a significant association between *H. pylori* infection associated GC and smoked food consumption. A strong association between extra salt consumption and *H. pylori* was found in several studies (Fox et al. 1999, Kato et al. 2006). Extra Salt-curing used to add flavor, allows the nitrites to penetrate the meat and is used to extract moisture from the food which in turn allows the smoke to penetrate more easily in the food. Generally, in the preparation of smoked meats, salt-curing is the first step. In Mizoram, it is common practice to make smoked foods rich in salt and turn it can create a favorable condition for *H. pylori* infection which will ultimately lead to the development of GC in this population. Further study with a large sample size is necessary to support this data. In the study, smoked food consumption was found to be a significant risk factor with EBV-infected GC cases. Smoking cigarettes and consumption of smoked food are significant contributing factors, for the development of carcinogenesis in GC patients, which might be amplified by the presence of EBV. It has been reported that there is a strong association of smoking with the risk of developing EBV-positive Hodgkin's lymphoma (Kamper-Jorgensen et al. 2013) and that tobacco, which is a risk factor for GC, may contain EBV-activating substances (Jia et al. 2012).

Two lifestyle factors, smoking and alcohol, were found to be the associated risk factors with GC development in this study. Studies have reported that smoking is a strong significant risk factor for developing GC (Bersten et al. 2013, Bonequi et al. 2013). The percentage of smokers (65%) was higher among GC patients, whereas more than 78% of the healthy controls were non-smokers in this study. Studies have reported that male smokers are a high-risk group for developing GC than women smokers (Li et al. 2019), while studies have shown that smoking is an independent risk factor for GC in both men and women (Nomura et al. 2012).

In this study, alcohol drinking was another lifestyle habit that was found as a significant risk factor with GC cases. The association between alcohol drinking and GC development is always a matter of conflict. Alcoholic beverages have been reported as a risk factor for developing several types of cancers by IARC (IARC monographs et al. 2010), but so far, there are no studies where the direct association of alcohol with GC has been established, as most of the studies were not showing consistent results. ALDH2 enzyme converts alcohol to acetate and any metabolic change in the enzyme activity will lead to the accumulation of Acetaldehyde (class I carcinogen). In Asian populations, there is a prevalence of particular mutations which can inactivate the ALDH2 enzyme (Ghosh et al. 2017). Studies have reported alcohol as an independent risk factor associated with GC in China (Moy et al. 2010). In a Korean study, it was reported that smoking had a significant association with GC in the upper third position of the stomach or cardia, and high alcohol consumption was associated with GC occurring in the distal part of the stomach (Sung et al. 2007). In our study, more than 36% of patients were drinkers whereas in healthy controls more than 97% of persons were non-drinkers. Figure 47 represents the hypothetical mechanism for developing GC with all the significant risk factors (after reviewing the current and previously published studies).

In this present study, another significant risk factor was tuibur consumption (tobacco-infused water) for the development of EBV-infected GC patients. A unique risk factor, tuibur is tobacco-infused water (smokeless tobacco) and it contains carbonyl compounds and polyaromatic hydrocarbons. Studies have reported that B-lymphocytes can be affected by smokeless tobacco (Malovichko et al. 2019), where latent EBV virus infection used to takes place (Hatton et al. 2014) and infected lymphocytes are responsible for tumorigenesis at a later stage. One study has reported that there is a positive association found between EBV type I and type II infections and smokeless tobacco (Jenson et al. 1999). One important aspect is EBV spreads by body fluids, like saliva. In rural villages, there are common practices of sharing the same tuibur bottle among the tuibur consumers for drinking and it can pass on to one healthy individual from an EBV infected person through saliva. As smoked food is prepared by exposing smoke and tuibur is prepared using the whole tobacco plant, it can also help to enhance the risk of EBV associated GC which needs to be explored by future studies

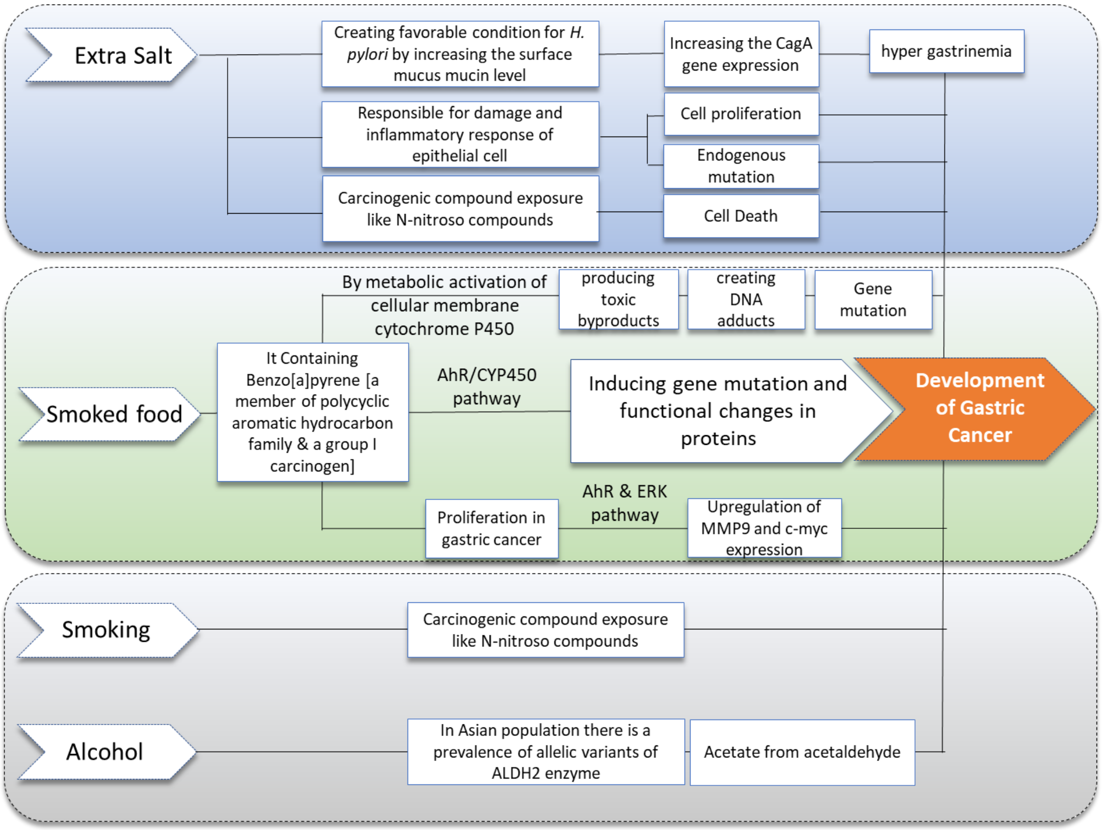

Extra Salt
Creating favorable condition for *H. pylori* by increasing the surface mucus mucin level
Increasing the CagA gene expression
hyper gastrinemia
Responsible for damage and inflammatory response of epithelial cell
Cell proliferation
Endogenous mutation
Carcinogenic compound exposure like N-nitroso compounds
Cell Death
By metabolic activation of cellular membrane cytochrome P450
producing toxic byproducts
creating DNA adducts
Gene mutation
Smoked food
It Containing Benzo[a]pyrene [a member of polycyclic aromatic hydrocarbon family & a group I carcinogen]
AhR/CYP450 pathway
Inducing gene mutation and functional changes in proteins
Development of Gastric Cancer
AhR & ERK pathway
Proliferation in gastric cancer
Upregulation of MMP9 and c-myc expression
Smoking
Carcinogenic compound exposure like N-nitroso compounds
Alcohol
In Asian population there is a prevalence of allelic variants of ALDH2 enzyme
Acetate from acetaldehyde

Figure 47: Review of a literature-based flow chart for understanding the mechanism of risk factors for developing GC

.

In this present study, two lifestyle factors, chewing tobacco and alcohol drinking were found as an associated significant risk factor with MMR deficient GC patients. Several studies have reported that MSI-H colorectal cancer cases are strongly associated with tobacco and alcohol drinking (Diergaarde et al. 2003; Eaton et al. 2005; Poynter et al. 2009; Warneke et al. 2003; Ghatak et al. 2016). In the present study, we found traditional tobacco (chewing tobacco) and alcohol drinking as a significant risk factor with MSI-associated GC.

In this study, we estimated the risk score of a panel of five epidemiological factors (BMI, extra salt, smoked food, alcohol consumption, and smoking habit) was used to estimate the chances of GC development (Figure 11B). In univariate analysis, the consumption of all the five factors achieved an independent association with GC risk whereas a significant *p*-value or association was not achieved for tuibur consumption in the multivariate model for GC risk. A risk score probability test between gastric cancer patients and healthy control with this panel of epidemiological factors was tested which can predict GC patients among the healthy controls. We found that the panel achieved a significant difference in risk score probability between gastric cancer patients and healthy controls (*p*-value < 0.0001, Figure 11C). This study achieved a panel of five epidemiology factors (BMI, Extra salt consumption, smoked food, drinking and smoking) with high AUC and sensitivity value (AUC = 0.946, sensitivity = 96.67, *p*-value < 0.0001) for detecting Gastric Cancer patients in early-stage (Figure 11E) for therapy implementation and prognosis.

In this study, 88.75% of cases were associated with pathogens indicating that EBV and *H. pylori* are playing a major role in developing Gastric cancer in this population. EBV-associated GC cases were found in 40% of patients. Studies have reported that worldwide 10-14% of cases were found to be EBV-associated GC (Shinozaki-Ushiku et al. 2015; Singh et al. 2017), while the prevalence of EBV is much higher in this population. Japanese, USA and German populations are reported to have high frequency (6.9 %, 16-18% respectively) of EBV infection than other countries (Takada et al 2000; van Beek et al. 2004). EBV enters the body through saliva or oral contacts and in the Mizo people have a common practice to share water glasses or cigarettes with each other, while drinking tuibur (tobacco infused water) or alcohol and smoking. This might be a cause for this high prevalence of EBV-associated GC cases in this population. Studies have reported that EBV infection promotes GC by its $BARF_1$ oncogene, which can develop proliferation in a gastric epithelial cell by upregulation of NF-κB signaling and also by reducing the cell cycle inhibitor P21 (Chang et al 2015; Tavakoli 2020). EBV type I genotype (32.25%) was more frequent than type II in this study. According to geographical regions, EBV type I genotype is more prevalent worldwide, mostly found in Europe, Asia and America while EBV type II is prevalent in Alaska, Central Africa and Papua New Guinea (Zanella et al. 2019). The present study also similar result as EBV type I genotype was prevalent in this study. Studies have reported that the transformation of B cells into lymphoblastoid cells can be done more efficiently by Type I genotype than Type II (Rickinson et al. 1987; Lucchesi et al. 20068).

In the Mizo population, *H. pylori* (+) cases were found in 63% of GC patients and similar data were reported from West Bengal, India (Saha et al. 2013). Prevalence of *H. pylori* cases can be found in developing countries (Aziz et al. 2014). *H. pylori* is a class I carcinogen that can lead to producing proinflammatory cytotoxins, oxidative stress and necrosis in the cells which in turn can develop chronic inflammation to lead to GC cancer (Singh et al. 2017; Carlos et al. 2019). One study from Asia has reported 60% prevalence of *H. pylori* CagA genotypes which is similar with our result as we have found 58% CagA genotypes associated GC cases in our study (Aziz et al. 2014). CagA translocates in gastric epithelial cells and can produce proinflammatory cytokines by activating cell signaling pathways. In our study, we found 21% of GC patients were VacA positive and 16% of cases were both positive. Studies have reported that prevalence of *H. pylori* VacA strain was found in presence of East-Asian type cagA genotype (Aziz et al 2015; Rasheed et al. 2011)

.

In this study, some patients were having only one pathogen infection and 13.75% of GC cases exhibited both pathogen infections. Abundance analysis showing that there are no particular trends of pathogen infection in GC, it is not evenly distributed. One study has reported that *H, pylori*-positive cases were showing a significantly higher load of EBV DNA which indicating that *H. pylori* are promoting the lytic phase of EBV (Dávila-Collado et al. 2020). Another study has reported that the co-infection of *H. pylori* and EBV can induce acute inflammation and hence, be used to increase the risk of developing intestinal-type Gastric Cancer (Cardenas-Mondragon et al. 2015). Co-infection of EBV and *H. pylori* can jointly act to induce the IL-17 expression level to promote severe inflammation on gastric mucosa (Carlos et al. 2019).

In this study, MSI cases were detected in 40% of GC samples and this frequency is slightly higher than in other studies. The frequency of MSI was higher in the Japanese than the American population (Theuer et al. 2002). Studies reported that MSI-H cases were associated with the distal location and intestinal-type Gastric Cancer (Cunningham et al. 2006; Smyth et al. 2017). In our study, the majority of the patients exhibited Gastric cancer in the distal region of the stomach. In this study, all the MSI-associated cases except one were positive for pathogens (either EBV or *H. pylori). H. pylori* might promote the development of gastric carcinoma at least in part through its ability to affect the DNA mismatch repair system and its deficiency resulted in MSI phenotype (Jiricny, 2006; Kim et al. 2002). Despite the association with *H. pylori* and MSI, there was no difference of presence of *EBV* level according to microsatellite state in our results. However, in gastric epithelial dysplasia with *H. pylori* infection, MSI tended the absence of EBV. It suggested that *H. pylori* might appear as a cofactor for inducing gastric carcinogenesis. It supports that progression of gastric epithelial dysplasia to true gastric cancer could be blocked after *H. pylori* eradication in the selected cases—MSI positive state.

This study has shown that *EBV*-infected GC patients are more aggressive with poor prognosis and the prognostic value of *EBV* infection was confirmed by multivariate analysis, even after adjustments for other clinical factors. The prognostic assessment for *EBV* (+) GC case is very much controversial in previous studies. One study reported that median survival time for *EBV*-negative tumor (5.3 years) is lower compared to *EBV* associated GC (8.5 years) (Camargo et al. 2014) and whereas, another study showed that the five years overall survival in *EBV* associated GC (71.4%) is higher compared to negative group (56.1%) (Song et al. 2011). The prognostic assessment for EBV infected GC is regionally and ethically confined with their food and lifestyle habits. Moreover, a higher prevalence of EBV-infected cases (40 %) was found in this cohort compared to worldwide status (10%) (Iizasa et al. 2012; Kim et al. 2020), while *H. pylori* infection does not have any significant change in survival rate with GC. One Chinese cohort reported a trend regarding a higher survival rate in GC patients with high-copies *H. pylori* infection compared to patients with low-copy infection (Qiu et al. 2010). In this study MMR proficient, GC cases were showing poor prognosis and considered as a high-risk group with more aggressive tumors, while MMR deficient GC patients exhibited good prognosis. The result is consistent with other studies which reported that MSI shows a better prognosis than MSS cases in gastric cancer (Beghelli et al. 2006; Kim et al. 2020; Choi et al. 2014; Smyth et al. 2017). We have used TCGA data for comparison of our prognostic assessments with *H. pylori* and MSI patient groups (Figure 4G and 4H). The present study has supports the fact that *H. pylori* infection does not affect the prognosis of GC patients in this population, might be in this population the strains of *H. pylori* are not up-regulating the expression of IL-8.

In this study, the top somatically mutated gene was *TP53* (47%) as reported in other studies in Gastric cancer (Park et al. 2016; Busuttil et al. 2014). *TP53* mutations are used to associate with late-stage or advance stage of Gastric cancer, similar to this study. Studies have also reported that *TP53* mutations are associated with the risk of developing distal GC (Perez-Perez et al. 2005; Bellini et al. 2012). In this study, most of the tumors occurred at the distal part of the stomach. Frequently mutated genes like *MUC6, FAT4 & APC* were also found to be mutated frequently in TCGA and ACRG studies, supporting our data. Other top genes like *RNF43, BCOR, PTPRC, ERBB2, CTNNB1, SOHLH2,* and *FBXW7* were also found to be associated with

Gastric cancer in TCGA and ACRG studies (Cancer Genome Atlas Research Network et al. 2014; Cristescu et al. 2015). In this study, *TP53* was the top mutated gene in all the subgroups (EBV +, MSI and MSS) of the samples.

APC gene was significantly mutated with EBV-associated gastric cases like other studies, though the percentage is much higher in this study (Shinozaki-Ushiku et al. 2015). A study reported hypermethylation of APC gene association for the development of EBV-infected non-cardiac GC cases (Geddert et al. 2010). Enrichment of *ARID1A* mutations was found in EBV (+) subtypes like other studies (Cancer Genome Atlas Research Network et al. 2014; Cristescu et al. 2015). Enrichment of *ERBB2* mutation was found only in EBV-associated cases. Studies have reported that the crosstalk between EBV and *HER-2* might play an important role to develop EBV-associated GC through receptor kinase signaling pathways (Gulley et al. 2015; Cyprian et al. 2018). ERBB2/HER2 signaling may be responsible for developing EBV-associated Gastric Cancer. Another important gene is *RNF43* which was frequently mutated in our study with EBV-associated cases that were not found in other studies. One study has reported that EBV infection inhibits the dsDNA break mechanism through *BKRF4* by inhibiting histone ubiquitylation at dsDNA breaks and restricting the mobilizing of *RNF168* (histone ubiquitin ligase). *RNF168* and *RNF43* belong to the same family and *RNF43*, DNA damage repair gene is also a histone ubiquitin ligase and takes part in Wnt signaling also (Ho et al. 2018; Degirmenci et al. 2018). EBV infection might be playing a role to develop cancer through RNF43 alterations. In the case of the MSI subgroup, the top mutated genes were *TP53, MUC6, FAT4, FAT3* and *APC*. One Korean cohort study reported that MSI-associated GC cases were significantly associated with MUC6 expression (KIM et al. 2013). In the TCGA study, *FAT4* was mutated significantly with the MSI subgroup. *TP53* and *APC* were mutated in the ACRG study. One study has reported that the Wnt signaling pathway can initiate both the MSI and MSS GC cases (Li et al. 2019) and in our study, we found that *APC* is associated with MSI-associated GC, which is a gene of the Wnt signaling pathway. In MSS cases, the top mutated gene was *TP53* similar to other major studies (Cancer Genome Atlas Research Network et al.

2014; Cristescu et al. 2015). *APC, FAT4 PTPRC* and *BCOR* were other two important genes of this study similar to the ACRG group and other studies (Cristescu et al. 2015; Li et al. 2019).

In this study, two molecular subtypes were found: one with TP53 mutation dominant group and another group was found with EBV infected cases. The higher frequency of high-grade tumors and enrichment of ERBB2 mutation in EBV-associated cases indicates that they are more aggressive tumors having poor prognosis (He et al, 2018). *TP53* mutations were less in EBV (+) group like in another study (Kim et al. 2016). *TP53* somatic mutation and EBV infections are the two drivers for developing Gastric cancer in the Mizo population. Luciya et al. This study identified seven known pathogenic somatic mutations (R306*, G245S, D769Y, V8421 E545K, H1047, RR876*) and 78 novel pathogenic mutations related to Gastric Cancer development in this study. Most of the variants were related to the TP53 pathway, RTK-RAS pathway, Wnt signaling pathway and Hippo signaling pathway. Studies have reported *TP53* alterations and its pathway as responsible for the development of Gastric cancer (Fenoglio-Preiser et al. 2003; Busuttil et al. 2014). Several studies have reported that alterations in RTK-RAS and Wnt signaling pathways can initiate the progression of Gastric cancer development and these genes can be targetable for therapeutic development (Gonzalez-Hormazabal et al. 2018; Deng et al. 2012; Chiurillo et al. 2015; Koushyar et al. 2020). FAT family genes (Hippo signaling pathway) are significantly linked with Gastric cancer, including another type of cancer (Katoh et al. 2012; Kang et al. 2016). These four pathways might be responsible for developing Gastric Cancer in the Mizo population.

To date, few studies gave us insights about germline mutated genes, except CDH1 in Gastric cancer. In this present study, frequent mutations in *MAP3K4, KMT2C, ATN1, MACF1, BRCA2, FAT4, FAT3, KMT2B, PLB1* and *APC* genes were obtained, except CDH1. Very few studies reported mutations in another gene besides CDH1 in the case of hereditary GC (Gaston et al. 2014; Villacis et al. 2016). MAP3K4 gene is a member of the mitogen-activated protein kinase (MAPK) pathway and plays an important role in Cancer development by activating the CSBP2, P38 and JNK MAPK pathways by phosphorylating MAP2K4 and MAP2K6 of the MAP3K family. It is an important gene that exhibited only one homozygous in-frame deletion with 91.66% in this population. Studies have reported the MAP3K6 gene as a novel predisposing factor due to getting germline mutations in unrelated individuals (Gaston et al. 2014). Here in this study, all the samples were collected from unrelated individuals, so MAP3K4 can be a novel predisposed gene in the Mizo population. The variant was found associated with the lung and upper aerodigestive tract in the COSMIC database, but for the first time, we are reporting the association of these variants with stomach cancer in our study.

MLL3, a member of the myeloid/lymphoid or mixed-lineage leukemia (MLL) family is a chromatin remodeling gene. The products of chromatin responsible can regulate the structure of chromatin for altering DNA accessibility and transcriptional efficiency and were observed as frequently mutated genes in GC (Cancer Genome Atlas Research Network. 2014). The frequently mutated chromatin remodeling gene is ARID1A (Wang et al. 2011). Studies have reported somatic alterations of *MLL (KMT2A)* and *MLL3 (KMT2C)* genes of mixed-lineage leukemia (MLL) family were also mutated in diffuse-type Gastric cancer (Zang et al. 2012; Kakiuchi et al. 2014). In this study, two subgroups one with patients with dominant KMT2C mutations and another with a combination of heterogeneous genes. It is a very important gene for this population with the novel pathogenic mutation.

Cell adhesion genes *CTNNA1, FAT3* and *FAT4* were mutated with 2%, 16.66% and 25% frequency, respectively. The genes of the FAT family, FAT3/4 were strongly significant (p-value = 0.003) with well and moderately differentiated cases. One study has reported that the tumor suppressor gene *FAT4* is a modulator of Wnt/β-catenin can be a novel therapeutic target for clinical development in GC (Cai et al. 2015). These Cadherin family genes besides *CDH1* might play an important role in developing Gastric Cancer in this population. Another important gene *MACF1* which maintains Cell

Motility and is involved in metastatic invasion by regulating the cytoskeleton structure was reported as a significantly mutated gene in Gastric cancer (Cancer Genome Atlas Research Network. 2014). This gene was also mutated in our study with a 31% occurrence frequency and significantly mutated with the group of advanced-stage patients. *MACF1* is showing poor prognosis as all the variants have aggressive effects and *BRCA1/2* was showing a good prognosis (*p*-value = 0.03). BRCA2 germline mutation was found in 20% of cases indicating the increased risk of Gastric cancer relation with BRCA1/2 mutations (Hiroshi et al, 2020). The variants were reported as pathogenic for Hereditary Breast and ovarian cancer.

Three genes *KMT2C* (*MLL3*), *PMS2* and *CTNNA1* mutations were found significant with familial GC cases. Among them, *PMS2* and *CTNNA1* were already present in the hereditary Gastric Cancer panel made by Chicago university. *KMT2C* or *MLL3* is the new gene that was significantly associated with familial gastric cancer samples in this study. KMT2C genes were showing a poor prognosis for familial gastric cancer patients. MLL3 gene mutation was associated with Lynch syndrome (Villacis et al. 2016) which is associated with GC development. Besides *CDH1, CTNNA1* of the cadherin family was also associated with hereditary diffuse GC development (Lauren et al. 1965). Alteration of these three genes can be screened for early detection of germline Gastric Cancer cases as this panel achieved a significant p-value and high accuracy rate for predicting GC in this present study. Studies have reported that *PMS2* mutations were significantly associated with Lynch syndrome (Lauren et al. 1965 and Fewings et al. 2018). L114F of *PMS2,* P309S and P922R of *KMT2C* gene and N283S of *CTNNA1* was the most pathogenic germline mutations which were significant for Familial Gastric Cancer in this population. Out of 78 non-synonymous variants, 9 were pathogenic and three novel pathogenic variants (P4952L and N2544S of *MACF1*, and G2606A of *FAT4*) were found in this study. The genes mutated in this study were found in RTK-RAS, Hippo, Wnt, PI3K, TP53 and NOTCH. RTK-RAS and Hippo pathway and these pathways were associated with developing Gastric Cancer in this study like reported in other studies also (Maganelli et al, 2020 and Qiao et al, 2018).

In whole exome germline analysis, immune system-related genes (*HLA-DRB1, HLAB* and *HLAC*) were the top mutated genes. One study has reported that EBV infects B lymphocytes to enter into the host body and HLA class II molecules used to act as a cofactor for initiation of this infection of B lymphocytes (Li et al. 1997). These results indicate that our immune-related genes were mutated frequently in this population due to the high prevalence of EBV infection, which is playing the prime role in developing GC in the Mizo population.

Thirty-four (34) genes were found to be mutated in more than 90% of samples, in the case of germline analysis through those variants might be polymorphic. But among them, six genes (*COL18A1, KCNJ18, CMYA5, FCGBP, HLA-DRB1* and *OR4M2*) exhibited pathogenic mutations. One study has reported that the FCGBP gene was significantly up-regulated in intestinal metaplasia (Lee et al. 2010). *COL18A1, KCNJ18, CMYA5* and *OR4M2* were not reported as associated with Gastric Cancer till date in any of the studies. Forty (40) novel genes were found in the case of germline mutation, which was not reported in other studies for association with Gastric Cancer. These novel genes might be following some different pathways for developing Gastric Cancer in this population.

Fifty-three (53) pathogenic germline variants were identified in WES analysis and out of the 14 were novel mutations. In this study, we found six variants, N2198Y and N2544S of *MACF1*, E319V of *TP53* and P857R of *KMT2B*, T1261L of *APC* and K253R of *EGFR* which were the most pathogenic germline mutation and significant for developing Gastric Cancer in this population as these variants were present in both targeted and whole-exome sequencing data.

In copy number analysis, we found that in tumor tissue *ERBB2* is having copy number gain compared to adjacent normal. A negative correlation was found between copy numbers among TP53 and ERBB2 genes. This supports that Tumor oncogenes have a gain in copy number and tumor suppressor genes have a deletion in copy number in cancer tissue samples (Lawrence et al, 2019). MS/Y781C was found to be responsible for gain in CNV in *ERBB2*. Previously, several study groups dealt with gastric cancer patients using HER2 ddPCR. Kinugasa et al. (2015) included 25 gastric cancer patients and showed that the concordance rate of circulating tumor DNA (ctDNA) and tissue DNA was 62.5%19. According to our study, in advanced gastric cancer, HER2 positive cases by tumor tissue DNA ddPCR may be candidates for survival and treatment prediction, but HER2 assessment by other methods using tissue samples should be done in HER2 negative cases by ddPCR of plasma cfDNA to overcome low sensitivity.

The expression of BAX, an apoptotic regulator gene, was higher in tumor cases compared to adjacent normal and the patients with positive expression of BAX were in a risk group for developing Gastric cancer (Liu et al, 1995). BAX expression was significantly higher in EBV (+) group. It has been reported in one study that BCL 2 expression was higher in EBV positive cases and BAX expression was comparatively higher in EBV negative group. The present study reported a contradictory report, which suggests that EBV infection might contribute to the apoptosis method (Lima et al. 2008). One study has reported that EBV infection might upregulate the expression of BCL family gene by the Notch signaling pathway (Fu et al. 2013). In this study, the alteration was found in Notch pathway genes which might also responsible for significant BAX expression in EBV infected cases. BAX expression was higher in Low-grade tumors than high-grade tumors while in the advanced stage the expression was lower (Golestani et al, 2014). Cell death might play a role to develop cancer at an early stage, as BAX is an apoptotic gene it is also following the same trend. Further study is necessary with a larger sample size to support our findings.

Positive TP53 expression was found more in later stages of GC. One study has reported that TP53 mutation occurs at a late stage, which converts the premalignant stage to GC (Busuttil et al. 2014). In this study, we selected TP53 mutated samples for IHC and as a result, they were showing more positive cases in later stages. During dysplasia, the last stage of disease progression might be due to some stress driving TP53 mutations, which contributes to the progression of GC. In this study, TP53 expression was not associated with H. pylori and EBV infection like other studies (Lan et al. 2003). TP53 expression was less in the EBV (+) group, which supports the sequence data of this study and other studies also reported that TP53 expression was more in EBV (-) subgroups (Kim et al. 2016). EBV infection might not alter the TP53 pathway for developing GC. TP53 expression was higher in microsatellite stable cases which have a similar trend of data like TCGA and ACRG studies (The Cancer Genome Atlas Research Network, 2014; Cristescu et al. 2015).

HER2 positive cases were found more in the early stages, indicating that HER2 can be targeted as a therapeutic marker for Gastric cancer in the Mizo population which can help to develop the treatment strategies. HER2-positive patients can be treated with trastuzumab (Bang et al. 2010). HER2 cases were showing positive expression on EBV positive and MSI cases, which supports the sequence data of enrichment of ERBB2 mutation on EBV subgroup. EBV infection might be affecting the receptor kinase signaling pathway (Gulley et al. 2015; Cyprian et al. 2018) for developing GC in this population.

ERCC1 positive expression cases were more in the early stages compared to late stages. ERCC1 can also be used as a prognostic marker in the Mizo population. ERCC1 positive expression cases were more in H. pylori-positive cases and MSI cases. This data shows that H. pylori infection might play role in DNA damage repair pathway (Kim et al. 2008; Wang et al. 2014; Kwon et al. 2007). The treatment for patients having a low expression of the ERCC1 gene can be followed by platinum-based adjuvant chemotherapy (Dosso et al. 2013).

Conclusion

The perspective of this present study is that a high incidence of Gastric Cancer in the Mizo population might be due to the effect of smoked food, tuibur, alcohol and smoking with EBV infection. The Mizo population being a homogeneous population has a unique set of driver genes with pathogenic alterations that may play a role to initiate the progression of Gastric cancer in this population. The panel of five epidemiological risk factors can predict early-stage GC cases, which is very necessary for the clinical field for deciding on patient treatment. The study will help clinicians to opt decision for the right therapy by applying the prognostic assessment of this study. Further study is necessary with a large cohort which would be beneficial to support our data.

This study reported some commonly mutated genes like *TP53*, *APC*, *MUC6*, *FAT4*, *FAT3*, *ERBB2*, *ARID1A*, *MACF1*, *KMT2C*, *MAP3K4*, *ABCA10*, *BRCA1*, *BRCA2*, and *RNF43* which are pathway-related genes and might be responsible for Gastric Cancer in this population. TP53, Wnt signaling, Hippo signaling, RTK-RAS, and Notch signaling pathways might be altering the progression of gastric cancer. This study reported a panel (*KMT2C*. *PMS2* and *CTNNA1*) for predicting familial gastric cancer.

The present study reported novel genes that were not earlier related to gastric cancer and some genes were mutated in 90% of the patients, among them some of the genes were identified as a biomarker for other cancer, like lungs, head and neck, colorectal, pancreatic, etc. and chronic diseases. *HLADRB1*, *HLA-C*, *HLA-B*, *MAML3*, *MUC6*, *CMAY5*, *FCGBP*, *SUSD2*, *TTN*, *SORBS1*, *ATIC*, *HSPA6*, *KRT6B*, *LIMS1*, *PDE4D*, *SIRPB1*, *LAMA5*, *ZNF638*, *OBSCN*, *CELSR1*, *USP8*, *SLCO2A1*, *XIRP2*, *ARHGEF10* and *CYP2C8* might be identified as biomarkers for Gastric cancer in this population.

This study reported that unique food habits and lifestyle factors along with pathogen and microsatellite status might be driving the novel driver mutations for developing Gastric Cancer in the Mizo population. A novel set of genes identified in this study might be the drivers for developing GC in this high-risk population. This study reporting new epidemiological markers as well as gene markers for detecting early Gastric cancer and familial GC cases, respectively which will help the clinicians in taking correct diagnostic and therapeutic decisions.

Summary

The present study was accomplished to find out the significant risk factors associated with Gastric cancer in this population along with pathogen infection and MSI status. This study was also carried out to find out the novel driver alterations and genes associated with GC development. Statistical analysis was performed to find out significant risk factors. Screening of *H. pylori*, EBV and MSI were performed for molecular subtyping. Targeted re-sequencing was performed for paired tumor and blood samples, to find out driver genes associated with GC in this population. Sequencing was performed on Illumine Hi-seq machine by capturing hybrids of interesting panel genes. Whole exome sequencing was also performed to find out a novel set of genes that might play for developing GC in this population. In addition, IHC was performed with tumor suppressor genes, oncogenes and apoptotic genes for studying their expression and prognosis on GC patients on the basis of clinical and mutation data.

The significant findings of this study are highlighted below:

- Food habits (extra salt and smoked meat consumption) and lifestyle factors (alcohol and smoking) are the significant risk factors with Gastric Cancer development in the Mizo population.
- Obese persons (high BMI) are the significant risk group for developing Gastric Cancer
- A panel of five epidemiological factors have been identified which can significantly detect Gastric cancer at an early stage
- This study might provide new opportunities in the clinical field for detection and prognosis of Gastric Cancer at an early stage in the high-risk Gastric Cancer population by using this panel of risk factors.
- In this population, EBV-associated Gastric Cancer cases were higher than in another part of the world. 78% of GC cases were positive for Pathogen infection indicating that the pathogens are playing a major role to develop Gastric Cancer.

- Smoked food and tuibur were associated with Pathogen infected Gastric Cancer and Chewed tobacco and alcohol drinking was significantly associated with MSI associated GC cases
- The top ten somatic mutated genes were *TP53* (47%), followed by *MUC6, FAT4, RNF43, BCOR, PTPRC, ERBB2, CTNNB1, SOHLH2,* and *FBXW7*
- APC gene mutations were associated with EBV (+) cases. Enrichment of ERBB2 mutation and the high-grade tumor was found in EBV associated GC cases
- TP53 mutations and EBV infections are the drivers of developing Gastric Cancer in the Mizo population.
- *MAP3K4* (92% cases) was the top germ-line mutated gene in this study followed by *KMT2C*, *ATN1*, *MACF1*, *BRCA2*, *FAT4, FAT3, KMT2B* and *PLB1*.
- The Mizo population may be genetically predisposed to pathogenic mutation of the MAP3K4 gene which can be responsible for developing Gastric Cancer.
- This population has a pathogenic mutation in the Chromatin remodeling gene (KMT2C). This gene might have to play an important role in developing GC in this population.
- FAT3/4 genes were significantly mutated with early gastric cancer cases which can be treated as a prognostic or therapeutic marker for GC in this population
- BRCA1/BRCA2 mutated patients show significantly good survival for GC patients, which can be used as a therapeutic marker for this population.
- Significant MACF1 gene mutations were found in an advanced stage of GC cases and showed a poor prognosis indicating that this gene should be targeted for therapeutic development in GC patients.
- There are germline pathogenic mutation in CTNNA1, PMS2 and KMT2C genes which were significantly associated with familial Gastric Cancer and showed poor prognosis, and these could be used as a panel for detecting familial GC in this population
- In whole exome germline analysis, immune system-related genes (*HLA-DRB1, HLAB* and *HLAC*) were mutated frequently which indicating that immune-related

genes were mutated frequently in this population due to the high prevalence of pathogenic association (EBV infection), and it might be the prime cause of GC development in this population.

- Twenty-six commonly mutated genes were found in this study which was related for other cancer also and might be playing an important role for developing GC in this population
- Thirty-four novel genes were found in this study reporting for the first time as a set of a gene associated with Gastric Cancer in this population and they might follow different pathways for driving GC.
- Twenty-six genes were identified from the frequently mutated gene (more than 90% patient) and novel genes, which were reported as a predictive, diagnostic, prognostic, or therapeutic biomarker for other cancers and chronic diseases, which might be reported as a biomarker for GC in this population.
- TP53, Hippo signaling, Wnt signaling, RTK-RAS, PI3K and Notch pathways were frequently altered, along with novel pathogenic somatic and germline alterations as well as pathogen association are responsible for developing Gastric cancer in the Mizo population.

Appendix

_______________*Questionnaire for Epidemiological Study of Gastric Cancer*

Referring Dr:_______________ Hospital Name/No ______________/_________
Referring Unit:______________ MZU Reg. No. /Date:_MZU/DBT/__________

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?*:* 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)*

PermanentAddres:____
PinCode______Tel/Mob.No.____
Email: _______________________________________
PresentAddres:____
PinCode__________Tel/Mob.No.____
Email: ____
Cancer Diagnosis/Treatment ____
Engtik kumah nge cancer I vei tih hmuhchhuah a nih? *(Year in which cancer was detected?):* _______

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

Syndrome Diagnosis:

Consent for sample collection: Yes/No Date: __________________________

Blood collected: Yes/No Date: __________Received on__________From_________

Second sample collected: Yes/No Date:________ Received by_________Through_________

Tumor Tissue Collected: Yes/No Date__________Biorepository: Genesis Lab/MZU/ MSCI

Samples transmitted to MZU (sample type/ Date/ Method of transfer etc.)

Samples transmitted to NIBMG (sample type/ Date/ Method of transfer etc.)

Details taken by: __________________________________

Date:___________________________

Pre- Test Counseling done by: _______________________

Date:__________________________

Post-Test Counseling done by: _______________________ Date: _________________________

FAMILY INFORMATION:

In chhungkua ah natna dang vei in awm em(cancer ni lo) ***(Any other type of diseases in the family (other than cancer)*****:**

Life style Habits						
Occupation						
Disease Information						
Sex/Age						

Name	Relation	Education

In chhungkua ah Cancer vei dang an awm em ***(Does anyone else in your family have cancer)***:

Name	Relation	Education	Sex/Age	Disease Information	Occupation	Life style Habits

Hereditary: Yes [ ] No [ ] Autosomal Dominant: Yes [ ] No [ ]
Autosomal Recessive: Yes [ ] No [ ] Sex linked: Yes [ ] No [ ]
[ ] Cannot ascertain/Not applicable [ ] Sporadic [ ] Early Onset [ ]
Routine RET [ ] Familial [ ] Others__________________

Chhungkaw member zat (***Number of deaths in the family due to disease***):
Boral tawh (*Decease number*): [ ]
Boral chhan (Reason) – Pumpui cancer (*Gastric cancer*): [ ]; Adang (*Other*): [

PEDIGREE

(Draw pedigree one degree above and below affected individuals and note consanguinity.)

GEOETHNIC ORIGIN

Sub tribe							
Tribe							
Occupation							
Family name/Surname							
Dist./State of origin							

	Place of birth (Dist./State)	Present place of stay (Dist./State) & duration
Index		
Father		
Mother		
Paternal Grandfather		
Paternal Grandmother		
Maternal Grandfather		
Remarks		

Environmental/ Lifestyle Factors

What has been your main occupation?___________________________________

Hengah te hian hna I thawk em? 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 (e.g. In a factory, laboratory/ medical setting)	Yes No Don’t Know				
Plastic factory/ burning/	Yes No Don’t Know				
Tobacco plants / Rubber plant	Yes No Don’t Know				
Pesticides/ Pest control / Mosquito Repellants	Yes No Don’t Know				

Chemical/Dyes/Fertilizer	Yes No Don't Know				
Any other exposure (Asbestos, Chromium or Lead)	Yes No Don't Know				
Tuibur/ Local Alcohol preparation	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 [ ]

Exercise I la ngai em? *How often do you exercise?* Ngai lo(N*ever)* [ ]; Karkhatah vawi khat aia tlem(*Less than once a week)* [ ]; Karkhatah vawi khat(*Once a week)* [ ]; Karkhatah vawi 2-3 (*2-3 times a week)* [ ]; Karkhatah vawi 4-6 (*4-6 times a week)* [ ]; Nitin(*Everyday)* [ ]

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 intake				
Boiled food				
Fried food				
Smoked food				
Salt brand/type (packed/raw)				
Oil brand/type				
Fibers food/fruits (Banana/Bamboo shoots)				

What type of utensils you normally use for your food items?): *Plastic* [ ] Aluminum [ ] Steel [ ] Other

Do you re-use oil for cooking/ frying: Aw/*Yes* [ ] Aih/*No* [ ]

Do you use Cosmetics/ Make up items: Regularly [ ] Occasionally [ ]

Tobacco & alcohol History:

Do you consume **(I ei ngai em)**	**0 (Never)**	**1 (Little)** **1 day in a week**	**2 (Average)** **2-4 days in a week**	**3 (Heavy)** **5-7 days in a week**	**Av. Quantity per day**
Hnamdang siam (*Branded Alcohol)*					
Mizo siam *(Local Alcohol)*					
*Tuibur[*Bazar a lei (*Local)/* Mahni a siam (*Self-made)]*					
Others					

Engtik atangin nge I in tan? (*When did you start taking alcohol?)* :
I nghei tawh anih chuan, engtik atangin? *(If quit already, since when?)*:
Engtik atangin nge I hmuam tan? (*When did you start taking tuibur?)*:
I nghei tawh anih chuan, engtik atangin? (*If quit already, since when?):*

If quantity of consumption of alcohol/ tuibur has changed during life time, the period of your highest consumption:

Beverage (Name)	Yes/No	From age	To age	Av. Quantity per day	Days/ Week

Mei I zu em? (*Do you smoke?): *Aw/*Yes* [ ] Aih/*No* [ ]
Engtik atangin nge I zuk tan? (*When did you start smoking?)*:
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 (Brand name)				
Biri				
Zozial				

Vaihlo a siam thil dang tih I nei em? (*Do you consume other tobacco products?):*Aw/Yes [   ]
Aih/No [   ]
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*)				
Kuhva *(Pan/Beetle nut)*				
Zarda Pan				
Supari				
Chewing without tobacco (eg. pan without tobacco)				
Adangte *(Others)*				

History of passive smoking:
Do any of your family member/colleagues smoke tobacco at home?   Yes/No
Frequency of exposure to passive smoking: Rarely/ Continuously

Medical History:

I blood group eng nge? *(What is your Blood group?)*
A+ [   ]   A- [   ]   B+ [   ]   B- [   ]   AB+ [   ]   AB- [   ]   O+ [   ]
O- [   ]
Ultrasonography:
Other: Region____________________________Report
Date:________________Impression____

CT scan: Region______________________________Report
Date________________________
Impression____
Colonoscopy/Endoscopy:
Regions___________________________________Date__________________
Impression____
Surgery:
Site/Procedure____
Pathological Staging-
pTNM___________________________________Date________________
Histopathological Report: Specimen_______________________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____________
Other:____
Treatment/Other
Remarks:____

<u>Syndromic features noted:</u>

Indigestion (Pum Puar)   Nausea or vomiting (Luakchhuak)   Dysphagia (Chawhelh)
Postprandial fullness (Hnawh ulh) Loss of appetite(Chaw ei tuilo)
Melena(Ek dum)
Hematemesis (Thi a luak)   Weight loss (Thla 6 chhunga kg8-10 vela tla hniam)
Pallor (Dawldang)   Anaemia (Thisen nei tlem)   Pain in abdomen(Pum na)
Natna/Damlohna dang I nei em? (*Do you have any other diseases?):* Aw/*Yes* [ ]
I neih chuan, eng natna nge? (*If yes, what type of disease?):*
____
H. pylori [ ]   Diabetes [ ]   obesity [ ]   HIV [ ]   HbsAg[ ]
HCV[ ] EBV [ ]   Gastric atrophy [ ]   Others_______________________

A hnuai ami te hmang hian enkawl I ni tawh em? History of taking HRT/Reflux /Proton Pump Inhibitors/ Others (Give details) ____________________

Obestric History :(Nau nei tawh zat)   Others
Gravity/Parity (Nau paizat)
Recurrent spontaneous abortions (Nau chhiat zat)

Still births/ Neonatal deaths (Thi a Hrin zat/Sen laia thi)
Congenital malformations (Fuke kim lova piang zat).

Remtihna *(Consent):*

Heng a chunga thu te hi ka hriatpui a, ka biological sample hi zir chian atan pek ka remti thlap e.

The information provided above was given with my full consent and I do not have any objection in providing my biological sample for research purposes. I have read and understood the consent information.

Hmun*(Plac*e):	Signature:
Date:	Hming (*Name)*:

KA LAWM E

(THANK YOU VERY MUCH FOR YOUR HELP)

Bibliography

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-249.

Allgayer H, Babic R, Gruetzner KU, Tarabichi A, Schildberg FW, Heiss MM. c-erbB-2 is of independent prognostic relevance in gastric cancer and is associated with the expression of tumor-associated protease systems. J Clin Oncol. 2000;18(11):2201-2209.

Almeida R, Silva E, Santos-Silva F, Silberg DG, Wang J, De Bolós C, David L. Expression of intestine-specific transcription factors, CDX1 and CDX2, in intestinal metaplasia and gastric carcinomas. J Pathol. 2003;199(1):36-40.

Ang TL, Fock KM. Clinical epidemiology of gastric cancer. Singapore Med J. 2014;55(12):621-628.

Aziz F, Chen X, Yang X, Yan Q. Prevalence and Correlation with Clinical Diseases of Helicobacter pylori cagA and vacA Genotype among Gastric Patients from Northeast China BioMed Res Int. 2014;2014:142980.

Aziz LM. Blood neutrophil-lymphocyte ratio predicts survival in locally advanced cancer stomach treated with neoadjuvant chemotherapy FOLFOX 4. Med Oncol. 2014;31(12):311….

Bae SW, Berlth F, Jeong KY, Suh YS, Kong SH, Lee HJ, Kim WH, Chung JK, Yang HK. Establishment of a [18F]-FDG-PET/MRI Imaging Protocol for Gastric Cancer PDX as a Preclinical Research Tool. J Gastric Cancer. 2020;20(1):60-71.

Bang YJ, Van Cutsem E, Feyereislova A, Chung HC, Shen L, Sawaki A, Lordick F, Ohtsu A, Omuro Y, Satoh T, Aprile G, Kulikov E, Hill J, Lehle M, Rüschoff J, Kang YK, ToGA Trial Investigators. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. Lancet. 2010;376(9742):687-97.

Beghelli S, de Manzoni G, Barbi S, Tomezzoli A, Roviello F, Di Gregorio C, Vindigni C, Bortesi L, Parisi A, Saragoni L, Scarpa A, Moore PS. Microsatellite instability in gastric cancer is associated with better prognosis in only stage II cancers. Surgery. 2006;139(3):347-356.

Bellini MF, Cadamuro AC, Succi M, Proença MA, Silva AE. Alterations of the TP53 gene in gastric and esophageal carcinogenesis. J Biomed Biotechnol. 2012;2012:891961.

Bersten DC, Sullivan AE, Peet DJ, Whitelaw ML. bHLH-PAS proteins in cancer. Nat Rev Cancer. 2013;13(12):827–841.

Birkman EM, Mansuri N, Kurki S, Ålgars A, Lintunen M, Ristamäki R, Sundström J, Carpén O. Gastric cancer: immunohistochemical classification of molecular subtypes and their association with clinicopathological characteristics. Virchows Arch. 2018;472(3):369-382.

Böger C, Warneke VS, Behrens HM, Kalthoff Het, Goodman SL, Becker T, Rocken C. Integrins αvβ3 and αvβ5 as prognostic, diagnostic, and therapeutic targets in gastric cancer. Gastric Cancer. 2015;18(4):784-795.

Boland CR, Yurgelun MB. Historical Perspective on Familial Gastric Cancer. Cell Mol Gastroenterol Hepatol. 2017;3(2):192-200.

Bonequi P, Meneses-González F, Correa P, Rabkin CS, Camargo MC. Risk factors for gastric cancer in Latin America: a meta-analysis. Cancer Causes Control. 2013;24(2):217-231.

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

Brooks-Wilson AR, Kaurah P, Suriano G, Leach S, Senz J, Grehan N, Butterfield YS, Jeyes J, Schinas J, Bacani J, Kelsey M, Ferreira P, MacGillivray B, MacLeod P, Micek M, Ford J, Foulkes W, Australie K, Greenberg C, LaPointe M, Gilpin C, Nikkel S, Gilchrist D, Hughes R, Jackson CE, Monaghan KG, Oliveira MJ, Seruca R, Gallinger S, Caldas C, Huntsman D. Germline E-cadherin mutations in hereditary diffuse gastric cancer: assessment of 42 new families and review of genetic screening criteria. J Med Genet. 2004;41(7):508-17.

Busuttil RA, George J, Tothill RW, Ioculano K, Kowalczyk A, Mitchell C, Lade S, Tan P, Haviv I, Boussioutas A. A signature predicting poor prognosis in gastric and ovarian cancer represents a coordinated macrophage and stromal response. Clin Cancer Res. 2014;20(10):2761-2772.

Busuttil RA, Zapparoli GV, Haupt S, Fennell C, Wong SQ, Fennell C, Wong SQ, Pang JM, Takeno EA, Mitchell C, Di Costanzo N, Fox S, Haupt Y, Dobrovic A, Boussioutas A. Role of p53 in the progression of gastric cancer. Oncotarget. 2014;5(23):12016-12026.

Cai J, Feng D, Hu L, Chen H, Yang G, Cai Q, Gao C, Wei D. FAT4 functions as a tumor suppressor in gastric cancer by modulating Wnt/β-catenin signalling. Br J Cancer. 2015;113(12):1720–1729.

Caldas C, Carneiro F, Lynch HT, Yokota J, Wiesner GL, Powell SM, Lewis FR, Huntsman DG, Pharoah PD, Jankowski JA, MacLeod P, Vogelsang H, Keller G, Park KG, Richards FM, Maher ER, Gayther SA, Oliveira C, Grehan N, Wight D, Seruca R, Roviello F, Ponder BA, Jackson CE. Familial gastric cancer: overview and guidelines for management. J Med Genet. 1999;36(12):873-80 .

Camargo MC, Goto Y, Zabaleta J, Morgan DR, Correa P, Rabkin CS. Sex hormones, hormonal interventions, and gastric cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev. 2012;21(1):20-38.

Camargo MC, Murphy G, Koriyama C, Pfeiffer RM, Kim WH, Herrera-Goepfert R, Corvalan AH, Carrascal E, Abdirad A, Anwar M, Hao Z, Kattoor J, Yoshiwara-Wakabayashi E, Eizuru Y, Rabkin CS, Akiba S. Determinants of Epstein-Barr virus-positive gastric cancer: an international pooled analysis. Br J Cancer. 2011;105(1):38-43.

Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014;513(7517):202-209.

Cárdenas-Mondragón MG, Torres J, Flores-Luna L, Camorlinga-Ponce M, Carreón-Talavera R, Gomez-Delgado A, Kasamatsu E, Fuentes-Pananá EM. Case–control study of Epstein–Barr virus and Helicobacter pylori serology in Latin American patients with gastric disease. Br J Cancer. 2015;112(12):1866-1873.

Carrasco-Avino G, Riquelme I, Padilla O, Villaseca M, Aguayo FR, Corvalan AH. The conundrum of the Epstein-Barr virus-associated gastric carcinoma in the Americas. Oncotarget. 2017;8(43):75687-75698.

Catalano V, Labianca R, Beretta GD, Gatta G, de Braud F, Van Cutsem E. Gastric cancer. Crit Rev Oncol Hematol. 2009;71(2):127-64.

Cavanagh H, Rogers KM. The role of BRCA1 and BRCA2 mutations in prostate, pancreatic and stomach cancers. Hered Cancer Clin Pract. 2015;13(1):16.

Chen K, Yang D, Li X, Sun B, Song F, Cao W, Brat DJ, Gao Z, Li H, Liang H, Zhao Y, Zheng H, Li M, Buckner J, Patterson SD, Ye X, Reinhard C, Bhathena A, Joshi D, Mischel PS, Croce CM, Wang YM, Raghavakaimal S, Li H, Lu X, Pan Y, Chang H, Ba S, Luo L, Cavenee WK, Zhang W, Hao X. Mutational landscape of gastric adenocarcinoma in Chinese: implications for prognosis and therapy. Proc Natl Acad Sci U S A. 2015;112(4):1107-1112.

Cheng XJ, Lin JC, Tu SP. Etiology and Prevention of Gastric Cancer. Gastrointest Tumors. 2016;3(1):25-36.

Chiurillo MA. Role of the Wnt/β-catenin pathway in gastric cancer: An in-depth literature review. World J Exp Med. 2015;5(2):84-102.

Choi KS, Jun JK, Park EC, Park S, Jung KW, Han MA, Choi IJ, Lee HY. Performance of different gastric cancer screening methods in Korea: a population-based study. PLoS One. 2012;7(11):e50041.

Choi YJ, Kim N. Gastric cancer and family history. Korean J Intern Med. 2016;31(6):1042-1053.

Choi YY, Bae JM, An JY, Kwon IG, Cho I, Shin, HB, Eiji T, Aburahmah M, Kim HI, Cheong JH, Hyung WJ, Noh SH. Is microsatellite instability a prognostic marker in gastric cancer? A systematic review with meta-analysis. J Surg Oncol. 2014;110(2):129-35

Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, Liu J, Yue YG, Wang J, Yu K, Ye XS, Do IG, Liu S, Gong L, Fu J, Jin JG, Choi MG, Sohn TS, Lee JH, Bae JM, Kim ST, Park SH, Sohn I, Jung SH, Tan P, Chen R, Hardwick J, Kang WK, Ayers M, Hongyue D, Reinhard C, Loboda A, Kim S, Aggarwal A. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med. 2015;21(5):449-56.

Cunningham D, Allum WH, Stenning SP, Thompson JN, Van de Velde CJ, Nicolson M, Scarffe JH, Lofts FJ, Falk SJ, Iveson TJ, Smith DB, Langley RE, Verma M, Weeden S, Chua YJ, MAGIC Trial Participants. Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. N Engl J Med. 2006;355(1):11–20.

Cyprian FS, Al-Antary N, Al Moustafa AE. HER-2/Epstein-Barr virus crosstalk in human gastric carcinogenesis: A novel concept of oncogene/oncovirus interaction. Cell Adh Migr. 2018;12(1):1-4.

Daniel J. Denis. SPSS Data Analysis for Univariate, Bivariate, and Multivariate Statistics 2018 DOI:10.1002/9781119465775.

Dávila-Collado R, Jarquín-Durán O, Dong LT, Espinoza JL. Epstein-Barr Virus and Helicobacter Pylori Co-Infection in Non-Malignant Gastroduodenal Disorders. Pathogens. 2020;9(2):104.

De Dosso S, Zanellato E, Nucifora M, Boldorini R, Sonzogni A, Biffi R, Fazio N, Bucci E, Beretta O, Crippa S, Saletti P, Frattini M. ERCC1 predicts outcome in patients with gastric cancer treated with adjuvant cisplatin-based chemotherapy. Cancer Chemother Pharmacol. 2013;72(1):159-65.

De Mandal S, Singh SS, Muthukumaran RB, Thanzami K, Kumar V, Kumar NS. Metagenomic analysis and the functional profiles of traditional fermented pork fat 'sa-um' of Northeast India. AMB Express. 2018;8(1):163.

De Martel C, Parsonnet J. Helicobacter pylori infection and gender: a meta-analysis of population-based prevalence surveys. Dig Dis Sci. 2006;51(12):2292–2301.

Degirmenci B, Hausmann G, Valenta T, Basler K. Wnt Ligands as a Part of the Stem Cell Niche in the Intestine and the Liver. Prog Mol Biol Transl Sci. 2018;153: 1-19.

Deng N, Goh LK, Wang H, Das K, Tao J, Tan IB, Zhang S, Lee M, Wu J, Lim KH, Lei Z, Goh G, Lim QY, Tan AL, Sin Poh DY, Riahi S, Bell S, Shi MM, Linnartz R, Zhu F, Yeoh KG, Toh HC, Yong WP, Cheong HC, Rha SY, Boussioutas A, Grabsch H, Rozen S, Tan P. A comprehensive survey of genomic alterations in gastric cancer reveals systematic patterns of molecular exclusivity and co-occurrence among distinct therapeutic targets. Gut. 2012;61(5):673-84.

Deng N, Goh LK, Wang H, Das K, Tao J, Tan IB, Zhang S, Lee M, Wu J, Lim KH, Lei Z, Goh G, Lim QY, Tan AL, Sin Poh DY, Riahi S, Bell S, Shi MM, Linnartz R, Zhu F, Yeoh KG, Toh HC, Yong WP, Cheong HC, Rha SY, Boussioutas A, Grabsch H, Rozen S, Tan P. A comprehensive survey of genomic alterations in gastric cancer reveals systematic patterns of molecular exclusivity and co-occurrence among distinct therapeutic targets. Gut. 2012;61(5):673-84.

Denova-Gutierrez E, Hernández-Ramírez RU, López-Carrillo L. Dietary patterns and gastric cancer risk in Mexico. Nutr Cancer. 2014;66(3):369–76.

Diergaarde B, Braam H, Muijen GNP, Ligtenberg MJL, Kok FJ, Kampman E. Dietary factors and microsatellite instability in sporadic colon carcinomas. Cancer Epidemiol Biomarkers Prev. 2003;12(11 Pt 1):1130-6.

Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R. CRAVAT: Cancer-Related Analysis of VAriants Toolkit Bioinformatics. 2013;29(5):647-648

El-Omar EM, Rabkin CS, Gammon MD, Vaughan TL, Risch HA, Schoenberg JB, Stanford JL, Mayne ST, Goedert J, Blot WJ, Fraumeni JF Jr, Chow WH. Increased risk of noncardia gastric cancer associated with proinflammatory cytokine gene polymorphisms. Gastroenterology. 2003;124(5):1193-201.

Eso Y, Seno H. Current status of treatment with immune checkpoint inhibitors for gastrointestinal, hepatobiliary, and pancreatic cancers. Therap Adv Gastroenterol. 2020;21;13:1756284820948773.

Fang X, Wei J, He X, An P, Wang H, Jiang L, Shao D, Liang H, Li Y, Wang F, Min J. Landscape of dietary factors associated with risk of gastric cancer: A systematic review and dose-response meta-analysis of prospective cohort studies. Eur J Cancer. 2015;51(18):2820-32.

Fassone L, Bhatia K, Gutierrez M, Capello D, Gloghini A, Dolcetti R, Vivenza D, Ascoli V, Lo Coco F, Pagani L, Dotti G, Rambaldi A, Raphael M, Tirelli U, Saglio G, Magrath IT, Carbone A, Gaidano G. Molecular profile of Epstein-Barr virus infection in HHV-8-positive primary effusion lymphoma. Leukemia. 2000;14(2):271-277.

Fenoglio-Preiser CM, Wang J, Stemmermann GN, Noffsinger A. TP53 and gastric carcinoma: a review. Hum Mutat. 2003;21(3):258-270.

Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136(5):E359-86.

Fewings E, Larionov A, Redman J, Goldgraben MA, Scarth J, Richardson S, Brewer C, Davidson R, Ellis I, Evans DG, Halliday D, Izatt L, Marks P, McConnell V, Verbist L, Mayes R, Clark GR, Hadfield J, Chin SF, Teixeira MR, Giger OT, Hardwick R, di Pietro M, O'Donovan M, Pharoah P, Caldas C, Fitzgerald RC, Tischkowitz M. Germline pathogenic variants in PALB2 and other cancer-predisposing genes in families with hereditary diffuse gastric cancer without CDH1 mutation: a whole-exome sequencing study. Lancet Gastroenterol Hepatol. 2018;3(7):489-498.

Figueiredo J, Söderberg O, Simões-Correia J, Grannas K, Suriano G, Seruca R. The importance of E-cadherin binding partners to evaluate the pathogenicity of E-cadherin missense mutations associated to HDGC. Eur J Hum Genet. 2013;21(3):301-309.

Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008;Chapter 10:Unit-10.11. doi: 10.1002/0471142905.hg1011s57.

Förster I, Brockmann M, Schildgen O, Schildgen V. Microsatellite instability testing in colorectal cancer using the QiaXcel advanced platform. BMC Cancer. 2018;18(1):484.

Fox JG, Dangler CA, Taylor NS, King A, Koh TJ, Wang TC. High-salt diet induces gastric epithelial hyperplasia and parietal cell loss, and enhances Helicobacter pylori colonization in C57BL/6 mice. Cancer Res. 1999;59(19):4823-4828.

Fu Q, He C, Mao ZR. Epstein-Barr virus interactions with the Bcl-2 protein family and apoptosis in human tumor cells. J Zhejiang Univ Sci B. 2013;14(1):8-24.

Fuccio L, Zagari RM, Eusebi LH, Laterza L, Cennamo V, Ceroni L, Grilli D, Bazzoli F. Meta-analysis: can Helicobacter pylori eradication treatment reduce the risk for gastric cancer? Ann Intern Med. 2009;151(2):121-128.

García I, Vizoso F, Martín A, Sanz L, Abdel-Lah O, Raigoso P, García-Muñiz JL. Clinical significance of the epidermal growth factor receptor and HER2 receptor in resectable gastric cancer. Ann Surg Oncol. 2003;10(3):234-41.

Gaston D, Hansford S, Oliveira C, Nightingale M, Pinheiro H, Macgillivray C, Kaurah P, Rideout AL, Steele P, Soares G, Huang WY, Whitehouse S, Blowers S, LeBlanc MA, Jiang H, Greer W, Samuels ME, Orr A, Fernandez CV, Majewski J, Ludman M, Dyack S, Penney LS, McMaster CR, Huntsman D, Bedard K. Germline mutations in MAP3K6 are associated with familial gastric cancer. PLoS Genet. 2014;10(10):e1004669.

Geddert H, Zur Hausen, A, Gabbert HE, arbia, M. EBV-infection in cardiac and non-cardiac gastric adenocarcinomas is associated with promoter methylation of p16, p14 and APC, but not hMLH1. Anal. Cell. Pathol. 2010;33(3):143–149.

Ghatak S, Yadav RP, Lalrohlui F, Chakraborty P, Ghosh S, Ghosh S, Das M, Pautu JL, Zohmingthanga J, Senthil Kumar N. Xenobiotic Pathway Gene Polymorphisms Associated with Gastric Cancer in High Risk Mizo-Mongoloid Population, Northeast India. Helicobacter. 2016;21(6):523-535.

Ghosh S, Bankura B, Ghosh S, Saha ML, Pattanayak AK, Ghatak S, Guha M, Nachimuthu SK, Panda CK, Maji S, Chakraborty S, Maity B, Das M. Polymorphisms in ADH1B and ALDH2 genes associated with the increased risk of gastric cancer in West Bengal, India. BMC Cancer. 2017;17(1):782.

Golestani Eimani B, Sanati MH, Houshmand M, Ataei M, Akbarian F, Shakhssalim N. Expression and prognostic significance of bcl-2 and bax in the progression and clinical outcome of transitional bladder cell carcinoma. Cell J. 2014;15(4):356-363.

Gonzalez-Hormazabal P, Musleh M, Bustamante M, Stambuk J, Pisano R, Valladares H, Lanzarini E, Chiong H, Rojas J, Suazo J, Castro VG, Jara L, Berger Z. Polymorphisms in RAS/RAF/MEK/ERK Pathway Are Associated with Gastric Cancer. Genes (Basel). 2018;10(1):20.

Gou HF, Chen XC, Zhu J, et al. Expressions of COX-2 and VEGF-C in gastric cancer: correlations with lymphangiogenesis and prognostic implications. J Exp Clin Cancer Res. 2011;30(1):14.

Gravalos C, Jimeno A. HER2 in gastric cancer: a new prognostic factor and a novel therapeutic target. Ann Oncol. 2008;19(9):1523-1529.

Gulley ML. Genomic assays for Epstein-Barr virus-positive gastric adenocarcinoma. Exp Mol Med. 2015;47(1):e134.

Gunathilake MN, Lee J, Choi IJ, Kim YI, Ahn Y, Park C, Kim J. Association between the relative abundance of gastric microbiota and the risk of gastric cancer: a case-control study. Sci Rep. 2019;9(1):13589.

Hamada T, Nowak JA, Masugi Y, Drew DA, Song M, Cao Y, Kosumi K, Mima K, Twombly TS, Liu L, Shi Y, da Silva A, Gu M, Li W, Nosho K, Keum N, Giannakis M, Meyerhardt JA, Wu K, Wang M, Chan AT, Giovannucci EL, Fuchs CS, Nishihara R, Zhang X, Ogino

S. Smoking and Risk of Colorectal Cancer Sub-Classified by Tumor-Infiltrating T Cells. J Natl Cancer Inst. 2019;111(1):42-51.

Hamidi EN, Hajeb P, Selamat J, Abdull Razis AF. Polycyclic Aromatic Hydrocarbons (PAHs) and their Bioaccessibility in Meat: a Tool for Assessing Human Cancer Risk. Asian Pac J Cancer Prev. 2016;17(1):15-23.

He SS, Wang Y, Bao Y, Cai XY, Yang XL, Chen DM, Chen Y, Lu LX. Dynamic changes in plasma Epstein-Barr virus DNA load during treatment have prognostic value in nasopharyngeal carcinoma: a retrospective study. Cancer Med. 2018;7(4):1110-1117.

Higashi H, Tsutsumi R, Fujita A, Yamazaki S, Asaka M, Azuma T, Hatakeyama M. Biological activity of the Helicobacter pylori virulence factor CagA is determined by variation in the tyrosine phosphorylation sites. Proc Natl Acad Sci U S A. 2002;99(22):14428-33.

Hirabayashi M, Inoue M, Sawada N, Saito E, Abe SK, Hidaka A, Iwasaki M, Yamaji T, Shimazu T, Tsugane S. Helicobacter pylori infection, atrophic gastritis, and risk of pancreatic cancer: A population-based cohort study in a large Japanese population: the JPHC Study. Sci Rep. 2019;15;9(1):6099.

Hofmann M, Stoss O, Shi D, Büttner R, van de Vijver M, Kim W, Ochiai A, Rüschoff J, Henkel T. Assessment of a HER2 scoring system for gastric cancer: results from a validation study. Histopathology. 2008;52(7):797-805.

Hu B, El Hajj N, Sittler S, Lammert N, Barnes R, Meloni-Ehrig A. Gastric cancer: Classification, histology and application of molecular pathology. J Gastrointest Oncol. 2012;3(3):251-61.

Hu SL, Kong XY, Cheng ZD, Sun YB, Shen G, Xu WP, Wu L, Xu XC, Jiang XD, Huang DB. Promoter methylation of p16, Runx3, DAPK and CHFR genes is frequent in gastric carcinoma. Tumori. 2010;96(5):726-33.

IARC monograms on the evolution of carcinogenic risks to human. 1994.

Ibrahim M, Gilbert K. Management of gastric cancer in Indian population. Transl Gastroenterol Hepatol. 2017;2:64.

Iizasa H, Nanbo A, Nishikawa J, Jinushi M, Yoshiyama H. Epstein-Barr Virus (EBV)-associated gastric carcinoma. Viruses. 2012;4(12):3420-3439.

Jenson HB, Baillargeon J, Heard P, Moyer MP. Effects of Smokeless Tobacco and Tumor Promoters on Cell Population Growth and Apoptosis of B Lymphocytes Infected with Epstein–Barr Virus Types 1 and 2. Toxicol Appl Pharmacol. 1999;160(2):171-82.

Jiricny J. The multifaceted mismatch-repair system. Nat Rev Mol Cell Biol. 2006;7(5):335-46.

Kakiuchi M, Nishizawa T, Ueda H, Gotoh K, Tanaka A, Hayashi A, Yamamoto S, Tatsuno K, Katoh H, Watanabe Y, Ichimura T, Ushiku T, Funahashi S, Tateishi K, Wada I, Shimizu N, Nomura S, Koike K, Seto Y, Fukayama M, Aburatani H, Ishikawa S. Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. Nat Genet. 2014;46(6):583-587.

Kamper-Jørgensen M, Rostgaard K, Glaser SL, Zahm SH, Cozen W, Smedby KE, Sanjosé S, Chang ET, Zheng T, La Vecchia C, Serraino D, Monnereau A, Kane EV, Miligi L, Vineis P, Spinelli JJ, McLaughlin JR, Pahwa P, Dosman JA, Vornanen M, Hjalgrim H. Cigarette smoking and risk of Hodgkin lymphoma and its subtypes: a pooled analysis from the International Lymphoma Epidemiology Consortium (InterLymph). Annals of oncology: official journal of the European Society for Medical Oncology. 2013;24(9):2245-55.

Kaneda A, Matsusaka K, Aburatani H, Fukayama M. Epstein-Barr virus infection as an epigenetic driver of tumorigenesis. Cancer Res. 2012;72(14):3445-50.

Kang Y, Hu W, Bai E, Zheng H, Liu Z, Wu J, Jin R, Zhao C, Liang G. Curcumin sensitizes human gastric cancer cells to 5-fluorouracil through inhibition of the NFκB survival-signaling pathway. Onco Targets Ther. 2016;9:7373-7384.

Karin M, Greten FR. NF-kappaB: linking inflammation and immunity to cancer development and progression. Nat Rev Immunol. 2005;5:749–759.

Kato S, Tsukamoto T, Mizoshita T, Tanaka H, Kumagai T, Ota H Katsuyama T, Asaka M, Tatematsu M. High salt diets dose-dependently promote gastric chemical carcinogenesis in Helicobacter pylori-infected Mongolian gerbils associated with a shift in mucin production from glandular to surface mucous cells. Int J Cancer. 2006;119(7):1558-66

Katoh M. Function and cancer genomics of FAT family genes (review). Int J Oncol. 2012;41(6):1913-1918.

Katona BW, Rustgi AK. Gastric Cancer Genomics: Advances and Future Directions. Cell Mol Gastroenterol Hepatol. 2017;14;3(2):211-217.

Kim JJ, Tao H, Carloni E, Leung WK, Graham DY, Sepulveda AR. Helicobacter pylori impairs DNA mismatch repair in gastric epithelial cells. Gastroenterology. 2002;123(2):542-53.

Kim JY, Shin NR, Kim A, Lee HJ, Park WY, Kim JY, Lee CH, Huh GY, Park DY. Microsatellite instability status in gastric cancer: a reappraisal of its clinical significance and relationship with mucin phenotypes. Korean J Pathol. 2013;47(1):28-35.

Kim SM, An JY, Byeon Sj, Lee J, Kim KM, Choi MG, Lee JH, Sohn TS, Bae JM, Kim S. Prognostic value of mismatch repair deficiency in patients with advanced gastric cancer, treated by surgery and adjuvant 5-fluorouracil and leucovorin chemoradiotherapy. Eur J Surg Oncol. 2020;46(1):189-194.

Kinugasa H, Nouso K, Tanaka T, Miyahara K, Morimoto Y, Dohi C, Matsubara T, Okada H, Yamamoto K. Droplet digital PCR measurement of HER2 in patients with gastric cancer. Br J Cancer. 2015;112(10):1652-1655.

Kobayashi D, Kodera Y, Fujiwara M, Koike M, Nakayama G, Nakao A. Assessment of quality of life after gastrectomy using EORTC QLQ-C30 and STO22. World J Surg. 2011;35(2):357-64.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568-576.

Koushyar S, Powell AG, Vincan E, Phesse TJ. Targeting Wnt Signaling for the Treatment of Gastric Cancer. Int J Mol Sci. 2020;21(11):3927.

Kumagai K, Yamamoto N, Miyashiro I, Tomita Y, Katai H, Kushima R, Tsuda H, Kitagawa Y, Takeuchi H, Mukai M, Mano M, Mochizuki H, Kato Y, Matsuura N, Sano T. Multicenter study evaluating the clinical performance of the OSNA assay for the molecular detection of lymph node metastases in gastric cancer patients. Gastric Cancer. 2014;17(2):273-280.

Kusters JG, van Vliet AH, Kuipers EJ. Pathogenesis of Helicobacter pylori infection. Clin Microbiol Rev. 2006;19(3):449-490.

Kwon HC, Roh MS, Oh SY, Kim SH, Kim MC, Kim JS, Kim HJ. Prognostic value of expression of ERCC1, thymidylate synthase, and glutathione S-transferase P1 for 5-fluorouracil/oxaliplatin chemotherapy in advanced gastric cancer. Ann Oncol. 2007;18(3):504-9.

Ladeiras-Lopes R, Pereira AK, Nogueira A, Pinheiro-Torres T, Pinto I, Santos-Pereira R, Lunet N. Smoking and gastric cancer: systematic review and meta-analysis of cohort studies. Cancer Causes Control. 2008;19(7):689-701.

Lalruatfela B, Zoremsiami J, Jagetia GC. In vitro effect of tuibur (tobacco brew) on the viability of human blood lymphocytes. Science Vision. 2017;17(1):19-24.

Lan J, Xiong YY, Lin YX, Wang BC, Gong LL, Xu HS, Guo GS. Helicobacter pylori infection generated gastric cancer through p53-Rb tumor-suppressor system mutation and telomerase reactivation. World J Gastroenterol. 2003;9(1):54-58.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980-D985.

Lauren P. The two Histological Main Types of Gastric Carcinoma: Diffuse And So-Called Intestinal-Type Carcinoma. An Attempt at a Histo-Clinical Classification. Acta Pathol Microbiol Scand. 1965;64:31-49.

Lauwers GY, Carneiro F, Graham DY. Gastric carcinoma. In: Bowman FT, Carneiro F, Hruban RH, eds. Classification of Tumours of the Digestive System. Lyon:IARC;2010.

Lazăr D, Tăban S, Sporea I, Dema A, Cornianu M, Lazăr E, Goldiş A, Vernic C. Ki-67 expression in gastric cancer. Results from a prospective study with long-term follow-up. Rom J Morphol Embryol. 2010;51(4):655-61.

Lee JH, Kim SH, Han SH, An JS, Lee ES, Kim YS. Clinicopathological and molecular characteristics of Epstein-Barr virus-associated gastric carcinoma: a meta-analysis. J Gastroenterol Hepatol. 2009;24:354–65.

Lee YC, Chiang TH, Chou CK, Tu YK, Liao WC, Wu MS, Graham DY. Association Between Helicobacter pylori Eradication and Gastric Cancer Incidence: A Systematic Review and Meta-analysis. Gastroenterology. 2016;150(5):1113-1124.e5.

Li H, Yao G, Zhai J, Hu D, Fan Y. LncRNA FTX Promotes Proliferation and Invasion of Gastric Cancer via miR-144/ZFX Axis. Onco Targets Ther. 2019;12:11701-11713.

Li N, Lu N, Xie C. The Hippo and Wnt signalling pathways: crosstalk during neoplastic progression in gastrointestinal tissue. FEBS J. 2019;286(19):3745-3756 .

Li Q, Spriggs MK, Kovats S, Turk SM, Comeau MR, Nepom B, Hutt-Fletcher LM. Epstein-Barr virus uses HLA class II as a cofactor for infection of B lymphocytes. J Virol. 1997;71(6):4657-4662.

Lima MA, Ferreira MV, Barros MA, Pardini MI, Ferrasi AC, Mota RM, Rabenhorst SH. Relationship between EBV infection and expression of cellular proteins c-Myc, Bcl-2, and Bax in gastric carcinomas. Diagn Mol Pathol. 2008;17(2):82-9.

Liu HF, Liu WW, Fang DC, Men RP. Expression and significance of proapoptotic gene Bax in gastric carcinoma. World J Gastroenterol. 1999;5(1):15-17.

Liu L, Song X, Li X, Xue L, Ding S, Niu L, Xie L, Song X. A three-platelet mRNA set: MAX, MTURN and HLA-B as biomarker for lung cancer. J Cancer Res Clin Oncol. 2019;145(11):2713-2723.

Liu X, Cai H, Huang H, Long Z, Shi Y, Wang Y. The prognostic significance of apoptosis-related biological markers in Chinese gastric cancer patients. PLoS One. 2011;6(12):e29670.

Loh JT, Torres VJ, Cover TL. Regulation of Helicobacter pylori cagA expression in response to salt. Cancer Res. 2007;67(10):4709-15.

Losso GM, Moraes Rda S, Gentili AC, Messias-Reason IT. Microsatellite instability--MSI markers (BAT26, BAT25, D2S123, D5S346, D17S250) in rectal cancer. Arq Bras Cir Dig. 2012;25(4):240-4.

Lu B, Li M. Helicobacter pylori eradication for preventing gastric cancer. World J Gastroenterol. 2014;20(19):5660-5.

Machlowska J, Baj J, Sitarz M, Maciejewski R, Sitarz R. Gastric Cancer: Epidemiology, Risk Factors, Classification, Genomic Characteristics and Treatment Strategies. Int J Mol Sci. 2020;21(11):4012.

Machlowska J, Pucułek M, Sitarz M, Terlecki P, Maciejewski R, Sitarz R. State of the art for gastric signet ring cell carcinoma: from classification, prognosis, and genomic characteristics to specified treatments. Cancer Manag Res. 2019;11:2151-2161.

Madathil S, Senthil Kumar N, Zodinpuii D, Muthukumaran RB, Lalmuanpuii R, Nicolau B. Tuibur: tobacco in a bottle-commercial production of tobacco smoke-saturated aqueous concentrate. Addiction. 2018;113:577-580.

Magnelli L, Schiavone N, Staderini F, Biagioni A, Papucci L. MAP Kinases Pathways in Gastric Cancer. Int J Mol Sci. 2020;21(8):2893.

Majewski IJ, Kluijt I, Cats A, Scerri TS, de Jong D, Kluin RJ, Hansford S, Hogervorst FB, Bosma AJ, Hofland I, Winter M, Huntsman D, Jonkers J, Bahlo M, Bernards R. An α-E-catenin (CTNNA1) mutation in hereditary diffuse gastric cancer. J Pathol. 2013;229(4):621-629.

Mathew A, Gangadharan P, Varghese C, Nair MK. Diet and stomach cancer: a case-control study in South India. Eur J Cancer Prev. 2000;9(2):89-97.

Mathur P, Sathishkumar K, Chaturvedi M, Das P, Sudarshan KL, Santhappan S, Nallasamy V, John A, Narasimhan S, Roselind FS; ICMR-NCDIR-NCRP Investigator Group. Cancer Statistics, 2020: Report From National Cancer Registry Programme, India. JCO Glob Oncol. 2020;6:1063-1075.

McDonald ST. Comparison of Health Risks of Smoked Foods as Compared to Smoke Flavorings: Are Smoke Flavors "Healthier"?. Advances In Food Technology And Nutritional Sciences Open Journal, 2015;1:5.

Moghimi-Dehkordi B, Safaee A, Zali MR. Comparison of colorectal and gastric cancer: survival and prognostic factors. Saudi J Gastroenterol. 2009;15(1):18-23.

Moy KA, Fan Y, Wang R, Gao YT, Yu MC, Yuan JM. Alcohol and tobacco use in relation to gastric cancer: a prospective study of men in Shanghai, China. Cancer Epidemiol Biomarkers Prev. 2010;19(9):2287-97.

Mukherjee S, Madathil SA, Ghatak S, Jahau L, Pautu JL, Zohmingthanga J, Pachuau L, Nicolau B, Kumar NS. Association of tobacco smoke-infused water (tuibur) use by Mizo people and risk of Helicobacter pylori infection. Environ Sci Pollut Res Int. 2020;27(8):8580-8585.

Murata-Kamiya N, Kurashima Y, Teishikata Y, Yamahashi Y, Saito Y, Higashi H, Aburatani H, Akiyama T, Peek RM Jr, Azuma T, Hatakeyama M. Helicobacter pylori CagA interacts with E-cadherin and deregulates the beta-catenin signal that promotes intestinal transdifferentiation in gastric epithelial cells. Oncogene. 2007;26(32):4617-26.

Murphy G, Pfeiffer R, Camargo MC, Rabkin CS. Meta-analysis shows that prevalence of Epstein-Barr virus-positive gastric cancer differs based on sex and anatomic location [published correction appears in Gastroenterology. 2011;140(3):1109.

Nemati A, Mahdavi R, Naghizadeh Baghi A. Case-control study of dietary pattern and other risk factors for gastric cancer. Health Promot Perspect. 2012;2(1):20-27.

Nishino Y, Inoue M, Tsuji I, Wakai K, Nagata C, Mizoue T, Tanaka K, Tsugane S; Research Group for the Development and Evaluation of Cancer Prevention Strategies in Japan. Tobacco smoking and gastric cancer risk: an evaluation based on a systematic review of epidemiologic evidence among the Japanese population. Jpn J Clin Oncol. 2006;36(12):800-807.

Odenbreit S, Püls J, Sedlmaier B, Gerland E, Fischer W, Haas R. Translocation of Helicobacter pylori CagA into gastric epithelial cells by type IV secretion. Science. 2000;287(5457):1497-1500.

Oki E, Kakeji Y, Zhao Y, Yoshida R, Ando K, Masuda T, Ohgaki K, Morita M, Maehara Y. Chemosensitivity and survival in gastric cancer patients with microsatellite instability. Ann Surg Oncol. 2009;16(9):2510-2515.

Oliveira C, Pinheiro H, Figueiredo J, Seruca R, Carneiro F. Familial gastric cancer: genetic susceptibility, pathology, and implications for management. Lancet Oncol. 2015;16(2):e60-70.

Park DI, Yun JW, Park JH, Oh SJ, Kim HJ, Cho YK, Sohn CI, Jeon WK, Kim BI, Yoo CH, Son BH, Cho EY, Chae SW, Kim EJ, Sohn JH, Ryu SH, Sepulveda AR. HER-2/neu amplification is an independent prognostic factor in gastric cancer. Dig Dis Sci. 2006;51(8):1371-9.

Park SR, Park YS, Ryu MH, Ryoo BY, Woo CG, Jung HY, Lee JH, Lee GH, Kang YK. Extra-gain of HER2-positive cases through HER2 reassessment in primary and metastatic sites in advanced gastric cancer with initially HER2-negative primary tumours: Results of GASTric cancer HER2 reassessment study 1 (GASTHER1). Eur J Cancer. 2016;53:42-50.

Pećina-Šlaus N, Kafka A, Salamon I, Bukovac A. Mismatch Repair Pathway, Genome Stability and Cancer. Front Mol Biosci. 2020;7:122.

Peleteiro B, Bastos A, Ferro A. Lunet N. Prevalence of Helicobacter pylori Infection Worldwide: A Systematic Review of Studies with National Coverage. Dig Dis Sci. 2014;59(8):1698–1709.

Perez-Perez GI, Bosques-Padilla FJ, Crosatti ML, Tijerina-Menchaca R, Garza-González E. Role of p53 codon 72 polymorphism in the risk of development of distal gastric cancer. Scand J Gastroenterol. 2005;40(1):56-60.

Pfeffer S, Zavolan M, Grässer FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C, Tuschl T. Identification of virus-encoded microRNAs. Science. 2004;304(5671):734-6.

Phukan RK, Zomawia E, Hazarlka NC, Baruah D, Mahanta J. High prevalence of stomach cancer among the people of Mizoram, India. Curr Sci. 2004;87:285-62004.

Phukan RK, Zomawia E, Narain K, Hazarika NC, Mahanta J. Tobacco use and stomach cancer in Mizoram, India. Cancer Epidemiol Biomarkers Prev. 2005;14(8):1892-6.

Piazuelo MB, Correa P. Gastric cáncer: Overview. Colomb Med (Cali). 2013;44(3):192-201.

Polom K, Marano L, Marrelli D, De Luca R, Roviello G, Savelli V, Tan P, Roviello F. Meta-analysis of microsatellite instability in relation to clinicopathological characteristics and overall survival in gastric cancer. Br J Surg. 2018;105(3):159-167.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Vander Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, Shakir K, Thibault J, Chandran S, Whelan C, Lek M, Gabriel S, Daly MJ , Neale B, MacArthur DG, Bankset E. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017:doi: https://doi.org/10.1101/201178.

Poynter JN, Haile RW, Siegmund KD, Campbell PT, Figueiredo JC, Limburg P, et al. Associations between smoking, alcohol consumption, and colorectal cancer, overall and by tumor microsatellite instability status. Cancer Epidemiol Biomarkers Prev. 2009;18:2745-50.

Qiao Y, Li T, Zheng S, Wang H. The Hippo pathway as a drug target in gastric cancer. Cancer Lett. 2018;28;420:14-25.

Rao DN, Ganesh B, Dinshaw KA, Mohandas KM. A case-control study of stomach cancer in Mumbai, India. Int J Cancer. 2002;99(5):727-31.

Rasheed F, Ahmad T, Ali M, Ali S, Ahmed S, Bilal R. High frequency of cagA and vacA s1a/m2 Genotype among *Helicobacter pylori* Infected Gastric Biopsies of Pakistani Children. Malaysian J Microbiol. 2011;7(3):167-170

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

Resende C, Gomes CP, Machado JC. Review: Gastric cancer: Basic aspects. Helicobacter. 2020;25 Suppl1:e12739.

Rickinson AB, Young LS, Rowe M. Influence of the Epstein-Barr virus nuclear antigen EBNA 2 on the growth phenotype of virus-transformed B cells. J. Virol. 1987;61(5):1310–7.

Rubin, H. Synergistic mechanisms in carcinogenesis by polycyclic aromatic hydrocarbons and by tobacco smoke: a bio-historical perspective with updates. Carcinogenesis. 2001;22(12):1903-30.

Sarrió D, Moreno-Bueno G, Hardisson D, Sánchez-Estévez C, Guo M, Herman JG, Gamallo C, Esteller M, Palacios J. Epigenetic and genetic alterations of APC and CDH1 genes in lobular breast cancer: relationships with abnormal E-cadherin and catenin expression and microsatellite instability. Int J Cancer. 2003;20;106(2):208-15.

Scheuner M, McNeel T, Freedman A. Population prevalence of familial cancer and common hereditary cancer syndromes. The 2005 California Health Interview Survey. Genet Med. 2010;12(11):726–735.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11(4):361-2.

Shannon-Lowe C, Rickinson A. The Global Landscape of EBV-Associated Tumors. Front Oncol. 2019;9:713.

Shinozaki-Ushiku A, Kunita A, Fukayama M. Update on Epstein-Barr virus and gastric cancer (review). Int J Oncol. 2015;46(4):1421-34.

Shinozaki-Ushiku A, Kunita A, Isogai M, Hibiya T, Ushiku T, Takada K, Fukayama M. Profiling of Virus-Encoded MicroRNAs in Epstein-Barr Virus-Associated Gastric Carcinoma and Their Roles in Gastric Carcinogenesis. J Virol. 2015;89(10):5581-91.

Singh P, Toom S, Huang Y. Anti-claudin 18.2 antibody as new targeted therapy for advanced gastric cancer. J Hematol Oncol. 2017;12;10(1):105.

Sitarz R, Skierucha M, Mielko J, Offerhaus GJA, Maciejewski R, Polkowski WP. Gastric cancer: epidemiology, prevention, classification, and treatment. Cancer Manag Res. 2018;10:239-248.

Smatti MK, Al-Sadeq DW, Ali NH, Pintus G, Abou-Saleh H, Nasrallah GK. Epstein-Barr Virus Epidemiology, Serology, and Genetic Variability of LMP-1 Oncogene Among Healthy Population: An Update. Front Oncol. 2018;8:211.

Smyth EC, Wotherspoon A, Peckitt C, Gonzalez D, Hulkki-Wilson S, Eltahir Z, Fassan M, Rugge M, Valeri N, Okines A, Hewish M, Allum W, Stenning S, Nankivell M, Langley R, Cunningham D. Mismatch repair deficiency, microsatellite instability, and survival: an exploratory analysis of the medical research council adjuvant gastric infusional chemotherapy (MAGIC) trial. JAMA Oncol. 2017;3(9):1197-203.

Steevens J, Schouten LJ, Goldbohm RA, van den Brandt PA. Alcohol consumption, cigarette smoking and risk of subtypes of oesophageal and gastric cancer: a prospective cohort study. Gut. 2010;59(1):39-48.

Sung NY, Choi KS, Park EC, Park K, Lee SY, Lee AK, Choi IJ, Jung KW, Won YJ, Shin HR. Smoking, alcohol and gastric cancer risk in Korean men: the National Health Insurance Corporation Study. Br J Cancer. 2007;97(5):700-704.

Suraweera N, Duval A, Reperant M, Vaury C, Furlan D, Leroy K, Seruca R, Iacopetta B, Hamelin R. Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR. Gastroenterology. 2002;123(6):1804-11.

Suzuki A, Katoh H, Komura D, Kakiuchi M, Tagashira A, Yamamoto S, Tatsuno K, Ueda H, Nagae G, Fukuda S, Umeda T, Totoki Y, Abe H, Ushiku T, Matsuura T, Sakai E, Ohshima T, Nomura S, Seto Y, Shibata T, Rino Y, Nakajima A, Fukayama M, Ishikawa S, Aburatani H. Defined lifestyle and germline factors predispose Asian populations to gastric cancer. Sci Adv. 2020;6;6(19):eaav9778.

Takada K. Epstein-Barr virus and gastric carcinoma. Mol Pathol. 2000;53(5):255-261.

Tan P, Yeoh KG. Genetics and Molecular Pathogenesis of Gastric Adenocarcinoma. Gastroenterology. 2015;149(5):1153-1162.e3.

Tatematsu M, Takahashi M, Fukushima S, Hananouchi M, Shirai T. Effects in Rats of Sodium Chloride on Experimental Gastric Cancers Induced by N-Methyl-N′-nitro-N-nitrosoguanidine or 4-Nitroquinoline-1-oxide2. J Natl Cancer Inst. 1975;55(1):101-6.

Tavakoli A, Monavari SH, Solaymani Mohammadi F, Kiani SJ, Armat S, Farahmand M. Association between Epstein-Barr virus infection and gastric cancer: a systematic review and meta-analysis. BMC Cancer. 2020;20(1):493.

Theuer CP, Campbell BS, Peel DJ, Lin F, Carpenter P, Ziogas A, Butler JA. Microsatellite Instability in Japanese vs European American Patients With Gastric Cancer. Arch Surg. 2002;137(8):960–966.

Ting-Hin Ho, Justine Sitz, Qingtang Shen, Ariane Leblanc-Lacroix, Eric I. Campos, Ivan Borozan, Edyta Marcon, Jack Greenblatt, Amelie Fradet-Turcotte, Dong-Yan Jin, Lori Frappier. A Screen for Epstein-Barr Virus Proteins That Inhibit the DNA Damage Response Reveals a Novel Histone Binding Protein. Journal of Virology. 2018;92(14):e00262-18.

Tramacere I, Negri E, Pelucchi C, Bagnardi V, Rota M, Scotti L, Islami F, Corrao G, La Vecchia C, Boffetta P. A meta-analysis on alcohol drinking and gastric cancer risk. Ann Oncol. 2012;23(1):28-36.

Truong CD, Feng W, Li W, Khoury T, Li Q, Alrawi S, Yu Y, Xie K, Yao J, Tan D. Characteristics of Epstein-Barr virus-associated gastric cancer: a study of 235 cases at a comprehensive cancer center in U.S.A. J Exp Clin Cancer Res. 2009;28(1):14.

Tsugane S, Sasazuki S. Diet and the risk of gastric cancer: review of epidemiological evidence. Gastric Cancer. 2007;10(2):75-83.

van Beek J, zur Hausen A, Klein Kranenbarg E, van de Velde CJ, Middeldorp JM, van den Brule AJ, Meijer CJ, Bloemena E. EBV-positive gastric adenocarcinomas: a distinct clinicopathologic entity with a low frequency of lymph node involvement. J Clin Oncol. 2004;15;22(4):664-70.

Van Cutsem E, de Haas S, Kang YK, Ohtsu A, Tebbutt NC, Ming Xu J, Peng Yong W, Langer B, Delmar P, Scherer SJ, Shah MA. Bevacizumab in combination with chemotherapy as first-line therapy in advanced gastric cancer: a biomarker evaluation from the AVAGAST randomized phase III trial. J Clin Oncol. 2012;10;30(17):2119-27.

van der Post RS, Vogelaar IP, Carneiro F, Guilford P, Huntsman D, Hoogerbrugge N, Caldas C, Schreiber KE, Hardwick RH, Ausems MG, Bardram L, Benusiglio PR, Bisseling TM, Blair V, Bleiker E, Boussioutas A, Cats A, Coit D, DeGregorio L, Figueiredo J, Ford JM, Heijkoop E, Hermens R, Humar B, Kaurah P, Keller G, Lai J, Ligtenberg MJ, O'Donovan M, Oliveira C, Pinheiro H, Ragunath K, Rasenberg E, Richardson S, Roviello F, Schackert H, Seruca R, Taylor A, Ter Huurne A, Tischkowitz M, Joe ST, van Dijck B, van Grieken NC, van Hillegersberg R, van Sandick JW, Vehof R, van Krieken JH, Fitzgerald RC. Hereditary diffuse gastric cancer: updated clinical guidelines with an emphasis on germline CDH1 mutation carriers. J Med Genet. 2015;52(6):361-74.

Varlet V, Knockaert C, Prost C, Serot T. Comparison of Odor-Active Volatile Compounds of Fresh and Smoked Salmon. Journal of Agricultural and Food Chemistry. 2006;54(9):3391-401.

Verma Y, Pradhan P, Gurung, N. Sapkot SD, Giri P, Sndas P, Bhattaria BN, Nadayil D, Ramnath T, Nandakumar A. Population-based cancer incidence in Sikkim, India: report on ethnic variation. Br J Cancer 106, 962–965 (2012).

Villacis RA, Miranda PM, Gomy I, Santos EM, Carraro DM, Achatz MI, Rossi BM, Rogatto SR. Contribution of rare germline copy number variations and common susceptibility loci in Lynch syndrome patients negative for mutations in the mismatch repair genes. Int J Cancer. 2016;138:1928–1935.

Wang F, Meng W, Wang B, Qiao L. Helicobacter pylori-induced gastric inflammation and gastric cancer. Cancer Lett. 2014;345(2):196-202.

Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data Nucleic Acids Research. 2010;38(16):e164.

Wang K, Liu R, Li J, Mao J, Lei Y, Wu J, Zeng J, Zhang T, Wu H, Chen L, Huang C, Wei Y. Quercetin induces protective autophagy in gastric cancer cells: involvement of Akt-mTOR- and hypoxia-induced factor 1α-mediated signaling. Autophagy. 2011;7(9):966-78.

Wang XQ, Terry PD, Yan H. Review of salt consumption and stomach cancer risk: epidemiological and biological evidence. World J Gastroenterol. 2009;15(18):2204-2213.

Wang Z, Zhang X, Hu J, Zeng W, Liang J, Zhou H, Zhou Z. Predictive factors for lymph node metastasis in early gastric cancer with signet ring cell histology and their impact on the surgical strategy: analysis of single institutional experience. J Surg Res. 2014;191(1):130-3.

Wang, PG, Li YT, Pan Y, Gao ZZ, Guan XW, Jia L, Liu FT. Lower expression of Bax predicts poor clinical outcome in patients with glioma after curative resection and radiotherapy/chemotherapy. J Neurooncol. 2019;141(1);71–81.

Warneke VS, Behrens HM, Böger C, Becker T, Lordick F, Ebert MPA, Rocken C. Her2/neu testing in gastric cancer: evaluating the risk of sampling errors. Ann Oncol. 2013;24(3):725-733.

Wong BC, Lam SK, Wong WM, Chen JS, Zheng TT, Feng RE, Lai KC, Hu WH, Yuen ST, Leung SY, Fong DY, Ho J, Ching CK, Chen JS; China Gastric Cancer Study Group. Helicobacter pylori eradication to prevent gastric cancer in a high-risk region of China: a randomized controlled trial. JAMA. 2004;291(2):187-94.

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

Wroblewski LE, Peek RM Jr. Helicobacter pylori in gastric carcinogenesis: mechanisms. Gastroenterol Clin North Am. 2013;42(2):285-298.

Wu CY, Kuo KN, Wu MS, Chen YJ, Wang CB, Lin JT. Early Helicobacter pylori eradication decreases risk of gastric cancer in patients with peptic ulcer disease. Gastroenterology. 2009;137(5):1641-8.e1-2.

Wu Y, Fan Y, Jiang Y, Wang Y, Liu H, Wei M. Analysis of risk factors associated with precancerous lesion of gastric cancer in patients from eastern China: a comparative study. J Cancer Res Ther. 2013;9(2):205-9.

Yusefi AR, Bagheri Lankarani K, Bastani P, Radinmanesh M, Kavosi Z. Risk Factors for Gastric Cancer: A Systematic Review. Asian Pac J Cancer Prev. 2018;27;19(3):591-603.

Zali H, Rezaei-Tavirani M, Azodi M. Gastric cancer: prevention, risk factors and treatment. Gastroenterol Hepatol Bed Bench. 2011;4(4):175-185.

Zanella L, Riquelme I, Buchegger K, Abanto M, Ili C, Brebi P. A reliable Epstein-Barr Virus classification based on phylogenomic and population analyses. Sci Rep. 2019;9(1):9829.

Zang ZJ, Cutcutache I, Poon SL, Zhang SL, McPherson JR, Tao J, Rajasegaran V, Heng HL, Deng N, Gan A, Lim KH, Ong CK, Huang D, Chin SY, Tan IB, Ng CC, Yu W, Wu Y, Lee M, Wu J, Poh D, Wan WK, Rha SY, So J, Salto-Tellez M, Yeoh KG, Wong WK, Zhu YJ, Futreal PA, Pang B, Ruan Y, Hillmer AM, Bertrand D, Nagarajan N, Rozen S, Teh BT, Tan P. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat Genet. 2012;44(5):570-4.

Zavros Y, Eaton KA, Kang W, Rathinavelu S, Katukuri V, Kao JY, Samuelson LC, Merchant JL. Chronic gastritis in the hypochlorhydric gastrin-deficient mouse progresses to adenocarcinoma. Oncogene. 2005;31;24(14):2354-66.

Zhang XL, Yang YS, Xu DP, Qu JH, Guo MZ, Gong Y, Huang J. Comparative study on overexpression of HER2/neu and HER3 in gastric cancer. World J Surg. 2009;33(10):2112-8.

Zhang Z, Liu Y, Liu P, Yang L, Jiang X, Luo D, Yang D. Non-invasive detection of gastric cancer relevant d-amino acids with luminescent DNA/silver nanoclusters. Nanoscale. 2017;9(48):19367-19373.

Zhu L, Li Z, Wang Y, Zhang C, Liu Y, Qu X. Microsatellite instability and survival in gastric cancer: A systematic review and meta-analysis. Mol Clin Oncol. 2015;3(3):699-705.

List of Abbreviations

Abbreviations	**Full forms**
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
GC content	Guanine and Cytosine content
DNA	Deoxyribonucleic Acid
PCR	Polymerase chain reaction
16S rRNA	16S-Ribosomal Ribonucleic Acid
pH	Negative Ion Of Hydrogen Ion Concentration
mM	Milli Molar
M	Molar
mm	Millimeter
h	Hour
d	Days
ng	Nanogram
pmol	Pico-Mol
µM	Micromolar
dNTPs	Deoxynucleotides
U/ µl	Unit Per MicroLiter
TAE	Tris Base, Acetic Acid and EDTA
TBE	Tris Base, Boric Acid and EDTA
UV	Ultra-Violet
EMBL	European Molecular Biology Laboratory
UCSC	University of California, Santa Cruz
IARC	International Agency for Research on Cancer
AJCC	American joint committee on cancer
ACRG	Asian Cancer Research Group
TCGA	Tumor Cancer Genome Atlas

HDGC	Hereditary diffuse gastric cancer
GAPPS	Gastric adenocarcinoma and proximal polyposis of the stomach
pTNM	Pathological tumor, node and metastasis
dbSNP	Database for SNP
bp	Base Pair
g	Gram
mg/ml	Milligram Per Milliliter
NaCl	Sodium Chloride
TP53	Tumor Protein 53
CDH1	Cadherin-1
GC	Gastric Cancer
HC	Healthy Control
cagA	cytotoxin-associated gene A
CNV	Copy Number Variation
ddPCR	Droplet Digital PCR
MMR	Mismatch repair
MSI	Microsatellite Instability
ROS	Reactive Oxygen Species
OS	Overall survival
DFS	Disease-free survival
HRs	Hazard ratio
ROC	Receiver-operating characteristic
AUC	Area under the curves
RBCs	Red Blood Cells
rpm	Rotation per Minute
IHC	Immunohistochemistry
FFPE	Formalin fixed Paraffin embedded
MSS	Microsatellite Stable

CIN	Chromosomal instability
FIGS	Familial intestinal gastric cancer
miRNA	Micro RNA
NGS	Next- generation sequencing
WES	Whole exome sequencing
EBV	*Epstein–Barr virus*
H. pylori	*Helicobacter pylori*
EBV	*Epstein Bar Virus*
ORs	Odds Ratios
CIs	95% Confidence Intervals
df	Degree of Freedom
SPSS	Statistical Package of the Social Sciences
LR	Logistic Regression
ATP	Adenosine Tri-Phosphate
del	Deletion
int	Insertion
Arg	Arginine
Pro	Proline
Ile	Isoleucine
Val	Valine
NCRP	National Cancer Registry Program
AAR	Age-Adjusted Rates
CVC	Cross Validation Consistency
MAP3K6	Mitogen-activated protein kinase kinase kinase 6
HER2	Epidermal growth receptor 2

BRIEF BIODATA

Payel Chakraborty
E-mail: pchakrabortybiotech86@gmail.com
Phone no: 6291237567

Career objective: To work in the field of cancer genomics research so as to advance discoveries in cancer treatment.

Area of interest: Cancer genetics, Gastric cancer, Cancer proteomics, Cancer biomarkers

Academic Qualifications

2017-	PhD (Biotechnology)	Ongoing	Mizoram University
2006 - 2008	M.Sc. Biotechnology	76.16%.	The Berhampur University, Orissa
2003 - 2006	B.Sc. Biotechnology	51.87%.	The Burdwan University, West Bengal
2002 - 2003	HSC	54.4%.	Tripura Board of Higher Secondary Education
2000 - 2001	SSC	68.25%.	Tripura Board of Secondary Education

Conference and Workshops

- ❑ Oral presentation on “Association of epidemiological factors with pathogen infection in gastric cancer patients” in National conference on “Microbes in Health, Agriculture & Environment”. Organized by Department of Biotechnology, School of Life Science, Mizoram University, during 20th- 21st June 2019.
- ❑ Oral presentation on “Panel of significant risk factors predicts gastric cancer at early stage” 2nd Annual Convention of North East (India) Academy of Science and Technology (NEAST) & International Seminar on Recent Advances in Science and Technology (ISRAST) during 16th – 18th November, 2020.
- ❑ Participated in the International workshop on “Molecular Phylogeny and Next-Generation Sequencing” organized by Department of Biotechnology, Mizoram University during 19th – 28th June, 2017.
- ❑ Participated in the 3rd Advanced Research Training workshop on “Understanding Human Disease and Improving Human Health using Genomics-Driven Approaches” organized by National Institute of Biomedical Genomics, Kalyani during 23rd – 31st July, 2018.

- ❑ Participated in the NER Training program on “Cell and Molecular Biology” organized by DBT – NER Biotechnology / Bioinformatics Training Centre, ACTREC Mumbai during 3rd – 7th November, 2020.

Research Experience:

I have worked in microsatellite instability study from gastric tumor samples in Genetic Analyzer 3500 Sanger sequencing platform through fragment analysis workflow.

I have worked on pathogen identifications (e.g. *Helicobacter pylori* and *Epstein-Barr virus*) from gastric tumor samples using multiplex PCR.

I have analysed the complete mitochondrial genome sequencing for gastric tumors and healthy control blood samples in Illumina Hiseq 2000 Next Generation Sequencing platform for mutation detection.

I have worked on the Illumina Hiseq 2000 Next Generation Sequencing platform for whole-exome and targeted re-sequencing analysis of gastric cancer.

I have also worked on NGS data analysis (e.g. BbB, Varscan2, ANNOVAR and GATK tool) for analysing the exome and targeted data of gastric cancer through various workflows available in the software.

I have worked in the immunohistochemistry analysis for specific protein using gastric cancer and adjacent normal formalin-fixed paraffin-embedded tissue samples.

I have also worked in analysing the copy number variation of tumor suppressor and oncogene associated with gastric cancer through droplet digital PCR.

I am also working in different bioinformatics pipelines for somatic and germline mutations detection and identified the recurrent driver mutations.

I have identified specific mutations to develop the prognostic markers associated with gastric cancer.

Publications:

Chakraborty P, Ghatak S**,** Mukherjee S, Chhakchhuak L, Chenkual S, Zomuana T, Lalruatfela ST, Maitra A, and Senthil Kumar N, Novel Somatic Mutations of the CDH1 Gene Associated with Gastric Cancer: Prediction of Pathogenicity Using Comprehensive *In silico.* Methods Current Pharmacogenomics and Personalized Medicine, 2020: DOI: 10.2174/1875692117999201109210911

Ghatak S, **Chakraborty P**, Sarkar SR, Chowdhury B, Bhaumik A, Senthil Kumar N. Novel APC gene mutations associated with protein alteration in diffuse type gastric cancer. BMC Med Genet. 2017 Jun 2;18(1):61. doi: 10.1186/s12881-017-0427-2.

Yadav RP, Ghatak S, **Chakraborty P**, Lalrohlui F, Kannan R, Kumar R, Pautu JL, Zomingthanga J, Chenkual S, Mutthkumaran R, Senthil Kumar N. Lifestyle chemical carcinogens associated with mutations in cell cycle regulatory genes increases the susceptibility to gastric cancer risk. Environ Sci Pollut Res. 2018 Nov;25(31):31691-31704.doi: 10.1007/s11356-018-3080-1.

Chakraborty P, Ghatak S, Chenkual S, Zomingthanga J, Bawihtlung Z, Khenglawt L, Pautu JL, Maitra A, Chhakchhuak L, Senthil Kumar N. Panel of significant risk factors predicts early stage gastric cancer and indication of poor prognostic association with pathogens and Microsatellite Stability. Genes and Environment.

Papers under Preparation

Chakraborty P, Ghatak S, Karkulang S, Palodi A, Zohmingthanga J, Khenglawt L, Pautu JL. Maitra A, Chhakchhuak L, Senthil Kumar N. TP53 mutaion and EBV infection driving Gastric Camcer in Mizo population. BMC Cancer.

Chakraborty P, Ghatak S, Karkulang S, Palodi A, Zohmingthanga J, Khenglawt L, Pautu JL, Maitra A, Chhakchhuak L, Senthil Kumar N. Association of recurrent pathogenic germline variant in chromatin remodeling gene in Gastric Cancer in Mizo population. Cancer Cell.

Personal details:

Date of Birth	: Junel 04, 1986
Gender	: Female
Nationality	: Indian
Permanent Address	: Vill-Siza, P.O- Khamargachi, Dist.-Hooghly Pin no. 712515, West Bengal

Declaration: I hereby declare that all the above information is true to the best of my knowledge and belief.

Name: Payel Chakraborty
Date: 26-02-2021

PARTICULARS OF THE CANDIDATE

NAME OF CANDIDATE: PAYEL CHAKRABORTY

DEGREE: PhD

DEPARTMENT: BIOTECHNOLOGY

TITLE OF THESIS: Identification of recurrent genomic alterations in Gastric adenocarcinoma in Mizo population

DATE OF ADMISSION: 17.08.2016

APPROVAL OF RESEARCH PROPOSAL:

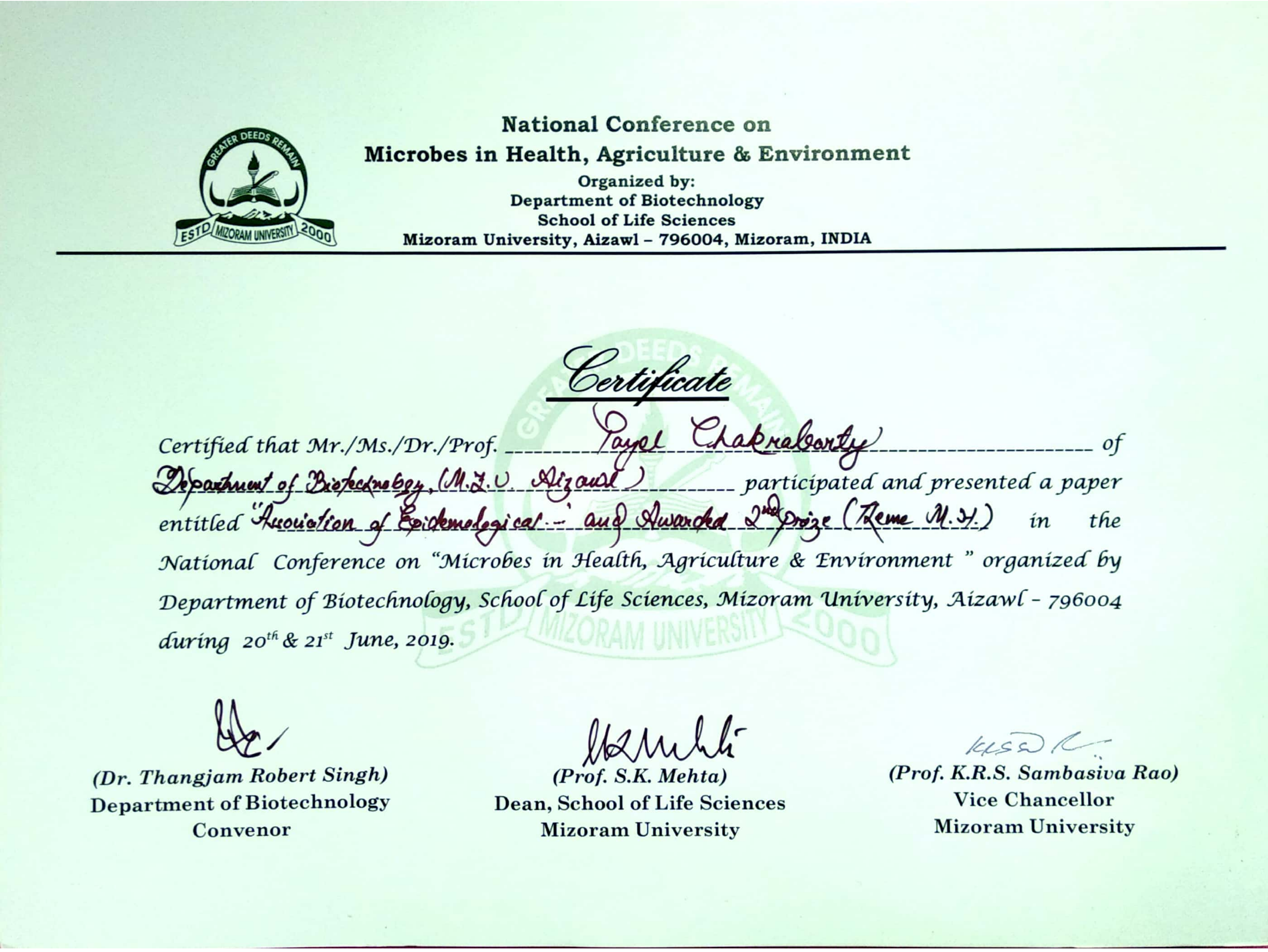
1. DRC: 11.05.2017
2. BOS: 12.05.2017
3. SCHOOL BOARD: 26.05.2017

MZU REGISTRATION NO.: 1600803

Ph.D. REGISTRATION NO. & DATE: MZU/PH.D./1045 OF 26.05.2017

Head

Department of Biotechnology



National Conference on

Microbes in Health, Agriculture & Environment

Organized by:
Department of Biotechnology
School of Life Sciences
Mizoram University, Aizawl – 796004, Mizoram, INDIA

Certificate

Certified that Mr./Ms./Dr./Prof. Payel Chakrabarty *of* Department of Biotechnology, (M.Z.U. Aizawl) *participated and presented a paper entitled* "Association of Epidemological ..." and Awarded 2nd prize (Theme M.H.) *in the National Conference on "Microbes in Health, Agriculture & Environment " organized by Department of Biotechnology, School of Life Sciences, Mizoram University, Aizawl – 796004 during 20th & 21st June, 2019.*

(Dr. Thangjam Robert Singh)
Department of Biotechnology
Convenor

(Prof. S.K. Mehta)
Dean, School of Life Sciences
Mizoram University

(Prof. K.R.S. Sambasiva Rao)
Vice Chancellor
Mizoram University

NEAST

2nd Annual Convention of

North East (India) Academy of Science and Technology (NEAST)

&

International Seminar on Recent Advances in Science and Technology (ISRAST)

(16th -18th November 2020)

(Virtual)

11215150233

Certificate

This is to certify that

Mr./Ms./Dr./Prof. Payel Chakraborty

Department of Biotechnology, Mizoram University, Aizawl–796004, Mizoram, India

has attended and presented an Oral Presentation entitled, "Panel of Significant Risk Factors Predicts Gastric Cancer at Early Stage" in the 2nd Annual Convention of North East (India) Academy of Science and Technology (NEAST) & International Seminar on Recent Advances in Science and Technology (IRSRAST) during 16th-18th November 2020 (Virtual) organized by NEAST, Mizoram University, Aizawl-796004, Mizoram (India) .

Diwak

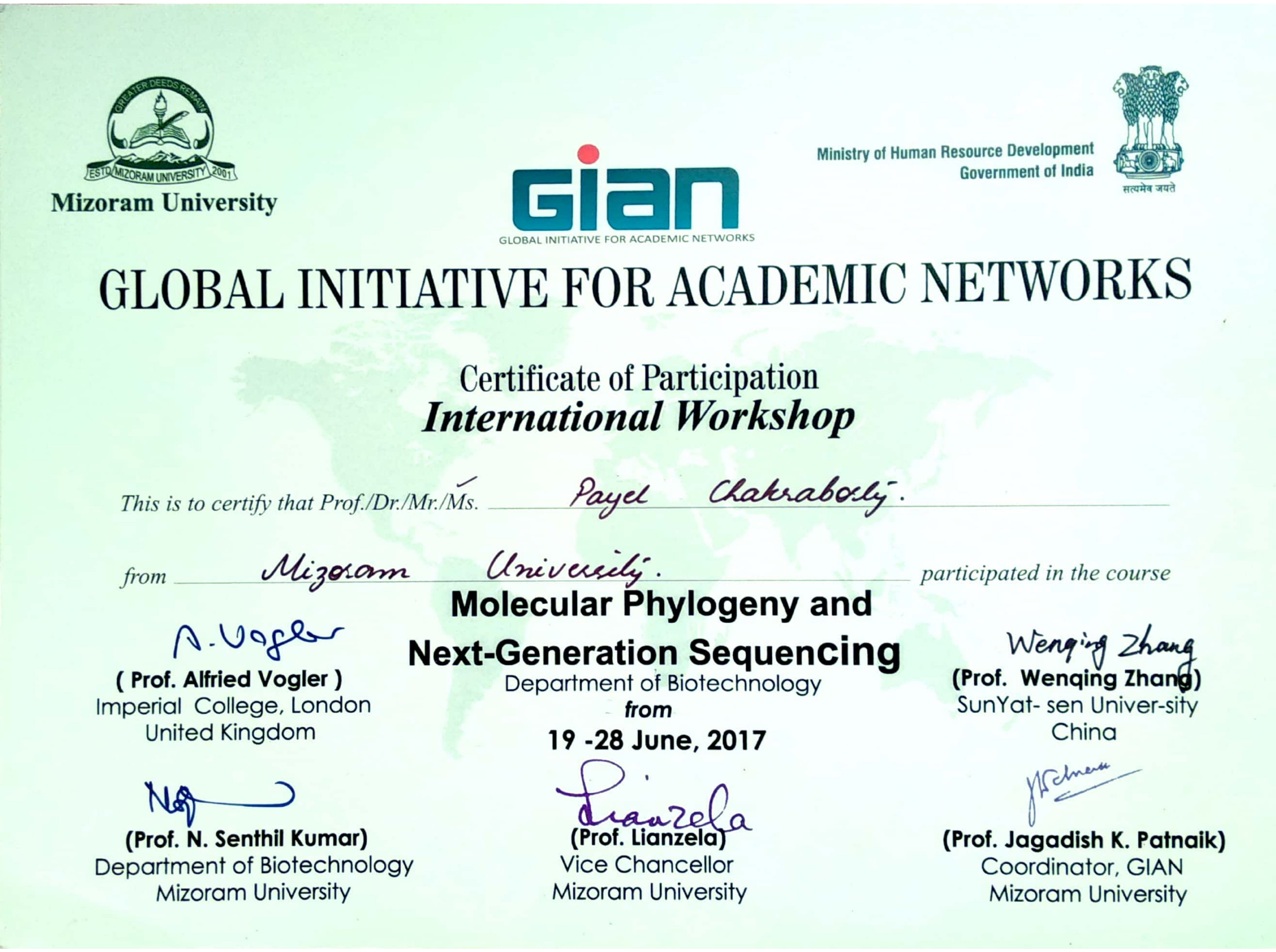
Prof. Diwakar Tiwari
General Secretary, NEAST

R. Lalthantluang

Prof. R. Lalthantluanga
President, NEAST

North East (India) Academy of Science & Technology (NEAST), Mizoram University, Aizawl-796004 (India)

Reg. No.: MSR 903 under Mizoram Societies Act-2018



Mizoram University

Ministry of Human Resource Development
Government of India

सत्यमेव जयते

GIAN
GLOBAL INITIATIVE FOR ACADEMIC NETWORKS

GLOBAL INITIATIVE FOR ACADEMIC NETWORKS

Certificate of Participation
International Workshop

This is to certify that Prof./Dr./Mr./Ms. Payel Chakraborty.

from Mizoram University. *participated in the course*

Molecular Phylogeny and
Next-Generation Sequencing
Department of Biotechnology
from
19 -28 June, 2017

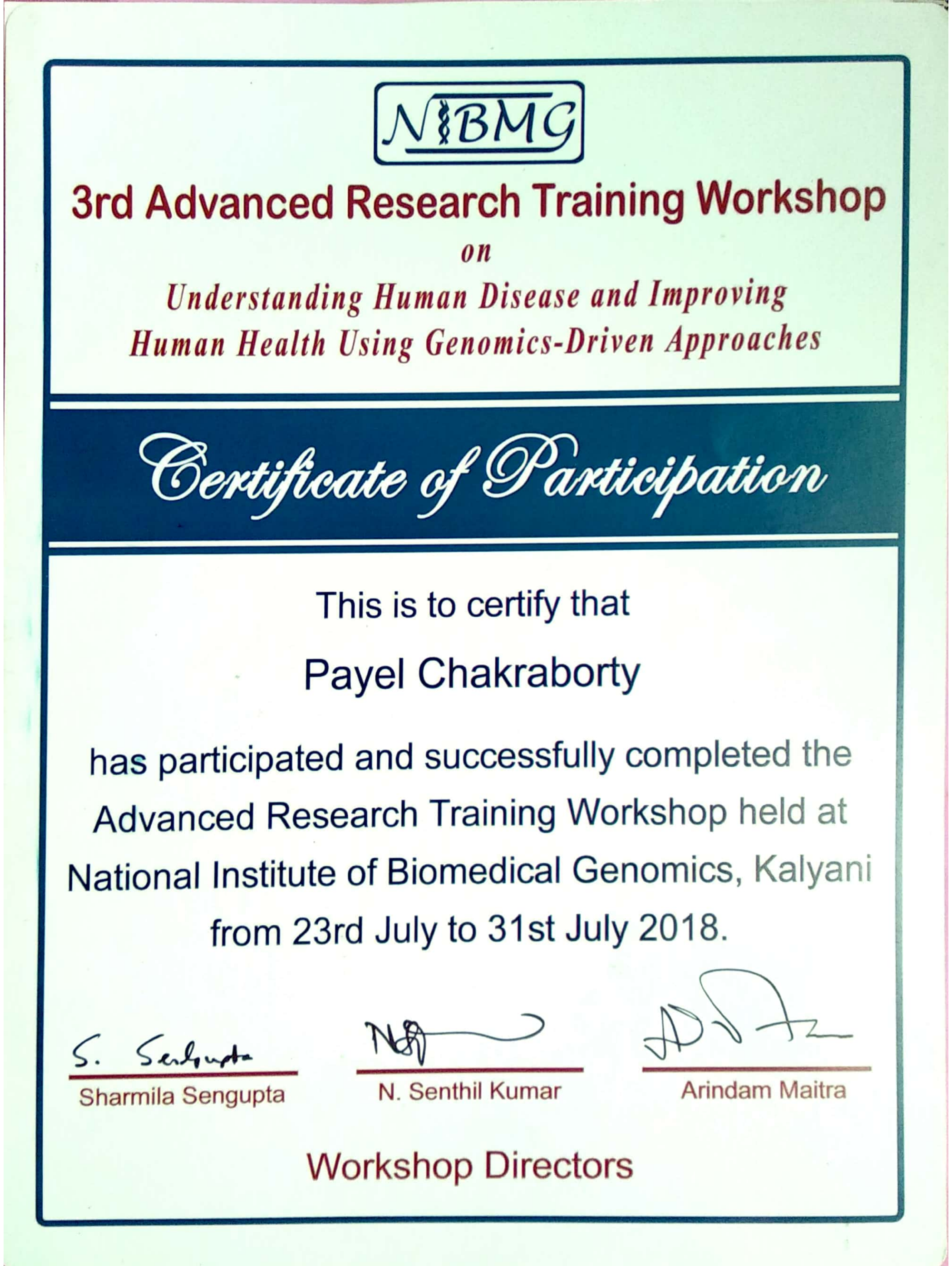
A. Vogler
(Prof. Alfried Vogler)
Imperial College, London
United Kingdom

Wenqing Zhang
(Prof. Wenqing Zhang)
SunYat- sen Univer-sity
China

(Prof. N. Senthil Kumar)
Department of Biotechnology
Mizoram University

Lianzela
(Prof. Lianzela)
Vice Chancellor
Mizoram University

(Prof. Jagadish K. Patnaik)
Coordinator, GIAN
Mizoram University



NIBMG

3rd Advanced Research Training Workshop

on

Understanding Human Disease and Improving Human Health Using Genomics-Driven Approaches

Certificate of Participation

This is to certify that

Payel Chakraborty

has participated and successfully completed the Advanced Research Training Workshop held at National Institute of Biomedical Genomics, Kalyani from 23rd July to 31st July 2018.

S. Sengupta

Sharmila Sengupta	N. Senthil Kumar	Arindam Maitra

Workshop Directors

TMC

Department
of Biotechnology
Govt. of India

Tata Memorial Centre
Advanced Centre for Treatment, Research and Education in Cancer (ACTREC),
Kharghar, Navi Mumbai - India

Certificate

This is to certify that

Payel Chakraborty

Has participated in the NER Training program on

Cell and Molecular Biology

Organised by
DBT – NER Biotechnology / Bioinformatics Training Centre,
Advanced centre for Treatment, Research & Education in cancer,
Kharghar, Navi Mumbai. India.

February 3rd – 7th, 2020

Varma

Dr. Ashok K. Varma
(Programme Coordinator)
Principal Investigator, ACTREC

Shirsat

Dr. Neelam Shirsat
Course Coordinator,ACTREC

Dr. Sudeep Gupta
Director, ACTREC

RESEARCH ARTICLE

Novel Somatic Mutations of the CDH1 Gene Associated with Gastric Cancer: Prediction of Pathogenicity Using Comprehensive *In silico* Methods

Payel Chakraborty[1,#], Souvik Ghatak[1,#], Ravi Prakash Yadav[1], Subhajit Mukherjee[1], Lalchhandama Chhakchhuak[2], Saia Chenkual[3], Thomas Zomuana[3], Sailo Tlau Lalruatfela[3], Arindam Maitra[4] and Nachimuthu Senthil Kumar[1,*]

[1]Department of Biotechnology, Mizoram University, Aizawl -796004, Mizoram, India; [2]Department of Pathology, Civil Hospital, Aizawl -796001, Mizoram, India; [3]Department of Surgery, Civil Hospital, Aizawl -796001, Mizoram, India; [4]National Institute of Biomedical Genomics, Kalyani - 741251, West Bengal, India

ARTICLE HISTORY

Received: May 15, 2020
Revised: September 26, 2020
Accepted: October 01, 2020

*DOI:
10.2174/1875692117999201109210911*

Abstract: ***Background***: Mutations in the *CDH1* and the role of E-cadherin proteins are well established in gastric cancer. Several insilico tools are available to predict the pathogenicity of the mutations present in the genes with varying efficiency and sensitivity to detect the pathogenicity of the mutations.

Objective: Our objective was to identify somatic pathogenic variants in *CDH1* involved in gastric cancer (GC) by Sanger sequencing as well as using insilico tools and to find out the best efficient tool for pathogenicity prediction of somatic missense variants.

Methods: Sanger sequencing of *CDH1* was done for 80 GC tumor and adjacent normal tissues. Synthetic data sets were downloaded from the COSMIC database for comparison of the known mutations with the discovered mutations from the present study. Different algorithms were used to predict the pathogenicity of the discovery and synthetic mutation datasets using various *in-silico* tools. Statistical analysis was done to check the efficiency of the tools to predict pathogenic variants by using MEDCALC and GraphPad.

Results: Six missense somatic variants were found in exons 3, 4, 7, 9, 12 and 15. Out of the 6 variants, 5 variants (chr16:68835618C>A, chr16:68845613A>C, chr16:68847271T>G, chr16:68856001T>G, chr16:68863585G>C) were novel and not reported in disease variant databases. PROVEAN, Polyphen 2 and PANTHER predicted the pathogenicity of the variants more efficiently in both the discovery and synthetic datasets. The overall sensitivity of predictions ranged from 60 to 80%, depending on the program used, with specificity from 55 to 100%.

Conclusion: This study estimates the specificity and sensitivity of prediction tools in predicting novel missense variants of CDH1 in Gastric Cancer. We report that PROVEAN, Polyphen 2 and PANTHER are efficient predictors with constant higher specificity and accuracy. This study will help the researchers to explore mutations with the best pathogenicity prediction tools.

Keywords: E-cadherin, software prediction, pathogenicity, mutations, cancer, *In-silico* method.

1. INTRODUCTION

E-cadherin is a transmembrane glycoprotein and has a role in cell-cell adhesion and cell differentiation [1]. The *CDH1* contains 16 exons and is located on chromosome 16q22.1. Mutations in *CDH1* are known to be involved in Hereditary Diffuse Gastric Cancer (HDGC) syndrome [2]. Most of the cancer cells originate from epithelial cells which are associated with the actin and intermediate cytoskeleton and is interconnected with tight, adherents-type and desmosome junctions. Ca2^{++}-dependent interactions by E-cadherin are the major molecules for the maintenance of these junctional complexes [3]. Calcium-dependent cell to cell adhesion is essential during the migration of epithelial cells. E-cadherin plays an important role in cell differentiation, morphogenesis [4] and oncogenesis [5] and hence is involved in epithelial cell adhesion. Studies have reported that germline and somatic mutations in CDH1 gene have an association in GC development [6, 7].

The study is based on the available *in-silico* methods and their efficiency in determining the pathogenicity of the mutations. Insilico approaches to predict the disease-causing SNPs have been established and their efficiency to classify the deleterious and disease-associated mutations, as well as envisaging their pathogenicity, effect on functional and structural properties, have been scientifically rationalized [8,

[*] Address correspondence to this author at the Department of Biotechnology, Mizoram University, Aizawl – 796004, Mizoram, India;
Tel: 09436352574; E-mail: nskmzu@gmail.com

[#] *Both the authors contributed equally.*

9]. *In-silico* tools are very useful to filter important variants from single-gene data or from a large scale data set to identify the pathogenic, deleterious and harmful SNPs which are responsible for diseases. Comparison studies with the prediction tools have reported that none of the available tools can predict the pathogenic variants efficiently [10]. In this study, we will be reporting the best efficient tools for pathogenicity prediction for the selected variations associated with gastric cancer. This analysis will help us to improve personalized medicines and lead to better diagnosis against genetic disorders. As newer genomes are sequenced, new variations are reported and are made available in the databases and hence, there is a continuous need to review and update the available next-generation tools and methods.

This study intended to determine the pathogenicity of somatic missense mutations found in the E-cadherin (CDH1) gene of Gastric Cancer patients using *in-silico* methods. Earlier studies have reported using the performance of few *in-silico* prediction tools and software like SIFT, PolyPhen-2, Align-GVGD and Mutation Taster 2 for rare and common missense variants [11, 12] Whereas, the Ensemble database catalogs the mutation pathogenicity using five prediction methods (CADD, SIFT, PolyPhen-2, REVEL, Mutation Assessor and MetalR) [13]. Therefore, it is essential and mandatory to identify the *in-silico* tools and algorithms to predict the accurate and true pathogenic information about the disease associated with genetic variations. To answer these questions, we analyzed the CDH1 gene mutation and their pathogenicity: 1) Is there a presence of any novel pathogenic mutations that contribute to gastric tumor proliferation? 2) Whether the available *in-silico* prediction software can compare the variants for estimating the pathogenicity and their effects on protein structure and function in Cancer? The present study can help in determining the reliability and use of this software for further functional laboratory validation. Herein, we performed a comprehensive analysis of *in-silico* pathogenicity estimation (gene and protein) from CDH1 gene sequence data and identified the novel pathogenic mutations which might contribute the gastric tumor proliferation and migration.

2. MATERIALS AND METHODS

2.1. Sample Description

A total of 80 tumors and matched adjacent normal specimens with primary gastric cancer were collected from the Civil Hospital Aizawl, Mizoram, Northeast India. Our inclusion and exclusion criteria's were patients with gastric cancer and without any other chronic diseases, and should not be pre-treated for any other type of cancer.

2.2. DNA Extraction, PCR Amplification and Sequencing

DNA was extracted from tumor tissue by the modified protocol of Ghatak *et al.* (2013) [14]. The CDH1 gene exonic regions with splicing sites were amplified by polymerase chain reaction (PCR) (Supplementary Table **1**). The reaction volume for 25 μl PCR consisted of template DNA (100 ng/ul), forward and reverse primers (0.2 pM each), PCR buffer (1X), $MgCl_2$ (1.5 mM), dNTPs (0.2 mM), Taq DNA polymerase (1U) (Fermentas, Germany). The reaction condition was: 94^0C- 5 min, followed by 30 cycles [each with 94^0C- 1 min, 50 to 60^0C- 1 min, 72^0C- 1 min] and final 72^0C- 5 min. All the amplified exons were sequenced from both the strands using the automated ABI HITACHI sequencer model 3500 Genetic Analyzer (PE Applied Biosystems).

2.3. Sequence Analysis

Chromas software version 2.13 and DNA baser were applied for screening of sequences and chromatograms and alignment was done by BLAST [www.ncbi.nlm.nih.gov/blast]. The evaluation of the variants was done by DNA baser version 3.5.4.2 and Codon Code aligner version V.4.2.2. All the exons of *CDH1* were checked from the E cadherin Gene card database [HGNC:17481, Ensembl Transcript ID ENST00000261769.5, NCBI: accession - Z13009, Database: Ensembl GRCh37, UniprotKB ID: P12830, Entrez Gene:9992, OMIM: 1920905] and the data obtained was used as the Discovery dataset. Detailed workflow design for the *In-silico* tools and their categorization to predict the pathogenicity of the selected non-synonymous single nucleotide variance (snSNVs) is represented in Fig. (**1**).

2.4. Synthetic Dataset

For validation of the results from the present study, ten known somatic variants were randomly selected from the COSMIC database of CDH1 gene found in Gastric Cancer. This includes four neutral (COSM4062185, COSM6918669, COSM2996742, and COSM8473222), two inconclusive (COSM20839 and COSM8166164), and four pathogenic variants (COSM2996747, COSM5576263, COSM4756921, and COSM4617690) (Table **3**).

2.5. *In silico* Tools for the Prediction of Mutation Effect

We used all the available tools for the prediction of missense variants found in the discovery dataset (our study) and the synthetic dataset (retrieved from the COSMIC database) for a meaningful comparison of best predictor tools. Mutation taster **(www.mutationtaster.org/)** predicts the disease-causing potential of a variant, and it combines information from ExAC, NCBI and 1000 genome databases and can analyze data for exonic and intronic regions (Supplementary Fig. **1**) [15]. PolyPhen**-2 (http://genetics.bwh.harvard.edu/pph2/)** was used to predict the non-synonymous substitutions based on the Bayesian classifier and sequence alignment (Supplementary Fig. **2**) [16]. It is used to classify variants as benign, probably damaging and possibly damaging based on a numerical score. SIFT/PROVEAN **(http://siftjcvi.org/)** was used to predict exonic variants can be based on a sequence homology-based algorithm [17]. It is a two-step method to predict missense SNPs based on homologous sequences and amino acid substitution matrix-based scores. SIFT classifies variants as tolerable or damaging based on conserved site and PROVEAN classifies as neutral or deleterious based on the score (Supplementary Fig. **3**) [18]. SNPs&GO **(http://snps-and-go.biocomp.unibo.it/snps-and-go/)** is based on the SVM (support vector machine) classifier and predicts depending on the type of mutation type and sequence information. This tool uses Panther for predictions [19] and functional information depending

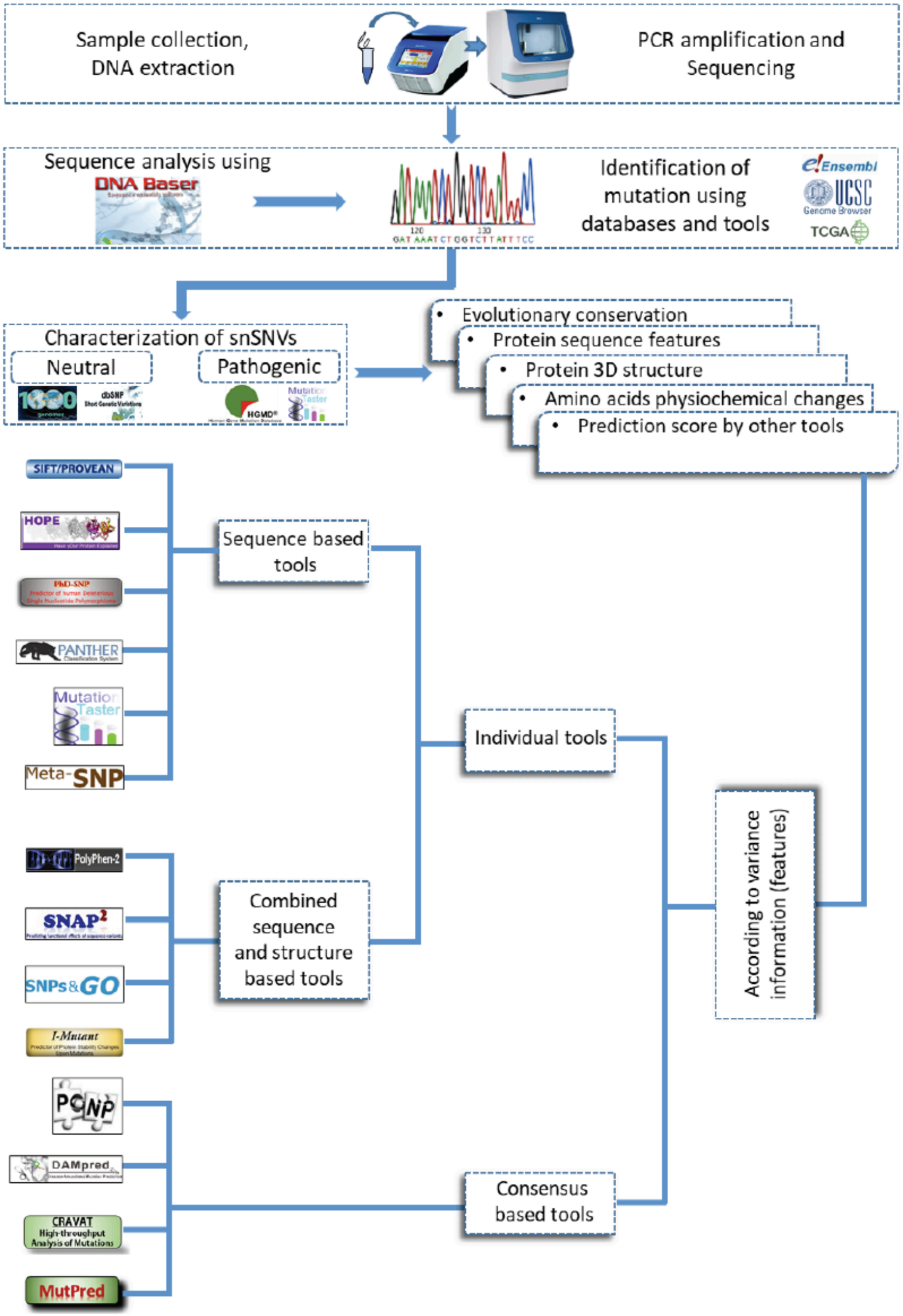


Fig. (1). The figure depicts the proposed methodology workflow which can be divided in the three main steps: data collection, *in silico* (quantitative) prediction of effects of mutations, filtering mutations by their predicted effect. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

on Gene Ontology (GO) terms [20]. It can predict whether a mutant protein is disease-causing or neutral based on the reliability index score (Supplementary Fig. **4**). The PANTHER **(http://pantherdb.org/tools/csnpScoreForm.jsp)** software predicts only coding SNPs. It is used to calculate the substitution position-specific conservation score to predict pathogenicity. The alignments are obtained from the PANTHER and are based on Hidden Markov Models (HMMs) library [21]. It will predict the alteration as deleterious or tolerated based on the above-mentioned score

Table 1. Pathogenicity prediction of discovery dataset by *In silico* prediction tools.

Exon Number	Mutation	Amino Acid Alteration	Mutation Taster	Polyphen 2.0	SIFT	PROVEAN	SNP &GO	PANTHER	PhD-SNP
				Prediction(Score)	Prediction(Score)	Prediction(Score)	Effect	Prediction	Effect
Exon 3	g.64491C>A* chr16:68835618C>A cDNA400C>A	S70Y	Polymorphisms	Probably damaging (0.976)	Damaging(0.001)	Deleterious(-3.25)	Disease(RI-1)	Probably damaging	Neutral
Exon 4	g.71240C>A chr16:68842367C>A cDNA.619C>A	P143H	Disease causing	Probably damaging (0.999)	Tolerated(0.058)	Deleterious(-4.30)	Neutral(RI-5)	Probably damaging	Disease
Exon 7	g.74486A>C* chr16:68845613A>C cDNA.1050A>C	T287P	Disease causing	probably damaging (1.000)	Damaging(0.011 ing(0.011)	Deleterious(-4.75)	Disease (RI-9)	Probably damaging	Neutral
Exon 9	76144T>G* chr16:68847271T>G cDNA.1384T>G	V398G	Disease causing	probably damaging (0.999)	Damaging(0.000)	Deleterious(-6.30)	Disease (RI-7)	Probably damaging	Neutral
Exon12	84874T>G* chr16:68856001T>G cDNA.2000T>G	C603W	Disease causing	probably damaging (1.000)	Damaging(0.000)	Deleterious(-10.70)	Disease (RI-7)	Probably damaging	Disease
Exon15	92458G>C* chr16:68863585G>C cDNA.2515G>C	G775A	Disease causing	probably damaging (1.000)	Tolerated(0.326)	Deleterious(-3.78)	Neutral (RI-1)	Probably damaging	Neutral

” *” = Novel mutations; RI - Reliability index

Polyphen2

Score between 0.0 - 0.15 = benign.

Score between 0.15 – 1.0 = possibly damaging

Score between 0.15 – 1.0 = probably damaging

SIFT

Score between 0.0 – 0.05 = damaging

Score between 0.05 – 1.0 = tolerated

PROVEAN

If the score is ≤ -2.5 = deleterious

If the score is > -2.5 = neutral

(Supplementary Fig. **5**). PhD-SNP **(http://snps.biofold.org/phd-snp)** works based on a sequence-based support vector machine and uses the variants from SwissProt [22]. It will predict the pathogenicity of missense variants as Disease causing or neutral, according to alteration of the neighboring sequence (Supplementary Fig. **6**). PON-P2 **(http://structure.bmc.lu.se/PON-P2/PON-P2)** is based on a machine learning method and is used to gather information from different tools. It can predict the pathogenicity of the variants based on substitution and gene ontology, evolutionary conservation and functional sites [23]. SNAP2 **(https://rostlab.org/services/snap/)** can predict the functional pathogenicity of substitutions without any alignment information using a specific method [24]. It is used to classify variants as effective or neutral and a heat map representation will be provided as an output (Supplementary Fig. **7**) [25]. Meta SNP **(http://snps.biofold.org/meta-snp)** is a multi-predictor tool and is used to access single prediction tools like PANTHER, PhD-SNP, SIFT, and SNAP to predict the pathogenicity of variants. Meta-SNP is a good tool for accessing the prediction from all the important prediction tools (Supplementary Fig. **11**) [26]. MutPred **(http://mutpred.mutdb.org/)** predicts an amino acid alteration as disease-causing or neutral [27]. The output of MutPred provides actionable, confident and very confident hypotheses according to variants (Supplementary Fig. **13**, Supplementary Table **1**). I-Mutant **(http://gpcr.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi)** is used to predict protein stability changes due to a single point mutation. Protein sequence and the variants are the input for this tool and it works according to free energy value (DDG value) and Gibbs free energy (DG) changes due to alteration in protein structure and sequence (Supplementary Fig. **10**). The algorithm depends on the support vector machine (SVM) system and according to this classifier, there will be a decrease in protein stability, if DDG value is or more than 20.5 and it will be increased in protein stability if the DDG value is or less than 0.5 [28]. DIM-Pred **(www.iitm.ac.in/bioinfo/DIM_Pred/)** depends on the SVM-based machine learning algorithm. It is used to predict only missense mutations and will predict the transition of a mutant residue position towards order to disorder (Supplementary Fig. **9**) [29].

2.6. HOPE (Have (y) Our Protein Explained - http://www.cmbi.ru.nl/hope.can)

HOPE can predict both the structural and functional effects of a mutation. UniPort ID or protein sequence and wild type, mutant residue and position are the inputs for this tool (Supplementary Fig. **12**). It will give a detailed report about Domain change, structure, variant, amino acid features, and hydrophobicity of the mutant residue and the 3D image of the altered protein. Here, we can get information on whether the variant is already reported in the ExPasy database [30]. HOPE analysis gives complete information about the substitutions.

2.7. CRAVAT (Cancer-related variants Annotation Toolkit - http://www.cravat.us)

CRAVAT is annotation software and it depends on the algorithms of CHASM and VEST training set [31]. CHASM dataset deals with the driver mutations retrieved from the COSMIC database and the VEST dataset deals with positive disease mutations from Human Gene Mutation Database (HGMD) and ESP6500 cohort dataset for negative variants. The chromosomal position and variants have to be given as input according to their format and the output file contains detailed information about the variants (Supplementary Fig. **14**). It will give information about both coding and non-coding variants and also their type (synonymous, non-synonymous, missense, nonsense, frameshift, insertion, deletion, frameshift). We can get information from 1000 genomes, Clinvar, gnomAD, COSMIC, dbSNP, ExAC and the variants can be classified as novel or reported earlier.

2.8. Assessment of Accuracy, Sensitivity and Specificity Prediction of *In-silico* Tools

MedCalc [32] and GraphPad [33] were used to estimate the efficiency of all the tools utilized in this study. The deleterious, pathogenic, or damaging mutations were considered as "+" class and neutral or benign as "−"class. The evaluation process of selected tools was influenced by input criteria, which was estimated according to different statistical parameters. We calculated the sensitivity and specificity for all the pathogenicity prediction tools based on the calculation of area under the curve (AUC) using the receiver operating curve (ROC) method. The area under the curve ranges between 0 and 1, with an AUC of 1 indicating a perfect classifier and 0.5 as the random classifier. Sensitivity is the likelihood that a test result will be positive (if the variance is pathogenic), and specificity will be negative (if the variance is not pathogenic).

3. RESULTS

3.1. CDH1 Mutation Pattern in Discovery Dataset

A total of 6 missense somatic variants were found in exons 3, 4, 7, 9, 12 and 15 (Fig. **2**). Exon 3 possessed one missense variant (chr16:68835618C>A), altering Serine to Tyrosine at position 70 with an 18% frequency of occurrence. Exon 4 exhibited one missense variant (chr16:68842367C>A), changing Proline to Histidine at position 143 with a 20 % frequency of occurrence. One missense variant (chr16:68845613A>C) was also observed in exon 7, altering Threonine to Proline at position 287 with 2.5 % frequency. Exon 9 exhibited one missense variant (chr16:68847271T>G), altering Valine to Glycine at position 398 with 26.25 % frequency. Exons 12 and 15 exhibited a single missense variant (chr16:68856001T>G and chr16:68863585G>C), each altering Cysteine to Tryptophan at position 603 and Glycine to Alanine at position 775 with 1.5 % and 2.5 % frequency of occurrence, respectively.

3.2. Pathogenicity Prediction of the Discovery Dataset by *In-silico* Tools

In Mutation taster, five variants (chr16:68842367C>A, chr16:68845613A>C, chr16:68847271T>G, chr16:68856001T>G & chr16:68863585G>C) were disease-causing and one variant (chr16:68835618C>A) was polymorphism /neutral. Out of the 6 missense variants, five (chr16:68835618C>A, chr16:68845613A>C, chr16:68847271T>G, chr16:68856001T>G & chr16:68863585G>C) were not reported in 1000 genome & ExAC database (Table **1**). The 6 missense mutations were further analyzed using the other software's. All the six variants were predicted as probably damaging in Polyphen 2.0 (0.976 - 1.000) (Table **1**). Four variants (chr16:68835618C>A, chr16:68847271T>G, & chr16:68856001T>G) were predicted as highly deleterious (score=0.00), one variant (chr16:68845613A>C) predicted as deleterious (≥ 0.05) and two variants (chr16:68842367C>A & chr16:68863585G>C) were tolerated in SIFT (Table **1**). In PROVEAN, all six variants were predicted as deleterious (Table **1**). All the variants were analyzed in the SNP&GO tool and the result was the same as SIFT, the same four variants were disease-causing and two were neutral. All 6 variants were predicted as damaging in PANTHER (Table **1**).

In PhD-SNP, only two variants (chr16:68842367C>A & chr16:68856001T>G) were predicted as disease-causing, and four variants were predicted as neutral. To further

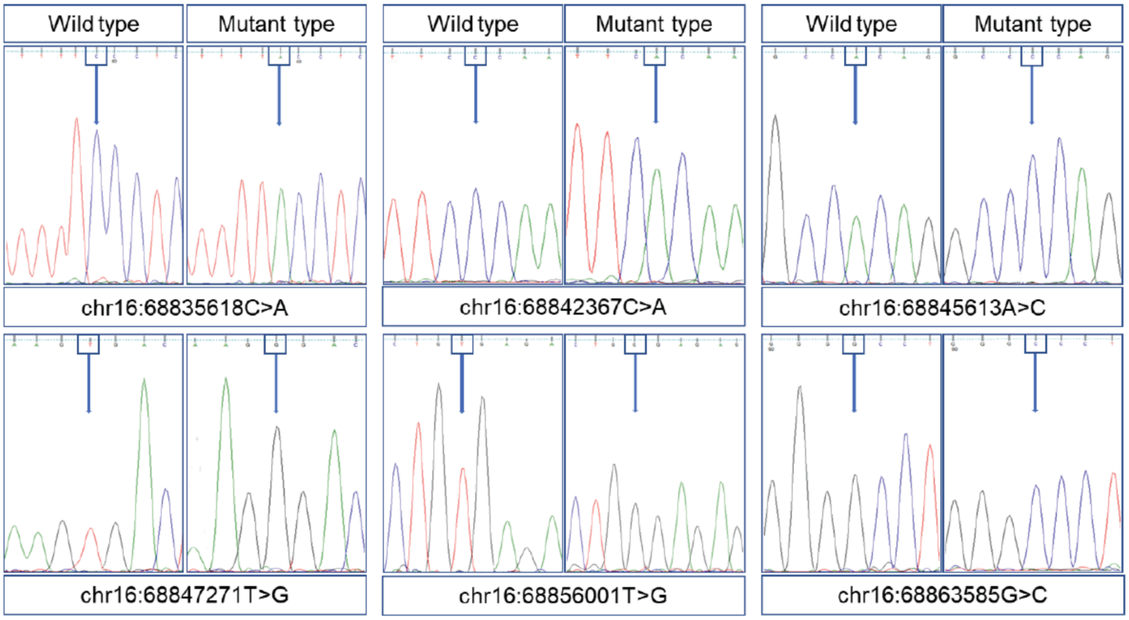


Fig. (2). Electropherogram of CDH1 gene mutations associated with gastric cancer in Mizo population.

➡ Representing the wild and mutant nucleotides and their positions in electropherogram. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 2. Prediction of pathogenicity and protein stability changes of discovery dataset by *In silico* tools and comparison with DIMPRED.

Exon Number	Mutation	Alteration	SNAP2	PON-P2	I-Mutant	DIMPRED
			Effect	**Effect**	**Stability**	**Direction**
Exon 3	g.64491C>A* chr16:68835618C>A cDNA400C>A	S70Y	Effect	Unknown	Increase	O→D
Exon 4	g.71240C>A chr16:68842367C>A cDNA.619C>A	P143H	Neutral	Unknown	Decrease	O→D
Exon 7	g.74486A>C* chr16:68845613A>C cDNA.1050A>C	T287P	Effect	Pathogenic	Increase	O→D
Exon 9	76144T>G* chr16:68847271T>G cDNA.1384T>G	V398G	Effect	Unknown	Decrease	O→O
Exon12	84874T>G* chr16:68856001T>G cDNA.2000T>G	C603W	Effect	Pathogenic	Decrease	O→D
Exon15	92458G>C* chr16:68863585G>C cDNA.2515G>C	G775A	Neutral	Pathogenic	Decrease	O→O

O = Order and D = Disorder; " *" = Novel mutations validate these results, we used SNAP 2 and PON-P2. Four variants (chr16:68835618C>A, chr16:68847271T>G, chr16: 68845613A>C & chr16:68856001T>G) were found to be having an effect on protein function in SNAP 2 tools like SIFT and SNP&GO (Table **2**). In PON-P2, three variants (chr16:68845613A>C, chr16:68856001T>G & chr16: 68863585G>C) were pathogenic (Table **2**). But, PON-P2 could not predict the effect of three variants and were reported as unknown. We further analyzed our variants using I-Mutant 3.0 and DIMPRED. Out of 6 variants, four variants (chr16:68842367C>A, chr16:68847271T>G, chr16: 68856001T>G & chr16:68863585G>C) were predicted to decrease the stability and two variants (chr16: 68835618C>A & chr16:68845613A>C) were predicted as increasing the stability of the protein in I- Mutant (Table **2**). In DIMPRED, all six variants caused order-disorder transition at the mutated position (Table **2**).

MetaSNP is a multipredictor, used to predict variants according to multiple tools like PANTHER, PhD-SNP, SIFT and SNAP in one way. Here, four variants (chr16:68835618C>A, chr16:68847271T>G, & chr16: 68856001T>G) were disease-causing and two variants (chr16:68842367C>A & chr16:68863585G>C) were neutral (Supplementary Fig. **11**). The result varied when these tools were used to predict individually. In MutPred, out of 6 variants, 2 were showing an actionable hypothesis (chr16: 68845613A>C & chr16:68847271T>G) (Supplementary Table **2**).

3.3. Detection of Novel Variants

After analysis with Mutation Taster, HOPE and CRAVAT, five novel variants (chr16:68835618C>A, chr16: 68845613A>C, chr16:68847271T>G, chr16:68856001T>G & chr16:68863585G>C) were obtained which were not reported in1000 Genome, COSMIC, dbSNP, ExAC, ExPasy, and gnomAD databases. Among the six variants, one variant (chr16:68856001T>G) was predicted as disease-causing in all of the used tools and also decreasing the stability of the protein. We further checked our variants with Ensemble, 1000 genome and HGVD variant table to ensure that they are novel variants.

3.4. Prediction in the Synthetic Dataset

Mutation Taster tool predicted neutral variants and inconclusive variants as polymorphisms, while all the four pathogenic variants were predicted as disease-causing (Table **3**). In the case of the Polyphen2 tool, all the neutral variants were predicted as Benign, whereas the two inconclusive were predicted as possibly damaging, and pathogenic variants were predicted as probably damaging (Table **3**). SIFT predicted two neutral variants as tolerated while the other two neutral variants were predicted as damaging. One of the inconclusive variants was predicted as tolerated, and another one was predicted as damaging, and all the four pathogenic variants were predicted as damaging (Table **3**). The prediction of PROVEAN and PANTHER was the same as all the neutral and inconclusive variants were predicted as neutral/probably benign, while all the four pathogenic variants were predicted as disease-causing/ probably damaging (Table **3**). SNP & GO predicted three of the neutral variants as neutral and another one was predicted as disease-causing, two inconclusive variants were predicted as neutral too, while all four pathogenic variants were predicted as disease-causing (Table **3**). SNAP2 predicted two of the neutral variants as the same and the other two neutral variants were predicted as effective, one of the inconclusive variants was predicted as effective while another one was predicted as neutral and all the four pathogenic variants were predicted as effective (Table **4**). In PON-P2, only the pathogenic variants were predicted as pathogenic, while neutral and inconclusive variants were predicted as unknown. Meta-SNP predicted three of the neutral variants as neutral and another one was predicted as effective; two inconclusive variants were predicted as neutral too, while all four pathogenic variants were predicted as effective (Table **4**). I-Mutant 3.0 predicted all the variants were showing decreasing the protein stability, except one pathogenic variant predicted as increasing the stability of the protein (Table **4**). In DIMPRED, three of the neutral variants as caused order-disorder transition and other one neutral variant caused order-order transition, one of the inconclusive variant caused order-disorder transition while another one was caused order-order transition and two of the pathogenic variants caused order-order transition and another two caused order-disorder transition at the mutated position (Table **4**). Functional prediction of variants was predicted by HOPE (Supplementary Table **4**).

3.5. Accuracy, Sensitivity and Specificity Prediction of Selected *In-silico* Tools

We evaluated seventeen tools to classify the mutations as pathogenic or neutral and to forecast the effects of non-synonymous variation on protein function. A summary of these tools is indicated in Fig. (**3**). The graphical area under the receiver operating characteristic (AUC - ROC) curve and the running time predicts the efficiency of the classifiers. When sensitivity and specificity increases, the AUC increases and achieves the highest performance. We have calculated the accuracy of the prediction tools according to input criteria and variance features. The ROC curves with high sensitivity and specificity values of the five servers in sequence-based tools: SIFT, PROVEAN, PANTHER, PhD-SNP, and MetaSNP are shown in (Fig. **3A**). The PANTHER (AUC = 0.72) and PROVEAN (AUC = 0.86) tool achieved the highest accuracy values as combined sequence-based tools (Fig. **3A**). Whereas the accuracy value of Polyphen2 (AUC = 0.72) and SNP2 (AUC = 0.75) were adequate for the pathogenicity estimation as a consensus-based tool (Fig. **3B**). The curves with sensitivity and specificity values of SIFT, PROVEAN, PANTHER, PhD-SNP and MetaSNP sequence-based classifiers were presented in Fig. **3D**. The sensitivity and specificity are significantly highest in PANTHER and PROVEAN than any other prediction tools (Fig. **3D**). Polyphen 2.0 achieved significant specificity and sensitivity among all the combined sequence and structure-based tools (Fig. **3E**). But there is no significant observation for consensus tools (Fig. **3F**).

We have analyzed the accuracy, sensitivity, and specificity prediction of the results of synthetic datasets prediction. The results are also supporting our prediction as ROC curves showing the same top five predictors: PROVEAN,

Table 3. Pathogenicity prediction of synthetic dataset by *In silico* prediction tools.

COSMIC ID & Prediction	**Mutation**	**Amino Acid Alteration**	**Mutation Taster**	**Polyphen 2.0**	**SIFT**	**PROVEAN**	**SNP&GO**	**PANTHER**	**PhD-SNP**
				Prediction(Score)	Prediction (Score)	Prediction (Score)	Effect	Prediction	Effect
COSM4062185 Neutral	cDNA.2262G>A g.86309G>A chr16:68857436G>A	A691T	Polymorphisms	Benign (0.000)	Tolerated (0.276)	Neutral (0.11)	Neutral (RI-5)	Probably benign	Neutral
COSM6918669 Neutral	cDNA.2295G>A g.86342G>A chr16:68857469G>A	E702K	Polymorphisms	Benign (0.014)	Tolerated (0.519)	Neutral (-0.17)	Neutral (RI-5)	Probably benign	Neutral
COSM2996742 Neutral	cDNA.459C>T g.64550C>T chr16:68835677C>T	R90W	Polymorphisms	Benign (0.001)	Damaging (0.039)	Neutral (-1.94)	Disease (RI-1)	Probably benign	Disease
COSM8473222 Neutral	cDNA.1431A>G g.76191A>G chr16:68847318A>G	T414A	Polymorphisms	Benign (0.008)	Damaging (0.018)	Neutral (-1.08)	Neutral (RI-4)	Probably benign	Neutral
COSM20839 Inconclusive	cDNA.1578G>C g.78357G>C chr16:68849484G>C	E463Q	Polymorphisms	Possibly Damaging (0.673)	Tolerated (0.052)	Neutral (-0.30)	Neutral (RI-7)	Probably benign	Disease
COSM8166164 Inconclusive	cDNA.561C>T g.64652C>T chr16:68835779C>T	R124C	Polymorphisms	Possibly Damaging (0.983)	Damaging (0.025)	Neutral (-2.05)	Neutral (RI-1)	Probably benign	Disease
COSM2996747 Pathogenic	cDNA.652G>C g.71273G>C chr16:68842400G>C	R154T	Disease causing	Probably Damaging (1.000)	Damaging (0.000)	Deleterious (-5.30)	Disease (RI-9)	Probably damaging	Disease
COSM5576263 Pathogenic	cDNA.657T>A g.71278T>A chr16:68842405T>A	W156R	Disease causing	Probably Damaging (1.000)	Damaging (0.000)	Deleterious (-12.11)	Disease (RI-8)	Probably damaging	Disease
COSM4756921 Pathogenic	cDNA.660G>C g.71281G>C chr16:68842408G>C	V157L	Disease causing	Probably Damaging (0.986)	Damaging (0.002)	Deleterious (-2.59)	Disease (RI-6)	Probably damaging	Disease
COSM4617690 Pathogenic	cDNA.1564A>T g.78343A>T chr16:68849470A>T	N458I	Disease causing	Probably Damaging (1.000)	Damaging (0.000)	Deleterious (-8.19)	Disease (RI-8)	Probably damaging	Neutral

RI - Reliability Index ;

Table 4. Prediction of pathogenicity and protein stability changes of synthetic dataset by *In silico* tools and comparison with DIMPRED.

COSMIC ID & Prediction	Mutation	Alteration	SNAP2	PON-P2	Meta-SNP	I-Mutant	DIMPRED
			Effect	**Effect**	**Effect**	**Stability**	**Direction**
COSM4062185 Neutral	cDNA.2262G>A g.86309G>A chr16:68857436G>A	A691T	Neutral	Unknown	Neutral	Decrease	O→D
COSM6918669 Neutral	cDNA.2295G>A g.86342G>A chr16:68857469G>A	E702K	Neutral	Unknown	Neutral	Decrease	O→D
COSM2996742 Neutral	cDNA.459C>T g.64550C>T chr16:68835677C>T	R90W	Effect	Unknown	Effect	Decrease	O→D
COSM8473222 Neutral	cDNA.1431A>G g.76191A>G chr16:68847318A>G	T414A	Effect	Unknown	Neutral	Decrease	O→O
COSM20839 Inconclusive	cDNA.1578G>C g.78357G>C chr16:68849484G>C	E463Q	Effect	Unknown	Neutral	Decrease	O→O
COSM8166164 Inconclusive	cDNA.561C>T g.64652C>T chr16:68835779C>T	R124C	Neutral	Unknown	Neutral	Decrease	O→D
COSM2996747 Pathogenic	cDNA.652G>C g.71273G>C chr16:68842400G>C	R154T	Effect	Pathogenic	Effect	Decrease	O→O
COSM5576263 Pathogenic	cDNA.657T>A g.71278T>A chr16:68842405T>A	W156R	Effect	Pathogenic	Effect	Decrease	O→O
COSM4756921 Pathogenic	cDNA.660G>C g.71281G>C chr16:68842408G>C	V157L	Effect	Pathogenic	Effect	Decrease	O→D
COSM4617690 Pathogenic	cDNA.1564A>T g.78343A>T chr16:68849470A>T	N458I	Effect	Pathogenic	Effect	Increase	O→D

O = Order and D = Disorder

PANTHER, SIFT, MetaSNP, and PhD-SNP, with high sensitivity and specificity values of sequence-based tools (Fig. **4A**). The PANTHER (AUC = 1.00) and PROVEAN (AUC = 0.91) tool achieved the highest accuracy values as combined sequence-based tools (Fig. **4A**). PANTHER and PROVEAN were the top prediction tools with the significantly highest sensitivity and specificity (Fig. **4D**). Polyphen2.0 achieved significant specificity and sensitivity among all the combined sequence and structure-based tools (Fig. **4E**).

4. DISCUSSION

Single missense somatic mutations in *CDH1* were found in Exons 3, 4, 7, 9, 12, and 15, which might result in loss of cadherin 2, 3, 4, 5 & TOPO domain protein functions. Due to mutation in exon 3 (chr16:68835618C>A), the amino acid alteration S70Y located in the signal peptide differs in the properties and might disturb the recognition of the signal peptide. This domain is annotated in Uniport as a calcium ion binding domain (Supplementary Tables **2** and **3**). In exon 4, P143H (chr16:68842367C>A) alteration can disturb the special conformation, which may be deleterious. Proline is known to have a very rigid structure, sometimes forcing the backbone in a specific conformation. Exon 7 exhibits T287P (chr16:68845613A>C) alteration, this mutation introduces an amino acid with different properties, which may abolish the function of cadherin 2 binding and extracellular binding TOPO domain (Supplementary Tables **2** and **3**).

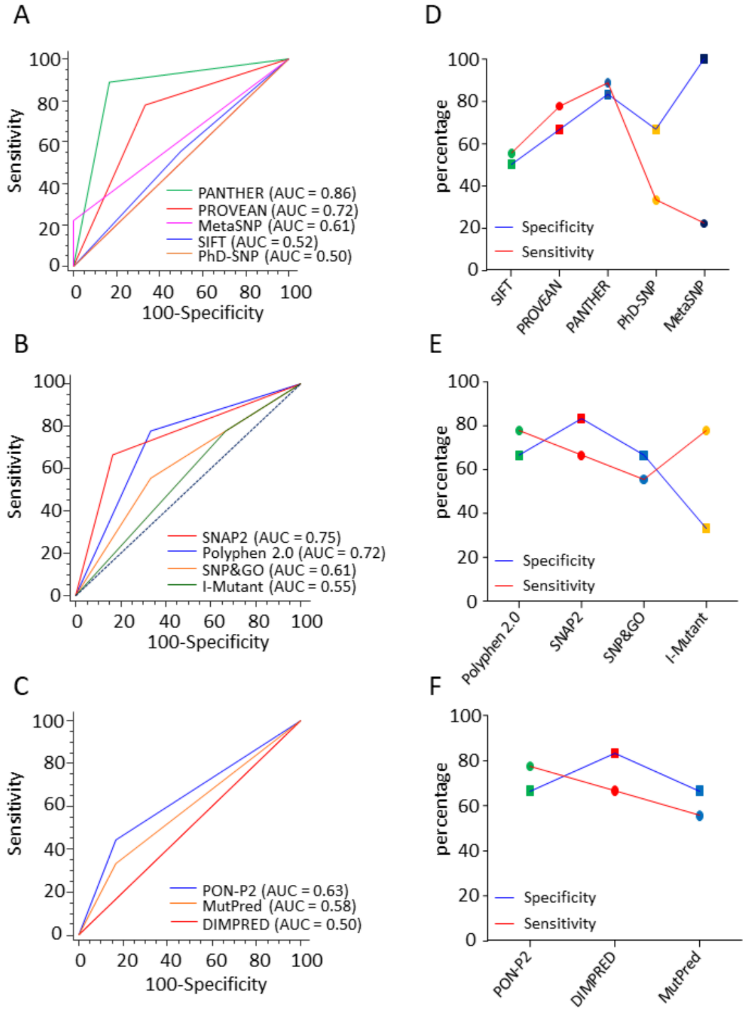


Fig. (3). Prediction accuracy, sensitivity and specificity of the selected *in-silico* tools for mutation pathogenicity estimation of Discovery dataset. (**A**) prediction accuracy of sequence-based tools, (**B**) combined sequence and structure-based tools, (**C**) Consensus based tools, (**D**) predicted specificity and sensitivity of sequence-based tools, (**E**) combined sequence and structurebased tools, (**F**) consensus based tools. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Exon 9 screened for V398G (chr16:68847271T>G) introduces a glycine at this position. Glycine's are very flexible and can disturb the required rigidity of the protein at this position, which might alter the function of cadherin 3, 4 and 5 binding Domains (Supplementary Tables **2** and **3**). Exon 12 exhibits C603W (chr16:68856001T>G) alteration, which was located in a conserved residue, and it may be damaging for the protein. Exon 15 exhibits G775A (chr16: 68863585G>C) alteration, which might change the function of the glycine. Due to these variants, the function of the cytoplasmic binding TOPO domain might be lost for E-cadherin protein (Supplementary Tables **2** and **3**). All the mutations except P143H (chr16:68842367C>A) have not been reported previously in any database.

Almost all the six variants that were predicted in *CDH1* as disease-causing, five were novel variants for Gastric cancer. One variant C603W (chr16:68856001T>G) exhibited in exon 12 was predicted as disease-causing in all the tools tested and was found to decrease the stability of the protein, emerging as the most pathogenic variant in this study. T287 (chr16:68845613A>C) and G775A (chr16:68863585G>C) were predicted as disease-causing or pathogenic variants in all the tools, except in PHD-SNP. V398G (chr16: 68847271T>G) was predicted as disease-causing or pathogenic variants in all the tools, except in PHD-SNP and PON-P2, decreasing the protein stability too.

All the six variants were predicted as disease-causing in PANTHER, PROVEAN and Polyphen 2.0. Polyphen2 is used to predict mutations based on protein function and structure by structural and comparative evolutionary conserved regions [16]. PANTHER also depends on the evolutionary preservation score [34]. PROVEAN depends on a sequence alignment-based approach like Polyphen2. These three prediction tools giving efficient results to predict disease-causing variants.

Four variants were commonly predicted as disease-causing in SIFT, SNP&GO and SNAP tools, which predict the pathogenicity based on protein function. PhD-SNP is based on a support vector machine algorithm [22]. PON-P2 is based on a machine learning-dependent classifier, dividing variants as pathogenic, neutral and unknown based on random forest probability score [23]. The prediction results

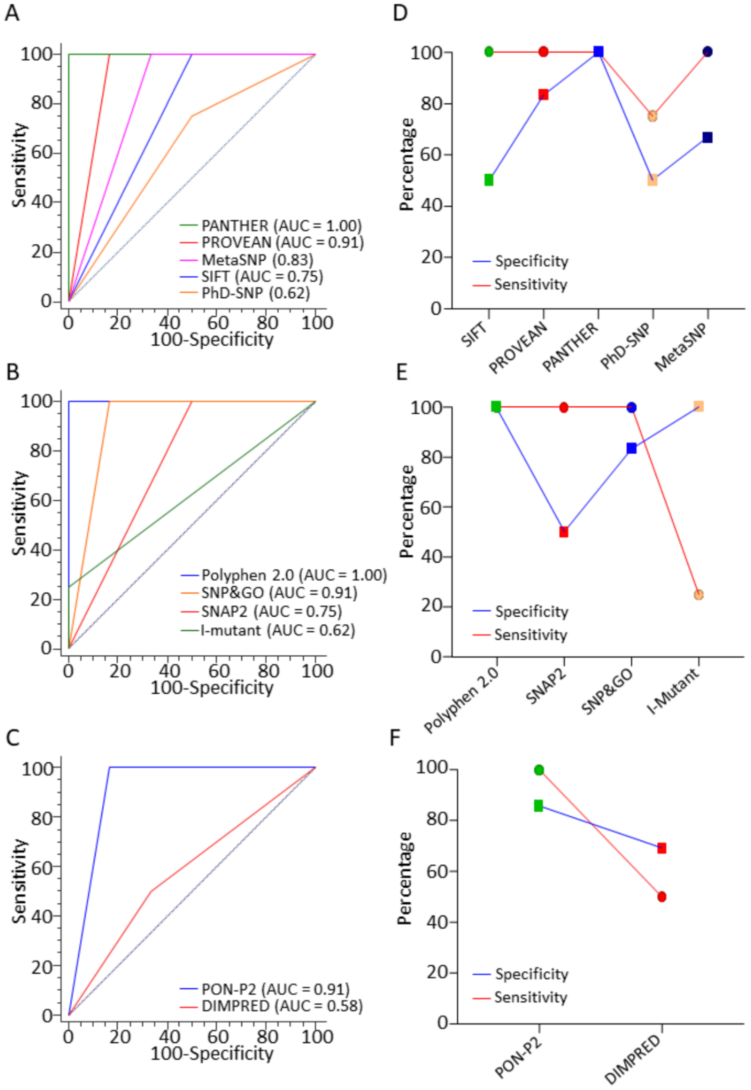


Fig. (4). Prediction accuracy, sensitivity and specificity of the selected *in-silico* tools for mutation pathogenicity estimation of Synthetic Dataset. (**A**) prediction accuracy of sequence-based tools, (**B**) combined sequence and structure-based tools, (**C**) Consensus based tools, (**D**) predicted specificity and sensitivity of sequence-based tools, (**E**) combined sequence and structurebased tools, (**F**) consensus based tools. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

of both the tools were different from other tools. I-Mutant provides information about the effect of the mutation on protein stability. Protein stability can increase or decrease due to mutation. MetaSNP is multi-tool predicting software and provides information from other single predictor tools, giving more strength to this tool. HOPE provides us the detail about the affected domain of the protein, and how amino acid alters the structure and function of the protein and their 3D view. CRAVAT is a very good tool for the annotation of cancer-related variants and includes databases like dbSNP, 1000 genome, gnomAD, CLINVAR, *etc*. CRAVAT is used to give a p-value of mutation depending on the VEST and CHIASM scores.

By comparing the classifiers for accuracy, sensitivity and specificity levels, it is obvious that PANTHER, PROVEAN and Polyphen 2.0 techniques outperformed other individual tools. The PANTHER classifier has 83.33% specificity, 88.89% sensitivity, 86.10% AUC around all variance, and in the sample dataset. The individual tool, PANTHER, and PROVEAN have high efficiency and accuracy over other *in-silico* pathogenicity prediction servers. Further, we compared our results with synthetic datasets prediction for a meaningful comparison of best predictor tools and, it gave more strength for supporting our prediction for best predicting *in-silico* tools. All statistical criteria such as sensitivity, specificity, AUC are comparatively better in a sequence and structure-based technique, respectively, compared to all evaluated consensus-based tools.

CONCLUSION

This study has provided a detailed survey and analysis of Non-synonymous Mutation detection using a range of software that differs from each other in their algorithm and output inference. This study concludes that Polyphen2, PROVEAN, and PANTHER as the best efficient missense somatic mutation predictor tools available. All statistical criteria such as sensitivity, specificity, AUC are comparatively better in a sequence and structure-based technique, respectively, compared to all evaluated consensus-based tools.

The present study also reports the unique variants (chr16:68835618C>A, chr16:68845613A>C, chr16: 68847271T>G, chr16:68856001T>G & chr16:68863585G>C) after comparing all the tools for the CDH1 gene in Gastric cancer samples. It has been shown that mutations are sufficient to disrupt the CDH1 function by altering the protein structure [35], which may lead to cancer. The mutations identified in the present study have not been previously reported and need further functional validation through molecular biology methods. The bioinformatics pipelines summarized in the present study can be used for all types of cancer mutations and disease phenotypes by interpreting from large datasets for understanding the role of mutations in the process of carcinogenesis.

LIST OF ABBREVIATIONS

VUS	=	Variance of Unknown Significance
HGDC	=	Hereditary Diffuse Gastric Cancer
dbSNP	=	Single Nucleotide Polymorphism Database
COSMIC	=	Catalogue Of Somatic Mutations In Cancer
ExAC	=	Exome Aggregation Consortium
gnomAD	=	Genome Aggregation Database
GVGD	=	Grantham Variation Grantham Deviation
CADD	=	Combined Annotation-Dependent Depletion
REVEL	=	Rare exome variant ensemble learner
HGNC	=	HUGO Gene Nomenclature Committee
NCBI	=	The National Center for Biotechnology Information
OMIM	=	Online Mendelian Inheritance in Man
HGMD	=	Human Gene Mutation Database
ExPASy	=	Expert Protein Analysis System
SVM	=	Support Vector Machines
HMMs	=	Hidden Markov Models
SIFT	=	Sorting Intolerant from Tolerant
PROVEAN	=	Protein Variation Effect Analyzer
HOPE	=	Have (y) Our Protein Explained
PANTHER	=	Protein Analysis Through Evolutionary Relationships
MUSTER	=	MUlti-Sources ThreadER
PhD-SNP	=	Predictor of human Deleterious Single Nucleotide Polymorphisms
GO	=	Gene Ontology
PSIPRED	=	PSI-blast based secondary structure PRE-Diction
DIM-Pred	=	Disorder Inducing Mutation Prediction
PhDSNP	=	Predictor of Human Deleterious Single Nucleotide Polymorphisms
CRAVAT	=	Cancer related Variants Annotation Tool
CHASM	=	Cancer-specific High-throughput Annotation of Somatic Mutations
VEST	=	Variant Effect Scoring Tool
AUC	=	Area under the Curve
ROC	=	Receiver Operating Curve

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study was approved by ethical committees of Civil Hospital, Aizawl (B.12018/1/13-CH(A)/IEC dtd. 18/04/2014) and Human Ethical Committee, Mizoram University (MZU/IHEC/2015/008 dtd. 14/12/15). Written informed consent was obtained from the participants for participate.

HUMAN AND ANIMAL RIGHTS

No Animals were used for studies that are base of this research. the reported experiments in accordance with the ethical standards of the committee responsible for human experimentation (institutional and national), and with the Helsinki Declaration of 1975, as revised in 2013 (http://ethics.iit.edu/ecodes/node/3931).

CONSENT FOR PUBLICATION

All participants signed the written informed consent for publication.

AVAILABILITY OF DATA AND MATERIALS

The data that support the findings of this study are available from the corresponding author upon reasonable request.

FUNDING

The work was supported by DBT-BTISNeT centre (BIF - BT/BI/12/060/2012) dated 29/09/2018), Mizoram University sponsored by Department of Biotechnology, New Delhi, Govt. of India.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

The authors are thankful to data collectors and patients.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's web site along with the published article.

REFFERENCES

[1] Pećina-Slaus N. Tumor suppressor gene E-cadherin and its role in normal and malignant cells. Cancer Cell Int 2003; 3(1): 17. http://dx.doi.org/10.1186/1475-2867-3-17 PMID: 14613514

[2] More H, Humar B, Weber W, *et al.* Guilford P Identification of seven novel germline mutations in the human E-cadherin (CDH1) gene. Hum Mutat 2017; 28(2): 1-9.

[3] Aberle H, Schwartz H, Hoschuetzky H, Kemler R. Single amino acid substitutions in proteins of the armadillo gene family abolish their binding to α-catenin. J Biol Chem 1996; 271(3): 1520-6. http://dx.doi.org/10.1074/jbc.271.3.1520 PMID: 8576147

[4] Del Buono R, Pignatelli M. The role of the E-cadherin complex in gastrointestinal cell differentiation. Cell Prolif 1999; 32(2-3): 79-84. http://dx.doi.org/10.1046/j.1365-2184.1999.32230079.x PMID: 10535354

[5] Caldas C, Carneiro F, Lynch HT, *et al.* Familial gastric cancer: overview and guidelines for management. J Med Genet 1999; 36(12): 873-80. PMID: 10593993

[6] Luo W, Fedda F, Lynch P, Tan D. CDH1 Gene and Hereditary Diffuse Gastric Cancer Syndrome: Molecular and Histological Alterations and Implications for Diagnosis And Treatment. Front Pharmacol 2018; 9: 1421. http://dx.doi.org/10.3389/fphar.2018.01421 PMID: 30568591

[7] Corso G, Carvalho J, Marrelli D, *et al.* Somatic mutations and deletions of the E-cadherin gene predict poor survival of patients with gastric cancer. J Clin Oncol 2013; 31(7): 868-75.
http://dx.doi.org/10.1200/JCO.2012.44.4612 PMID: 23341533

[8] Kamaraj B, Rajendran V, Sethumadhavan R, Purohit R. *In-silico* screening of cancer associated mutation on PLK1 protein and its structural consequences. J Mol Model 2013; 19(12): 5587-99.
http://dx.doi.org/10.1007/s00894-013-2044-0 PMID: 24271645

[9] Kamaraj B, Purohit R. Mutational analysis of oculocutaneous albinism: a compact review. BioMed Res Int 2014; 2014: 905472.
http://dx.doi.org/10.1155/2014/905472 PMID: 25093188

[10] Schiemann AH, Stowell KM. Comparison of pathogenicity prediction tools on missense variants in RYR1 and CACNA1S associated with malignant hyperthermia. Br J Anaesth 2016; 117(1): 124-8.
http://dx.doi.org/10.1093/bja/aew065 PMID: 27147545

[11] Corinna E, Eric H, Christoph E, *et al.* Performance of *in silico* prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. BMC Med Genomics 2018; 11(1): 1-10.
PMID: 29329538

[12] Pshennikova VG, Barashkov NA, Romanov GP, *et al.* Comparison of Predictive *In Silico* Tools on Missense Variants in *GJB2*, *GJB6*, and *GJB3* Genes Associated with Autosomal Recessive Deafness 1A (DFNB1A). ScientificWorldJournal 2019; 2019(20): 5198931.
http://dx.doi.org/10.1155/2019/5198931 PMID: 31015822

[13] Cunningham F, Achuthan P, Akanni W, *et al.* Ensembl 2019. Nucleic Acids Res 2019; 47(1): 745-51.
http://dx.doi.org/10.1093/nar/gky1113 PMID: 30407521

[14] Ghatak S, Muthukumaran RB, Nachimuthu SK. A simple method of genomic DNA extraction from human samples for PCR-RFLP analysis. J Biomol Tech 2013; 24(4): 224-31.
http://dx.doi.org/10.7171/jbt.13-2404-001 PMID: 24294115

[15] Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 2014; 11(4): 361-2.
http://dx.doi.org/10.1038/nmeth.2890 PMID: 24681721

[16] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 2013; 7(1): 20.
http://dx.doi.org/10.1002/0471142905.hg0720s76 PMID: 23315928

[17] Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 2015; 31(16): 2745-7.
http://dx.doi.org/10.1093/bioinformatics/btv195 PMID: 25851949

[18] Hu J, Ng PC. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. PLoS One 2013; 8(10): 77940.
http://dx.doi.org/10.1371/journal.pone.0077940 PMID: 24194902

[19] Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 2009; 30(8): 1237-44.
http://dx.doi.org/10.1002/humu.21047 PMID: 19514061

[20] Kamaraj B, Rajendran V, Sethumadhavan R, Kumar CV, Purohit R. Mutational analysis of FUS gene and its structural and functional role in amyotrophic lateral sclerosis 6. J Biomol Struct Dyn 2015; 33(4): 834-44.
http://dx.doi.org/10.1080/07391102.2014.915762 PMID: 24738488

[21] Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 2013; 41(Database issue): 377-86.
PMID: 23193289

[22] Anoosha P, Sakthivel R, Gromiha MM. Prediction of protein disorder on amino acid substitutions. Anal Biochem 2015; 491: 18-22.
http://dx.doi.org/10.1016/j.ab.2015.08.028 PMID: 26348538

[23] Niroula A, Urolagin S, Vihinen M. PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS One 2015; 10(2): 0117380.
http://dx.doi.org/10.1371/journal.pone.0117380 PMID: 25647319

[24] Nailwal M, Chauhan JB. Computational Analysis of High-Risk SNPs in Human DBY Gene Responsible for Male Infertility: A Functional and Structural Impact. Interdiscip Sci 2019; 11(3): 412-27.
http://dx.doi.org/10.1007/s12539-018-0290-7 PMID: 29520635

[25] Duarte AJ, Ribeiro D, Moreira L, Amaral O. *In Silico* Analysis of Missense Mutations as a First Step in Functional Studies: Examples from Two Sphingolipidoses. Int J Mol Sci 2018; 19(11): 1-10.
http://dx.doi.org/10.3390/ijms19113409 PMID: 30384423

[26] Capriotti E, Altman RB, Bromber Y. Collective judgment predicts disease-associated single nucleotide variants. Mutations in proteins. BMC Genomics 2013; 14(S2): 1-9.

[27] Pejaver V, Urresti J, Lugo-Martinez JK, *et al.* MutPred2: inferring the molecular and phenotypic impact of amino acid variants. bioRxiv 2017; 1-28.
http://dx.doi.org/10.1101/134981

[28] Capriotti E, Fariselli P, Casadio R. R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure Nucleic Acids Res 2005; 33(Web server issue): 306-10.

[29] Mahdieh N, Rabbani B. An overview of mutation detection methods in genetic disorders. Iran J Pediatr 2013; 23(4): 375-88.
PMID: 24427490

[30] Venselaar H, Te Beek TAH, Kuipers RKP, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics 2010; 11(548): 548.
http://dx.doi.org/10.1186/1471-2105-11-548 PMID: 21059217

[31] Douville C, Carter H, Kim R, *et al.* CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics 2013; 29(5): 647-8.
http://dx.doi.org/10.1093/bioinformatics/btt017 PMID: 23325621

[32] MedCalc Statistical Software version 16.4.3 https://www.medcalc.org2016.

[33] One-way ANOVA followed by Dunnett's multiple comparisons test was performed using GraphPad Prism version 7.00 for Windows GraphPad Software, www.graphpad.com

[34] Tang H, Thomas PD. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. Bioinformatics 2016; 32(14): 2230-2.
http://dx.doi.org/10.1093/bioinformatics/btw222 PMID: 27193693

[35] Ozawa M, Ringwald M, Kemler R. Uvomorulin-catenin complex formation is regulated by a specific domain in the cytoplasmic region of the cell adhesion molecule. Proc Natl Acad Sci USA 1990; 87(11): 4246-50.
http://dx.doi.org/10.1073/pnas.87.11.4246 PMID: 2349235

DISCLAIMER: The above article has been published in Epub (ahead of print) on the basis of the materials provided by the author. The Editorial Department reserves the right to make minor modifications for further improvement of the manuscript.

Genes and Environment

Panel of significant risk factors predicts early stage gastric cancer and indication of poor prognostic association with pathogens and Microsatellite Stability

--Manuscript Draft--

Manuscript Number:	GENV-D-20-00041R2
Full Title:	Panel of significant risk factors predicts early stage gastric cancer and indication of poor prognostic association with pathogens and Microsatellite Stability
Article Type:	Research
Funding Information:	Department of Biotechnology , Ministry of Science and Technology ((BT/551/NE/U-Excel/2014) and (DBT-NER/health/46/2015)) | Dr. Nachimuthu Senthil Kumar
Abstract:	Background: There are very few studies covering the epidemiological risk factors associated with Epstein Barr Virus (EBV) and Microsatellite stability for Gastric Cancer (GC) cases. Early diagnosis of GC through epidemiological risk factors is very necessary for the clinical assessment of GC. The aim of this study was to find out the major risk factors to predict GC in early stage and the impact of pathogen infection and MSI on survival rate of patients. GC samples were screened for Helicobacter pylori, Epstein Barr Virus , and Mismatch repair (MMR) gene status (microsatellite stable or instable). Chi-square and logistic regression analysis of Odd ratio and 95% confidence interval (OR, 95% CI) were performed to find out the association between epidemiological factors and the risk of gastric cancer. The pathogen and MMR gene status were analysed to predict their effect on overall survival and the risk score and hazard ratio was calculated for prognostic assessment. Results: Excess body weight, consumption of extra salt, smoked food, alcohol, and smoking were the major risk factors for GC development. This study achieved a high area under the curve (AUC = 0.94) for the probable GC patients in early-stage using the five-panel epidemiological risk factors. H. pylori infected cases were significant with smoked food, while EBV was found to be associated with tuibur intake and smoked food. In overall survival analysis EBV infected and microsatellite stable (HR: 1.32 and 1.34 respectively) GC cases were showing poor prognosis. Conclusion: This study might provide new opportunities for personalized treatment options using this epidemiological factor risk score and clinicopathological factors assessment for early detection and prognosis in high-risk GC populations.
Corresponding Author:	Nachimuthu Senthil Kumar Mizoram University Aizawl, Mizoram INDIA
Corresponding Author E-Mail:	nskmzu@gmail.com
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Mizoram University
Corresponding Author's Secondary Institution:	
First Author:	Payel Chakraborty
First Author Secondary Information:	
Order of Authors:	Payel Chakraborty
	Souvik Ghatak
	Saia Chenkual
	Lalawmpuii Pachuau
	Jhon Zohmingthanga
	Zothankima Bawihtlung

	Lalfakzuala Khenglawt
	Jeremy L. Pautu
	Arindam Maitra
	Lalchhandama Chhakchhuak
	Nachimuthu Senthil Kumar
Order of Authors Secondary Information:	

Response to Reviewers:

Answers to Reviewer’s comments

The authors are thankful to the editor and the reviewer’s for the valuable comments and suggestions on the manuscript. The authors have carried out all the comments as per the suggestions of the reviewer’s.

Reviewer #1:

Line 110; What is criterion for separating the participants into consumers and non-consumers for each lifestyle factor or food factor?
Lifestyle habits such as: a) smoking, categorised as smokers (who used to smoke at least once a week for three months or more) and non-smokers (if the person never smoked before or left smoking for more than 5 years), b) chewing tobacco in smokeless form, categorised as consumer (who used to take atleast once a week for six months or more) and non-consumer (if the person never consume before or left more than 5 years before), c) tuibur or tobacco infused water, categorised as drinkers (if the person used to drink at least once in a week) and non-drinkers (if the person never drink) and d) alcohol, categorised as drinkers (if the person used to drink at least one day in a week) and non-drinkers (if the person never drink).
Food habits such as: a) extra salt intake, categorised as consumers (if the person takes extra salt at least for once in their meal in a week) and as non-consumers (if the person never takes extra salt with their daily food for once), b) smoked food, categorised as consumers (if the person ate at least for one day in a week) and as non-consumers (if the person did not ate even for a single day in a week) and c) sa-um or fermented pork fat, categorised as consumers (if the person ate at least for once in a week) and as non-consumers (if the person did not ate even for once in a week).

The authors thank the reviewer for the valuable comment. These points on categorisation have been now mentioned in the Statistical analysis section in the revised manuscript.

Line 115; The questionnaire must be presented as a supplemental material.
Yes, we have now included the questionnaire as a supplementary file (file name: Questionnaire).
Line 231; P-value of smokeless tobacco consumer was 0.06. Was this categorized as a risk group, because p < 0.05 was considered as statistically significant in this study as described on line167?
The authors thank the reviewer for pointing our mistake. The smokeless tobacco consumer was not a risk group for this study. We have re-written the sentence as “Smokeless tobacco (tuibur) consumers (p-value = 0.06) were at low risk for developing EBV associated GC”.
Table 1; Odd ratios for food factors with statistical significance were lower than 1, but those for lifestyle factors were higher than 1. What means a difference in data presentation between food factors and lifestyle factors? Please explain a meaning of this difference.
In this study, most of the controls and patients have same food habits, so the frequency are not much different between control and patients groups and hence, the Odd ratios is less than 1. With respect to life style habits, the control group does not follow the same lifestyle habits like patients, and that may be the reason for higher Odd ratio in lifestyle factors.

Authors must discuss limitations of this epidemiological study; e.g. numbers of cases were low this study.
The authors thank the reviewer for the valuable suggestion. We have included the statement about the limitation and strength of this study in the end paragraph of the

discussion section.

Reviewer #2:

It seems unexpected that microsatellite stable GC cases show poor prognosis. Usually microsatellite instability is associated with increased cancer risk. This unexpected result should be checked again and discussed in the Discussion.
We have found in several studies that MSI patients had better prognosis in gastric cancer patients than MSS patients.
We have now mentioned new references in the discussion section (like MAGIC trial with large number of patient’s cohort) and included in the references also, which have greatly strengthened the revised manuscript. The authors thank the reviewer for the valuable comment.
If the bulk of the cases were late stage GC then the results obtained most likely relate to late stage GC rather than early stage GC. Discuss whether the study actually relates to late stage GC rather than early stage GC.
We have used all the stages for the analysis in the manuscript. First, we have established the model using all the stages (Stage I, N = 20; Stage II, N = 14; Stage III, N = 40; Stage IV, N = 6). Then, we have used only the early stage samples (Stage I, N = 20; Stage II, N = 14;) to establish the model and calculate the risk score.

We have added these results in Figure 3 and also in Result and Discussion section in the revised manuscript.

Explain "MMR" in abstract and main text.
We have mentioned the full form of MMR in abstract and in the introduction section also, as per the suggestion of the Reviewer.
Line 65. Is the "Mizo" population, the population in Mizoram? Explain.
Mizo are the tribes of Mizoram and Mizo population means the people of Mizoram.

Line 78. Which early stage GC biomarkers were used in the reported study? What are their strengths and weaknesses?
We have used Patient’s Sex, BMI, Extra salt intake, Smoked food consumption, alcohol drinking and smoking habits were used as an early stage GC epidemiological factors. Most of the markers were the strong risk factors for gastric cancer. We have mentioned about all the strength and weaknesses in the discussion section.

Line 93 and line 97. In line 93 it states that a 2:1 ratio of controls: cases was planned but in lines 93 and 97 it indicates that only 120 controls and 80 cases were recruited. What is the reason for a deficit in 40 controls? There must be an error because in line 97 it indicates "120 healthy controls (79 male, 81 female)".

The authors thank the reviewer for pointing out the typographical mistake. The total number of healthy controls was 160 for this study. We have corrected this error in the revised manuscript.
Line 112. Explain "saum".
Saum is a fermented pork fat, which is a common food habit for this studied population. Sa-um preparation takes place on a cottage-industrial scale in households which does not have firmly established procedures and as a result the production process fluctuates on a seasonal basis. Sa-um, has peculiar sensorial attributes (smell and taste) due to ripening process besides the enzymatic lipolytic activities of the microbial populations present in it.
We have mentioned about it in the Introduction section in the revised manuscript.
Line 113. Explain "TNM staging".
"TNM staging" is used to describe the stage of Cancer: Tumor (T) - How deeply has the primary tumor spread into the stomach wall? Node (N) - Has the tumor spread to the lymph nodes? If so, where and how many? Metastasis (M): Has the cancer spread to other parts of the body? The results are combined to determine the stage of stomach cancer: Stages I, II, III and IV.
We have mentioned this in the Methodology - Data collection section of the revised manuscript.
Line 142. The PCR temperature should show a "degree" symbol not a box
Yes, we have corrected the symbol throughout the manuscript.
Line 183. How many of the cases were early stage gastric cancer?

A total of 34 cases (Stage I, N = 20 and Stage II, N = 14) of early stage gastric cancer.

Lines 193-197. Effect of micronutrient deficiencies and/or obesity should also be reported/discussed even if not significant.

For all the analysis (univariate and multivariate), we have included BMI index and it is highly significant factor for the advance and early stage gastric cancer patients.

We have included the BMI index for the analysis in the revised manuscript, according to the reviewer comments. We have mentioned in the result (Table 1, Figure 3) and also in the revised manuscript.

Lines 220-222. How many of the cases were positive for both H. pylori and EBV?

A total of 11 (13.75%) gastric cancer patients were positive for both H. pylori and EBV. We have mentioned this in the Result section of the revised manuscript.

Line 225. change "were as" to "whereas"

Yes, we have changed "were as" to "whereas" in mentioned sentences.

Did gender have an impact on GC risk??

Yes, gender has significant impact for GC for this population. We have found a large number of male gastric cancer patients (66.25%) in this studied population than female patients.

We have mentioned this in the Result section (Table 1) of the revised manuscript.

Lines to 276-277. There is no evidence that salt is a source of N-nitroso compounds. It does not make sense. reference 26 does not suggest this…….it suggests that salt may increase susceptibility to the carcinogenic affects of N-nitroso compounds. Please delete the sentence or amend it as indicated.

The authors thank the reviewer for pointing out the mistake. We have re-written the sentence according the reference 26 in the discussion section of the revised manuscript.

Lines 303-305. This sentence is vague. It should state that ALDH2 is required to detoxify acetaldehyde which is a Class I carcinogen derived from alcohol, by converting it to acetate. Mutations that inactivate ALDH2 are more prevalent in some Asian countries. Reference: Ghosh S, Bankura B, Ghosh S, Saha ML, Pattanayak AK, Ghatak S, Guha M, Nachimuthu SK, Panda CK, Maji S, Chakraborty S, Maity B, Das M. Polymorphisms in ADH1B and ALDH2 genes associated with the increased risk of gastric cancer in West Bengal, India. BMC Cancer. 2017 Nov 22;17(1):782. doi: 10.1186/s12885-017-3713-7. PMID: 29166882; PMCID: PMC5700676.

We have rewritten the sentence according to the reviewer suggestions in the discussion section of the revised manuscript.

The discussion should include a paragraph on the strengths and weaknesses of the study.

We have included a paragraph about the strengths and weaknesses of the study in the last paragraph of the discussion section of the revised manuscript.

Table 1. Replace "ODD Ratio" with "ODDS Ratio"

Yes, we have replaced the "ODD Ratio" with "ODDS Ratio" throughout the manuscript.

Table 3. Explain what the ORs relate to.

We have explained the ORs in the Table 3 (as foot note) of the revised manuscript.

Figure 2 legend. Explain the significance of the width of the ribbons.

We have explained the significance of the width of the circus plot ribbons in the Figure 2 legend in the revised manuscript.

Figure 5 replace "acetate from acetaldehyde" to "alcohol conversion to acetaldehyde, a Class I carcinogen"

We have replaced the "acetate from acetaldehyde" to "alcohol conversion to acetaldehyde, a Class I carcinogen" in Figure 5 of the revised manuscript.

Additional Information:	
Question	**Response**

Panel of significant risk factors predicts early stage gastric cancer and indication of poor prognostic association with pathogens and Microsatellite Stability

Payel Chakraborty[1], Souvik Ghatak[1], Saia Chenkual[2] , Lalawmpuii Pachuau[3], John Zohmingthanga[3], Zothankima Bawihtlung[4], Lalfakzuala Khenglawt[4], Jeremy L. Pautu[5], Arindam Maitra[6], Lalchhandama Chhakchhuak[3] and Nachimuthu Senthil Kumar[1]*

[1]Department of Biotechnology, Mizoram University, Aizawl -796004, Mizoram, India
[2]Department of Surgery, Civil Hospital, Aizawl -796001, Mizoram, India
[3]Department of Pathology, Civil Hospital, Aizawl -796001, Mizoram, India
[4]Department of Radiation Oncology, Mizoram State Cancer Institute, Zemabawk, Aizawl, Mizoram, India
[5]Department of Oncology, Mizoram State Cancer Institute, Zemabawk, Aizawl, Mizoram, India
[6]National Institute of Biomedical Genomics, P.O. NSS, Kalyani, District Nadia - 741251, West Bengal, India

Corresponding Author

Nachimuthu Senthil Kumar
Professor, Department of Biotechnology
Mizoram University *(A Central University)*
Aizawl – 796 004, Mizoram, India
Email: nskmzu@gmail.com

Abstract

Background: There are very few studies covering the epidemiological risk factors associated with *Epstein Barr Virus (EBV)* and Microsatellite stability for Gastric Cancer (GC) cases. Early diagnosis of GC through epidemiological risk factors is very necessary for the clinical assessment of GC. The aim of this study was to find out the major risk factors to predict GC in early stage and the impact of pathogen infection and MSI on survival rate of patients. GC samples were screened for *Helicobacter pylori, Epstein Barr Virus*, and Mismatch repair (MMR) gene status (microsatellite stable or instable). Chi-square and logistic regression analysis of Odd ratio and 95% confidence interval (OR, 95% CI) were performed to find out the association between epidemiological factors and the risk of gastric cancer. The pathogen and MMR gene status were analysed to predict their effect on overall survival and the risk score and hazard ratio was calculated for prognostic assessment.

Results: Excess body weight, consumption of extra salt, smoked food, alcohol, and smoking were the major risk factors for GC development. This study achieved a high area under the curve (AUC = 0.94) for the probable GC patients in early-stage using the five-panel epidemiological risk factors. *H. pylori* infected cases were significant with smoked food, while *EBV* was found to be associated with tuibur intake and smoked food. In overall survival analysis *EBV* infected and microsatellite stable (HR: 1.32 and 1.34 respectively) GC cases were showing poor prognosis.

Conclusion: This study might provide new opportunities for personalized treatment options using this epidemiological factor risk score and clinicopathological factors assessment for early detection and prognosis in high-risk GC populations.

Keywords: Gastric Cancer; Risk factors; *H. pylori*; *EBV*; MSI; Clinicopathological Data

Introduction

Gastric Cancer (GC) is a heterogeneous disease and varies widely based on etiological factors and genetic architecture. Histologically, most of the GC are adenocarcinoma and can be further classified as diffuse (poorly differentiated) or intestinal (well-differentiated) types [1, 2], with unique epidemiological influence and genetic signatures. GC, being the fifth most commonly occurring cancer, is prevalent in the eastern and central parts of Asia and is the third most common cancer as per the mortality rate [3, 4]. Mizoram, Northeastern tribal state of India has the highest incidence rate of gastric cancer in India [5] and globally occupies the fifth position for GC [6].

Several studies reported that dietary, behavioral, and lifestyle habits significantly increase GC risk, and every population/ race has unique dietary and lifestyle habits. Mizo population also have unique traditional food and habit which might play role for developing GC. Mizo ethnic food, sa-um (fermented pork fat) is rich in fat content and has been shown to retain pathogens which can have an adverse effect on human health [7]. Sa-um preparation takes place on a cottage-industrial scale in households which does not have firmly established procedures and as a result the production process fluctuates on a seasonal basis. Another unique traditional habit of use of alkaline tobacco infused water (tuibur) containing polyaromatic hydrocarbons and carbonyl compounds [8, 9] may also have an effect on pathogen incidence as well as GC. Various studies suggest that there is an association between pathogens (*Helicobacter pylori* and *Epstein Barr Virus*) and microsatellite instablilty (MSI) for GC development [10-12]. The prevalence of pathogens and MSI associated GC cases varies depending on different populations [13, 14]. Few studies have highlighted about the risk factors of MSI in other cancers [15-19].

Till date, there is no in-depth study on the epidemiological risk factors associated with *Epstein Barr Virus (EBV)* and MSI for GC cases. Therefore, this study was carried out to find the unique risk factors which might be involved for developing pathogen and MSI associated GC. There is a lack of epidemiological markers to predict GC at an early stage and there is not much information available about the pathogen specific risk factors and their prognosis assessment on associated GC cases.

The aim of the present study is to: i) to find the predictive epidemiological factors which can aid to estimate the GC risk at an early stage, ii) to assess the significant risk factors which can elevate GC risk in presence of pathogens and MSI, and iii) also to assess the effect of pathogen infection and Mismatch repair (MMR) gene status on survival rate of patients. We hypothesize that the exposure to major risk factors can predict GC at an early stage and individuals with pathogens or MSI have increased risk of developing GC that might affect the survival rate of the patients.

Materials and Methods

Study population

This is a case-control study consisting of GC patient samples collected from different hospitals (Civil Hospital Aizawl, Ebenezer Hospital, Aizawl Hospital, and Green Wood Hospital) in Mizoram, Northeast India from September 2016 to January 2019. The controls and cases were randomly selected at 2:1 ratio by age and sex, respectively. A total of 80 patients (53 male and 27 female) were selected after conforming histologically as a case of stomach adenocarcinoma by the pathologists. Their age ranged from 31 to 86 years. Patients who had any chronic diseases without GC, history or present record of gastritis and pre-treated for any other type of cancer were not eligible for this study. A total of 160 healthy controls (79 male and 81 female) were randomly selected from the same ethnic group from where the patients were selected and belong with almost similar age from 31 to 85 (57.96 ± 11.48). Patients who had any chronic diseases, gastritis, and cancer were not eligible as control. The work was approved by ethical committees of Civil Hospital, Aizawl (B.12018/1/13-CH(A)/IEC dtd. 18/04/2014), and Human Ethical Committee, Mizoram University (MZU/IHEC/2015/008 dtd. 14/12/15). The details study design was represented in Figure 1.

Data collection

All the participants of this study were interviewed using a well-designed and informative questionnaire with a duly informed consent form. A telephonic interview was also done for the follow-up study, with the patient group and respective clinicians. The questionnaire contains demographic information (age, sex and BMI), lifestyle habits like smoking (categorized as smokers and non-smokers), smokeless tobacco Chewed tobacco, Paan with betel nut, tuibur (tobacco infused water), a unique habit of Mizoram (categorized as consumers and non-consumers), alcohol (categorized as drinkers and non-drinkers), food habits like extra salt consumption, smoked food and saum, fermented pork fat (categorized as consumers and non-consumers). The clinical data like tumor size, anatomy, pathological TNM staging (American Joint Committee on Cancer, 8th edition: Tumor (T)- How deeply has the primary tumor spread into the stomach wall?; Node (N)- has the tumor spread to the lymph nodes? where and how many?; Metastasis (M)- has the cancer spread to other parts of the body?), tumor Grade, family history and overall survival status were also recorded using a structured questionnaire.

DNA isolation from Tissue and Blood sample

Fresh gastric tumor tissue and peripheral blood samples (3 ml in EDTA Vial) for each patient and peripheral blood sample for healthy controls were collected. Genomic DNA was extracted from the cancerous tumor tissue and corresponding blood samples using commercially QIAamp® DNA Tissue Kit and QIAamp® Blood DNA mini kit. The extracted DNA was electrophoresised with 0.8% agarose gel and quantified using Picogreen dye in Qubit Fluorimeter (Invitrogen).

Pathogen Genotyping

The presence of *Helicobacter pylori* infection was determined in GC patients by PCR amplification of specific 16SrRNA region, *UraC* genes. The presence of *Epstein Barr Virus (EBV)* type1/ type 2 infection was carried out using a standard PCR assay by type-specific region (EBNA3C - Epstein–Barr virus nuclear antigen 3C) gene using specific primer sets [19]. The PCR reaction volume of 10 µl contained: 1x PCR buffer with, 1 unit of Taq DNA Polymerase, 0.2 mM dNTPs (All from the Thermo Scientific, USA), and 0.2 Pico mol primer (Active Oligo-ILS, Bangalore, India). The reaction mixture (10 µl) was PCR amplified for initial denaturation at 95°C for 5 min, followed by 35 cycles at 95°C for 1 min., n°C (depending on primer) for 40 s, 72°C for 40 sec/1 min followed by extension at 72°C for 5 min. (Supplementary Table 1). *H. pylori* and *EBV* positive and negative control samples were used in all the PCR amplification for confirmation.

PCR amplification of microsatellite loci

MSI was determined by comparison of the allelic profiles of the mononucleotide repeat markers BAT-25, BAT-26, and Dinucleotide Markers D2S123, D17S250, D16S752, D16S265, D16S398, D16S496, D18S58, and D16S3057 in tumor and corresponding blood and control blood 33-34

(Supplementary Table 1). The forward primers for the markers were labelled with fluorescent dye 6-FAM, VIC, NED, and PET. The PCR reaction volume of 10 µl contained: 1x PCR buffer, 1 unit of Taq DNA Polymerase, 0.2 mM dNTPs, and 0.15 Pico mol primers (Thermo Scientific). PCR was performed with a Master cycler (Eppendorf, nexus GX2). The following cycling regime was used as a "standard" PCR protocol: initial denaturation at 95°C for 10 min, followed by 35 cycles at 94°C for 1 min, 55°C for 40 sec and 72°C for 40 sec and the final extension step of 7 min at 72°C (Supplementary Table 1).

Fragment Analysis

The amplified loci were analyzed using the automated ABI sequencer model 3500 Genetic Analyzer (Applied Biosystems, Singapore). In brief, 8.7 µl deionized formamide was combined with 0.3 µl GeneScan^{Tm}-600 size standards (Applied Biosystems, V-2.0) and 1 µl PCR product in a Genetic Analyzer sample plate. After adding samples, the plate was sealed by septa, and mixing was done by mild vortexing. The denaturation step was done at 90°C for 2 min, followed by keeping the plate on ice, and a mini-centrifugation for 1 min. The MSI of the investigated loci was defined as allele shift or (and) appearance of novel peaks. Samples were classified as MSI or MMR deficient if at least two or more than two markers were having instability and the instability was found only in BAT-26 Maker. If instability was not found in any of the markers, then the sample was classified as MSS [20] (Supplementary Figure 1).

Statistical Analysis

Distribution of demographic and lifestyle characteristics between the control and case groups were compared by chi-square test [21]. The odd ratio (OR) and 95% confidence intervals (CIs) were estimated for determining association in each group of factors among case-control subjects by binary logistic regression (Univariate and Multivariate analysis) [19]. All the demographic factors were grouped as follows: excess body weight [body mass index (BMI) ⩾ 25], Lifestyle habits such as: a) smoking, categorised as smokers (who used to smoke at least once a week for three months or more) and non-smokers (if the person never smoked before or left smoking for more than 5 years); b) chewing tobacco in smokeless form, categorised as consumer (who used to take atleast once a week for six months or more) and non-consumer (if the person never consume before or left more than 5 years before); c) tuibur or tobacco infused water, categorised as drinkers (if the person used to drink at least once in a week) and non-drinkers (if the person never drink); and d) alcohol, categorised as drinkers (if the person used to drink at least one day in a week) and non-drinkers (if the person never drink).It has detailed information on food habits such as: a) extra salt intake, categorised as consumers (if the person takes extra salt at least for once in their meal in a week) and as non-consumers (if the person never takes extra salt with their daily food for once); b) smoked food, categorised as consumers (if the person ate at least for one day in a week) and as non-consumers (if the person did not ate even for a single day in a week); and c) sa-um or fermented pork fat, categorised as consumers (if the person ate at least for once in a week) and as non-consumers (if the person did not ate even for once in a week).

The independent impact of hazard components was further explored in a multivariate model (presenting all factors and terms of connections) keeping only those statistically significant or demonstrating a confounding effect on the contemplated elements. The likelihood test was utilized to choose whether to hold each covariate in the model. BMI, Cigarette smoking, alcohol, smoked food (meat or vegetable consumption), high intake of salt were considered altogether in the estimated risk model as potential confounders to assess the relationship of hazard factors and susceptibility to gastric cancer. For all tests, a two-sided p-value <0.05 was considered as statistically significant. Circos plot was generated using circos software for association demographic factors between GC patients and healthy control. Another association approach was done within the patients between the risk factors and clinical data among the subgroups of with or without *H. pylori*, *EBV* infection and MMR deficient (MSI)/MMR proficient (MSS) were estimated by calculating odds ratio (OR) and 95% confidence intervals (CIs) using binary logistic regression method and representing by forest plot using R software. Overall survival was determined using the Cox proportional-hazards regression model (using three years cut-off). The log-rank test, Kaplan-Meier survival analyses were used to assess the impact of the variables on survival. Variable used for survival analysis were *H. pylori* status, *EBV* status, MSI status, and anatomical site.

Results

The baseline characteristics of the total GC patient cohort are presented in Supplementary Table 2. The age group interval of 40-69 years shows the highest number of GC patients (75%) in this cohort. About 32.5% of patients were having a first-degree family history of all types of cancer. Among the 80 GC patients, 50% of the cases were found in stage III and 8.75% were graded as well-differentiated, 46.25% were moderately differentiated and 32.5% were poorly differentiated cases. Most of the tumor was located in the distal part of the stomach and the prevalence of GC was high in male patients in this cohort (Supplementary Table 2).

Supplementary Table 3 presents the distribution of demographic and lifestyle habits among GC patient and controls. Extra salt consumption was a significant risk factor (*p* value = 0.0001) along with Smoked food consumption (*p* value = 0.01), Smoking (*p* value = 0.0001) and alcohol drinking (*p* value = 0.0001) which are also high risk factors for developing GC. The frequency and association of demographic factors and lifestyle habits between GC patients and healthy control (HC) were presented as Circos plot (Figure 2).

The univariate binary logistic regression analysis was performed for sex, BMI, dietary and lifestyle habits.). Sex (*p*-value = 0.019) and BMI (*p*-value = 0.0001) were significant factors for the gastric cancer patients (Table 1). Among the dietary factors, extra salt consumption (*p*-value = 0.007) and smoked food consumption (*p*-value = 0.0001) were the major risk factors for the GC patients. Smokeless tobacco (tuibur) intake (*p*-value = 0.011), smoking (*p*-value = 0.0001) and alcohol consumption (*p*-value = 0.0001) became significant lifestyle risk factor for GC risk (Table 1).

We further performed the multivariate analysis with these seven significant factors for finding out the major risk factors and confounding factors which were associated with GC development. Five factors were predicted as significantly associated with GC risk with high OR and 95% CI in multivariate analysis. BMI (*p*-value = 0.0001), Extra salt consumers (*p*-value = 0.042), smoked food consumers (*p*-value = 0.001), smokers (*p*-value = 0.0007) and alcohol drinkers (*p*-value = 0.001) were the high-risk groups associated with GC development (Table 1, Figure 3A,). A risk score was estimated with the 5 factors using a logistic model and validated the risk score in the GC clinical cohort (Stage I, N = 20; Stage II, N = 14; Stage III, N = 40; Stage IV, N = 6) (Figure 3A). The exposer of five-panel epidemiological factors might be successful in predicting the GC risk with different early symptoms (area under the curve – AUC = 0.91; *p*-value < 0.0001) (Figure 3B). This 5-panel epidemiological factor achieved high-risk core with significant-high positive probability values for GC patients with high sensitivity (79.45%) and specificity (91.72%) (Figure 3C).

For predicting GC at early-stage, a risk score was estimated with the 5 factors using a logistic model and was validated in the early stage (Stage I, N = 20 and II, N = 14) GC clinical cohort (Figure 3D). The exposer of five-panel epidemiological factors might be successful in predicting the GC risk during the premalignant stage with different early symptoms with higher AUC value (0.946; *p*-value < 0.0001) (Figure 3E). This 5-panel epidemiological factor achieved high-risk core with significant-high positive probability values for GC patients with high sensitivity (96.67%) and specificity (80.89%) (Figure 3F). The estimated significant factors (BMI, extra salt consumption, smoked food, alcohol drinking, and smoking) were the major risk factors associated with GC development.

We have subdivided our GC patient cohort for pathogen infections and mismatch repair (MMR) gene deficiency with molecular identification of *H. pylori*, *EBV*, and MSI. Out of 80 patients, 71 (88.75%) cases were positive for the pathogens. Fifty cases (70.04%) were detected positive for *H. pylori*, *EBV* positive cases were 32 (45.07%) and a total of 11 (13.75%) gastric cancer patients were positive for both *H. pylori* and EBV. Moreover, 40% of cases were detected as MMR deficient (microsatellite instable-MSI) (Table 2, Supplementary Figure 1A) and 60% cases were detected as MMR proficient (microsatellite stable-MSS) (Table 2, Supplementary Figure 1B).

We categorized our cases as *H. pylori* (+), *H. pylori* (-), *EBV* (+), *EBV* (-), MMR deficient, and MMR proficient and compared each group with clinical, demographic and lifestyle habit data to find out significant factor with each subgroup of GC patients. Table 2 presents the frequency distribution of clinical factors among the subgroups of GC patients. The tumor was located at high frequency in the distal portion of the stomach for the MMR deficient (87.5%) and *H. pylori*-positive (70%) patients group whereas less frequency was observed for *EBV* positive

(65.62%) patient group. MMR deficient, *EBV* positive and *H. pylori*-positive cases were high for the poorly differentiated adenocarcinoma group whereas MMR proficient, *EBV* negative, and *H. pylori*-negative cases were high for the moderately differentiated adenocarcinoma patient group.

Smoked food consumption was the only significant risk factor associated with *H. pylori* positive GC patients and *EBV* infected patient group with respective *p* value (*p*-value= 0.006 and *p*-value= 0.002). Smokeless tobacco (tuibur) consumers (*p*-value = 0.06) were at low risk for developing *EBV* associated GC (table 3). Tobacco chewing and Alcohol drinking were found as significant risk factor for MMR deficient patients group with high OR, 95% CI (*p*-value = 0.04) and (*p*-value= 0.03), respectively (Table 3).

For further verification, we performed binary logistic regression for determining the odd ratio and 95% CI. A significant association was found between *H. pylori*-infected GC patients with consumption of smoked food (*p*-value = 0.007) (Figure 4A, Supplementary Table 4). Smoked food consumption (*p*-value=0.003) and tuibur intake (*p*-value = 0.05) were significant factors for *EBV* infected GC patients and tuibur consumption (Figure 4C, Supplementary Table 4). Significant association was observed with chewing tobacco (p-value = 0.04) and alcohol drinking (p-value = 0.03) for the MMR deficient (MSI) patient group (Figure 4E, Supplementary Table 4). Factors such as smoked food and tuibur consumption are found to be the major risk for pathogen infection in GC patients and chewing tobacco, alcohol drinking as lifestyle factors became the risk factors for MMR deficient GC patients (Figure 4E).

We further studied the overall survival (OS) rate of patients with the subgroup of with and without *H, pylori*, *EBV* infections, and MMR deficient and proficient patients to find out prognostic risk factors by unadjusted analysis after a follow-up of 36 months using the Kaplan Meier curve. A univariate Cox proportional hazards model demonstrated that there was no relation between *H. pylori* infections and GC patient's prognosis with stage I, II, and III (HR: 1.13, 95% CI: 0.86 - 1.73; *p*-value = 0.13; Figure 4B). *EBV* infections and MSI were an independent prognostic predictor for GC patients with stage I, II, and III (Figures 4D and 4F). The *EBV* infected GC patients with stage I, II, and III was poor prognosis and high-risk patients (HR: 2.22, 95% CI: 0.92 - 2.97; *p*-value = 0.05). The comparison between MMR deficiency and proficiency exhibited significant prognostic predictor for stage I, II, and III GC patient groups (HR: 3.43; 95% CI: 0.95 - 4.08; *p*-value = 0.03). MSI/MMR deficient cases showed a good prognosis, whereas MSS/MMR proficient cases showed poor prognosis for GC patients (Figure 4F). We have compared the *H. pylori*, *EBV*, and MMR gene status as independent prognostic factors for stage I, II, and III gastric cancer patients group in the TCGA-STAD cohort. Cox proportional-hazards regression model showed that there was no significant log-rank value *p*-value with *H. pylori* status (Figure 4G) whereas MMR gene status an independent prognostic factor in TCGA-STAD cohort (HR: 1.60; 95% CI: 1.04 – 1.91; *p*-value = 0.03) (Figure 4H).

Discussion

To the best of our knowledge, this is the first retrospective study in Southeast Asia designed to assess the potential role of *H. pylori* / *EBV* infections, MMR gene status and epidemiological risk in the prognosis of GC patients. The results of the present study indicate that smoked food is the major risk factor that is significant in most of the subgroups of GC patients and the unique risk factor (tuibur) is found to be significantly associated with EBV infection. *EBV* infection is a strong risk factor for poor prognosis of GC in this Indian high-risk population.

Gender has significant impact for GC for this population. A large number of male gastric cancer patients (66.25%; OR = 0.50; 95% CIs = 0.28 – 0.89; *p*-value = 0.019) was found in our study. Excess body weight (BMI ⩾ 25) was associated with an increased risk of gastric cancer (OR = 0.63; 95% CIs = 0.56 – 0.72; *p*-value = 0.0001). Specifically, a multivariate stratified analysis showed that excess body weight was associated with an increased risk of gastric cancer [overweight and obese (BMI ⩾ 25), OR = 0.69, 95% CI = 0.60 – 0.79; *p*-value = 0.0001)]. Consumption of extra salt was found as dietary risk factor for GC. Extra salt can increase the mucin level of the surface mucus in the stomach which provides the possible condition for colonization of *H. pylori*, a significant risk factor of stomach cancer [22, 23]. It can significantly increase the carcinogenic A (CagA) gene expression in *H. pylori* which in turn alters the function of the epithelial cells and induces the hypergastrinemia in GC patients [24]. Extra salt intake could induce the inflammatory response of epithelial cells which may be responsible for cell proliferation and endogenous mutation [25] and moreover, it may increase susceptibility to the carcinogenic effects of N-nitroso compounds which can cause cell death [26]. Considering the present and past literature, we hypothesized that salt is a promoter of gastric adenocarcinoma, particularly in combination with *H. pylori* infection and that optimum quantity of salt consumption is significant for avoiding the gastric adenocarcinoma.

In this study, smoked food was found as another significant dietary risk factor with more than 60% of patients consuming the smoked food. Smoked food is generated by smoking or grilling method (a technique for cooking food on an oven or smoke generating system like burning of wood or charcoal) [27], and could produce both good antioxidants and antimicrobial properties, as well as carcinogenic chemicals like Polycyclic Aromatic Hydrocarbons (PAH) [28]. Benzo[a]pyrene (BaP), a member of PAH family, is a group I carcinogen which plays a role in the progression of GC and other cancers as well (Figure 5). BaP accumulates in our body by metabolic activation of cellular membrane cytochrome P450 followed by producing toxic byproducts that will bind with DNA to create DNA adducts leading to gene mutation [29] and functional changes in proteins through AhR/CYP450 pathway [30]. BaP causes proliferation in GC cell lines and upregulation *via*. MMP9 and c-myc expression [31]. Studies have reported that smoked-dried or processed foods are strongly associated with GC development [19, 32] which is supporting our results.

In this study, smoking and alcohol consumption were also found as significant risk factors. Several studies have reported that smoking is an associated risk factor with GC [30, 33]. In this cohort, 65% of GC patients were smokers, whereas more than 78% were non-smokers in the healthy control group. Studies have reported that smoking has more impact on developing GC in men than in women [34], while another study has suggested that smoking is significant for both (men and women) to develop GC [35].

The effect of alcohol drinking on GC is always a matter of conflict. IARC has reported alcoholic beverages as a risk factor for humans in case of several cancers, but for GC no direct relations has been established so far [36], as most of the study showed uncertain results. ALDH2 is required to detoxify acetaldehyde which is a Class I carcinogen derived from alcohol by converting it to acetate. Mutations that inactivate ALDH2 are more prevalent in some Asian countries [37]. China has reported alcohol as an independent risk factor for GC in their population [38]. One Korean cohort study has reported that GC in the stomach cardia or upper third position had a significant association with smoking, and GC occurring in the distal part was associated with high alcohol consumption [39]. In our current study, more than 97% of healthy controls were non-drinkers, whereas more than 36% of patients were drinkers. One of the hypothetical mechanisms from the current and previous published study with all the significant risk factors for developing GC has been represented in Figure 5.

We estimated the chances of GC development using the estimated risk score of the 5 different epidemiology factors (BMI, extra salt, smoked food, alcohol consumption, and smoking habit, Figure 3A). The consumption of all the five factors is independently associated with GC risk in univariate analysis, whereas tuibur consumption did not achieve significance *p*-value in the multivariate model for GC risk. We found a significant difference in risk score probability between gastric cancer patients and healthy control (*p*-value < 0.0001, Figure 3B). This study achieved a high area under the curve (AUC = 0.946) value for detecting the probable GC patients from the population using the 5-panel epidemiology factors at an early stage (Figure 3C).

In the current study, a significant association was observed between smoked food consumption and *H. pylori* infection associated GC. Several studies reported the strong association between *H. pylori* and extra salt [22, 23]. Extra Salt-curing or brining adds flavor, allows the nitrites to penetrate the flesh and most important, extracts moisture from the food, allowing the smoke to penetrate more easily. Most cold-smoked meats are generally salt-cured or brined first. In this population, smoked foods are also rich in salt which can make a favorable condition for *H. pylori* infection and lead to developing the GC. Further studies with prospectively collected GC samples are necessary to support our data. In the present study, the consumption of smoked food was a significant risk factor for GC with *EBV* infection. Consumption of smoked food, smoking cigarettes are significant contributing factors, for developing carcinogenesis of GC, which might be amplified by the presence of *EBV*. It has been reported that smoking has a strong association with the risk of *EBV*-positive Hodgkin’s lymphoma [40] and that tobacco a risk factors for GC, may contain *EBV*-activating substances [41]. In the current study, tuibur consumption (tobacco infused water) was also a significant risk factor for *EBV* infected GC patients. Tuibur, a unique risk factor is tobacco infused water, so we can categorize it as smokeless tobacco and it contains polyaromatic hydrocarbons and carbonyl compounds. Studies have reported that smokeless tobacco affects B-lymphocytes [42], where latent *EBV* virus infection takes place [43] and infected lymphocytes at a later stage are responsible for tumorigenesis. Studies reported a positive association between smokeless tobacco and *EBV* type I and type II infections [44]. Another important aspect is *EBV* spreads by body fluids, especially saliva. In rural villages, the tuibur consumers share the same tuibur bottle for drinking and it can pass on from an *EBV* infected person to others through saliva. As smoked food preparation is done by exposing smoke and whole tobacco plant is used for tuibur preparation, so it can also help to increase the risk of *EBV* infected GC which needs to be revealed by further study.

This study has reported two lifestyle factors, chewing tobacco and alcohol drinking as a significant risk factor associated with MMR deficient GC patients. Studies have reported that tobacco and alcohol drinking are strongly associated with MSI-H colorectal cancer cases [15-19]. In our study, we found alcohol drinking and traditional tobacco (chewing tobacco) as a significant risk with MSI associated GC.

This study has claimed that *EBV* infected GC patients are showing poor prognosis and multivariate analysis has confirmed the prognostic value of *EBV* infection, even after adjustments for other clinical factors. The prognostic assessment for *EBV* associated GC is very much controversial as previous studies reported that median survival times for *EBV* associated GC (8.5 years) is higher, compared to *EBV*-negative tumor (5.3 years) [45] and five year OS in *EBV* associated GC (71.4%) is higher, compared to negative group (56.1%) [46]. The prognostic assessment for *EBV* associated GC is regionally and ethically restricted with their food and lifestyle habits. Moreover, there is a prevalence of *EBV* infected cases (45.07%) in this cohort compared to worldwide status (10%) [47], while *H. pylori* infection does not exhibit any significant change in survival rate associated with GC. In a study performed in china, a trend towards a higher survival rate in patients with high-copies *H. pylori* infection was observed compared to patients with low-copy infection [48]. MMR deficient GC patients exhibited good prognosis, while MMR proficient GC cases were considered as a high-risk group and more aggressive. The result is consistent with several studies reporting that MSI shows a better prognosis than MSS in gastric cancer [49, 50, 51, 52]. Our prognostic assessments were comparable with TCGA data for the *H. pylori* and MSI patient groups (Figure 4G and 4H). Our study also supports the fact that *H. pylori* infection is not a prognostic factor for GC patients for this population.

The prospective of this study is the panel of epidemiological risk factors which can predict GC at early stage, which is very necessary for making clinical decision on patient treatment. The prognostic assessment of this study will help clinicians to opt for the right therapy. Other strength of this study is the consistency in result obtained for the positive association between excess body weight (BMI), extra salt intake, smoked food and alcohol consumptions, smoking and gastric cancer across high-quality studies with different patient populations. The limitation of this study is the smaller sample size and further studies with large cohort would be beneficial to support our data.

Conclusion

This study reported significant etiological factors associated with GC development through multidisciplinary approaches. The study has augmented the literature on unique lifestyle risk factors associated with *EBV* infected patients and has identified a panel of five epidemiological risk factors to predict GC in early stages, which is necessary for better diagnosis and treatment of patients. This study gave an assessment on the survival of GC patients associated with pathogen and MMR gene status.

In conclusion, most of the cases are reported at an advanced stage which decreases the scope of treatment and resulting in poor survival rate. The risk score for 5-panel epidemiological factors, from the present study, could be used for predicting gastric cancer risk in the pre-malignant stage with an early symptom. Smoked food emerged as a major exposure for GC development and we can conclude that *EBV* infection is also the strong risk factor for gastric cancer mortality. In clinical practices, patients with curatively resected gastric cancer who are positive for *EBV* may need more careful follow-up and more aggressive antitumor treatment to prolong life expectancy. Further research is required to elucidate the exact mechanisms of inflammation and tumor suppression with or without pathogen infection, which might provide new opportunities for personalized treatment options using this risk score and clinicopathological factors.

Abbreviation

GC: gastric cancer; *EBV*: *Epstein bar virus*; *H.pylori*: *Helicobacter pylori* MSI: Microsatellite Instability; MSS: Microsatellite stability; MMR: Mismatch repair; pTNM: Pathological. Tumour. Node. Metastasis; PAH: Polycyclic Aromatic Hydrocarbons; BaP: Benzo amino pyrene; TCGA-STAD: The Cancer Genome Atlas-Stomach Adenocarcinoma; ACRG: Asian Cancer Research Group; IARC: International Agency for Research on Cancer; AhR pathway: Aryl hydrocarbon receptor pathway; ERK pathway: Extracellular-signal-regulated kinase pathway; EMT: Epithelial-mesenchymal transition. HR: Hazard Ratio; CI: Confidence Interval.

Declarations

Ethics approval and consent to participate

The work was approved by ethical committees of Civil Hospital, Aizawl (B.12018/1/13-CH(A)/IEC dtd. 18/04/2014), and Human Ethical Committee, Mizoram University (MZU/IHEC/2015/008 dtd. 14/12/15).

Consent for participate

All participants signed the written informed consent.

Consent for publication

All participants signed the written informed consent.

Availability of data and material

The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Competing Interest

The authors declare that they have no competing interests.

Funding

This project was supported by research funds from the Department of Biotechnology (DBT), New Delhi sponsored projects on Gastric Cancer (BT/551/NE/U-Excel/2014) and (DBT-NER/health/46/2015) to Mizoram University

Author’s contribution

AM, NSK, SC, JLP conceived the basic concept and study design. SC, LP, JZ, BZ, KL, JLP, CL examined the patients and analyzed the pathological and clinical data. PC, SG, AM, and NSK performed the data analysis and interpreted the results. PC, SG, CL, AM, and NSK wrote the manuscript. All the authors reviewed the final draft of the manuscript.

Acknowledgments

The authors are thankful to data collectors and Lab Technicians, Ms. Nubawihi khenglawt, Mr. Jonathan Lalramhluna, Ms. Baby Lalrintluangi for collecting the clinical information and samples.

Reference

1. Crew KD, Neugut AI. Epidemiology of gastric cancer. World J Gastroenterol. 2006;12:354-62.
2. Lauren P. The two histological main types of Gastric cancer: Diffuse and so-called Intestinal–type carcinoma. An attempt at a histo-clinical classification. Acta Pathol Microbiol Scand. 1965;64:31-49.
3. Rawla P, Barsouk A. Epidemiology of gastric cancer: global trends, risk factors and prevention. Gastroenterology Rev. 2019;14:26–38.

4. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2018;68:394-424.

5. Ibrahim M, Gilbert K. Management of gastric cancer in Indian population. Transl gastroenterol hepatol. 2017;2:1-8.

6. Phukan AC, Borah PK, Mahanta J. Japanese encephalitis in Assam, northeast India. Southeast Asian J Trop Med Public Health. 2004;35:618-22.

7. De Mandal S, Singh SS, Muthukumaran RB, Thanzami K Kumar V, Kumar NS. Metagenomic analysis and the functional profiles of traditional fermented pork fat 'sa-um' of Northeast India. AMB Express. 2018;8:163.

8. Phukan RK, Narain K, Zomawia E, Hazarika NC, Mahanta J. Dietary habits and stomach cancer in Mizoram. Indian J Gastroenterol. 2006;41:418-24.

9. Madathil S, Senthil Kumar N, Zodinpuii D, Muthukumaran RB, Lalmuanpuii R, Nicolau B. Tuibur: tobacco in a bottle-commercial production of tobacco smoke-saturated aqueous concentrate. Addiction. 2018;113:577-80.

10. Moss SF. The Clinical Evidence Linking Helicobacter pylori to Gastric Cancer. Cellular and molecular gastroenterology and hepatology. 2017;3:183-91.

11. Shannon-Lowe C, Rickinson A. The Global Landscape of EBV-Associated Tumors. Front Oncol. 2019;9:713.

12. Zhu L, Li Z, Wang Y, Zhang C, Liu Y, Qu X. Microsatellite instability and survival in gastric cancer: A systematic review and meta-analysis. Mol Clin Oncol. 2015;3(3):699-705.

13. Sitarz R, Skierucha M, Mielko J, Offerhaus GJA, Maciejewski R, Polkowski WP. Gastric cancer: epidemiology, prevention, classification, and treatment. Cancer Manag Res. 2018;10:239-48.

14. Liu X, Fan J, Liaw K-L, Xu M, Zhou Y, Amonkar M, et al. Abstract 616: Literature review and meta-analyses of the prevalence of microsatellite instability high (MSI-H) and deficient mismatch repair (dMMR) for colorectal (CRC), gastric (GC), endometrial (EC) and ovarian cancers (OC) in Chinese population. Cancer Res. 2019;79:616.

15. Diergaarde B, Braam H, Muijen GNPv, Ligtenberg MJL, Kok FJ, Kampman E. Dietary factors and microsatellite instability in sporadic colon carcinomas. Cancer Epidemiol Biomarkers Prev. 2003;12:1130-6.

16. Eaton AM, Sandler R, Carethers JM, Millikan RC, Galanko J, Keku TO. 5,10-methylenetetrahydrofolate reductase 677 and 1298 polymorphisms, folate intake, and microsatellite instability in colon cancer. Cancer Epidemiol Biomarkers Prev. 2005;14:2023-9.

17. Poynter JN, Haile RW, Siegmund KD, Campbell PT, Figueiredo JC, Limburg P, et al. Associations between smoking, alcohol consumption, and colorectal cancer, overall and by tumor microsatellite instability status. Cancer Epidemiol Biomarkers Prev. 2009;18:2745-50.

18. Warneke VS, Behrens H-M, Haag J, Balschun K, Böger C, Becker T, et al. Prognostic and Putative Predictive Biomarkers of Gastric Cancer for Personalized Medicine. Diagn Mol Pathol. 2003;22:127-37.

19. Ghatak S, Yadav RP, Lalrohlui F, Chakraborty P, Ghosh S, Ghosh S, et al. Xenobiotic Pathway Gene Polymorphisms Associated with Gastric Cancer in High Risk Mizo-Mongoloid Population, Northeast India. Helicobacter. 2016;21:523-35.

20. Slattery ML, Anderson K, Curtin K, Ma KN, Schaffer D, Samowitz W. Dietary intake and microsatellite instability in colon tumors. Int J Cancer. 2001;93:601-7.

21. Gunathilake MN, Lee J, Jang A, Choi IJ, Kim Y-I, Kim J. Physical Activity and Gastric Cancer Risk in Patients with and without Helicobacter pylori Infection in A Korean Population: A Hospital-Based Case-Control Study. Cancers. 2018;10:369.

22. Fox JG, Dangler CA, Taylor NS, King A, Koh TJ, Wang TC. High-Salt Diet Induces Gastric Epithelial Hyperplasia and Parietal Cell Loss, and Enhances <em>Helicobacter pylori</em> Colonization in C57BL/6 Mice. Cancer Res. 1999;59:4823-8.

23. Kato S, Tsukamoto T, Mizoshita T, Tanaka H, Kumagai T, Ota H, et al. High salt diets dose-dependently promote gastric chemical carcinogenesis in Helicobacter pylori-infected Mongolian gerbils associated with a shift in mucin production from glandular to surface mucous cells. Int J Cancer. 2006;119(7):1558-66.

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

25. Wang XQ, Terry PD, Yan H. Review of salt consumption and stomach cancer risk: epidemiological and biological evidence. World J Gastroenterol. 2009;15:2204-13.

26. Tatematsu M, Takahashi M, Fukushima S, Hananouchi M, Shirai T. Effects in Rats of Sodium Chloride on Experimental Gastric Cancers Induced by N-Methyl-N′-nitro-N-nitrosoguanidine or 4-Nitroquinoline-1-oxide2. J Natl Cancer Inst. 1975;55(1):101-6..

27. McDonald ST. Comparison of Health Risks of Smoked Foods as Compared to Smoke Flavorings: Are Smoke Flavors “Healthier”? ADVANCES IN FOOD TECHNOLOGY AND NUTRITIONAL SCIENCES Open Journal, 2015;1:5.

28. Varlet V, Knockaert C, Prost C, Serot T. Comparison of Odor-Active Volatile Compounds of Fresh and Smoked Salmon. Journal of Agricultural and Food Chemistry. 2006;54(9):3391-401.

29. Rubin, H. Synergistic mechanisms in carcinogenesis by polycyclic aromatic hydrocarbons and by tobacco smoke: a bio-historical perspective with updates. Carcinogenesis. 2001;22:1903-30.

30. Bersten DC, Sullivan AE, Peet DJ, Whitelaw ML. bHLH–PAS proteins in cancer. Nat Rev Cancer. 2013;13(12):827-41.

31. Wei Y, Zhao L, He W, Yang J, Geng C, Chen Y, et al. Benzo[a]pyrene promotes gastric cancer cell proliferation and metastasis likely through the Aryl hydrocarbon receptor and ERK-dependent induction of MMP9 and c-myc. Int J Cancer. 2016;49:2055-63.

32. Phukan RK. Tobacco use and stomach cancer in Mizoram, India. Cancer Epidemiol Biomarkers Prev. 2005;14:1892-6.

33. Bonequi P, Meneses-González F, Correa P, Rabkin CS, Camargo MC. Risk factors for gastric cancer in Latin America: a meta-analysis. Cancer Causes Control. 2013;24(2):217-31.

34. Li W-Y, Han Y, Xu H-M, Wang Z-N, Xu Y-Y, Song Y-X, et al. Smoking status and subsequent gastric cancer risk in men compared with women: a meta-analysis of prospective observational studies. BMC Cancer. 2019;19:377.

35. Nomura AMY, Wilkens LR, Henderson BE, Epplein M, Kolonel LN. The association of cigarette smoking with gastric cancer: the multiethnic cohort study. Cancer Causes Control. 2012;23:51-8.

36. Alcohol consumption and ethyl carbamate. IARC monographs on the evaluation of carcinogenic risks to humans. 2010;96:3-1383.

37. Ghosh S, Bankura B, Ghosh S, Saha ML, Pattanayak AK, Ghatak S, Guha M, Nachimuthu SK, Panda CK, Maji S, Chakraborty S, Maity B, Das M. Polymorphisms in ADH1B and ALDH2 genes associated with the increased risk of gastric cancer in West Bengal, India. BMC Cancer. 2017;17(1):782.

38. Moy KA, Fan Y, Wang R, Gao YT, Yu MC, Yuan JM. Alcohol and tobacco use in relation to gastric cancer: a prospective study of men in Shanghai, China. Cancer Epidemiol Biomarkers Prev. 2010;19:2287-97.

39. Sung NY, Choi KS, Park EC, Park K, Lee SY, Lee AK, et al. Smoking, alcohol and gastric cancer risk in Korean men: the National Health Insurance Corporation Study. Br J Cancer. 2007;97(5):700-4.

40. Kamper-Jorgensen M, Rostgaard K, Glaser SL, Zahm SH, Cozen W, Smedby KE, et al. Cigarette smoking and risk of Hodgkin lymphoma and its subtypes: a pooled analysis from the International Lymphoma Epidemiology Consortium (InterLymph). Ann Oncol. 2013;24:2245-55.

41. Jia W-H, Qin H-D. Non-viral environmental risk factors for nasopharyngeal carcinoma: A systematic review. Seminars in Cancer Biology. 2012;22:117-26.

42. Malovichko MV, Zeller I, Krivokhizhina TV, Xie Z, Lorkiewicz P, Agarwal A, et al. Systemic Toxicity of Smokeless Tobacco Products in Mice. Nicotine Tob Res. 2019;21:101-10.

43. Hatton OL, Harris-Arnold A, Schaffert S, Krams SM, Martinez OM. The interplay between Epstein-Barr virus and B lymphocytes: implications for infection, immunity, and disease. Immunol Res. 2014;58:268-76.

44. Jenson HB, Baillargeon J, Heard P, Moyer MP. Effects of Smokeless Tobacco and Tumor Promoters on Cell Population Growth and Apoptosis of B Lymphocytes Infected with Epstein–Barr Virus Types 1 and 2. Toxicol Appl Pharmacol. 1999;160:171-82.

45. Camargo MC, Kim WH, Chiaravalli AM, Kim KM, Corvalan AH, Matsuo K, et al. Improved survival of gastric cancer with tumour Epstein-Barr virus positivity: an international pooled analysis. Gut. 2014;63:236-43.

46. Song HJ, Kim KM. Pathology of epstein-barr virus-associated gastric carcinoma and its relationship to prognosis. Gut liver. 2011;5:143-8.

47. Nishikawa J, Yoshiyama H, Iizasa H, et al. Epstein-barr virus in gastric carcinoma. Cancers (Basel). 2011;6:2259-2274.

48. Qiu H-B, Zhang L-Y, Keshari R-P, Wang G-Q, Zhou Z-W, Xu D-Z, et al. Relationship between H.Pylori infection and clinicopathological features and prognosis of gastric cancer. BMC Cancer. 2010;10:374.

49. Beghelli S, de Manzoni G, Barbi S, Tomezzoli A, Roviello F, Di Gregorio C, et al. Microsatellite instability in gastric cancer is associated with better prognosis in only stage II cancers. Surgery. 2006;139:347-56.

50. Kim SM, An JY, Byeon Sj, Lee J, Kim KM, Choi MG, et al. Prognostic value of mismatch repair deficiency in patients with advanced gastric cancer, treated by surgery and adjuvant 5-fluorouracil and leucovorin chemoradiotherapy. Eur J Surg Oncol. 2020;46:189-94.

51. Choi YY, Bae JM, An JY, Kwon IG, Cho I, Shin, HB, et al. Is microsatellite instability a prognostic marker in gastric cancer? A systematic review with meta-analysis. J Surg Oncol. 2014;110:129-35.

52. Smyth EC, Wotherspoon A, Peckitt C, Gonzalez D, Hulkki-Wilson S, Eltahir Z, et al. Mismatch repair deficiency, microsatellite instability, and survival: an exploratory analysis of the medical research council adjuvant gastric infusional chemotherapy (MAGIC) trial. JAMA Oncol. 2017;3:1197-203.

Factors	**ODDS ratio (95% CI)**	***p* value**
	Univariate analysis	
Sex	0.50 (0.28 – 0.89)	0.019
Age	1.01 (0.99 – 1.04)	0.07
BMI	0.63 (0.56 – 0.72)	0.0001
Extra Salt	0.59 (0.41 – 0.86)	0.007
Sa-um	0.75 (0.50 – 1.13)	0.180
Smoked Food	0.49 (0.34 – 0.70)	0.0001
Tuibur	1.48 (1.09 – 2.00)	0.011
Alcohol drinking	3.11 (1.96 – 4.92)	0.0001
Smoking	7.50 (4.03 – 13.94)	0.0001
Paan with betel nut	0.99 (0.56 – 1.76)	0.984
Multivariate analysis (logistic model)		
Sex	0.58 (0.24 – 1.40)	0.230
BMI	0.69 (0.60 – 0.79)	0.0001
Extra Salt	0.68 (0.41 – 1.14)	0.042
Smoked Food	0.64 (0.40 – 1.04)	0.001
Tuibur	1.30 (0.80 – 2.12)	0.285
Alcohol drinking	1.83 (1.03 – 3.26)	0.001
Smoking	4.41 (1.86 – 10.43)	0.0007

Table 1: Univariate and multivariate analysis of the risk factors compared between Gastric Cancer patients (n = 73) and Healthy Controls (n = 153).

Factors	***H. pylori* (+) cases (n = 50)**	***H. pylori* (-) cases (n = 30)**	***EBV* (+) cases (n = 32)**	***EBV* (-) cases (n = 48)**	**MMR gene deficient (n = 32)**	**MMR gene proficient (n = 48)**
Anatomy						
Proximal	8 (16%)	3 (10%)	4 (12.5%)	7 (14.58%)	3 (9.37%)	8 (16.66%)
Distal	35 (70%)	24 (80%)	21 (65.62%)	38 (79.16%)	28 (87.5%)	31 (64.58%)
Data Not available	7 (14%)	3 (10%)	7 (21.87%)	3 (6.25%)	1 (3.12%)	9 (18.75%)
TNM Stage						
I	11 (22%)	9 (30%)	8 (25%)	12(25%)	9(28.12%)	11(22.91%)
II	9 (18%)	5 (16.66%)	5 (15.62%)	9 (18.75%)	5 (15.62%)	9 (18.75%)
III	24 (48%)	16 (53.33%)	17 (53.12%)	23 (47.91%)	17 53.12%)	23 (47.91%)
IV	2 (4%)	0	1 (3.12%)	1 (2%)	0	2(4.16%)
Data Not available	4 (8%)	0	1 (3.12%)	3(6.25%)	1(3.12%)	3(6.25%)
Grade						
[a]WD	4 (8%)	3 (10%)	2 (6.25%)	5(10.41%)	2(6.25%)	5(10.41%)
[b]MD	23 (46%)	15 (50%)	12 (37.5%)	26 (54.16%)	12 (37.5%)	26 (54.16%)
[c]PD	20 (40%)	11 (36.66%)	16 (50%)	15 (31.25%)	17 (53.12%)	14 (29.16%)
Data Not available	3 (6%)	1 (3.3%)	2 (6.25%)	2(4.16%)	1(3.12%)	3(6.25%)
Family history of Cancer						
Yes	13 (26%)	14 (46.66%)	12 (37.5%)	15 (31.25%)	13 (40.62%)	14 (29.16%)
No	37 (74%)	16 (53.33%)	20 (62.5%)	33 (68.75%)	19 (59.37%)	34 (70.83%)

Table 2: Distribution of clinical factors among the various sub-groups in the gastric cancer patients' cohort (n = 80). [a]WD - Well Differentiated, [b]MD - Moderately Differentiated, [c]PD - Poorly Differentiated.

Factors	*H. pylori* (+) cases (n = 50)	*H. pylori* (-) cases (n = 30)	EBV (+) cases (n = 32)	EBV (-) cases (n = 48)	MMR gene deficient (n = 32)	MMR gene proficient (n =48)
Age (mean)	59.5 ± 12.37	59.5 ± 9.76	59.5 ± 9.94	59.5 ± 12.36	56.5 ± 12.31	60 ± 10.60
Sex						
Male	34 (68%)	19 (63.33%)	20 (62.5%)	33 (68.75%)	12 (37.5%)	31 (64.58%)
Female	16 (32%)	11 (36.66%)	12 (37.5%)	15 (31.25%)	20 (62.5%)	17 (35.41%)
Extra salt						
Consumers	36 (72%)	20 (66.66%)	20 (62.5%)	36 (75%)	22 (68.74%)	34 (70.83%)
Non-consumers	14 (28%)	10 (33.33%)	12 (37.5%)	12 (25%)	10 (31.25%)	14 (29.16%)
ORs (95% CI), *p* value	1.32 (0.49 – 3.51); 0.57		0.55 (0.21 – 1.46); 0.23		0.90 (0.34 – 2-39); 0.84	
Sa-um						
Consumers	42 (84%)	24 (80%)	25 (78.12%)	41 (85.41%)	29 (90.62%)	37 (77.08%)
Non- consumers	8 (16%)	6 (20%)	7 (21.87%)	7 (14.58%)	3 (9.37%)	11 (22.91%)
ORs (95% CI), *p* value	1.31 (0.40 – 4.23); 0.64		0.60 (0.19 – 1.94); 0.40		2.87 (0.73 – 11.26); 0.12	
Smoked food						
Consumers	26 (52%)	25 (83.33%)	27 (84.37%)	24 (50%)	22 (68.74%)	29 (60.41%)
Non-consumers	24 (48%)	5 (16.66%)	5 (15.62%)	24 (50%)	10 (31.25%)	19 (39.58%)
ORs (95% CI), *p* value	**0.21 (0.07 – 0.65); 0.006**		**5.40 (1.78 – 16.37); 0.002**		1.44 (0.56 – 3.70); 0.44	
Paan with betel nut						
Consumers	30 (60%)	20 (66.66%)	21 (65.62%)	29 (60.41%)	23 (71.87%)	27 (56.25%)
Non-consumers	20 (40%)	10 (33.33%)	11 (34.37%)	19 (39.58%)	9 (28.12%)	21 (43.75%)
ORs (95% CI), *p* value	0.75 (0.29 – 1.93); 0.55		1.25 (0.49 – 3.17); 0.63		1.98 (0.76 – 5.18); 0.16	
Chewed tobacco						
Consumers	26 (52%)	15 (50%)	15 (46.87%)	26 (54.16%)	12 (37.5%)	29 (60.41%)
Non- consumers	24 (48%)	15 (50%)	17 (53.12%)	22 (52.08%)	20 (62.5%)	19 (39.58%)
ORs (95% CI), *p* value	1.08 (0.43 – 2.67); 0.86		0.74 (0.30 – 1.83); 0.52		**0.39 (0.15 – 0.98); 0.04**	
Tuibur						
Consumers	13 (26%)	8 (26.66%)	12 (37.5%)	9 (18.75%)	8 (25%)	13 (27.08%)
Non- consumers	37 (74%)	22 (73.33%)	20 (62.5%)	39 (81.25%)	24 (75%)	35 (72.91%)
ORs (95% CI), *p* value	0.96 (0.34 – 2.69); 0.94		2.60 (0.93 – 7.20); 0.06		0.89 (0.32 – 2.49); 0.83	
Smoking						
Smokers	35 (70%)	17 (56.66%)	19 (59.37%)	33 (68.75%)	21 (65.62%)	31 (64.58%)
Non-smokers	15 (30%)	13 (43.33%)	13 (40.62%)	15 (31.25%)	11 (34.37%)	17 (35.41%)
ORs (95% CI), *p* value	1.78 (0.69 – 4.57); 0.22		0.66 (0.26 – 1.68); 0.39		1.04 (0.40 – 2.67); 0.92	
Alcohol drinking						
Drinkers	17 (43%)	12 (40%)	10 (31.25%)	19 (39.58%)	16 (50%)	13 (27.08%)
Non-drinkers	33 (66%)	18 (60%)	22 (68.75%)	29 (60.41)	16 (50%)	35 (72.91%)
ORs (95% CI), *p* value	0.77 (0.30 – 1.97); 0.58		0.69 (0.26 – 1.78); 0.44		**2.69 (1.05 – 6.89); 0.03**	

Table 3: Distribution of demographic factors and lifestyle habits among the various sub-groups in the gastric cancer patients' cohort (n = 80), ORs - ODDS Ratios.

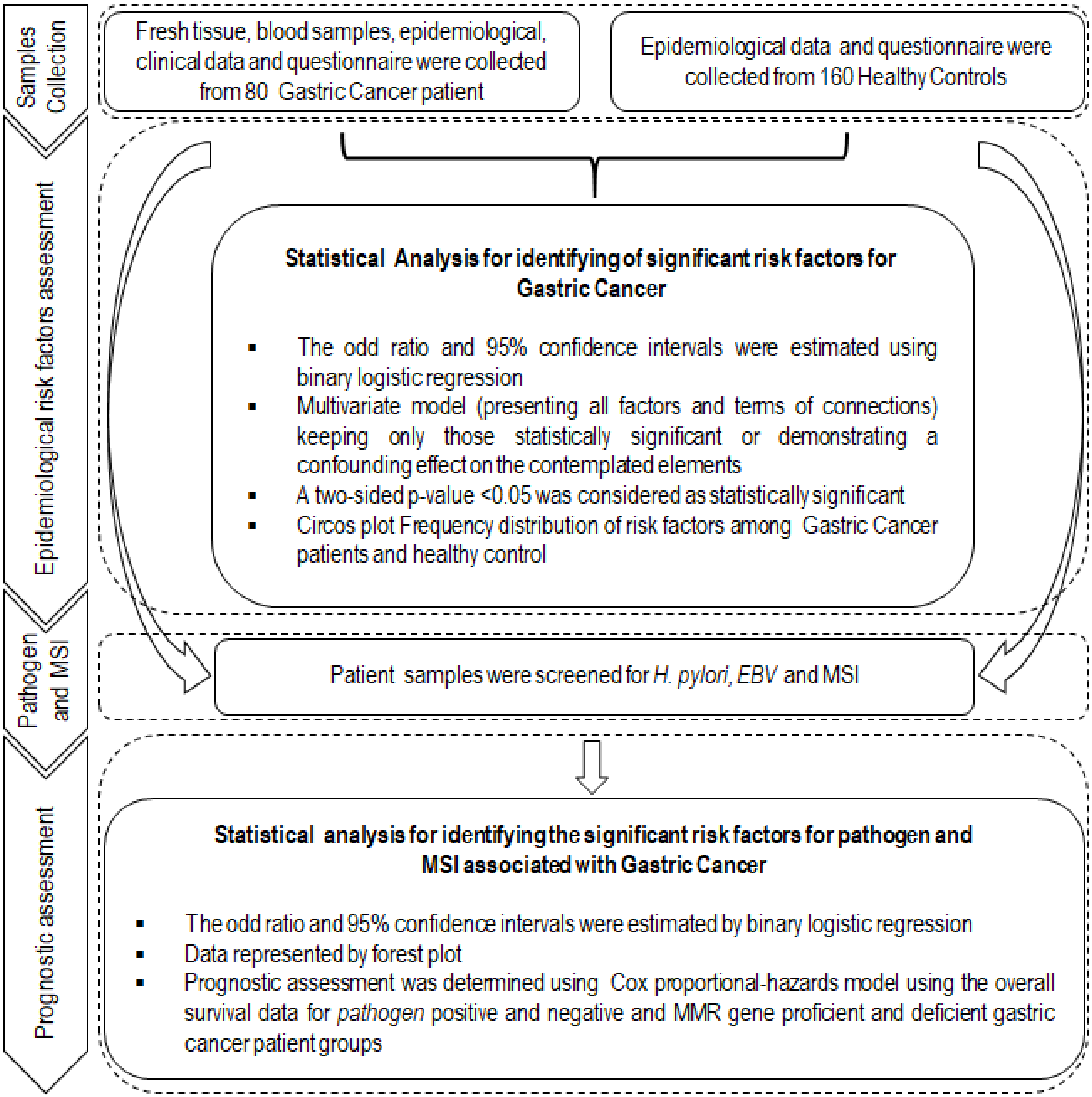

Samples Collection
Epidemiological risk factors assessment
Pathogen and MSI
Prognostic assessment
Fresh tissue, blood samples, epidemiological, clinical data and questionnaire were collected from 80 Gastric Cancer patient
Epidemiological data and questionnaire were collected from 160 Healthy Controls
Statistical Analysis for identifying of significant risk factors for Gastric Cancer
▪ The odd ratio and 95% confidence intervals were estimated using binary logistic regression
▪ Multivariate model (presenting all factors and terms of connections) keeping only those statistically significant or demonstrating a confounding effect on the contemplated elements
▪ A two-sided p-value <0.05 was considered as statistically significant
▪ Circos plot Frequency distribution of risk factors among Gastric Cancer patients and healthy control
Patient samples were screened for H. pylori, EBV and MSI
Statistical analysis for identifying the significant risk factors for pathogen and MSI associated with Gastric Cancer
▪ The odd ratio and 95% confidence intervals were estimated by binary logistic regression
▪ Data represented by forest plot
▪ Prognostic assessment was determined using Cox proportional-hazards model using the overall survival data for pathogen positive and negative and MMR gene proficient and deficient gastric cancer patient groups

Figure 1: Study design for the epidemiological risk factors and prognostic assessments for H. pylori, EBV and MMR gene status among gastric cancer patient group.

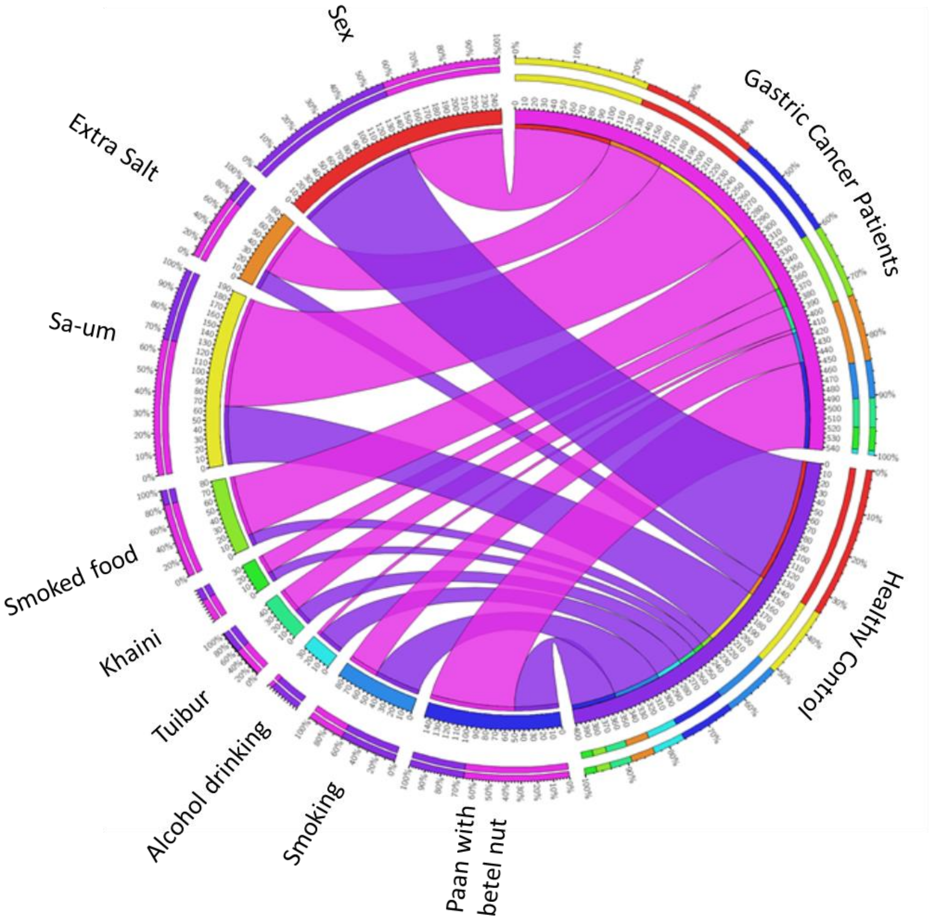


Figure 2: Frequency distributions of each demographic factors in the gastric cancer patients (pink ribbon) and healthy control (blue ribbon) groups in study cohort. The data were visualized via Circos software. The frequency of occurrence of different demographic factors association with gastric cancer and heathy control groups is depicted in the outer ring. The inner ring of circos plot depicts the subject number exposed with different demographic risk factors. Each factor has been assigned a specific color. The arc originates from gastric cancer and healthy control groups and terminates at different demographical factors to compare the association between the origin and terminating factors. The area of each colored ribbon depicts the frequency of the samples.

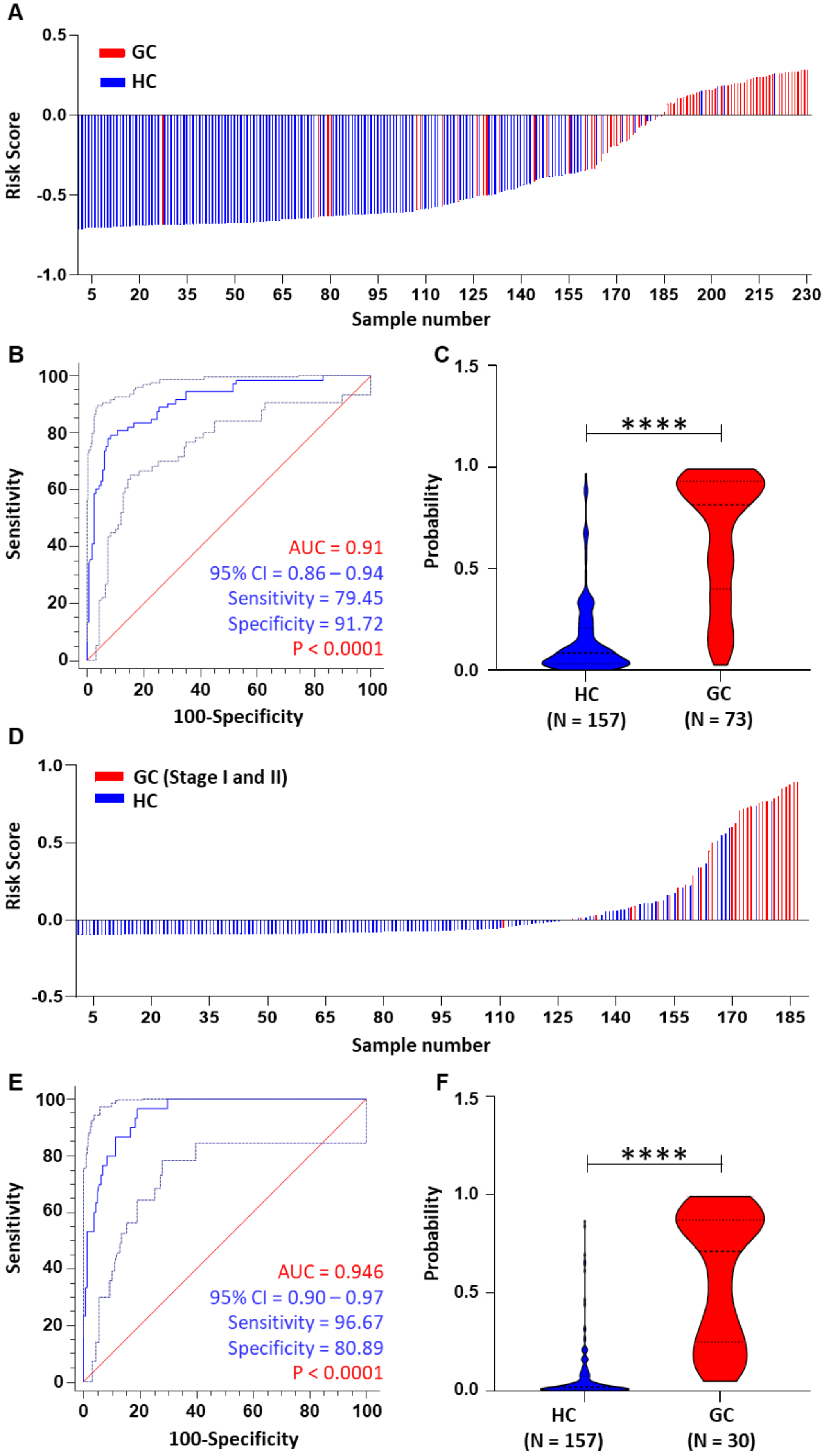

A
0.5
0.0
-0.5
-1.0
Risk Score
GC
HC
5
20
35
50
65
80
95
110
125
140
155
170
185
200
215
230
Sample number
B
100
80
60
40
20
0
Sensitivity
0
20
40
60
80
100
100-Specificity
AUC = 0.91
95% CI = 0.86 – 0.94
Sensitivity = 79.45
Specificity = 91.72
P < 0.0001
C
1.5
1.0
0.5
0.0
Probability
****
HC
(N = 157)
GC
(N = 73)
D
1.0
0.5
0.0
-0.5
Risk Score
GC (Stage I and II)
HC
5
20
35
50
65
80
95
110
125
140
155
170
185
Sample number
E
100
80
60
40
20
0
Sensitivity
0
20
40
60
80
100
100-Specificity
AUC = 0.946
95% CI = 0.90 – 0.97
Sensitivity = 96.67
Specificity = 80.89
P < 0.0001
F
1.5
1.0
0.5
0.0
Probability
****
HC
(N = 157)
GC
(N = 30)

Figure 3: Estimation of accuracy value of the significant epidemiological factors based on the logistic model between gastric cancer and healthy control samples (A) Water fall plot and risk score estimation for stage-I, II, III and IV samples, (B) Receiver operating curve (ROC) and accuracy estimation of epidemiological factors panel (BMI, extra salt consumptions, smoked food consumptions, alcohol drinking and smoking) (C) Significant association of the estimated probability values of the epidemiological factors panel between gastric cancer (n = 73) and healthy controls (n = 157), (D) Water fall plot and risk score estimation for stage-I and II samples, (E) Receiver operating curve (ROC) and accuracy estimation of epidemiological factors panel. (F) Significant association of the estimated probability values of the epidemiological factors panel between stage-I and II gastric cancer (n = 30) and healthy controls (n = 157).

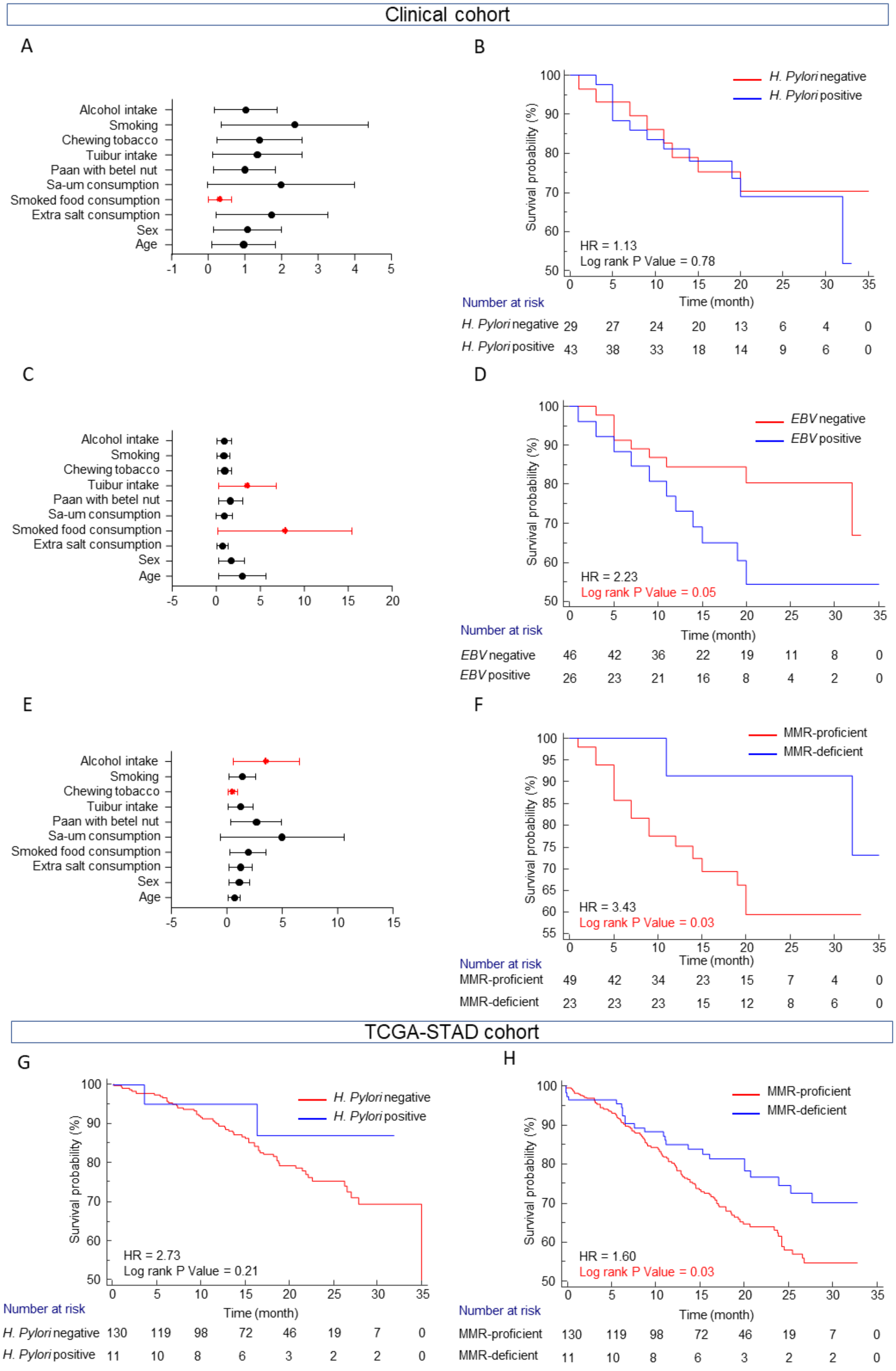

Clinical cohort
A
Alcohol intake
Smoking
Chewing tobacco
Tuibur intake
Paan with betel nut
Sa-um consumption
Smoked food consumption
Extra salt consumption
Sex
Age
-1 0 1 2 3 4 5
B
H. Pylori negative
H. Pylori positive
Survival probability (%)
100 90 80 70 60 50
HR = 1.13
Log rank P Value = 0.78
0 5 10 15 20 25 30 35
Time (month)
Number at risk
H. Pylori negative 29 27 24 20 13 6 4 0
H. Pylori positive 43 38 33 18 14 9 6 0
C
Alcohol intake
Smoking
Chewing tobacco
Tuibur intake
Paan with betel nut
Sa-um consumption
Smoked food consumption
Extra salt consumption
Sex
Age
-5 0 5 10 15 20
D
EBV negative
EBV positive
Survival probability (%)
100 90 80 70 60 50
HR = 2.23
Log rank P Value = 0.05
0 5 10 15 20 25 30 35
Time (month)
Number at risk
EBV negative 46 42 36 22 19 11 8 0
EBV positive 26 23 21 16 8 4 2 0
E
Alcohol intake
Smoking
Chewing tobacco
Tuibur intake
Paan with betel nut
Sa-um consumption
Smoked food consumption
Extra salt consumption
Sex
Age
-5 0 5 10 15
F
MMR-proficient
MMR-deficient
Survival probability (%)
100 95 90 85 80 75 70 65 60 55
HR = 3.43
Log rank P Value = 0.03
0 5 10 15 20 25 30 35
Time (month)
Number at risk
MMR-proficient 49 42 34 23 15 7 4 0
MMR-deficient 23 23 23 15 12 8 6 0
TCGA-STAD cohort
G
H. Pylori negative
H. Pylori positive
Survival probability (%)
100 90 80 70 60 50
HR = 2.73
Log rank P Value = 0.21
0 5 10 15 20 25 30 35
Time (month)
Number at risk
H. Pylori negative 130 119 98 72 46 19 7 0
H. Pylori positive 11 10 8 6 3 2 2 0
H
MMR-proficient
MMR-deficient
Survival probability (%)
100 90 80 70 60 50
HR = 1.60
Log rank P Value = 0.03
0 5 10 15 20 25 30 35
Time (month)
Number at risk
MMR-proficient 130 119 98 72 46 19 7 0
MMR-deficient 11 10 8 6 3 2 2 0

Figure 4: Association of overall survival probability and demographic factors with *H. pylori* infection, *EBV* infection and MMR gene status in gastric cancer patients. Odd ratios and 95% confidence interval of the demographic factors presented for the *H. pylori* (A), *EBV* (C) and MMR gene status (E). Association between overall survival and the *H. Pylori* (G) and MMR gene status (H) in TCGA-STAD cohort. *EBV* status could not be analyzed due to less sample size in TCGA-STAD dataset.

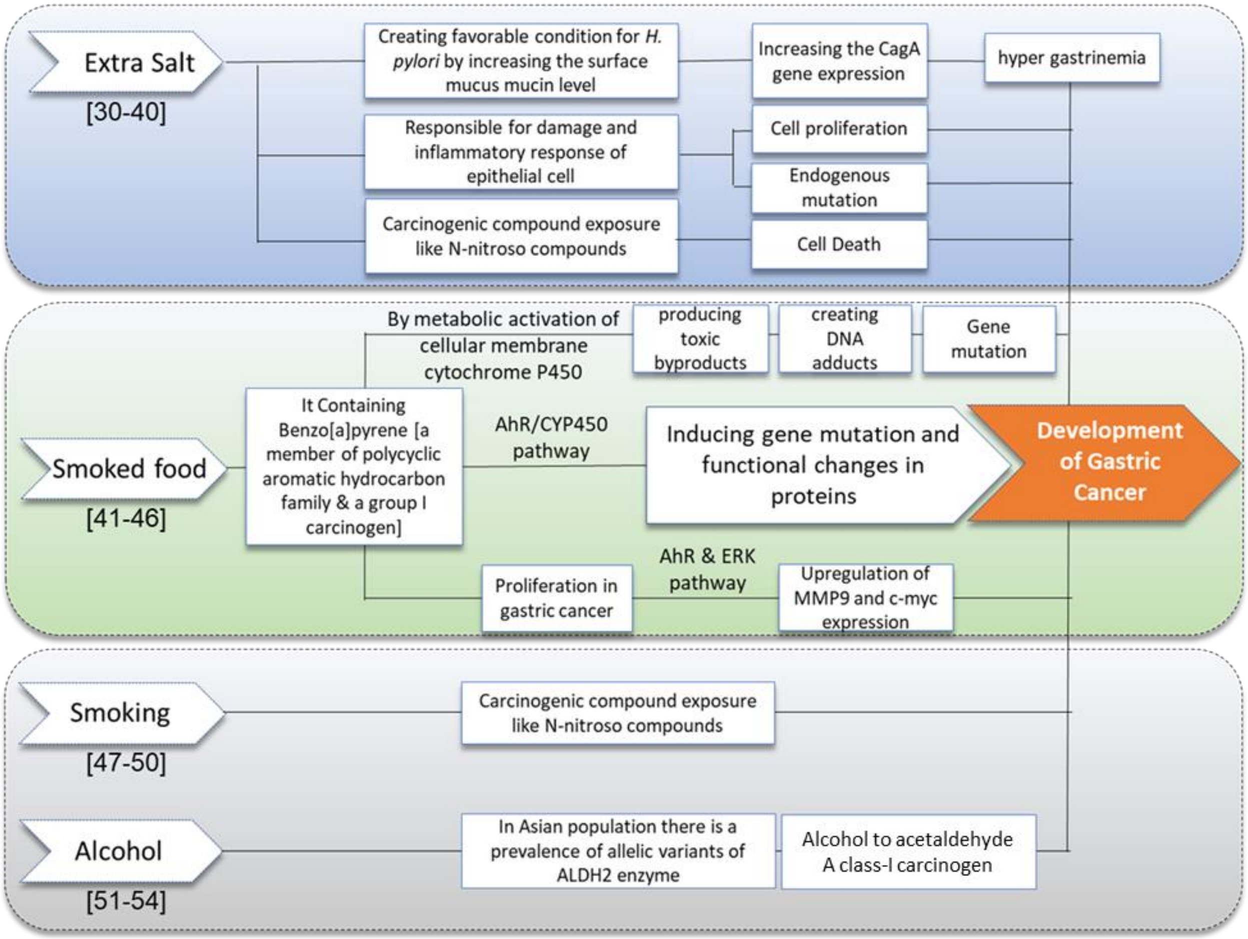

Extra Salt
[30-40]
Creating favorable condition for *H. pylori* by increasing the surface mucus mucin level
Increasing the CagA gene expression
hyper gastrinemia
Responsible for damage and inflammatory response of epithelial cell
Cell proliferation
Endogenous mutation
Carcinogenic compound exposure like N-nitroso compounds
Cell Death
By metabolic activation of cellular membrane cytochrome P450
producing toxic byproducts
creating DNA adducts
Gene mutation
Smoked food
[41-46]
It Containing Benzo[a]pyrene [a member of polycyclic aromatic hydrocarbon family & a group I carcinogen]
AhR/CYP450 pathway
Inducing gene mutation and functional changes in proteins
Development of Gastric Cancer
Proliferation in gastric cancer
AhR & ERK pathway
Upregulation of MMP9 and c-myc expression
Smoking
[47-50]
Carcinogenic compound exposure like N-nitroso compounds
Alcohol
[51-54]
In Asian population there is a prevalence of allelic variants of ALDH2 enzyme
Alcohol to acetaldehyde A class-I carcinogen

Figure 5: Flow chart depicting the Major risk factors in the present study and their mechanism of Gastric Carcinogenesis from Literature review. The numerical in parenthesis [ ] represents the bibliographic information.

Revised Supplementary Tables

Questionnaire

Ghatak *et al. BMC Medical Genetics* (2017) 18:61
DOI 10.1186/s12881-017-0427-2

BMC Medical Genetics

RESEARCH ARTICLE Open Access

Novel APC gene mutations associated with protein alteration in diffuse type gastric cancer

Souvik Ghatak[1], Payel Chakraborty[1], Sandeep Roy Sarkar[2], Biswajit Chowdhury[2], Arup Bhaumik[2] and Nachimuthu Senthil Kumar[1*]

Abstract

Background: The role of adenomatous polyposis coli (APC) gene in mitosis might be critical for regulation of genomic stability and chromosome segregation. APC gene mutations have been associated to have a role in colon cancer and since gastric and colon tumors share some common genetic lesions, it is relevant to investigate the role of APC tumor suppressor gene in gastric cancer.

Methods: We investigated for somatic mutations in the Exons 14 and 15 of APC gene from 40 diffuse type gastric cancersamples. Rabbit polyclonal anti-APC antibody was used, which detects the wild-type APC protein and was recommended for detection of the respective protein in human tissues. Cell cycle analysis was done from tumor and adjacent normal tissue.

Results: APC immunoreactivity showed positive expression of the protein in stages I, II, III and negative expression in Stages III and IV. Two novel deleterious variations (g.127576C > A, g.127583C > T) in exon 14 sequence were found to generate stop codon (Y622* and Q625*)in the tumor samples. Due to the generation of stop codon, the APC protein might be truncated and all the regulatory features could be lost which has led to the down-regulation of protein expression. Our results indicate that aneuploidy might occurdue to the codon 622 and 625 APC-driven gastric tumorigenesis, in agreement with our cell cycle analysis. The APC gene function in mitosis and chromosomal stability might be lost and G1 might be arrested with high quantity of DNA in the S phase. Six missense somatic mutations in tumor samples were detected in exon 15 A-B, twoof which showed pathological and disease causing effects based on SIFT, Polyphen2 and SNPs & GO score and were not previously reported in the literature or the public mutation databases.

Conclusion: The two novel pathological somatic mutations (g.127576C > A, g.127583C > T) in exon 14 might be altering the protein expression leading to development of gastric cancer in the study population. Our study showed that mutations in the APC gene alter the protein expression and cell cycle regulation in diffuse type gastric adenocarcinoma.

Keywords: Adenomatous polyposis coli (APC), Gastric adenocarcinomas, Immunoreactivity, Mutation, Cell cycle

* Correspondence: nskmzu@gmail.com
[1]Department of Biotechnology, Mizoram University, Aizawl 796004, Mizoram, India
Full list of author information is available at the end of the article

Background

Gastric cancer is one of the most common cancer worldwide and there are more than 100 new cases per year in Tripura, Northeast India with a 5-year survival rate < 10% [1]. A number of genetic abnormalities have been identified in gastric cancer, including mutations in tumor suppressor gene [2].However, the abnormalities individually exhibit frequencies of less than 50% in gastric tumors, and are variable depending on the population and number of the samples analysed.

The human APC (adenomatous polyposis coli) gene is a tumor suppressor gene located on the long (q) arm of chromosome 5 and it encodes a protein of 312 kDa with 2843 amino acids. Inactivation of the APC geneis thought to be an initiating event for carcinogenesis [3]. Germline mutations of the APC gene are responsible for familial adenomatous polyposis (FAP) [4, 5]. About 700 mutations in the APC gene have been identified and most of these mutations lead to the production of short and abnormalprotein which cannot suppress the cellular overgrowth,leading to the formation of polyps and become cancerous [6]. The APC gene inhibits the members of Wnt signalling pathway that promotes β-catenin expression as a stimulator of cell division within the intestinal crypts [7]. A functioning APC protein is thus vital in maintaining low levels of cytosolic β-catenin, thereby preventing excessive cell proliferation [8]. APC controls metaphase-anaphase transition and mitotic exit and regulates G1 phase [9, 10]. Over-expression of APC in fibroblasts and colon cancer cell lines leads to arrest of G1 phase in the cell cycle [11, 12]. Role of APC in mitosis is critical for regulation of genomic stability and chromosome segregation [13]. Somatic mutations in the APC gene have been described in several tumour types such as pancreatic cancer [14], oral squamous-cell carcinoma [15] and oesophageal cancer [16]. APC mutations have been reported in gastric adenomas [17, 18] and in differentiated and signet-ring cell carcinomas [19]. Furthermore, frequent loss of heterozygosity on chromosome 5q has been detected in gastric carcinomas, particularly in well-differentiated type [20].Moreover, some differentiated types of gastric carcinoma are thought to originate from the intestinal metaplastic regions in gastric mucosa [21].

Since gastric and colon tumors share some common genetic lesions [22], it is relevant to investigate the role of APC tumor suppressor gene in the case of diffuse type gastric cancerwhich is not well characterized. Exons 14 and 15 are the most frequently mutated region for colorectal and gastric cancer as well as patients with familial adenomatous polyposis [23]. To clarify the role of APC gene mutations in the development of diffuse type gastric adenocarcinoma, we have investigated the mutations in the exons 14 and 15 of APC gene in a North East Indian population.

Methods

Subjects

The study design and data collection methods have been described in detail previously [24]. For this study, a total of 62 gastric cancer (GC) patients with or without a family history of cancer (median age 58 years; range 37–79) who received treatment between September 2012 and February 2014 at Agartala Govt. Medical College, Tripura,Northeast India and 40 healthy volunteers (median age 52 years; range 31–73) were recruited. Individuals less than age 45 were classified as younger and those age 45 and older were classified as older. From the 62 samples, 40 diffuse type gastric tumor samples were selected and the patients with gastric neoplasms other than adenocarcinoma (MALT lymphoma, stromal or carcinoid tumors), secondary or recurrent GC, previous history of other malignancies or refusal to participate were excluded. The healthy control samples wereage and sex adjusted, and selected from same ethnic group, free of any other chronic diseases, not having any record of gastritis and not pre-treated for any other type of cancer. The tumour and adjacent normal tissue of the patients were grossed properly by a trained histopathology technician followed by preparation of paraffin block. Histologic assessment of tumor type and grade were performed routinely on 4 to 5 μm thick hematoxyline & eosin stained sections of formalin-fixed paraffinembedded tumors according to the criteria outlined in the World Health Organization Classification of Tumors. After staining, the cytopathological data was obtained from microscopic observationsto confirm that all the adjacent normal tissues were devoid of tumor cells.The blood samples were collected by an experienced laboratory technician using vacu-puncture procedure. The peripheral blood samples of the patients were kept in EDTA rinsed microcentrifuge tubes and 50 μl of blood samples were processed for DNA isolation. Medical charts were reviewed to obtain information on cancer treatment, clinical stage, previous disease history and weight history. All participants gave written informed consent to the study protocol which was approved by the Ethical Committee of the Civil Hospital, Mizoram and Mizoram University, India (B.12018/1/13-CH(A)/IEC).The study protocol was also approved by the Institutional Review Board of all institutes involved in the study.

Immunohistochemical analysis

For the immunohistochemical study, 4-μm histological fragments were obtainedfrom the tumor tissue and adjacent normal tissue of the cases and placed on glass slides pre-treated with poly L-lysine (Sigma Chemical Co, MO, USA). Initially, histological slides were placed in an oven at 60 °C for 24 hours to obtain better tissue adhesion and deparaffinization. Deparaffinization was performed in three xylene baths at room temperature for 15 min and placed in three baths of absolute ethanol baths for 1 min each. The slides were washed in running water for 5 min and submitted to heat induced antigen recovery by steam in a 10 mM citrate buffer solution with pH 6.0 for 30 min. After cooling for 20 min at room temperature, the slides were washed in running water for 5 min and endogenous peroxidase blocking was performed using a hydrogen peroxide solution at 3% in four baths of 5 min each. The slides were again washed in running water for 5 min and then washed with phosphate buffered saline (PBS) (pH 7.2–7.6) for 5 min.

Rabbit polyclonal anti-APC antibody(ab52223) (Abcam, Japan) was used, which detects the wild-type APC protein and is recommended for detection of the respective protein in human tissue. Incubation was carried out at aconcentration of 1:100 in a humidified chamber at 4 °C for at least 16–18 hours (overnight). Subsequently, after three washes in PBS at pH 7.2 – 7.6, the incubation was performed with the streptavidin-biotin peroxidase kit (LSAB, DakoCytomation, CA, USA) in a humidified chamber at room temperature for 30 min. This step was followed by washes with PBS at pH 7.2–7.6 and development with liquid DAB (Sigma Chemical Co, MO, USA) at room temperature for 5 min. After washing in running water for 3 min, counter-staining was performed with Harris hematoxylin for 1 min. The sections were dehydrated in three baths with absolute ethanol and three baths of xylene and then mounted using cover slips with Entellan resin (Sigma Chemical Co., MO, USA) for analysis by optical microscopy. As positive control, slides with histological sections previously demonstrated as being positive for these antibodies were used. A similar slide was used as a negative control, subtracting the primary antibody from the reaction [25]. Staining was recorded as either present or absent. Presence of staining was not rated according to the intensity of staining. Extent of staining was graded as: 0, 0–10% of cells positive; 1, 10–50% of cells positive; 2, greater than 50% of cells positive for APC. Staining was considered positive, if the extent of staining was graded as 2. Staining was considered reduced, if the extent was graded as 1 and 0.

DNA extraction from the blood sample

The lymphocytes from patients' blood and unaffected control blood were separated by lysing the RBCs using a hypotonic buffer (ammonium bicarbonate and ammonium chloride, Hi-media) with minimal lysing effect on lymphocytes. Three volumes of RBC lysis buffer were added to the blood sample, mixed by vortexing and inverting thoroughly for 5 min and centrifuged (Eppendorf 5415R, Germany) at 2,000 × g for 10 min. The lymphocytes were used for DNA extraction by modified protocol of Ghatak et al. [26].

DNA extraction from the tissue samples

Deparaffinization was carried out by adding 1 ml of xylene to the tumor and adjacent normal tissue section in each microfuge tube, followed by vigorous vortexing for 10 mins. and centrifuged at 12000 rpm for 10 mins. The supernatant was discarded and the deparaffinization steps were repeated once again, followed by rehydration through subsequent washings with 100%, 90 and 70% absolute ethanol diluted in RNase free DEPC treated water, respectively. The deparaffinised tumor and adjusted normal tissue from the cases was used for the DNA extraction by the modified protocol of Ghatak et al. [27].

PCR amplification of exons14 and 15AB of APC gene

PCR was performed with the DNA from tumor, adjacent normal tissue, patient's blood and unaffected control blood samples. The APC gene exon 14 was amplified by PCR using primers Exon14-F (5'- ACATAGAAGTTAAT GAGAGAC -3') and Exon14-R (5'- TTGCTTACAAT TAGGTCTTTTTGA G -3'). The primers were designed for known polymorphic sites by using the IDT primer quest software. Polymerase chain reaction (PCR) was carried out in 25 μl total reaction volume, each containing 100 ng of template DNA, 0.2 pM of each primer, 2.5 μl of 10X PCR buffer, 1.5 mM $MgCl_2$, 200 mMdNTPs, and 1 U of Taq DNA polymerase (Fermentas, Germany). The reaction mixture was heated to 94 °C for 5 minutes, followed by 30 cycles each consisting of 40 sec denaturation at 94 °C, 40 sec annealing at 54 °C, 1 min of extension at 72 °C and a final 5 min extension at 72 °C. The APC exon 15A-B region was amplified by using Exon 15A-BF (5'- GGCAAGACCCAAACACATAATAG-3') and 15A-BR (5'- GGAGATTTCGCTCCTGAAGAA -3').The polymerase chain reaction (PCR) was carried out in 25 μl total reaction volume, each containing 100 ng of template DNA, 0.2 pM of each primer, 2.5 μl of 10X PCR buffer, 1.5 mM MgCl2, 200 mMdNTPs, and 1 unit of Taq DNA polymerase. The reaction mixture was heated to 94 °C for 5 minutes, followed by 35 cycles each consisting of 30 sec denaturation at 94 °C, 30 sec annealing at 59 °C, 1 min and 30 sec of extension at 72 °C and a final 7 min extension at 72 °C. The PCR amplification products (10 μl) was subjected to electrophoresis in a 1.2% agarose gel in 1X TAE buffer at 80 V for 30 min, stained with (0.5ug/ml) Ethidium Bromide and images were obtained in GBOX gel documentation system (UK). PCR products were purified with a Qiagen gel extraction kit (Qiaquick columns; Qiagen, Chatsworth, CA) and stored at -20 °C until sequenced using ABI 3500 Genetic Analyzer (Singapore) in Department of Biotechnology, Mizoram University, India.

Cell cycle estimation

0.1 g of grossly gastric tumor and adjacent normal gastric mucosa tissue from the caseswere used for cell cycle analysis. Cells were harvested by mechanical dis-aggregation and fine-needle aspiration. Two separate aliquots of 6×10^6 tumor cells were prepared for each sample. Pellets were incubated with 250 mL of 0.1% RNAse (Sigma, St Louis, MO, USA) and 50 mg/mL Propidium iodide (presence of Sodium citrate and TritonX-100) for 30 min at 37 °C and flow cytometric analysis was performed by Facs Canto and DIVA software (BD, Germany). Four distinct phases could be recognized in a proliferating cell population: the G1, S- (DNA synthesis phase), G2- and M-phase (mitosis). G2- and M-phase could not be discriminated because of the presence of identical DNA content [28]. The data obtained was analyzed using the ModFit LT software (DNA Modeling System) version 2.0 (Verity Software House, Inc.) and single parameter histograms were obtained.

Single-strand conformation polymorphism (SSCP) analysis

The 5’ half of exon 15 (codons 654- 1700) of APC gene was amplified using primer set (15A-B). An aliquot of 0.75 μl of each PCR product from tumor and adjacent normal gastric mucosa tissue were diluted with an equal volume of water and mixed with 1.5 μ1 of 95% formamide. This mixture was denatured at 95 °C for 5 min, cooled on ice and 2 μl was used for loading on SSCP gel (8% non-denaturing polyacrylamide gels). SSCP Gels were pre-run at 400 V, 20 mA, 2 W, for 10 or 50 volt-hours (Vh). Electrophoresis was performed at 400 V, 20 mA, 2 W, for 200–300 Vh. Electrophoresis was carried out at either 4, 10, 15 or 20 °C depending on the optimal temperature for a given PCR fragment [29]. The gels were ethidium bromide stained, and gel documented using Syngen-G-BOX (USA).

Sequence analysis

The samples exhibiting polymorphism and instability after SSCP analysis was taken for further sequencing and mutation analysis. All PCR products from the tumor, adjacent normal tissue, blood and unaffected control blood were sequenced from opposite directions to ensure reading accuracy. Sequences and chromatograms obtained were examined by chromas software version 2.13, DNA baser and align by BLAST [www.ncbi.nlm.nih.gov/blast]. The APC exons 14 and 15 were checked from Gene card database [HGNC - 583, Entrez Gene - 324, Ensembl - ENSG00000134982, OMIM - 611731, UniProtKB - P25054]. The sequences of tumor, adjacent normal was compared and sequence variation in tumor tissues from adjacent normal was recorded as somatic mutations. Further, it was confirmed that the sequence of patient’s adjacent normal, blood and healthy control blood samples are 100% identical. All the sequences containing the mutation were evaluated for their potential pathogenicity using the following algorithms: DNA baser version 3.5.4.2, Codon Code aligner version V.4.2.2, Mutation taster [www.mutationtaster.org/], PolyPhen-2 [http://genetics.bwh.harvard.edu/pph2/index.shtml.], SIFT [http://sift.jcvi.org], Mutation Assessor [http://mutationassessor.org/]. The MEGA Align algorithm was used at two depths of alignment [Cancer to Normal and Normal to database sequences]. The results of PolyPhen-2 was retrieved from the original webpage [version 2.2.2] but also from version 2.0.22 run by PON-P and version 1 run by Condel, which use them for weighted average scores. Circos plot [30] was generated to visualize the mutations in exons 14 – 15, protein expression and their association with gastric tumor stages and ploidy levels based on the observed data. This cross representation between mutations, APC protein expression and ploidy level explains the consequences of altered cell cycle regulation.

Reconfirmation of mutations by restriction digestion

Codon 622 – 625 mutations in exons 14 alter the recognition site of restriction enzyme. The specific mutation detected together with restriction enzyme used and size of fragments expected after digestion of PCR products are given in Table 1. Digestion products were analysed by electrophoresis in 8% polyacrylamide gels which were stained with ethidium bromide and documented under UV light. Restriction digestion of PCR products was performed with the DNA from tumor, adjacent normal tissue and unaffected control blood samples.

Statistical analysis

Chi-square and Fisher’s exact tests were used to assess the association ofAPC protein expression and cell cycle distribution with APC gene mutation status in relation to the stage of gastric cancer. For all tests, a two-sided *P*-value <0.05 was considered statistically significant. All analyses were performed using R statistical package ver3.3.0 [31].

Results

Gastric cancer was more prevalent in males (55%) in the Tripura population. The median age in the younger age group was 36 years (range16–45), and this group contained a lesser proportion of patients (35%) than the older age group (65%) (Table 2). The most common symptoms were abdominal pain followed by weight loss and vomiting in the case of older age patient group. Most of the gastric cancer patients were operated with stage II tumor. The symptoms at recruitment in both groups are shown in Table 2.

The tumor samples used in the present study were diffuse type gastric adenocarcinoma as confirmed after H&E staining. Our data showed that 47.5% samples were in stage II, 32.5% in stage III and 20% in stage IV. In the normal control gastric mucosa, APC immunoreactivity was positive in all the 40 samples examined. Rabbit polyclonal

Table 1 Somatic mutational profiling of APC gene exon 14 using PCR-RFLP

Codon	Enzymes	Size of normal alleles (bp)	Mutation	Amino acids	Size of mutant alleles	Sample Frequency
622[b]	*MspI*	189,163	TAC > TAA	Y >[a]	352	10%
625[b]	*MaeI*	266, 86	CAG > TAG	Q >[a]	135, 131, 86	5%

[a]represents stop codon

[b]represents Novel mutations (unreported in the database)

anti-APC antibody (ab52223) (Abcam, Japan) specificity was reported for endogenous levels of total APC protein and is expressed in a variety of tissues (http://www.abcam.com/apc-antibody-ab52223.html).Maximal APC immunoreactivity was present in the cytoplasm of the cell, but staining was not present in the mucus vacuoles. In 10% of the adenocarcinoma sample, APC immunoreactivity was completely absent despite the abundant expression of the protein in the adjacent normal mucosa. Four samples (10%) were negative for APCprotein expression in adenocarcinoma and 36 (90%) werepositive (Table 3, Fig. 1).In gastric tumour Stage III, 7.5% of the samples showed negative protein expression.After performing the Fisher exact test, the APC expression was not significantlycorrelated with the Stages of gastric cancer ($p = 0.077$).APC immunoreactivity showedpositive expression of the protein in the stage I (47.5%), stage II (25%) and stage III (17.5%) gastric adenocarcinoma and Stage III (7.5%) and stage IV (2.5%) showed negative expression of the protein.

Table 2 Clinicopathological features of gastric cancer patients (Stratified by age)

Parameters	Younger age group (Age ≤ 45 years)	Older age group (Age 46–79 years)	*P*-value
Gender			
Male	06 (15%)	16 (40%)	0.326
Female	08 (20%)	10 (25%)	
BMI (Mean ± SD)	21.4 kg/m^2 ± 3.6	22.1 kg/m^2 ± 2.9	0.058
Tumor size (cm), (mean ± SD)	4.6 ± 2.8	4.9 ± 3.1	0.922
Tumor location			
Upper	08 (20%)	13 (32.5%)	0.869
Middle	02 (5%)	05 (12.5%)	
Lower	04 (10%)	06 (15%)	
Whole	0	02 (5%)	
Type of gastrectomy			0.186
Total	10 (25%)	12 (30%)	
Subtotal	04 (10%)	14 (35%)	
Stage			
Stage I	0	0	
Stage II	7 (17.5%)	12 (30%)	0.828
Stage III	5 (12.5%)	8 (20%)	
Stage IV	2 (5%)	6 (15%)	
Abdominal pain	9 (22.5%)	16 (40%)	0.161
Weight loss	5 (12.5%)	12 (30%)	0.05
Hemorrhage	7 (17.5%)	3 (7.5%)	0.205
Dysphagia	6 (15%)	5 (12.5%)	0.76
Early satiety	3 (7.5%)	3 (7.5%)	1.00
Vomiting	4 (10%)	12 (30%)	0.045
Increased Abdominal girth	1 (2.5%)	0	0.317

Values in parenthesis indicates percentage of that sample represented from the total number of studied samples

We analysed the complete 352 bp coding region of exon 14 in the APC gene and found two novel deleterious sequence variations (g.127576C > A, g.127583C > T) changing the codons 622 and 625 to stop codons (Y622* and Q625*) in 10% of tumor samples. But, thesesomatic mutations were not observed in adjacent normal tissues and blood samples of patients as well as in healthy control blood samples (Table 1, Figs. 2 and 3). The mutation was reconfirmed at codons 622 and 625 by performing restriction digestion with *MspI* and *MsaI* (Additional file 1: Figure S1A). The wild type 622 codon (TAC) produced two digested products (189 bp and 163 bp), whereas mutant type codon (TAA) showed an uncut 352 bp band after *MspI* digestion. And, the 625 wild type codon (CAG) produced two digested products (266 bp and 86 bp), whereas mutant type codon (TAG) showed three distinct digested band (135 bp, 131 bp, 86 bp) in the polyacrylamide gel.

Samples containing mutations in codon 622 and codon 625 ofexon 14 showed abnormal cell cycle stages and indicated that aneuploidy occurs due to Apc-driven gastric tumorigenesis. Samples with well differentiated diffuse type gastric adenocarcinoma showed a nonsense mutation from TAC (Y) to TAA (stop codon) at codon 622 and

Table 3 Immunohistochemical staining of APC protein in different gastric cancer stages

Tissue Type	APC immunohistochemistry	
	Positive	Negative
Adjacent Normal cell	40 (100%)	0
Tumor cell		
Stage II	19 (47.5%)	0
Stage III	10 (25%)	3 (7.5%)
Stage IV	7 (17.5%)	1 (2.5%)

P value = 0.07

Values in parenthesis indicates percentage of that sample represented from the total number of samples

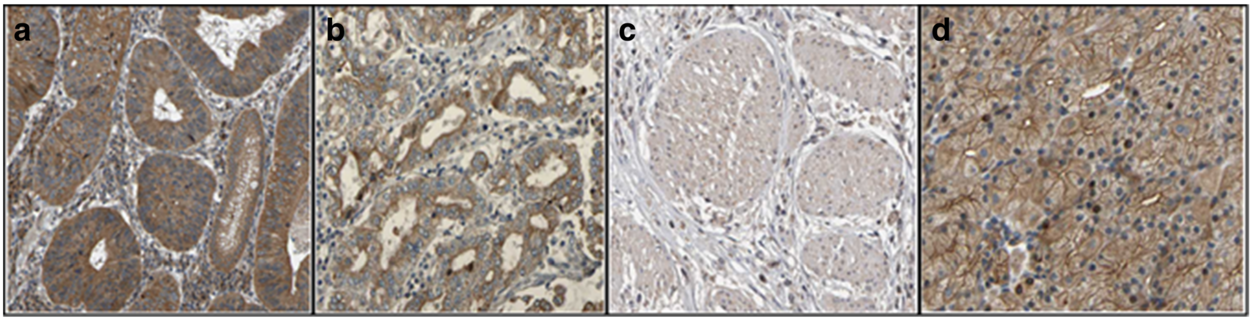


Fig. 1 Microscopic view of well differentiated adenocarcinoma of gastric tumor cells. **a** Positive high immunoexpression of anti-APC antibody in cancer cell (**b**) Positive moderate immunoexpression of anti-APC antibody in cancer cell (**c**) Negative immunoexpression of anti APC antibody in cancer cell (**d**) Positive moderate immunoexpression of anti-APC antibody in adjacent normal cell (from negative immunoexpression cancer cell), represented by the brownish colour in the cytoplasm and membrane

samples with poorly differentiated diffuse type gastric adenocarcinoma had a change from CAG (tryptophan) to TAG (stop codon) at codon 625 (Fig. 2) resulting in a truncated gene product. The tumor samples with Y622* and Q625* mutations exhibited G1 phase arrest with high S phase DNA (*p* value = 0.071) leading to loss of the role of APC in mitosis and chromosome stability [32].Most of the gastric cancer samples showed diploidy, except in samples containing 622 and 625 codon change where aneuploidy resulted in less DNA content in G2/M phase and high DNA content in S phase (Fig. 4).

The 936 bp coding region of exon 15 in the APC gene and somatic variants were detected in the gastric cancersamples (Table 4). We observed a change of exon 15 A-B region by SSCP (Additional file 1: Figure S1B). These tumour samples showedan instability banding pattern,unlike thematched adjacent normal tissue and blood of the patient's sample as well as the healthy control blood samples. Further, these samples were sequenced and six missense somatic mutations (g.131270A > G, AA1058D > G; g.131346 T > G, AA1083D > E; g.131420A > G, AA1108N > S; g.131836G > A, AA1247A > T; g.132017 T > A, AA1307I > K; g.132046G > C, AA1317E > Q) were detected randomly in a total of 40% of tumor samples which causes abnormal protein products(Table 4, Fig. 3, Additional file 1: Figure S2). Among the six missense mutations, two (1058D > G and 1307I > K) were not previously reported in the literature or the public ensemble mutation databases. Both the mutations were pathological and disease causing based on SIFT, Polyphen2 and SNPs & GO scores. Most of the exon 15 mutations were found in the compositional bias region of the APC protein.

Based on the Circos plot analysis, stage III and IV tumor samples were associated with the absence(negative) of APC protein expression, whereas Y622* and Q625* mutations were associated with stage II, III and IV tumor samples. Y622* mutated and negative APC protein expressing gastric tumor samples had a high concordance with aneuploid cells (Fig. 3). Fisher's exact test exhibited a significant statistical association of Y622* and A1247T mutations with negative APC protein expression ($P = 0.0002$; 0.005), whereas, a positive protein

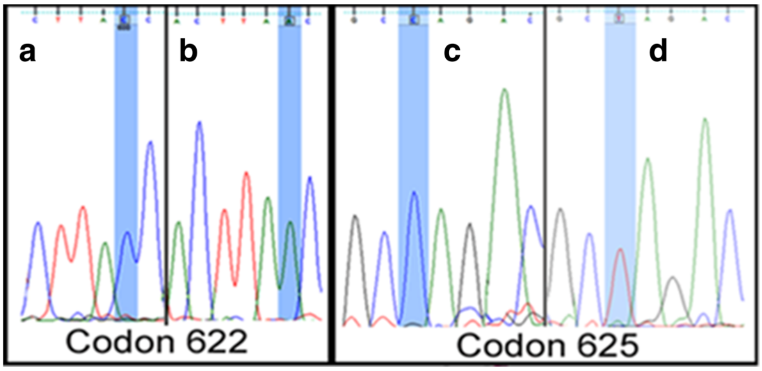


Fig. 2 Different Mutation in the exon 14 (g.127576C > A, g.127583C > T) of APC protein. **a** Wild type codon 622 (TAC) in adjacent normal sample, **b** Mutant type codon 622 (TAA) in tumor sample, **c** Wild type codon 625 (CAG) in adjacent normal sample, **d** Mutant type codon (TAG) in tumor sample

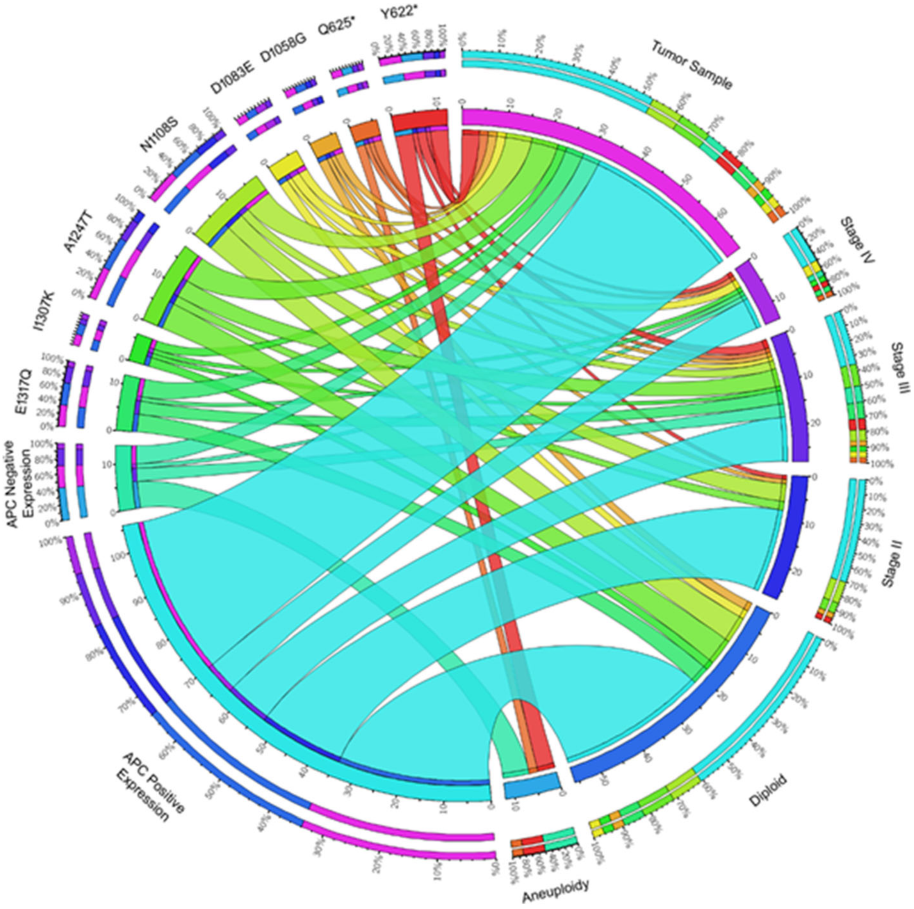


Fig. 3 Circos plot of representative APC mutation in gastric tumor sample and their association with cancer stages, cell cycle, and APC protein expression. The frequency of occurrence of different factors such as mutations, APC protein expression pattern, ploidy level and tumor stages is depicted in the outer ring. The inner ring of circos plot depicts the association between the mutations, APC protein expression pattern, ploidy level and tumor stage involved in gastric cancer. Each factor has been assigned a color. The arc originates from mutations and APC protein expression status and terminates at tumor staging and ploidy level to compare the association between the origin and terminating factors. The area of each colored ribbon depicts the frequency of the samples related with the particular mutations and APC protein expression

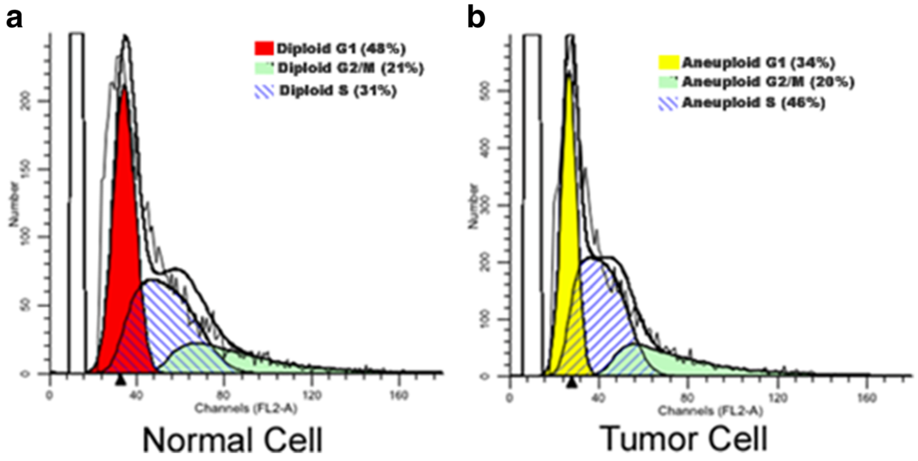


Fig. 4 Histogram of Cell cycle analysis of (**a**) adjacent control gastric cell and (**b**) Tumor gastric adenocarcinoma cell (*p* value = 0.071)

Table 4 Detection of somatic mutations in APC gene exon 15

Codon	Mutation	Amino acid change	Sift score	Polyphen2 score	SNPs & GO Effect/RI	Sample Frequency	Motifs	Domains	Amino acid property Change
1058	GAT > GGT[a]	Asp(D) > Gly(G)	Pathological	0.37 (Pathogenic)	Disease/3	5%	–	Beta-Catenin Binding	• The charge of the wild-type residue will be lost, this can cause loss of interactions with other molecules or residues • The mutation introduces a more hydrophobic residue at this position. This can result in loss of hydrogen bonds and/or disturb correct folding.
1083	GAT > GAG	Asp(D) > Glu(E)	Natural	0.08 (Benign)	Natural/1	5%	–	–	• The mutant residue is bigger, this might lead to bumps.
1108	AAT > AGT	Asn(N) > Ser(S)	Natural	0.08 (Benign)	Disease/0	15%	–	Beta-Catenin Binding	• The mutation introduces a more hydrophobic residue at this position. This can result in loss of hydrogen bonds and/or disturb correct folding.
1247	GCC > ACC	Ala(A) > Thr(T)	Natural	0.10 (Benign)	Natural/3	15%	GSK3 phosphorylation site	Beta-Catenin Binding	• The hydrophobicity of the wild-type and mutant residue differs. • Hydrophobic interactions, either in the core of the protein or on the surface, will be lost.
1307	ATA > AAA[a]	Ile(I) > Lys(K)	Pathological	0.72 (Pathogenic)	Disease/7	5%	WDR5 WD40 repeat (blade 5,6)-binding ligand	Beta-Catenin Binding	• The mutation introduces a charge, this can cause repulsion of ligands or other residues with the same charge.
1317	GAA > CAA	Glu(E) > Gln(Q)	Pathological	0.41 (Pathogenic)	Disease/4	10%	Glycosaminoglycan attachment site	Beta-Catenin Binding	• The charge of the wild-type residue will be lost, this can cause loss of interactions with other molecules or residues.

[a]- represents Novel mutations (unreported in the database); *RI* - Reliability Index

expression ($P = 0.0003$) was observed in association with N1108S mutation. The mutated region was responsible for down-regulation through a process mediated by direct ubiquitination which will affect the protein function and alter the cell cycle regulation.

Discussion

Mutations of APC gene has been shown to play an important role in colorectal tumorigenesis [33]. In the current study, we have found a significant relationship between the APC mutation, cell cycle regulation and protein expression indicating a positive role of the mutations in diffuse type gastric adenocarcinomas. We have found frequent pathogenic mutations at codon 622 of exon14 APC in gastric tumors that generates stop codon (Y622*). All the samples containing codon 622 mutation showed abnormal cell cycle regulation. All the samples containing 622 and 625 codon mutation coded for a truncated protein and resultant cells were aneuploid with high S phase. Tumor samples with codon 622,625 and 1307 mutations were strongly associated with the negative expression of the APC protein in cytoplasm as shown in immunohistochemistry analysis. Previous study showed that truncations in APC eliminate microtubule binding contributing to chromosome instability (the CIN phenotype) in colon cancer cells because they directly affect chromosome-spindle attachment [13]. Phosphorylation of APC by Bub kinases may be an important aspect of CIN phenotype, explaining why the loss of Bub1 kinase activity is a common feature of colon cancer cell lines [13]. Loss of APC function results in microtubule plus-end attachment defects during mitosis and consequent chromosome misalignment and CIN [34]. Due to exon 14 mutation, APC protein might be truncated and the same phenomenon might occurin gastric cancer leading to aneuploidy and G1 phase arrest followed by high S phase in cell cycle for diffuse type gastric cancer. It is evident that the two mutations in exon 14 of APC gene were independent of each other and are responsible for loss of protein function based on our data analysis and from the APC Mutation Database.

Our results are in agreement with the finding that mutations of APC in sporadic cases have been detected in coding amino acids within a short range from 1058 to 1317, which exists inside exon15 and is called the mutation cluster region (MCR) [35]. Mutations in exon 15 of APC gene were detected in 40% of gastric cancers and are similar to the previous studies [17]. Our results imply that APC plays a crucial role in gastric carcinogenesis as was observed in colorectal carcinogenesis [33]. The mutations detected were located in relatively small part of exon 15. Since exon 15 is extremely large, covering codons 654 through 2843 (77% of the whole coding region), it is probable that this exon encodes one of the important domains of the gene product. Alternatively, this region may be a hot spot of mutation for targeting by carcinogens. One particular missense variant, I1307K, is found in Tripura population, and carriers of this allele are at several fold higher risk of developing multiple gastric adenomas and colorectal cancer [36]. As the I1307K variant consist of T-A substitution producing a poly (A) tract, it was assumed that the variant precipitated polymerization error during DNA replication, and thus indirectly predisposition to cancer [36]. E1317Q codes for a mutation in the MCR region of the APC gene at β-catenin binding site and this mutation acts like I1307K, by a dominant negative effect on the APC/β-catenin pathway, thus leading to adenoma [37]. E1317Q mutation has been detected in colorectal polyps or cancer like as in the case of the present study [38].

The antisera that react with the specific epitope of the APC protein were used for immuno-histochemical staining to demonstrate the protein expression in gastric tumors. The subcellular localization of the APC protein is reported to be predominantly cytoplasmic in normal tissues, though mammary epithelium has been reported to show equal distribution of cytoplasmic and nuclear APC [25].In the present study, APC protein was detected more in the cytoplasmic staining in gastric tumors when compared to normal tissues. An antibody for the C-terminal region of the APC protein that detects only full-length or wild-type APC protein was used in the present study. The mutated APC protein loses its binding site in the β-catenin destruction complex resulting in low expression of APC in the cytoplasm and nucleus, which ultimately results in decreased membrane expression [39].The gastric cancer tumor in stage III and stage IV showed negative expression of the protein in cytoplasm and nucleus.

Sequencing analysis confirmed that the mutations in exons 14 and 15 of APC gene resulted in truncation of the gene products or in an amino acid change. APC gene encodes a large protein with multiple cellular functions and mutations in this gene lead to alterations in signal transduction, differentiation, intercellular adhesion, cytoskeletal stabilization, cell cycle and apoptosis [40, 41]. Truncating mutations in exons 14 and 15 were strongly associated with gastric and colorectal cancer [42]. A wealth of data shows that almost all colorectal tumors with APC mutations lose the SAMP (connexion/actin/β-catenin binding) repeats and all,or otherwise one or two of the seven β -catenin binding/degradation sites. Colorectal tumour retains a truncated APC protein to control the transcriptional activity of β -catenin and avoids it to reach too high levels, which is detrimental for tumour growth, in agreement with the “just right signalling” model [43]. The truncated APC can influence the transcriptional activity of β -catenin by at least two different mechanisms: a stimulation of the transcriptional activity of β -catenin upon APC downregulation without any obvious increase of the β-catenin level [44] and alternatively, truncated APC might be required for tumour development independently of its control over the transcriptional activity of β -catenin as previously discussed [45]. The APC protein might be truncated and all the regulatory features might be lost, especially the feature responsible for down-regulation through a process mediated by direct ubiquitination.

In the present study, eight APC mutations in exons 14 and 15 were all detected in diffuse type gastric cancer. The codon 622 and 625 mutations are significantly associated with cell cycle abnormality. This result indicates that APC gene is mutational target for gastric cancer tumor cells and supports the hypothesis that APC mutation-positive tumors may identify an alternative pathway which is probably different from the normal pathway. Our study showed that mutations in APC can contribute to development of diffuse type gastric adenocarcinomas by altering the APC protein expression and cell cycle regulation and additional genetic changes could account for the differences in pathology.

Conclusion

The present study suggests the implication of novel APC gene alterations in gastric cancer related with cell cycle abnormalities and APC protein expression in diffuse type gastric cancer. Our findings need to be confirmed by a larger cohort study, however, we reduced the risk of false-positive diagnosis of patients with other diseases by enrolling the patients with only diffuse type gastric cancer.

Additional file

Additional file 1: Figure S1. (A) PCR-RFLP of APC gene Exon 14 (M – 100 bp marker; S1, S2 – Normal alleles; S3, S4 – Mutant alleles); (B) SSCP analysis of APC gene exon 15A.b. (1, 2 – Healthy Control; 3,4,5,6 – Tumour tissues; 7, 8 – Matched tissues). **Figure S2.** Exon 15 non-synonymous mutation positions in the APC protein (DOCX 856 kb).

Abbreviations

μl: Microliter; μm: Micro meter; APC: Adenomatous Polyposis Coli; DNA: Deoxyribonucleic acid; H: Hour; kDa: Kilo Dalton; MCR: Mutation cluster region; Min: Minute; mM: Mile Moller; PBS: Phosphate Buffer Saline; PCR: Polymerase chain reaction; RNA: Ribonucleic acid; SNP: Single Nucleotide Polymorphism; SSCP: Single stranded conformation polymorphism; Vh: Volt-hours

Acknowledgements

The authors also thank the Department of Pathology, Agartala Gov. Medical College for the sample collection and DBT- infrastructural facility (BT/Med/NE-SFC/2009 Dated: 24.09.2009) and Mizoram University DBT- Uexcel project (BT/551/NE/U-Excel/2014/43) for the sequencing facility.

Funding

The authors thank Department of Biotechnology, New Delhi, India for support in the form of Advanced level State Biotech Hub (BT/04/NE/2009 Dated: 29.08.2014) and Bioinformatics Infrastructural Facility (BT/BI/12/060/2012 (NERBIF-MUA), Mizoram University which provided all the essential facilities to carry out the work.

Availability of data and materials

Materials, data, code, and associated protocols are promptly available to readers without undue qualifications. There are no restrictions on the availability of materials, data and associated protocol. Study data and materials will be available in Department of Biotechnology, Mizoram University, Aizawl, Mizoram. All the nucleotide sequences are available in EMBL-EBI (GenBank accession codes: LT855206 - LT855224).

Author's contributions

Participated in research design: AB and NSK. Conducted experiments: SG, PC, SRS and BC. Performed data analysis: SG, PC and SRS. Wrote or contributed to the writing of the manuscript: SG, AB and NSK. All authors read and approved the final manuscript.

Publisher's Note

Author details

[1]Department of Biotechnology, Mizoram University, Aizawl 796004, Mizoram, India. [2]Department of Pathology, Agartala Government Medical College, Tripura, India.

Received: 10 April 2017 Accepted: 24 May 2017
Published online: 02 June 2017

References

1. Gandhi AK, Rath GK. National cancer control and registration program in India. Indian J Med Paediatr Oncol. 2014;35:288–90.
2. Lee HS, Lee HK, Kim HS, Yang HK, Kim WH. Tumour suppressor gene expression correlates with gastric cancer prognosis. J Pathol. 2003; 200:39–46.
3. Powell SM, Zilz N, Beazer-Barclay Y, Bryan TM, Hamilton SR, Thibodeau SN, Vogelstein B, Kinzler KW. APC mutations occur early during colorectal tumorigenesis. Nature. 1992;359:235–7.
4. Nishisho I, Nakamura Y, Mivoshi Y, Miki Y, Ando H, Horii A. Mutations of chromosomes 5q21 genes in FAP and cokorctal cancer patients. Science. 1991;253:665–9.
5. Groden J, Thliveris A, Samovitz WS, Carlson MI, Gilbert L, Albertsen H, Joslyn G, Stevens J, Spirio L, Robertson M, Sargeant L, Krapcho K, Wolff E, Burt R, Hughes JP, Warrington J, McPherson J, Wasmuth J, Paslier DL, Abderrahim H, Cohen D, Leppert M, White R. Identification and characterization of the familial adenomatous polyposis coli gene. Cell. 1991;66:589–600.
6. Smith KJ, Johnson KA, Bryan TM, Hill DE, Markowitz S, Willson JKV, Paraskeva C, Petersenii GM, Hamilton SR, Vogelstein B, Kinzler KW. The APC gene product in normal and tumor cells. Proc Nat Acad Sci. 1993;90:2846–850.
7. Behrens J, Von-Kries JP, Kuhl M, Bruhn L, Wedlich D, Grosschedl R, Birchmeier W. Functional interaction of β-catenin with the transcription factor LEF-1. Nature. 1996;382:638–42.
8. Dumas YR, He X. Wnt signaling: What the X@# is WTX! EMBO J. 2011; 30:1415–7.
9. Vodermaier HC. APC/C and SCF: controlling each other and the cell cycle. Curr Biol. 2004;14:R787–96.
10. Nakayama KI, Nakayama K. Ubiquitin ligases: cell-cycle control and cancer. Nat Rev Cancer. 2006;6:369–81.
11. Ishidate T, Matsumine A, Toyoshima K, Akiyama T. The APC-hDLG complex negatively regulates cell cycle progression from the G0/G1 to S phase. Oncogene. 2000;19:365–72.
12. Heinen CD, Goss KH, Cornelius JR, Babcock GF, Knudsen ES, Kowalik T, Groden J. The APC tumor suppressor controls entry into S-phase through its ability to regulate the cyclin D/RB pathway. Gastroenterol. 2002;123:751–63.
13. Kaplan KB, Burds AA, Swedlow JR, Bekir SS, Sorger PK, Nathke IS. A role for the Adenomatous Polyposis Coli protein in chromosome segregation. Nat Cell Biol. 2001;3:429–32.
14. Horii A, Nakatsuru S, Miyoshi Y, Ichii S, Nagase H, Ando H, Yanagisawa A, Tsuchiya E, Kato Y, Nakamura Y. Frequent somatic mutations of the APC gene in human pancreatic cancer. Cancer Res. 1992;52:6696–8.
15. Wang XL, Uzawa K, Imai FL, Tanzawa H. Localization of a novel TumorSuppressor gene associated with human oral cancer on chromosome 4q25. Oncogene. 1999;18:823–5.

16. Rumpel CA, Powell SM, Moskaluk CA. Mapping of genetic deletions on the LongArm of chromosome 4 in human esophageal adenocarcinomas. Am J Path. 1999;154:1329–33.

17. Nakatsuru S, Yanagisawa A, Ichii S, Tahara E, Kato Y, Nakamura Y, Horii A. Somatic mutation of the APC gene in gastric cancer: Frequent mutations in very well differentiated adenocarcinoma and signet-ring cell carcinoma. Hum Mol Genet. 1992;1:559–63.

18. Tamura G, Maesawa C, Suzuki Y, Ogasawara S, Terashima M, Saito K, Satodata R. Primary gastric carcinoma cells frequently lose heterozygosity at the APC and MCC genetic loci. Jpn J Cancer Res. 1993;84:1015–8.

19. Maesawa C, Tamura G, Suzuki Y, Ogasawara S, Sakata K, Kashiwaba M, Satodate R. The sequential accumulation of genetic alterations characteristic of the colorectal adenoma-carcinoma sequence does not occur between gastric adenoma and adenocarcinoma. J Pathol. 1995;176:249–8.

20. Sano T, Tsujino T, Yoshida K, Nakayama H, Haruma K, Ito H, Nakamura Y, Kajiyama G, Tahara E. Frequent loss of heterozygosity on chromosomes lq, 5q, and 17p in human gastric carcinomas. Cancer Res. 1991;51:2926–31.

21. Jarvi O, Lauren P. On the role of heterotopies of the intestinal epithelium in the pathogenesis of gastric cancer. Acta pathol micobiol scand. 1951;29:26.

22. Uchino S, Noguchi M, Ochiai A, Saito T, Kobayashi M, Hirohashi S. p53 mutations in gastric cancer: a genetic model for carcinogenesis is common to gastric and colorectal cancer. Int J Cancer. 1993;54:759–64.

23. Liu Q, Li X, Ma H, Li S, Yang L. Three novel mutations of APC gene in Chinese patients with familial adenomatous polyposis. Tumir Biol. 2016;37:11421–7.

24. Ghatak S, Lallawmzuali D, Lalmawia, Sapkota R, Zothanpuia, Pautu JL, Muthukumaran RB, Senthil-Kumar N. Mitochondrial D-loop and Cytochrome Oxidase C subunit I polymorphisms among the breast cancer patients of Mizoram, Northeast India. Curr Genet. 2014;60(3):201–12.

25. Grace A, Butler D, Gallagher M, Al-Agha R, Xin Y, Leader M, Key E. APC gene expression in gastric carcinoma: an immunohistochemical study. Appl Immunohistochem Mol Morphol. 2002;10(3):221–4.

26. Ghatak S, Muthukumaran RB, Senthil-Kumar N. A simple method of genomic DNA extraction from human samples for PCR-RFLP analysis. J Biomol Tech. 2013;24:224–31.

27. Ghatak S, Sanga Z, Pautu JL, Kumar NS. Coextraction and PCR based analysis of nucleic acids from formalin-fixed paraffin-embedded specimens. J Clin Lab Anal. 2014;DOI: 10.1002/jcla.21798.

28. Blanco R, Rengifo CE, Cedeño M, Frómeta M, Rengifo E. Flow cytometric measurement of aneuploid DNA content correlates with high S-phase fraction and poor prognosis in patients with Non-small-cell lung cancer. ISRN Biomarkers. 2013;2013:1–8.

29. Gayther SA, Warren W, Mazoyer S, Russell PA, Harrington PA, Chiano M, Seal S, Hamoudi R, Rensburg EJV, Dunning AM, Love R, Evans G, Easton D, Clayton D, Stratton MR, Ponder BAJ. Germline mutations of the BRCA1 gene in breast and ovarian cancer families provide evidence for a genotype-phenotype correlation. Nat Genet. 1995;11:428–33.

30. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.

31. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016. URL https://www.R-project.org/.

32. Caldwell CM, Kaplan KB. The role of APC in mitosis and in chromosome instability. Adv Exp Med Biol. 2009;656:51–64.

33. World Cancer Research Fund/American Institute for Cancer Research. The associations between food, nutrition and physical activity and the risk of stomach cancer and underlying mechanisms. Leeds, UK: University of Leeds; 2006.

34. Green RA, Kaplan KB. Chromosome instability in colorectal tumor cells is associated with defects in microtubule plus-end attachments caused by a dominant mutation in APC. J Cell Biol. 2003;163:949–61.

35. Miyoshi Y, Nagase H, Ando H, Nishisho I, Horii A, Aoki IST, Miki Y, Mori T, Nakamura Y. Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. Hum Mol Gen. 1992;4:229–33.

36. Laken SJ, Petersen GM, Gruber SB, Oddoux C, Ostrer H, Giardiello FM, Hamilton SR, Hampel H, Markowitz A, Klimstra D, Jhanwar S, Winawer S, Offit K, Luce MC, Kinzler KW, Vogelstein B. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. Nat Genet. 1997;17:79–83.

37. Frayling IM, Beck NE, Ilyas M, Dove-Edwin I, Goodman P, Pack K, Bell JA, Williams CB, Hodgson SV, Thomas HJW, Talbot IC, Bodmer WF, Tomlinson IPM. The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. Proc Natl Acad Sci. 1998;95:10722–7.

38. Popat S, Stone J, Coleman G, Marshall G, Peto J, Frayling I, Houlston R. Prevalence of the APC E1317Q variant in colorectal cancer patients. Cancer Lett. 2000;149(1–2):203–6.

39. Luo L, Shen GQ, Stiffler KA, Wang QK, Pretlow TG, Pretlow T. Loss of the heterozygosity in human aberrant crypt foci (ACF), a putative precursor of colon cancer. Carcinogenesis. 2006;27:1153–9.

40. Zhang F, White RL, Neufeld KL. Phosphorylation near nuclear localization signal regulates nuclear import of adenomatous polyposis coli protein. Proc Natl Acad Sci. 2000;97:12577–82.

41. Fearnhead NS, Britton MP, Bodmer WF. The ABC of APC. Hum Mol Genet. 2001;10(7):721–33.

42. Wang J, Wang X, Gong W, Mi B, Liu S, Jiang B. Increased expression of ß-catenin, phosphorylated glycogen synthase kinase 3ß, cyclin d1, and c-mycin laterally spreading colorectal tumors. J Histochem Cytochem. 2009;57:363–71.

43. Albuquerque C, Breukel C, Van-der LR, Fidalgo P, Lage P, Slors FJM, Leitão CN, Fodde R, Smits R. The 'just-right' signaling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade. Hum Mol Genet. 2002;11:1549–60.

44. Chandra SHV, Wacker I, Appelt UK, Behrens J, Schneikert J. A common role for various human truncated adenomatous polyposis coli isoforms in the control of beta-catenin activity and cell proliferation. PLoS ONE. 2012;7(4):e34479.

45. Schneikert J, Behrens J. The canonical Wnt signalling pathway and its APC partner in colon cancer development. Gut. 2007;56:417–25.

Environmental Science and Pollution Research (2018) 25:31691–31704
https://doi.org/10.1007/s11356-018-3080-1

RESEARCH ARTICLE

Lifestyle chemical carcinogens associated with mutations in cell cycle regulatory genes increases the susceptibility to gastric cancer risk

Ravi Prakash Yadav[1] · Souvik Ghatak[1] · Payel Chakraborty[1] · Freda Lalrohlui[1] · Ravi Kannan[2] · Rajeev Kumar[2] · Jeremy L. Pautu[3] · John Zomingthanga[4] · Saia Chenkual[5] · Rajendra Muthukumaran[6] · Nachimuthu Senthil Kumar[1]

Received: 16 February 2018 / Accepted: 27 August 2018 / Published online: 12 September 2018

Abstract

In the present study, we correlated the various lifestyle habits and their associated mutations in cell cycle (*P21* and *MDM2*) and DNA damage repair (*MLH1*) genes to investigate their role in gastric cancer (GC). Multifactor dimensionality reduction (MDR) analysis revealed the two-factor model of oral snuff and smoked meat as the significant model for GC risk. The interaction analysis between identified mutations and the significant demographic factors predicted that oral snuff is significantly associated with *P21* 3′UTR mutations. A total of five mutations in *P21* gene, including three novel mutations in intron 2 (36651738G > A, 36651804A > T, 36651825G > T), were identified. In *MLH1* gene, two variants were identified viz. one in exon 8 (37053568A > G; 219I > V) and a novel 37088831C > G in intron 16. Flow cytometric analysis predicted DNA aneuploidy in 07 (17.5%) and diploidy in 33 (82.5%) tumor samples. The G2/M phase was significantly arrested in aneuploid gastric tumor samples whereas high S-phase fraction was observed in all the gastric tumor samples. This study demonstrated that environmental chemical carcinogens along with alteration in cell cycle regulatory (*P21*) and mismatch repair (*MLH1*) genes may be stimulating the susceptibility of GC by altering the DNA content level abnormally in tumors in the Mizo ethic population.

Keywords Chemical carcinogens · Gastric cancer · Flow cytometer · Cell cycle · Mutation · Mizo population

Introduction

Cooking over an open flame is an ancient practice, starting from the early days of evolution of Homo sapiens. The consumption of grilled and smoked meat/vegetables seems to have increased in the modern food habits and population-wide changes in dietary style have dramatically increased stomach cancer rates in different parts of the world. There are so many delicacies prepared by smoking the food. The *N*-nitroso polycyclic aromatic hydrocarbons (NPAHs) are generated in the grilled and smoked meat/vegetables which will be the most harmful carcinogen causing stomach cancer (Correa et al. 1985; Ghatak et al. 2016). During the burning of wood by smoking and direct heat drying process, many harmful chemicals are formed, such as formaldehyde, polycyclic aromatic hydrocarbons (PAH), nitrogen and sulfur oxides, dioxins, heavy metals, etc. The PAHs are well known to cause several types of cancer in lab animals, such as liver, skin, and stomach (Stepanov et al. 2005, 2010).

Gastric cancer (GC) is the fifth most common cancer and is the third leading cause of cancer-related death worldwide

Ravi Prakash Yadav and Souvik Ghatak contributed equally to this work.

Responsible editor: Philippe Garrigues

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11356-018-3080-1) contains supplementary material, which is available to authorized users.

✉ Nachimuthu Senthil Kumar
nskmzu@gmail.com

[1] Department of Biotechnology, Mizoram University, Aizawl, Mizoram 796004, India

[2] Cachar Cancer Hospital and Research Centre, Silchar, Assam 788015, India

[3] Mizoram State Cancer Institute, Zemabawk, Aizawl, Mizoram 796017, India

[4] Department of Pathology, Civil Hospital, Aizawl, Mizoram 796001, India

[5] Department of Surgery, Civil Hospital, Aizawl, Mizoram 796001, India

[6] Department of Chemistry, Mizoram University, Aizawl, Mizoram 796004, India

(Ferlay et al. 2015). Globally, incidence of GC shows a wide geographic and ethnic variation, being particularly high in East Asian countries (Torre et al. 2015). In India, there are high GC incidence regions like Mizoram with an age-adjusted rate (AAR) of 50.6 and 23.3 per 10^5 populations in male and female, respectively (NCRP 2013). The general population in Mizoram is socially and ethnically unique from any other tribes and groups of India. Mizoram comprises of a distinct ethnic population with peculiar dietary habits such as consumption of smoked meat/vegetables, sa-um (fermented pork fat), nitroso salts, tobacco products; a tobacco smoke-saturated aqueous concentrate—tuibur, hand-rolled locally made cigarette—meiziol, freshly cut areca nut, slaked lime with half-betel leaf—kuhva, oral snuff and betel quid ("fresh" areca nut, slaked lime, condiments and coarse tobacco/pan masala wrapped in betel leaf for chewing), and alcohol (Phukan et al. 2006). A unique addiction of use of "tuibur" (tobacco smoke-infused aqueous answer) has been noticed in Mizoram (Madathil et al. 2018). Moist snuff is used by placing it between the lower lip or cheek and gum, and the nicotine in the snuff is absorbed through the tissues of the mouth. Moist snuff also comes in small, teabag-like pouches or sachets that can be placed between the cheek and gum. These are designed to be both "smoke-free" and "spit-free" and are marketed as a discreet way to use tobacco (Madathil et al. 2018). In Mizoram, manufactured smokeless tobacco products which are preferred as tobacco product packed in tear packs (gutka, khaini) and handmade cottage product packed in plastic packets (coarse tobacco mixed with slaked lime enriched water and trace levels of molasses known locally as "sahdah"). Hospital-based data from Mizoram have shown GC to be the most common cancer accounting for 30% of all cancer cases (Phukan et al. 2004). Hence, tobacco consumption may correlate with the high incidence of GC in Mizoram.

Gastric tumorigenesis is a multistep and multifactorial process associated with various genetic and epigenetic alterations including the activation of various oncogenes and inactivation of tumor suppressor genes and mismatch repair (*MMR*) genes (Igaki et al. 1994). According to multistep model of gastric carcinogenesis, the most common and principle pathway affected is cell cycle by genetic aberrations. Cell cycle progression is a highly ordered biological process; hence, alterations in the cell cycle genes has been suggested to contribute the underlying the tumorigenesis of GC (Decesse et al. 2001).

P21 has been reported to play multiple roles within the cell including cell cycle regulation, senescence, apoptosis, DNA repair, and differentiation (Parker et al. 1995; Ciccarelli et al. 2005; Jung et al. 2010). Although mutations in *P21* are infrequent in human cancers (Shiohara et al. 1994), previous studies have shown that *P21* may act to either promote or suppress in various cancers. *P21* is a putative tumor suppressor gene and its mutations have been studied as a risk factor in various cancers (Watanabe et al. 1995), including GC (Mousses et al. 1995; Bahl et al. 2000). The murine double minute 2 (*MDM2*) is one of the central nodes in the p53 pathway and can control p53 protein levels and activity. *MDM2* gene encodes an important negative regulating protein which promotes ubiquitin-dependent proteosomal degradation of p53 by functioning as an E3 ubiquitin ligase (Oren et al. 2002; Bouska et al. 2008). The mismatch repair (*MMR*) system plays an essential role in identifying and rectifying replication errors as well as additional errors in DNA which may arise through physical or chemical damage. The pathogenic alterations are scattered in the carboxyterminus domain of MLH1 protein and the position annotated as pms2, mlh3, and pms1 interaction domain (Guerrette et al. 1999; Lipkin et al. 2000; Kondo et al. 2001).

In Mizo tribal population, there is limited evidence for the genetic and environmental risk factors that may be associated with stomach cancer (Ihsan et al. 2011; Malakar et al. 2012). In the present study, a case–control study for the high prevalence of GC in Mizoram has been attempted in order to identify the mutations in cell cycle genes, *P21* and *MDM2*, besides DNA damage repair gene, *MLH1*. Further, the correlation between these mutations and the environmental as well as dietary factors that seem to play an important role in GC etiology are also described.

Material and methods

Subjects

This study included a cohort of patients with pathologically confirmed gastric tumor. A total of 40 gastric tumor tissues (28 males and 12 females) and their matched adjacent normal gastric mucosa were collected. Peripheral blood samples were also collected from the patients and sex-age matched healthy individuals. All the gastric tumor and adjacent normal samples were collected from Civil Hospital and Genesis Laboratory and Diagnostics, Aizawl, Mizoram. The demographic information such as age, gender, dietary habits, familial incidence of cancer, smoking habits, and alcohol consumption were obtained after getting informed consent using a structured questionnaire. All the study participants received written information and gave consent for the publication of the required results. Ethics approval for this study was obtained from the Institutional Review Board (IRB) of the Civil Hospital, Aizawl (B.12018/1/13-CH(A)/IEC). Hematoxylin and eosin-stained slides were prepared from paraffin block of tumor tissues after micro-dissection to determine the type gastric adenocarcinoma and TNM staging.

DNA extraction and PCR amplification

Genomic DNA was isolated from the blood following the standard protocol (Ghatak et al. 2013). The genomic DNA was extracted from the paraffin embedded tumor tissue and adjacent normal by the modified protocol of Ghatak et al. (2014). All exons and adjacent intronic regions of the *P21*, *MDM2*, and *MLH1* genes were screened. PCR (Eppendorf, USA) was carried out in 25 μl total reaction volumes (containing 100 ng template DNA, 0.2 pM of each primer, 1× PCR buffer, 1.5 mM $MgCl_2$, 200 mM dNTPs, 1 unit Taq DNA polymerase (Fermentas Inc., Glen Burnie, MD). The reaction mixture was heated to 94 °C for 7 min, followed by 40 cycles, each consisting of 1 min denaturation at 94 °C, 1 min annealing at 63 °C, 1 min extension at 72 °C, and a final 7 min extension at 72 °C (Supplementary Table 1). The PCR amplification products (3 μl) were subjected to electrophoresis on 1.2% agarose gel in 1× Tris-acetate-EDTA buffer at 80 V for 30 min and stained with ethidium bromide (Himedia, India) and images were obtained in gel documentation (G-Box; Syngene, UK) system (Fig. 1).

Sequencing and sequence analysis

Sequencing of the exons of these genes is most useful when using primers that include a portion of the intron/exon boundary. This allows the entire exon to be sequenced as well as the splice sites where mutations are known to occur (Boland and Goel 2010). The impact of amino acid allelic variants on protein structure/function can be predicted after performing multiple sequence alignments and protein 3D structures. The Sorting Intolerant from Tolerant (SIFT) algorithm was applied. SIFT is a program that predicts the effect of amino acid substitutions on protein function, on the basis of sequence conservation during evolution and the nature of the amino acids substituted in a gene of interest. SIFT scores were calculated online (http://sift.jcvi.org/). If the value is less than 0.05, the amino acid substitution was predicted as intolerant, while those with a value greater than or equal to 0.05 were classified as tolerated. Human Splicing Finder (http://www.umd.be/HSF/) was used for finding the splice site region.

Splicing donor/acceptor sites and branch point sequences

To predict potential 5′ss and 3′ss (splicing site), we used matrices-derived splicing finder database. A potential splice site is defined as an n-mer sequence. For each “n” position, a weight is given to each nucleotide, based on its frequency and the relative importance of its position in the sequence motif (position weight matrices, PWM). Only n-mer sequences with consensus values (CV) higher or equal to a given threshold are considered as potential 5′ or 3′ss. The human branch point (BP) consensus sequence is YNYCRAY, and the threshold for BP sequences was fixed at 67. Since many intronic sequences match the BP consensus sequence, hence, we

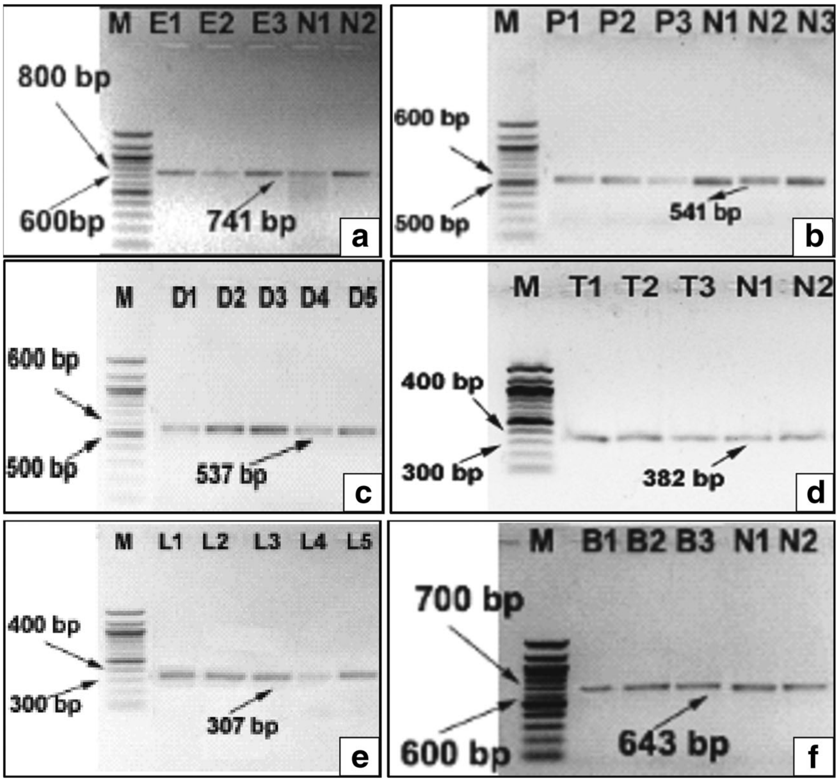


Fig. 1 PCR products of mdm2 gene. **a** Exon 2, **b** Exon 8. Lanes M: 100 bp DNA ladder (Invitrogen, USA), D1, D2, D3, D4, D5 and T1, T2, T3 = GC samples, N1, N2 = controls. Gel images of the PCR products of p21gene. **c** Exon 2, **d** Exon 3. Lanes M: 100 bp DNA ladder (Invitrogen, USA), E1, E2, E3 and P1, P2, P3 = GC samples, N1, N2, N3 = controls. Gel images of the PCR products of mlh1 gene. **e** Exon 8, **f** Exon 16. Lanes M: 100 bp DNA ladder (Invitrogen, USA), L1, L2, L3, L4, L5 and B1, B2, B3 = GC samples, N1, N2 = controls

included the AG-Exclusion Zone algorithm (Gooding et al. 2006) to predict BP candidates. Splicing donor/acceptor sites and BP was splice site finding was estimated by Human Splicing Finder (http://www.umd.be/HSF/). HSF searches all AG dinucleotides that are included in a 3′ss candidate sequence (threshold of 67) and therefore define the exclusion zones for a given intronic sequence and its intron–exon boundary. HSF annotates the functional BP as the strongest candidate without a 3′-exclusion zone before the natural 3′ss because it has been shown that the BP allows the recognition of the first downstream 3′ss. Additionally, to take into account the steric obstruction caused by the spliceosome, we excluded BP sequences located at less than 12 nt from the exon. Finally, as most BP sequences are located between − 21 and − 34 nt from the exon and only a window of 100 bp is processed.

DNA content analysis

Flow cytometry measures DNA contents (ploidy) of cancer cells and rate of proliferation, indicating the proportion of cells under DNA synthesis (S-phase fraction) and has been shown to yield prognostic information in many of the human malignancies (Merkel and Mcguire 1990). S-phase fraction (SPF) is also a well-known independent prognostic factor in some human malignancies, such as breast, prostate cancer, and gynecological malignancies (David and Hedley 1994).

For the analysis of cell cycle distribution, paraffin-embedded tumor tissue was used following modified method described by David and Hedley (1994); 10^6 cells were harvested by centrifugation, washed in phosphate-buffered saline (PBS) (Sigma P4170), fixed with ice cold 70% ethanol, and treated with 1 mg/ml RNAse for 30 min. Intracellular DNA was labeled with propidium iodide (50 mg/ml) and incubated at 4 °C in the dark. Samples were then analyzed using flow cytometer FACSCalibur (BD, Germany). The data obtained was analyzed using the ModFit LT software (DNA Modeling System) version 2.0 (Verity Software House, Inc.) and single parameter histograms was obtained.

Multifactor dimensionality reduction analysis

The multifactor dimensionality reduction (MDR) is a nonparametric, genetic model-free statistical approach to identify high-order gene–gene and gene–environment interactions associated with GC risk (Hahn et al. 2003; Cattaert et al. 2011). It is applied for overcoming the sample size limitations. In the present study, MDR software package (MDR 3.0.2) was used to generate the best one-dimensional multifactor model to classify and predict GC susceptibility. The best model was selected based on maximum cross-validation consistency (CVC) and testing balance accuracy (TBA). The MDR permutation results were considered to be statistically significant at the 0.05 level (Ritchie et al. 2003; Manuguerra et al. 2007).

Interaction entropy graph

Interaction graphs were built to visualize and interpret the results obtained from MDR. Entropy estimates were used to determine the information gain about a class variable (e.g., case–control status) from merging two variables together. Entropy estimates are useful for building interaction graphs facilitating the interpretation of relationships between variables (Choudhury et al. 2015).

Statistical analysis

The polymorphism and demographic factor in each group were estimated for their association with GC using odds ratios (ORs) and 95% confidence intervals (CIs) in the logistic regression (LR) model adjusted with multivariable analysis. Each polymorphism was checked by the presence and absence of the SNPs. Additionally, logistic regression analyses were conducted to compute the influence of both genetic and environmental factors for GC. For all tests, a two-sided p value < 0.05 was considered statistically significant. All statistical analyses were performed using SPSS 20.0 program (SPSS Ibérica, Madrid, Spain) and SYSTAT 13.0. (Systat Software Inc., USA). Hardy–Weinberg equilibrium by a chi-square (χ^2) test with one degree of freedom (df) was performed between case and control subjects. Fisher's exact test also used for comparing the demographic and habits between patients and controls. Correlation between clinicopathological features and DNA content was estimated by SPSS 20.0 program.

Results

Characteristics of study subjects

The frequency distributions and selected characteristics of the patients and controls are presented in Table 1. The median age was 58.7 years for the patients and 52.18 years for the controls. The analysis indicates that smoked meat/vegetable (OR 16.214; 95% CI 2.746–95.749; p 0.002) and oral snuff (OR 10.496; 95% CI 2.410–45.710; p 0.002) are the major risk factor for GC among our study population (Table 1).

MDR and interaction entropy analysis

MDR analysis was performed to explore the potential gene–gene-environment interaction. In the present study for the entire dataset, smoked meat is the best one-factor model found statistically significant ($p < 0.0001$) with a CVC of 8/10 and testing accuracy of 0.6653. The combination of smoked meat and oral

Table 1 Demographic characteristics of the cases and control samples

Demographic factor	HC (*n* = 40)	GC (*n* = 40)	ORs (95% CI)	*p* value
Age years ± SD (range)	52.18 ± 12.35	58.7 ± 9.76	–	
Male Female	12 (30%) 28 (70%)	28 (70%) 12 (30%)	–	
Sa-um	29 (72.5%)	37 (92.5%)	0.979 (0.346–2.770)	0.968
High salt intake	30 (75%)	31 (77.5%)	0.507 (0.077–3.340)	0.480
Smoked meat/vegetable	25 (62.5%)	38 (95%)	16.214 (2.746–95.749)	0.002
Pickle	22 (55%)	23 (57.5%)	0.340 (0.108–1.072)	0.065
Tuibur consumption	14 (35%)	17 (42.5%)	0.755 (0.350–1.631)	0.475
Cigarette smoking	24 (60%)	29 (72.5%)	2.091 (0.810–5.400)	0.127
Oral snuff	6 (15%)	20 (50%)	10.496 (2.410–45.710)	0.002
Tiranga/Gutkha	8 (20%)	4 (10%)	8.954 (0.816–98.308)	0.073
Kuhva(betel nut, slaked lime wrapped in betel leaf)	15 (37.5%)	17 (42.5%)	1.094 (0.394–3.036)	0.864
Family history of gastric cancer	5 (12.5%)	7 (17.5%)	2.148 (0.579–7.970)	0.253
Family history of other cancers	5 (12.5%)	3 (7.5%)	1.011 (0.369–2.769)	0.983

HC healthy control, *GC* gastric cancer, *OR* odd ratio, *95% CI* 95% confidence interval

snuff was found to be the best two-factor model which was also the best overall model with a CVC of 10/10 and TBA of 0.7389 ($p < 0.0001$). The combination of sa-um, smoked meat, and tuibur was found to the best three-factor model with a CVC of 3/10 and TBA of 0.4042 ($p < 0.0001$) (Fig. 2a). The previous statistical analysis results were reproduced in MDR analysis also.

Interaction entropy graphs were created using MDR results, for better verification and visualization of interactions between gene–environment factors. In interaction entropy graph, smoked meat showed the highest independent effect (20.56%) and also had moderate synergistic interaction with sa-um (0.43%). Oral snuff (8.19%) also explained considerable entropy independently (Fig. 2b).

The association between gene mutations and GC risk

A binary logistic regression model was applied to estimate the association between gene mutations and risk of GC (Table 2). The mutations in *P21* gene at 3′ UTR were associated with oral snuff consumption in GC patients (OR 9.256; 95% CI

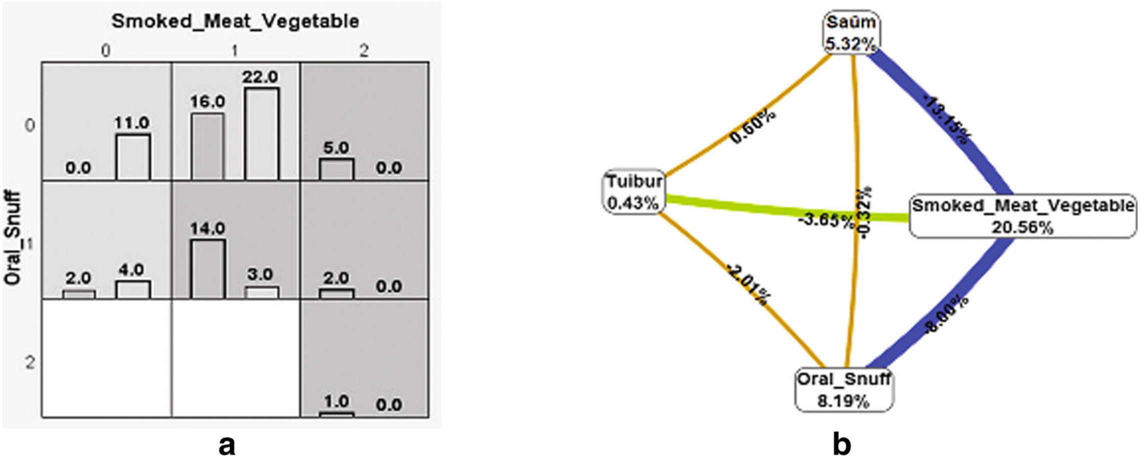


Fig. 2 Multifactor dimensionality reduction (MDR) analysis. **a** The summary of the two-factor model (smoked meat and oral snuff) predicted by MDR is represented in the graph. The distribution of high-risk (dark shading) and low-risk (light shading) combinations associated with GC risk. For smoked meat and oral snuff, 0 represents less consumption, 1 represents moderate consumption, and 2 represents high consumption. **b** Interaction entropy graph. The percent of the entropy for independent factors as well as their interactions are represented in the graph where positive percentage of entropy denotes synergistic interaction while negative percentage denotes redundancy. The red color indicates a high degree of synergistic interaction and orange a lesser degree, whereas gold represents midpoint and blue represents the highest level of redundancy followed by green

Table 2 Interaction between mutations and significant demographic factors

Factor	Gene name	Position	ORs (95% CI)	*p* value
Oral snuff	p21	Intron 2	1.025 (0.253–4.150)	0.972
		3′ UTR	9.256 (1.842–46.509)	0.007
	mlh1	Exon 8	0.956 (0.183–4.986)	0.958
		Intron 16	1.732 (0.143–20.956)	0.666
Smoked meat/vegetable	p21	Intron 2	4.149 (0.970–17.738)	0.050
		3′ UTR	0.728 (0.182–2.909)	0.653
	mlh1	Exon 8	1.510 (0.324–7.043)	0.600
		Intron 16	1.386 (0.109–17.543)	0.801

OR odds ratio, *95% CI* 95% confidence interval

1.842–46.509; $p < 0.007$). Similarly, smoked meat/vegetable and *P21* intron 2 (OR 4.149; 95% CI 0.970–17.738; $p < 0.050$) were also found to be significantly associated with increased risk for GC (Table 3). However, the other genes did not show any association with the demographic factors.

Sequence variations of *P21*, *MDM2*, and *MLH1* genes and their consequence

The molecular analysis revealed a total of five mutations in P21 gene (Table 2, Fig. 2). In intron 2, mutations 36651738G > A, 36651804A > T, and 36651825G > T were identified in 5% of total GC samples and were found to be novel (not reported previously in the database). The mutations were of single base change type. Splice site changes were identified as a result of 36651738G > A mutation, and it was also predicted that this might affect the protein folding and/or functional features. Splice site donor is marginally increased (wt 0.8042/mu 0.8719). No variation in potential splice site changes were identified due to 36651804A > T, whereas, 36651825G > T was found to increase marginally at the splice site donor (wt 0.8725/mu 0.9361). 3′UTR region of *P21* showed two known (previously reported) mutations

Table 3 Polymorphism and mutations in cell regulatory genes of gastric cancer samples

Gene name	Position	Nomenclature of mutation	Frequency of mutation (%)	AA change	PolyPhen-2/SIFT / PROVEAN	Novel/ reported	Effect of mutation by mutation taster
p21	Intron 2	36651738G > A	5	–	–	Novel	Polymorphism (single base change) Protein features (might be) affected Splice site changes (donor marginally increased wt 0.8042/mu 0.8719)
		36651804A > T	5	–	–	Novel	Polymorphism (single base change) No abrogation of potential splice site
		36651825G > T	5	–	–	Novel	Polymorphism (single base change) Donor marginally increased (wt 0.8725/ mu 0.9361)
	3′UTR	36653580C > T	10	–	–	Reported in the database	Polymorphism (single base change) Splice site changes (splice site change occurs after stop codon, acceptor marginally increased, wt 0.53/ mu 0.64)
		36653597C > T	5	–	–	Reported in the database	Polymorphism (single base change) Splice site changes (splice site change occurs after stop codon, acceptor marginally increased, wt 0.5311/ mu 0.5459)
mlh1	Exon 8	37053568A > G	10	219I > V (ATC > GTC)	(0.018) Benign/ neutral/tolerated	Reported n the database	In the protein structure helix (212–220) might be lost Splice site changes (wt 0.7064/mu 0.7505, acceptor marginal increased)
	Intron 16	37088831C > G	5	–	–	Novel	Polymorphism (wt 0.5187/mu 0.5635, acceptor marginal change)

36653580C > T in 10% and 36653597C > T 5% of GC samples. These mutations are affecting the splice site change by acting after stop codon. Splice site acceptor is marginally increased in both the cases. Annotated sequences were deposited in EBI repository with accession number (LN997431-LN997630).

Whereas in *MLH1* gene, two variants were identified one in exon 8 (37053568A > G; 219I > V) and other in intron 16 (37088831C > G) of which intronic variant is novel (37088831C > G) (Table 2, Fig. 3). The SIFT score for the *MLH1* exonic variant ((37053568A > G; 219I > V)) demonstrated that it is tolerated and may elicit only minor effect on the protein structure. 37088831C > G variant in intron 16 of *MLH1* gene was identified in 5% of GC samples which is also novel (not reported in database) and found to affect splice site by increasing the acceptor marginally (wt 0.5187/mu 0.5635). In the present study, we expected to find common pathogenic mutations, but for this cohort, none were found in *MDM2* gene. This result is indicative of the fact that there may not be any common or founder mutations for *MDM2* gene in Mizo population. Prediction of structural variation between wild- and mutant-type amino acids for *MLH1* exonic variant ((37053568A > G; 219I > V)) was carried out by HOPE analysis which indicated that the mutant type is smaller in size than the wild type, affecting the intramolecular and external interactions due to clashes (Fig. 4).

DNA content analysis

Flow cytometric analysis (FCM) analysis can provide not only the kinetic estimates such as the fraction of cells in S-phase (SPF) but also capable of subdividing neoplasms into DNA diploid or DNA aneuploid tumors based on the presence of different sub-populations in different phases of cell cycle. FCM analysis predicted DNA aneuploid in 07 (17.5%) and diploid in 33 (82.5%) diploid tumor samples (Fig. 5). Significantly higher S-phase fraction (SPF) was observed in all the GC samples (51.24–72.09) compared to controls (32.45–44.12). The G2/M phase was found arrested in gastric tumor samples. The G2/M phase is found to be arrested in the gastric tumor samples besides more DNA content in S-phase (Table 4, Fig. 2).

Splicing abnormality due to mutations

The difference between wild-type (wt) active sites and mutant-type inactive sites was predicted by the HSF algorithm and was calculated by the consensus values (CV) of 50ss or 30ss. The CV higher than 80 represent as a stronger relation with active sites and between 70 to 80 represent a weaker relation with active site. Mutations can create a new cryptic splice site rather than disruption of a 50ss or a 30ss active sites which was correctly predicted by HSF. 36653580C > T mutation in 3′UTR of *P21* gene showing 1.53% CV variation between wild type (86.39) and mutant (84.86) type. Due to the potential CV change, the branch point motifs also potentially changed for wild type (CGCCCAC) and mutant type (TGCCCAC) (Table 5).

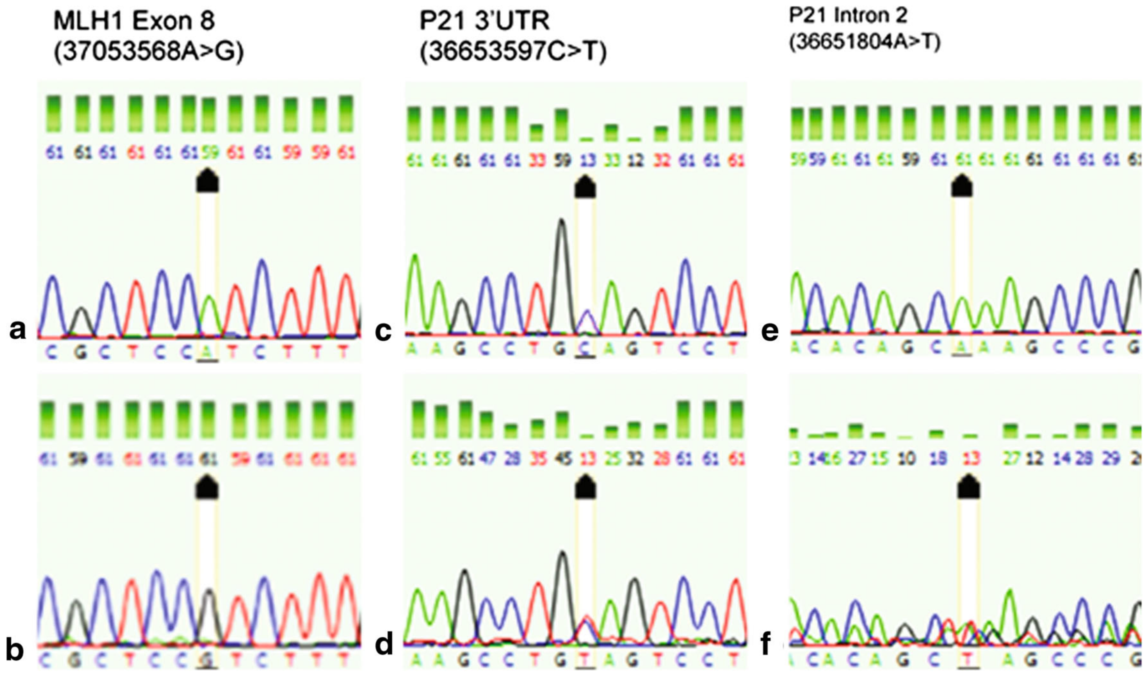


Fig. 3 Electropherogram of the genes from GC samples (**b**, **d**, **f**) compared with healthy samples (**a**, **c**, **e**)

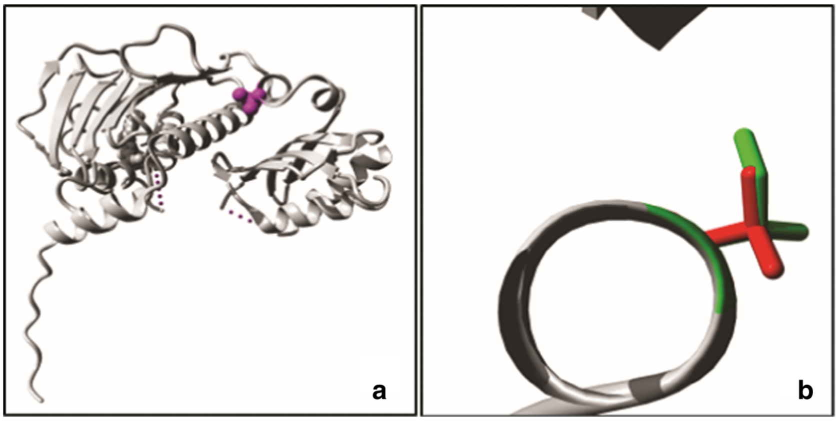


Fig. 4 3D structure of the mutation (37053568A > G; 219I > V) in mlh1 gene. **a** Complete chain of protein and the pink spheres and rest of the protein is shown in gray representing the site of mutation (219I > V). **b** Mutation of isoleucine to valine at position 219 due to 37053568A > G. Wild-type and mutant-type side chains are shown in green and red, respectively

Discussion

Our attempt was to accumulate evidence to identify the relationship between chemical carcinogens and GC risk in relation to gene mutations. GC has positive association with consumption of smoked and salted meat/vegetables (Correa et al. 1985; Ghatak et al. 2016). *N*-nitroso compounds can be generated in meat during smoke drying and preservation which are highly carcinogenic due to the reaction between nitrite with amines and amides, which is found in meat and other proteins (Correa et al. 1985; Ghatak et al. 2016). The nitrite present in smoked meat play a secondary role in the progress of chronic atrophic gastritis, which can develop as stomach cancer in the later stages (Nomura et al. 1990). Previous studies have reported positive association of high intake of smoked meat as potential confounder for GC (Kneller et al. 1992; Appelman et al. 1992; Ward and Lopez-Carrillo 1999). In the present study, smoked meat and high salt intake was positively associated with GC. Intra-gastric high salt accumulation causes the expansion of surface mucous prompting for aggravation and damage, for example, diffuse erosion, atrophic gastritis, and diminished corrosiveness of the stomach which creates a condition supporting *H. pylori* infection (Tsugane et al. 2004; Tsugane and Sasazuki 2007). Gastric mucus can also be damaged by smoked meat intake with extra salt, leading to increased epithelial cell proliferation as part of the repair process (Campos et al. 2006).

In Mizoram, mostly pork, beef, fresh water fish, birds, and/or animal meat along with the seasonal vegetables are used for heat drying (traditional wood-burning) for preservation. Higher concentrations of polycyclic heterocyclic amines (PAHs) formed during the preparation of food at higher temperature conditions such as frying, roasting, and/or grilling (Phillips 1999). Interestingly, only modest levels of PAHs are formed, while cooking the food by steaming/stewing/boiling. Along with lean meat, heat drying of "meat with fat" besides "skin and fat" of pork meat is preferred in Mizoram for preservation. Few of the PAHs and nitrosamines are carcinogenic, while some of the PAHs, HAAs, and volatile nitrosamines are indeed pro-carcinogens. These pro-carcinogenic species are metabolically activated, after being ingested and metabolized subsequently as carcinogens (Hecht 2003). In addition, soda (sodium

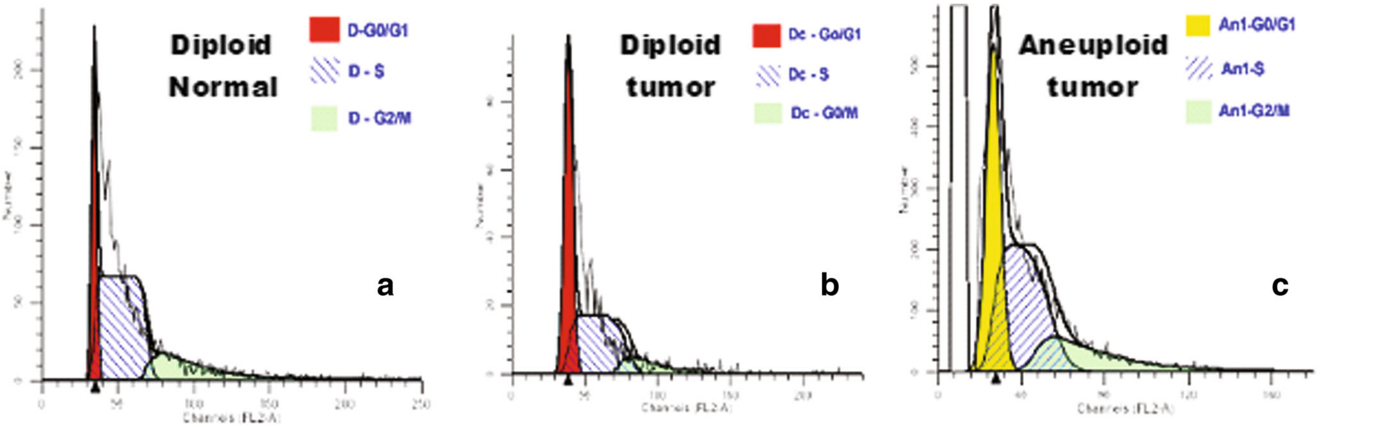


Fig. 5 DNA content analysis in gastric cancer samples measured by the flow cytometric profile. Histogram of **a** diploid normal sample with G0/G1 peak (red), S-phase (shaded peak) and G2/M peak (green). **b** Diploid gastric cancer tumor sample. **c** Aneuploid gastric cancer tumor sample with aneuploidy peak (yellow)

Table 4 Distribution of different phase fraction according to DNA content

Sample no.	Ploidy status	Ploidy (%)	G0/G1 phase			S-phase			G2/M phase	
			Low (%)	Moderate (%)	High (%)	Low (%)	Moderate (%)	High (%)	Low (%)	High (%)
Healthy (40)	Diploid	40 (100)	1 (2.5)	13 (32.5)	26 (65)	38 (95)	2 (5)	0	3 (7.5)	37 (92.5)
Cancer (40)	Diploid	33 (82.5)	18 (54.5)	15 (45.4)	0	0	3 (9.09)	30 (90.9)	28 (84.8)	5 (15.1)
	Aneuploid	7 (17.5)	0	1 (14.3)	6 (85.7)	0	1 (14.3)	6 (85.7)	7 (100)	0

bicarbonate)—an alkaline preparation, frequently used as food additives was significantly associated with increased risk of stomach cancer in Mizoram (Phukan et al. 2006).

During consumption of oral snuff and chewing tobacco, the harmful contaminant can mix with saliva and enter inside the stomach, although the association between oral snuff and stomach cancer incidence was negatively reported for some study and one study clearly reported the positive associations (Chao et al. 2002; Furberg et al. 2006). Oral snuff production in other parts of the world involves heat processing rather than curing and fermentation as is done in the Mizoram, northeast India. Fermentation of tobacco can generate higher amounts of tobacco-specific *N*-nitrosamines and volatile *N*-nitrosamines due to the high *N*-nitrosation of nicotine.

It is important to note that higher concentration levels of PAHs (Phillips 1999) and lower concentration levels of HAAs in pork meat with higher fat content was observed (Chen et al. 1990). PAHs such as benzo(a) pyrene formed in smoked food have been correlated in many areas of the world with high stomach cancer rates (Yeh et al. 2009). Furthermore, oral snuff and other tobacco products contain carcinogens like the nitrosamines which acts as cofactor for pathogenesis of GC (Curado et al. 2007). Gutkha, khaini, and sahdah demonstrably contains relatively high concentration levels of tobacco-specific nitrosamines (TSNAs) (some of which are proven carcinogens to human, while some of which are pro-carcinogens), nitrate, and/or nitrite as well as PAHs (Stepanov et al. 2005, 2010). It may be possible that PAHs besides nitrite and TSNAs may play an important role in the tumorigenesis of GC patients in Mizoram, due to the higher level of PAHs, nitrite, and TSNAs exposed in the lifestyle habits.

It is widely accepted that genetic and environmental factors are major etiological factors for GC. In the present study, we investigated the spectrum of mutations in genes *P21*, *MDM2*, and *MLH1* among GC patients. Among the genetic factors, cell cycle regulatory genes besides DNA mismatch repair gene were found to be associated with various cancers including GC (Gartel and Tyner 1999). Cell cycle control is crucial for the normal cell growth and differentiation. The present study also attempted to establish a potential association of alteration in these gene variations with demographic and dietary factors. In this study, we have tried to establish a significant relation between driver gene mutations and epidemiological factor for diffuse type GC. To the best of our knowledge, the present study results are the first report of the association of environmental factors with mutations in cell cycle regulatory genes suggesting the implication of genetic alterations and its correlation with cell cycle in GC development.

The high prevalence of GC in Mizoram has been attributed to peculiar dietary habits (viz. high consumption of smoked meat, salt-preserved foods, dietary nitrite, traditional fermented food, and heavy addiction to various tobacco products and alcohol (Phukan et al. 2005, 2006). The demographic factors play major role in pathogenesis of GC. A hospital-based matched case–control study showed an elevated risk of stomach cancer in case of frequent consumption of sa-um (fermented pork fat) and smoked dried salted meat and fish (Phukan et al. 2006). Smoking as a variable risk factor for

Table 5 Mutations leading to splicing defects

Gene	Mutation	Position	WT branch point motif	Mutant branch point motif	WT CV	Mutant CV	CV variation (%)
p21	Intron 2	36651738G > A	AAGCAGG	AAGCAAG	9.95	39.57	29.62
		36651804A > T	AGCAAG	AGCTAG	61.53	65.52	3.99
		36651825G > T	ATAGTGT	ATATTGT	6.23	21.31	15.08
	3′UTR	36653580C > T	CGCCCAC	TGCCCAC	86.39	84.86	− 1.53
		36653597C > T	CTGCAGT	CTGTAGT	16.53	25.56	9.03
mlh1	Exon 8	37053568A > G	CTCCATC	CTCCGTC	46.15	53.04	6.89
	Intron 16	37088831C > G	TTGACAG	TTGAGAG	47.68	42.74	− 4.94

A new site was created by the mutation; the motif was abolished by the mutation stomach cancer has also been reported from India (Dikshit et al. 2012). In the present study, the results of demographic study indicated that males are more susceptible to GC than females. The putative association between the risk of GC and unique dietary habits has been controversial for decades. In this study, we found that high intake of smoked meat/vegetable (16.214 (2.746–95.749); $p = 0.002$) and oral snuff (10.496 (2.410–45.710); $p = 0.002$) are significant risk factors for the high incidence of GC (Table 1). However, the reason for disparities is still not yet known, although recent research has suggested that genetic factors may be the reason for differences in GC susceptibility of various populations (Yan et al. 2015). Also, family history of other cancers are found to be associated with the increased risk of GC in this population which may be because of their genetic make-up and inheritance of faulty genes which renders them predisposed to cancer (Yaghoobi et al. 2010). The Mizo population is mongoloid in origin and distinct from the rest of India in terms of their diet, lifestyle, and geographical distribution (Ghatak et al. 2013). Also, in other mongoloids like Japanese, daily consumption of meat among women was found to increase the risk of GC by 6.5-fold (Santarelli et al. 2008). In another study, the potential causal role of tobacco was observed in high-risk area of China, where smoking was found to nearly double the risk of transition to gastric dysplasia (Piazuelo and Correa 2013). In the present study, the interaction between identified mutations and the significant demographic factors, smoked meat/vegetable and oral snuff were found to be associated with risk of GC, with a significant association between oral snuff and *P21* 3′UTR mutations similarly association between smoked meat/vegetable and *P21* Intron 2 mutations. Consumption of oral snuff ($p = 0.002$) was significantly associated with GC followed by Tiranga/Gutkha consumptions ($p = 0.073$) (Table 3).

Cell cycle deregulation is common pathway in pathogenesis of human cancer, and alteration of *P21*, the cell cycle regulator, is involved in the development of many human malignancies (Gartel 2005; Lin et al. 2011).The molecular analysis revealed in total five mutations in P21 gene. In the intron 2 of *P21* gene, 36651738G > A, 36651804A > T, and 36651825G > T novel mutations were identified in 5% of total GC samples. Splice site changes were identified as a result of 36651738G > A and 36651825G > T mutation, and it was also predicted that this might affect the protein structural features as the splice site donor is marginally increased. The *P21* CDK inhibitor gene is located at 6q21.2, and its expression has been shown to be regulated largely at the transcriptional level by both p53-dependent and independent mechanisms by a variety of transcription factors that are induced by a number of different signaling pathways (An et al. 2014). Previous studies demonstrated that FOXA2, transcription factor activation of *P21* transcription via direct binding to the P21 promoter and affects the activity of *P21* gene, which results in cell cycle arrest at the G1 phase and inhibition of cell proliferation in p53-deficient cell (Wang et al. 2012). 3′UTR region of *P21* showed two known (previously reported) mutations 36653580C > T in 10% and 36653597C > T 5% GC samples. 36653580C > T polymorphism is thought to cause a functional change in *P21* due to generation of a cryptic spice site by acting after stop codon, and as this polymorphism lies in a crucial region for cell differentiation, proliferation may increase cancer risk by altering messenger RNA stability, which, in turn, may affect protein expression and activity (Campbell et al. 2009). Mutations or single nucleotide polymorphisms (SNPs) in the *P21* gene may result in alteration of P21 expression and/or activity, thereby modulating susceptibility to cancer (Keshava et al. 2002; Gravina et al. 2009; Ma et al. 2011).

In *MLH1* gene, a known exonic variant 37053568A > G (rs1799977) was observed in 10% of the study participants with the replacement of isoleucine to valine in codon 219 (219I > V) in exon 8 (Mathonnet et al. 2003). The polymorphism, I219V (A655G), was reported to be associated with childhood acute lymphoblastic leukemia (Listgarten et al. 2004). Studies also found a significant association between breast cancer and homozygous GG variant (Raptis et al. 2007). The homozygous or heterozygous G allele at nucleotide position 655 in *MLH1* gene was commonly reported for western populations (Christensen et al. 2008; Mann et al. 2008). However, in the current study, it was detected only in 10% of GC patients. In GC patients, the G allele frequency was 10%, higher than in controls which demonstrated a frequency of 0.5%, similar to data reported in Eastern Asians where the G allele frequency is reported to be approximately 2% (Trojan et al. 2002). The nucleotide position 655 is in conserved region thorough all mammals in exon 8. Earlier report published from the result of functional analyses that the homozygous or heterozygous G allele has efficient DNA repair activity (Raevaara et al. 2005; Kondo et al. 2003) and binding properties to PMS2 (Kim et al. 2004). It was well documented that 655A > G is also associated with reduced MLH1 protein expression in sporadic CRCs in Korean population (Marchetti et al. 1995). The HOPE analysis (in silico study) showed that the mutant type is smaller in size than the wild type, which might affect the intramolecular and external interactions due to clashes. 37088831C > G variant in intron 16 of *MLH1* gene was identified in 5% of GC samples. It is found to affect splice site by increasing the acceptor marginally (wt 0.5187/mu 0.5635) which might affect the normal splicing leading to abrupt transcription of this gene.

In the present study, *MDM2* gene had no significant mutations in GC samples. According to the previous studies, mutations and polymorphisms were identified in various exons of *MDM2* gene in esophageal and GC (Sauli et al. 2015). The known variant of *MDM2*, rs2279744 was found to influence independently the susceptibility to GC in Chinese population (Cho et al. 2008). According to Cho et al. (2008), SNPs of

MDM2 gene were not associated with increased GC risk in Korean population, and is consistent with our study. This can be explained due to the difference in the genetic pool and other demographic and dietary factors between the different populations.

The effect of mutations in the splice site was identified by Human Splicing Finder (HSF). We used all the intronic mutations and polymorphism that disturb the active site of 5′ss and 3′ss for validating the splicing effect and new cryptic splice site. The sequence of branch point represents another essential splicing signal. The analysis further revealed that 36653580C > T in 3′UTR of *P21*, branch point is changing with a CV variation of − 1.53% due to which the motif might get abolished leading to splicing defect (Table 5). 36653580C > T polymorphism is thought to cause a functional change in P21, and as this polymorphism lies in a crucial region for cell differentiation, and its proliferation may increase cancer risk by altering messenger RNA stability, which in turn may affect subsequent protein activity. Similarly, for *MLH1* intron 16, 37088831C > G mutation, the CV variation is − 4.94% conferring to splicing defect. Mutations located in the introns of mismatch repair genes can interfere with splicing and cause alternately spliced mRNA transcripts leading to non-functional mismatch repair proteins (Petersen et al. 2013).

According to the MDR analyses, the best model for GC risk in Mizo population is combination of smoked meat/vegetable and oral snuff consumption after performing the gene–environment interaction (Table 1, Fig. 2). Consumption of smoked meat showed the highest independent effect (20.56%) and also had modest synergistic interaction with sa-um (0.43%). Oral snuff (8.19%) also explained considerable entropy independently in GC risk and thus validated the results of gene-environment interaction. In India, many epidemiological studies reported significant positive association of tobacco and dietary habits containing harmful carcinogen such as *N*-nitroso compounds with GC (Sumathi et al. 2009). PAHs generated during preparation of heat-dried smoked food have been significantly associated in different geographical population in the world with high stomach cancer rates (Yeh et al. 2009). Soda-an alkaline preparation, frequently used as food additives, was significantly associated with increased risk of stomach cancer in Mizoram (Phukan et al. 2006). Different tobacco products such as oral snuff and smoking contains high amount of carcinogens like the nitrosamines which acts as cofactor for development of GC (IARC 2007).

We hypothesized that the balance in cell cycle control is disrupted by lifestyle habits (environmental risk factors) leading to the alteration (mutations) in the regulatory genes. These mutations affect the normal cell cycle by inducing abnormal distribution of DNA content that ultimately results in tumorigenesis. The aberrant content of DNA, or aneuploidy, is a hallmark of tumorigenesis (Giam and Rancati 2015). Thus, the flow cytometric study was conducted to evaluate the DNA content and S-phase fraction. Flow cytometric analysis of tumor samples showed 17.5% of aneuploid and 82.5% diploid for gastric tumor samples. A high S-phase fraction (SPF) was observed in GC samples (51.24–72.09) compared to controls (32.45–44.12) (Table 2, Fig. 4). The G2/M phase was significantly arrested in most of the GC tumor samples. In previous studies, DNA aneuploidy has been reported in 40–50% of GC tumors (Brito et al. 1993; Malumbres and Carnero 2003). Flow cytometric analysis predicted DNA aneuploid in 07 (17.5%) and diploid in 33 (82.5%) tumor samples (Table 4) followed by a high S-phase fraction (SPF). Deregulation of cell cycle events leads to uncontrolled cell proliferation and a high S-phase fraction which is a hallmark of GC (Baba et al. 2002). According to an earlier study, the *P21* variant genotypes have been demonstrated to play an important role in cell cycle control. The disruption in cell cycle control due to DNA damage is probably caused by carcinogens present in tobacco-related product (Flejou et al. 1993). Arrest of G2/M phase was observed in case of aneuploidy. DNA aneuploidy has been reported in 40–50% of tumors (Nanus et al. 1989; Quirke et al. 2005). Previous study reported aneuploidy in 76% (25 of 33) of adenocarcinomas arising in the gastric cardia, compared with 30% (8 of 27) of adenocarcinomas arising in the gastric antrum (Gleeson et al. 1998).

Conclusion

In conclusion, our findings indicate that mutations in cell cycle regulatory (*P21*) and mismatch repair (*MLH1*) genes are more predisposed to higher incidence of GC in Mizo population and may play an important role in tumorigenesis by inducing the aberrant distribution of DNA content during different phases of cell cycle in tumor cells. Ethnicity and dietary habits are acting as crucial covariates, suggesting that the mutations have different penetrance according to ethnicity, dietary, and lifestyle habits. This study could afford early detection of patients who are at risk of developing micro- or macroscopic, pathological lesions as well as the introduction of appropriate preventive measures. Due to the complexity as well as the correlations of multiple genetic and environmental factors in the development of GC, large population studies are required in order to overcome the limitation of sample size and encompass virtually all variables including the exposure to environmental factors, ethnic and demographic features besides the association with mutations in genes for DNA repair genes, cell cycle regulatory genes, and cell cycle study.

Funding This work was supported by DBT—Advanced Level State Biotech Hub (BT/04/NE/2009 dt. 29.08.2014), Bioinformatics Infrastructure Facility (No. BT/BI/12/060/2012 (BTISNeT), and DBT eLibrary Consortium (DeLCON) sponsored by the Department of

Biotechnology (DBT), New Delhi, Govt. of India, Mizoram University which provided all the essential facilities to carry out the work.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest to report.

Ethics, consent, and permissions All participants gave written informed consent to the study protocol which was approved by the Ethical Committee of the Civil Hospital, Mizoram and Mizoram University, India (B.12018/1/13-CH(A)/IEC), to conduct and publish the research work. The study protocol was also approved by the Institutional Review Board of all institutes involved in the study.

References

An J-H, ASM J, Kim JW, Kim CH, Choi KH (2014) The expression of P21 is upregulated by forkhead box A1/2 in p53-null H1299 cells. FEBS Lett 588:4065–4070

Appelman HD, Mclaughlin JK, Blot WJ, Fraumeni JF (1992) A cohort study of stomach cancer in a high-risk American population. Cancer 69:2867–2868

Baba H, Korenaga D, Kakeji Y, Haraguchi M, Okamura T, Maehara Y (2002) DNA ploidy and its clinical implications in GC. Surgery 131(1 Suppl):S63–S70

Bahl R, Arora S, Nath N, Mathur M, Shukla NK, Ralhan R (2000) Novel polymorphism in P21(waf1/cip1) cyclin dependent kinase inhibitor gene: association with human esophageal cancer. Oncogene 19:323–328

Boland CR, Goel A (2010) Microsatellite instability in colorectal cancer. Gastroenterology 138(6):2073–2087

Bouska A, Lushnikova T, Plaza S, Eischen CM (2008) MDM2 promotes genetic instability and transformation independent of p53. Mol Cell Biol 28:4862–4874

Brito MJ, Filipe MI, Williams GT, Thompson H, Ormerod MG, Titley J (1993) DNA ploidy in early gastric carcinoma (T1): a flow cytometric study of 100 European cases. Gut 34:230–234

Campbell PT, Curtin K, Ulrich CM, Samowitz WS, Bigler J, Velicer CM, Caan B, Potter JD, Slattery ML (2009) Mismatch repair polymorphisms and risk of colon cancer, tumour microsatellite instability and interactions with lifestyle factors. Gut 58:661–667

Campos F, Carrasquilla G, Koriyama C, Serra M, Carrascal E, Itoh T, Nomoto M, Akiba S (2006) Risk factors of GC specific for tumor location and histology in Cali, Colombia. World J Gastroenterol 12:5772–5779

Cattaert T, Calle ML, Dudek SM, Mahachie JJM, Van-Lishout F et al (2011) Model-based multifactor dimensionality reduction for detecting epistasis in case–control data in the presence of noise. Ann Hum Genet 75(1):78–89

Chao A, Thun MJ, Henley SJ, Jacobs EJ, Mccullough ML, Calle EE (2002) Cigarette smoking, use of other tobacco products and stomach cancer mortality in us adults: the cancer prevention study II. Int J Cancer 101:380–389

Chen C, Pearson AM, Gray JI (1990) Meat mutagens. Adv Food Nutr Res 34:387–449

Cho YG, Choi BJ, Song JH, Kim CJ, Cao Z, Nam SW, Lee JY, Park WS (2008) No association of MDM2 T309G polymorphism with susceptibility to Korean GC patients. Neoplasma 55:256–260

Choudhury JH, Singh SA, Kundu S, Choudhury B, Talukdar FR, Srivasta S et al (2015) Tobacco carcinogen-metabolizing genes CYP1A1, GSTM1, and GSTT1 polymorphisms and their interaction with tobacco exposure influence the risk of head and neck cancer in northeast Indian population. TumorBiol 36(8):5773–5783

Christensen LL, Madsen BE, Wikman FP, Wiuf C, Koed K, Tjonneland A (2008) The association between genetic variants in hMLH1 and hMSH2 and the development of sporadic colorectal cancer in the Danish population. BMC Med Genet 9:52–63

Ciccarelli C, Marampon F, Scoglio A, Mauro A, Giacinti C, Cesaris PD (2005) P21WAF1 expression induced by MEK/ERK pathway activation or inhibition correlates with growth arrest, myogenic differentiation and onco-phenotype reversal in rhabdomyosarcoma cells. Mol Cancer 13(4):14

Correa P, Fontham E, Pickle LW, Chen V, Lin YP, Haenszel W (1985) Dietary determinants of GC in south Louisiana inhabitants. J Natl Cancer Inst 75:645–654

Curado MP, Edwards B, Shin HR, Storm H, Ferlay J, Heanue M, et al. (eds) (2007) Cancer incidence in five continents, vol. IX, IARC Scientific Publications No. 160. IARC, Lyon

Decesse JT, Medjkane S, Datto MB, Crémisi CE (2001) RB regulates transcription of the P21/WAF1/CIP1 gene. Oncogene 20(8):962–971

Dikshit R, Gupta PC, Ramasundarahettige C, Gajalakshmi V, Aleksandrowicz L, Badwe R (2012) Million death study collaborators. Cancer mortality in India: a nationally representative survey. Lancet 379:1807–1816

Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M (2015) Cancer incidence and mortality worldwide, sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136:359–386

Flejou JF, Potet F, Muzeau F, Le-Pelletier F, Fekete F, Henin D (1993) Overexpression of p53 protein in Barrett's syndrome with malignant transformation. J Clin Pathol 46:330–333

Furberg H, Lichtenstein P, Pedersen NL, Bulik C, Sullivan PF (2006) Cigarettes and oral snuff use in Sweden: prevalence and transitions. Addiction 101(10):1509–1515

Gartel AL (2005) The conflicting roles of the CDK inhibitor P21(CIP1/WAF1) in apoptosis. Leuk Res 29:1237–1238

Gartel AL, Tyner AL (1999) Transcriptional regulation of the P21 (WAF/CIP1) gene. Exp Cell Res 246:280–289

Ghatak S, Muthukumaran RB, Nachimuthu SK (2013) A simple method of genomic DNA extraction from human samples for PCR-RFLP analysis. J Biomol Tech 24:224–231

Ghatak S, Yadav RP, Lalrohlui F, Chakraborty P, Ghosh S, Ghosh S (2016) Xenobiotic pathway gene polymorphisms associated with GC in high risk Mizo-mongoloid population, northeast India. Helicobacter 21:523–535

Ghatak S, Zothansanga PJL, Nachimuthu SK (2014) Co-extraction and PCR based analysis of nucleic acids from formalin-fixed paraffin-embedded specimens. J Clin Lab Anal 29(6):485–492

Giam M, Rancati G (2015) Aneuploidy and chromosomal instability in cancer: a jackpot to chaos. Cell Div 10:3

Gleeson CM, Sloan JM, McManus DT, Maxwell P, Arthur K, McGuigan JA (1998) Comparison of p53 and DNA content abnormalities in adenocarcinoma of the oesophagus and gastric cardia. Br J Cancer 77(2):277–286

Gooding C, Clark F, Wollerton MC, Grellscheid SN, Groom H, Smith CW (2006) A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. Genome Biol 7:23–31

Gravina S, Lescai F, Hurteau G, Brock GJ, Saramaki A, Salvioli S (2009) Identification of single nucleotide polymorphisms in the P21 (CDKN1A) gene and correlations with longevity in the Italian population. Aging (Albany NY) 1:470–480

Guerrette S, Acharya S, Fishel R (1999) The interaction of the human MutL homologues in hereditary nonpolyposis colon cancer. J Biol Chem 274:6336–6341

Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19(3):376–382

Hecht SS (2003) Tobacco carcinogens, their biomarkers and tobacco-induced cancer. Nat Rev Cancer 3:733–744

Hedley DW (1994) DNA analysis from paraffin-embedded blocks. Flow cytometry Second Edition, Part A PP: 231–240

Igaki H, Sasaki H, Kishi T, Sakamoto H, Tachimori Y, Kato H (1994) Highly frequent homozygous deletion of the p16 gene in esophageal cancer cell lines. Biochem Biophys Res Commun 203:1090–1095

Ihsan R, Devi TR, Yadav DS, Mishra AK, Sharma J, Zomawia E (2011) Investigation on the role of p53 codon 72 polymorphism and interactions with tobacco, betel quid, and alcohol in susceptibility to cancers in a high-risk population from north east India. DNA Cell Biol 30:163–171

Jung YS, Qian Y, Chen X (2010) Examination of the expanding pathways for the regulation of P21 expression and activity. Cell Signal 22(7): 1003–1012

Keshava C, Frye BL, Wolff MS, McCanlies EC, Weston A (2002) Waf-1 (P21) and p53 polymorphisms in breast cancer. Cancer Epidemiol Biomark Prev 11:127–130

Kim JC, Roh SA, Koo KH, Ka IH, Kim HC, Yu CS (2004) Genotyping possible polymorphic variants of human mismatch repair genes in healthy Korean individuals and sporadic colorectal cancer patients. Fam Cancer 3:129–137

Kneller RW, Guo WD, Hsing AW, Chen JS, Blot WJ, Li JY et al (1992) Risk factors for stomach cancer in sixty-five Chinese counties. Cancer Epidemiol Biomark Prev 1:113–118

Kondo E, Horii A, Fukushige S (2001) The interaction domains of three MutL heterodimers in man: hMLH1 interacts with 36 homologous amino acid residues within hMLH3, hPMS1 and hPMS2. Nucleic Acid Res 29:1695–1708

Kondo E, Suzuki H, Horii A, Fukushige S (2003) A yeast two-hybrid assay provides a simple way to evaluate the vast majority of hMLH1 germ-line mutations. Cancer Res 63:3302–3308

Lin G, Fang F, Yu XJ, Yu L (2011) Meta-analysis of the relationship between P21 Ser31Arg polymorphism and lung cancer susceptibility. Genet Mol Res 10:2449–2456

Lipkin SM, Wang V, Jacoby R, Banerjee-Basu S, Baxevanis AD, Lynch HT (2000) MLH3: a DNA mismatch repair gene associated with mammalian microsatellite instability. Nat Genet 24:27–35

Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A (2004) Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clin Cancer Res 10:2725–2737

Ma H, Zhou Z, Wei S, Wei Q (2011) Association between P21 Ser31Arg polymorphism and cancer risk: a meta-analysis. Chin J Cancer 30: 254–263

Madathil S, Senthil Kumar N, Zodinpuii D, Muthukumaran RB, Lalmuanpuii R, Nicolau B (2018) TUIBUR: tobacco in a bottle—commercial production of tobacco smoke-saturated aqueous concentrate. Addiction 113:577–580. https://doi.org/10.1111/add.14117

Malakar M, Devi KR, Phukan RK, Kaur T, Deka M, Lalhriat-Puia L (2012) Genetic polymorphism of glutathione S-transferases M1 and T1, tobacco habits and risk of stomach cancer in Mizoram, India. Asian Pac J Cancer Prev 13(9):4725–4732

Malumbres M, Carnero A (2003) Cell cycle deregulation: a common motif in cancer. Prog Cell Cycle Res 5:5–18

Mann A, Hogdall E, Ramus SJ, DiCioccio RA, Hogdall C, Quaye L (2008) Mismatch repair gene polymorphisms and survival in invasive ovarian cancer patients. Eur J Cancer 44:2259–2265

Manuguerra M, Matullo G, Veglia F, Autrup H, Dunning AM, Garte S (2007) Multi-factor dimensionality reduction applied to a large prospective investigation on gene-gene and gene-environment interactions. Carcinogenesis 28(2):414–422

Marchetti A, Buttitta F, Pellegrini S, Merlo G, Chella A, Angeletti C (1995) MDM2 gene amplification and overexpression in non-small cell lung carcinomas with accumulation of the p53 protein in the absence of p53 gene mutations. Diagn Mol Pathol 4:93–97

Mathonnet G, Krajinovic M, Labuda D, Sinnett D (2003) Role of DNA mismatch repair genetic polymorphisms in the risk of childhood acute lymphoblastic leukaemia. Br J Haematol 123:45–48

Merkel DE, Mcguire WL (1990) Ploidy, proliferative activity and prognosis. Cancer 65:1194–1205

Mousses S, Ozcelik H, Lee PD, Malkin D, Bull SB, Andrulis IL (1995) Two variants of the CIP1/WAF1 gene occur together and are associated with human cancer. Hum Mol Genet 4:1089–1092

Nanus DM, Kelsen DP, Niedzwiecki D, Chapman D, Brennan M, Cheng E (1989) Flow cytometry as a predictive indicator in patients with operable GC. J Clin Oncol 7(8):1105–1112

National Cancer Registry Programme (2013) Three-year report of the population based cancer registries 2011–2013. National cancer registry programme, Indian Council of Medical Research (ICMR), Bangalore, India. Available from, http://www.pbcrin-dia.org. Accessed 16 Oct 2014

Nomura A, Grove JS, Stemmermann GN, Severson RK et al (1990) A prospective study of stomach cancer and its relation to diet, cigarettes, and alcohol consumption. Cancer Res 50:627–631

Oren M, Damalas A, Gottlieb T, Michael D, Taplick J, Leal JF (2002) Regulation of p53: intricate loops and delicate balances. Biochem Pharmacol 64:865–871

Parker SB, Eichele G, Zhang P, Rawls A, Sands AT, Bradley A (1995) p53-independent expression of P21Cip1 in muscle and other terminally differentiating cells. Science 267(5200):1024–1027

Petersen SM, Dandanell M, Rasmussen LJ, Gerdes AM, Krogh LN, Bernstein I (2013) Functional examination of MLH1, MSH2, and MSH6 intronic mutations identified in Danish colorectal cancer patients. BMC Med Genet 14:103–114

Phillips DH (1999) Polycyclic aromatic hydrocarbons in the diet. Mutat Res 443:139–147

Phukan RK, Hazarika NC, Baruah D, Mahanta J (2004) High prevalence of stomach cancer among the people of Mizoram, India. Curr Sci 87: 285–286

Phukan RK, Narain K, Zomawia E, Hazarika NC, Mahanta J (2006) Dietary habits and stomach cancer in Mizoram, India. J Gastroenterol 41:418–424

Phukan RK, Zomawia E, Narain K, Hazarika NC, Mahanta J (2005) Tobacco use and stomach cancer in Mizoram, India. Cancer Epidemiol Biomark Prev 14:1892–1896

Piazuelo MB, Correa P (2013) Gastric cáncer: overview. Colomb Med (Cali) 44(3):192–201

Quirke DP, Dixon MF, Clayden AD, Durdey P, Dyson JED, Williams NS (2005) Prognostic significance of DNA aneuploidy and cell proliferation in rectal adenocarcinomas. J Pathol 151(4):285–291

Raevaara TE, Korhonen MK, Lohi H, Hampel H, Lynch E, Lönnqvist KE (2005) Functional significance and clinical phenotype of nontruncating mismatch repair variants of MLH1. Gastroenterology 129:537–549

Raptis S, Mrkonjic M, Green RC, Pethe VV, Monga N, Chan YM (2007) MLH1 -93G>A promoter polymorphism and the risk of microsatellite-unstable colorectal cancer. J Natl Cancer Inst 99:463–474

Ritchie MD, Hahn LW, Moore JH (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24(2):150–157

Santarelli RL, Pierre F, Corpet DE (2008) Processed meat and colorectal cancer: a review of epidemiologic and experimental evidence. Nutr Cancer 60(2):131–144

Sauli E, Hongna L, Vedastus AK, Weiyue H, Song L, Nongyue H (2015) Polymorphisms in NEIL-2, APE-1, CYP2E1 and MDM2 genes are independent predictors of GC risk in a northern Jiangsu population (China). J Nanosci Nanotechnol 15(7):4815–4828

Shiohara M, Deiry WE, Wada M, Nakamaki T, Takeuchi S, Yang R (1994) Absence of WAF1 mutations in a variety of human malignancies. Nature 84:3781–3784

Stepanov I, Hecht SS, Ramakrishnan S, Gupta PC (2005) Tobacco-specific nitrosamines in smokeless tobacco products marketed in India. Int J Cancer 116(1):16–19

Stepanov I, Villalta PW, Knezevich A, Jensen J, Hatsukami DK, Hecht SS (2010) Analysis of 23 polycyclic aromatic hydrocarbons in smokeless tobacco by gas chromatography–mass spectrometry. Chem Res Toxicol 23(1):66–73

Sumathi B, Ramalingam S, Navaneethan U, Jayanthi V (2009) Risk factors for GC in South India. Singap Med J 50:147–151

Torre L, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A (2015) Global Cancer Statistics, 2012. CA Cancer J Clin 65:87–108

Trojan J, Zeuzem S, Randolph A, Hemmerle C, Brieger A, Raedle J (2002) Functional analysis of hMLH1 variants and HNPCC-related mutations using a human expression system. Gastroenterology 122:211–219

Tsugane S, Sasazuki S (2007) Diet and the risk of GC: review of epidemiological evidence. Gastric Cancer 10:75–83

Tsugane S, Sasazuki S, Kobayashi M, Sasaki S (2004) Salt and salted food intake and subsequent risk of GC among middle-aged Japanese men and women. Br J Cancer 90:128–134

Wang X, Wang J, Huang V, Place RF, Li LC (2012) Induction of NANOG expression by targeting promoter sequence with small activating RNA antagonizes retinoic acid-induced differentiation. Biochem J 443:821–828

Ward MH, Lopez-Carrillo L (1999) Dietary factors and the risk of GC in Mexico City. Am J Epidemiol 149:925–392

Watanabe H, Fukuchi K, Takagi Y, Tomoyasu S, Tsuruoka N, Gomia K (1995) Molecular analysis of the Cip1/Waf1 (P21) gene in diverse types of human tumors. Biochim Biophys Acta 1263:275–280

Yaghoobi M, Bijarchi R, Narod SA (2010) Family history and the risk of GC. Br J Cancer 102(2):237–242

Yan L, Fang L, Sha Z, Suqing S, Li L, Lifeng L (2015) Genetics and GC susceptibility. Int J Clin Exp Med 8(6):8377–8383

Yeh JM, Kuntz KM, Ezzati M, Goldie SJ (2009) Exploring the cost-effectiveness of *Helicobacter pylori* screening to prevent GC in China in anticipation of clinical trial results. Int J Cancer 124:157–166

ABSTRACT

IDENTIFICATION OF RECURRENT GENOMIC ALTERATIONS IN GASTRIC ADENOCARCINOMA IN MIZO POPULATION

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

PAYEL CHAKRABORTY

MZU REGN NO.: 1600803

PH.D REGN NO.: MZU/PH.D./1045 OF 26.05.2017

DEPARTMENT OF BIOTECHNOLOGY

SCHOOL OF LIFE SCIENCES

FEBRUARY 2021

ABSTRACT

Worldwide stomach cancer occupies fifth position according to incidence wise and is recorded as third most lethal cancer according to the mortality rate (Bray et al. 2018). National Cancer Registry Programme in India has reported stomach cancer as the third most prevalent cancer among males, according to incidence and it is the fourth most prevalent cancer in the North Eastern region of India (Mathur et al. 2020). Mizoram, a north eastern state of India, recorded the highest occurrence rate of Gastric Cancer in India (Mohammed et al. 2017) and globally holds fifth position (Phukan et al. 2004).

In Lauren classification, GC can be classified into two main types: i) intestinal and ii) diffuse (Lauren et al. 1965). Besides Lauren classification, WHO classified GC into four main types: i) Tubular adenocarcinomas, ii) Papillary adenocarcinomas, iii) Mucinous adenocarcinomas and iv) Signet-ring cell carcinomas (Lauwers et al. 2010). American joint committee on cancer (AJCC), 8th edition on cancer staging classified Pathological tumor, node and metastasis information (pTNM) in four stages: I, II, III and IV. AJCC also classified grading as: well differentiated, moderately differentiated and poorly differentiated types (Lauwers et al. 2010).

Etiologically, Gastric cancer is a heterogeneous disease which develops due to multitude of risk factors like environmental factors, *H. pylori* infection, diet, smoking, alcohol drinking and genomic as well as epigenetic alterations (Patrick Tan et al. 2015). Epstein Bar Virus (EBV) is also associated with GC development in 5-10% cases (Shinozaki-Ushiku et al. 2015). Salt and salted food specially salted fish, pickled vegetables, cured meat and other salt-preserved are always the major risk factors for atrophic gastritis (Tsugane et al. 2007). Consuming smoked food in excess amount is also a risk factor for GC (Wu et al. 2013; Strumylaite et al. 2006). In several studies, it has reported that tobacco consumption or smoking is also a risk factor for GC (Ladeiras

et al. 2008). In some studies, it has been recorded that alcohol consumption is also a risk factor for developing GC (Phukan et al. 2005; Steevens et al. 2010; Verma et al. 2012). Mizo people have their own unique food (smoked and fermented) and smokeless tobacco (tuibur) habits which can increase the risk of developing Gastric Cancer. Saum, is a fermented pork fat, and was reported as a habitat for pathogens which may affect human health (Mandal et al. 2018). Tuibur, tobacco infused water has been reported as risk factor for Gastric Cancer in many studies (Phukan et al. 2005; Mukherjee et al. 2020).

Helicobacter pylori (H. pylori), a class I carcinogen has been considered as a significant environmental risk factor for developing Gastric Cancer (IARC 1994; Steven et al. 2007; Lu and Li, 2014). Studies has proved extra salt consumption in presence of *H. pylori* infection make the condition favourable for developing Gastric Cancer (Loh et al. 2007; Zhang et al. 2017). *H. pylori* has two virulence factors: cytotoxin-associated gene A (CagA) and vacuolating cytotoxin A (VacA). Epstein Bar Virus is a high risk factor for Cancers (Claire et al. 2019), especially with diffuse type of Gastric cancer (10%) (Gonzalo et al. 2017). Epstein Bar Virus (EBV) infection is very common in worldwide (90%), but very rarely infects epithelial cells (Alexandre et al. 2016). It has been noticed that in EBV infected GCs, methylation of CpG Island in cancer associated genes is a prime feature of this subgroup due to the expression of viral protein LMP2A (Kaneda et al. 2012).TCGA and Asian Cancer Research Group (ACRG) groups have categorised MSI and EBV as molecular subtype of GC (The Cancer Genome Atlas Research Network, 2014; Cristescu et al. 2015). Studies have reported that MSI-H cases have significant association with overall survival rate of Gastric Cancer (Zhu et al. 2015).

Hereditary GC can be related with three main syndromes: hereditary diffuse gastric cancer (HDGC), gastric adenocarcinoma and proximal polyposis of the stomach (GAPPS), and familial intestinal gastric cancer (FIGC) (Lauren et al. 1965). Out of these three syndromes, germline *CDH1* mutations are related with HDGC and *CTNNA1* also reported as significant for this type. Germline mutation of *APC* gene has association with GAPPS and molecular characterization of FIGC syndromes are not yet understood properly. Other cancer associated syndromes which are associated with GC are as follows: Lynch (genes related to this syndrome are *MLH1*, *MS2*, *MSH6*, *PMS2* and *EPCAM*), LiFraumeni (TP53), Peutz-Jeghers (*STK11*), hereditary breast–ovarian cancer syndromes (*BRCA1* and *BRCA2*), familial adenomatous polyposis (*APC*), and juvenile polyposis (*BMPR1A* and *SMAD4*) (Lauren et al. 1965; Bryson W et al. 2017).

TCGA and ACRG identified distinct molecular subtypes of Gastric Cancer by NGS technology on the basis of somatic mutations. TCGA study has described four distinct molecular classification associated with GC. In case of MSI subgroup: *PIK3CA, TP53, PTAN, KRAS, ERBB3* and *ARID1A* were the hyper-mutated genes. EBV (+) subgroup had significant mutations in *PIK3CA, ARID1A* and *BCOR*. Hyper mutation was observed in *ARID1A, RHOA* and *CDH1*a genes in genomically stable (GS) cases. TP53 gene was mutated in case of chromosomal instability (CIN) cases (The Cancer Genome Atlas Research Network, 2014). ACRG group again reported four subgroups associated with GC. In case of MSI subgroup, the hypermutated genes were common like TCGA study (*PIK3CA, KRAS* and *ARID1A*). *PIK3CA, KRAS, ARID1A* and *APC* were highly mutated in MSS/TP53^{+} subgroup. *ARID1A* gene was mutated in MSS/EMT subgroup and *TP53* were mutated in MSS/TP53^{-} subgroup (Cristescu et al. 2015).

Immunohistochemistry (IHC) is a staining method of Formalin fixed Paraffin embedded (FFPE) tissues extensively used in Pathology lab for obtaining histological information from cancer tissue. IHC helps to get into deep in the area of tumour classification, pathology, multi-lineage expression, pathogenic infection status and disease progression. Further for developing biomarkers, IHC is a commonly used technique by which the behaviour and progression of tumor can be understand easily, which in turn can provide information on the biological behaviour and prognosis of a tumor. This present study was used BAX, TP53, ERBB2 and ERCC1 proteins to check their expression on GC patients in this population.

In this present study, we hypothesized that high incidence of Gastric Cancer in Mizo population might be due to the effect of environmental risk exposure including unique dietary habits and lifestyle factors along with pathogen (*H. pylori* and EBV) infection. Furthermore, as the Mizo tribal population is homogeneous, a unique set of driver genes with pathogenic alterations may play a role to initiate the progression of Gastric cancer. Very few studies have been reported from India to understand the Gastric Cancer genomics and from Mizoram as well. This study is important to find out the significant etiological factors and prediction of driver gene alterations related to GC in this population to find out whether "the population is genetically predisposed with pathogenic mutation related to GC?". This results obtained from this study can be translated to clinical field for therapeutic improvement for this high risk Gastric Cancer population.

Materials method

Sample Description

The ethical committees of Civil Hospital, Aizawl (B.12018/1/13-CH(A)/IEC dtd. 18/04/2014), and Human Ethical Committee, Mizoram University (MZU/IHEC/2015/008 dtd. 14/12/15) approved the study. Samples (Tumor tissues, blood samples and paraffin embedded blocks) from 80 patients were collected from four different hospitals: Civil Hospital Aizawl, Ebenezer Hospital, Aizawl Hospital and Green Wood Hospital, Aizawl during September 2016 to January 2019. A total of 160 controls (79 males and 81 females) were randomly selected from the same ethnic group from where the patients were selected and belong with an almost similar age range.

The inclusion and exclusion criteria for selecting patient samples to conduct the study were:

Subject inclusion criteria

- Patients with Gastric Cancer and without any pre-treatment for cancer were included.
- Cases clinically diagnosed by oncologist and confirmed by pathologist.
- Samples were collected only from Mizo ethnic tribe.

Subject exclusion criteria

- Gastric cancer patients with other chronic diseases were excluded.
- Patients who were pre-treated for any other type of cancer were excluded.

Data collection

A well-designed and informative questionnaire was collected from each participant with a duly informed consent form. The patient group and healthy controls were interviewed by a telephonic interview for the follow-up study. The questionnaire was included lifestyle habits: smoking, chewing tobacco in smokeless form, tuibur or tobacco infused water and alcohol The questionnaire also had detailed information on food habits such as: extra salt intake, smoked food and sa-um or fermented pork fat. All the factors were categorized as consumers and non-consumers. The excess body weight [body mass index (BMI) ≥25] was categorised as obese.

DNA isolation from Tumor Tissue and Blood samples

Genomic DNA was extracted from the tissue using commercially available QIAamp® DNA Tissue Kit and DNA was extracted from blood samples using commercially available QIAamp® Blood DNA mini kit. Genomic DNA from tissue and blood was also isolated by conventional method using the phenol-chloroform method according to Ghatak et al. (2013). DNA visualization was done in electrophoresis by using 0.8% agarose gel and quantification was done by using Picogreen dye in Qubit 2.0 Fluorimeter (Invitrogen).

Pathogen Genotyping

Detection of *Helicobacter pylori* infection in GC patients was performed by PCR amplification of specific 16SrRNA region and *UraC* gene. Genotyping of *H. pylori* was done by PCR amplification of CagA and VacA genes. The detection and genotyping of *Epstein Barr Virus* (*EBV*) type1/ type 2 infections was determined by using a standard PCR assay of EBNA3C - Epstein–Barr virus nuclear antigen 3C gene using distinct primer sets according to Fassone et al. (2000).

PCR amplification of microsatellite loci

The determination of MSI/MSS associated GC cases were carried out by allele comparison of the mononucleotide repeat markers BAT-25, BAT-26, and dinucleotide repeat markers D2S123, D17S250, D16S752, D16S265, D16S398, D16S496, D18S58, and D16S3057 (Suraweera et al. 2002; Sarrio et al. 2003; Losso et al. 2012; Pećina-Šlaus et al. 2017; Forster et al. 2018) in tumor and corresponding blood samples and also in healthy control blood samples.

Fragment Analysis

Fragment analysis was performed using the Automated ABI sequencer model 3500 Genetic Analyzer (Applied Biosystems, Singapore) to analyze the amplified loci. In brief, 8.7 µl deionized formamide was combined with 0.3 µl GeneScan^{Tm}-600 size standards (Applied Biosystems, V-2.0) and 1 µl PCR product in a Genetic Analyzer sample plate.

Targeted re-sequencing approach to find out driver gene alterations

Forty-eight patients were selected for targeted re-sequencing based on the pathogen and MSI status. Among them 42% (20) and 65% (31) of patients were found to be infected with *H. pylori* and EBV, respectively and 42% (20) of patients were Microsatellite Instable. Paired tumor and blood samples were used for sequencing. A panel of 60 genes of 284.262 kb region size was designed with 401.060 kb probes size and 100% converge by Agilent SureSelect to cover the interested region of panel genes.

Wet lab method of NGS sequencing

The method employed by the Agilent Sure Select™ Target Enrichment System extracts target regions from genomic libraries by hybridization to in-solution biotinylated cRNA probes, or "baits". This hybrid capture-based library preparation helps in elimination of amplification and sequencing artifacts that limit the sensitivity of sequencing. Capture hybrids of this panel of genes and paired tumour and blood DNA samples from each patient were amplified, pooled and sequenced in HiSeq-2500 (Illumina). A mean coverage depth of 1000X was achieved for GC tumor DNA, and 600X for matched normal blood cells. Data were analysed for finding both somatic and germline variants.

Bioinformatics pipeline for analyzing somatic variants

The sequences reads obtained were mapped to hg19 reference sequence with BWA MEM aligner. Variant calling was done by 2 variant callers, VarScan2 (Koboldt et al. 2012) and Base by Base (BBB) in house (NIBMG) developed pipelines (India Project Team of the International Cancer Genome Consortium. 2013). Both the vcf files were annotated by CRAVAT annotation tool (Douville et al. 2013).Then, union of coding variants of BbB and Verscan 2 were considered and three filters were applied: i) removal of somatic variants with VAFs ≤ 0.05 (Tumor) or ≥ 0.02 (Blood). ii) Selection of variants =< 0.01 allele frequency in 1000 genome database, and iii) exclusion of synonymous variants, respectively to get the discovery set.

Bioinformatics pipeline for analysing germline variants

The sequences reads obtained were mapped to hg19/GRCH37 reference genome using BWA-MEM. Sequence and variant calls were identified using GATK v3.8.0 suite's Haplotype Caller and annotation was done by ANNOVAR database (Wang et al. 2010). After annotation, five filters were applied to get unique variants for the study population. First, only the exonic variants were selected and the Second filter was to discard off target genes beyond the gene panel. Then, the third filter was to find out the unique variants (by excluding the common variants in other populations) by selecting variants with ≤ 0.01 allele frequency in 1000 Genomes. Then we have excluded the synonymous variants (Figure 6). Lastly, the variants which were present (mutation) in all the patients were excluded, as it is a germ line analysis (Suzuki et al. 2020).

Whole Exome Sequencing (WES)

Whole exome sequencing was done for 37 patient samples and 4 healthy controls. Seventeen samples were taken from previous batch of targeted re-sequencing and 20 new samples were selected for this analysis. Paired-end sequencing was performed for matched blood and tumour samples on Illumina Hiseq-2500 at an average depth of 90 X. BWA-MEM was used for alignment and mapping of reads with hg19 reference genome. GATK v3.8.0 suite's Haplotype Caller was used for variant calling (Poplin et al. 2017). The variants were annotated by ANNOVAR tool (Wang et al. 2010). After annotation, five filters were applied to get unique variants for the study population. Only the exonic variants were selected by applying first filter. Second filter was to exclude the common variants in other populations by selecting variants with ≤ 0.01 allele frequencies in 1000 Genomes to find out the unique variants of the population. The synonymous variants and the variants which were present in healthy controls were excluded (Figure 7). Finally, the variants which were present (mutation) in all the patients were excluded as it is a germ line analysis (Suzuki et al. 2020).

Pathogenicity prediction

Prediction of pathogenicity of known variants was done by ClinVar (Landrum et al. 2014) and COSMIC database (Forbes et al. 2008). Prediction of novel missense variants were done by Mutation taster (Schwarz et al. 2014), Polyphen 2 (Adzhubei et al. 2010), PROVEAN (Choi et al. 2012), and PANTHER (Thomas et al. 2003). Variants were classified as i) pathogenic and ii) benign.

Copy Number Variation Analysis

The copy number variation (**CNV**) is defined as the variation in the number of copies of a particular gene from one individual to the other. As every gene has two copies, there will be a change in copy number if there is a duplication or deletion. Seventeen (17) samples were selected for this analysis on the basis of mutation data derived from targeted re-sequencing using Droplet Digital PCR (ddPCR). Two genes were targeted, ERBB2 (Oncogene) and TP53 (Tumour suppressor gene), to compare the copy number status with mutated patient samples and EFTUD2 was used as reference gene.

Protein Expression study using Immunohistochemistry (IHC)

This application was applied to see the expression of apoptotic and cell proliferating gene BAX (ab32503), TP53 (ab80645), ERBB2 (D8F12) XP – 4290T and ERCC1 (D6G6) XP – 12345T in tumor and adjacent normal tissues. TP53 was raised in mouse, while BAX ERBB2 and ERCC1 were rabbit monoclonal antibody. Two types of secondary antibody [Anti mouse, HRP linked Secondary antibody- 7076P2 and HRP Rabbit (8114S), Cell Signaling] was used in this study.

Statistical Analysis

The association of demographic factors among case–control subjects was tested for Hardy–Weinberg equilibrium by a chi-square test with one degree of freedom (df) (Gunathilake et. al. 2018). Non-parametric T test was also performed. The odd ratio (OR) and 95% confidence intervals (CIs) were estimated for determining association in each group of factors among case-control subjects and among each subgroup and factors by binary logistic regression (Univariate and Multivariate analysis) (Denis et al. 2018). For all tests, a two-sided p-value <0.05 was considered statistically significant. Then, the independent impact of hazard components was explored in a multivariate model keeping only those statistically significant or demonstrating a confounding effect on the contemplated elements. Overall survival was determined using the Cox proportional-hazards regression model (using three years cut-off). The log-rank test, Kaplan-Meier survival analysis was used to assess the impact of the variables on survival rate (Moghimi-Dehkordi et al. 2009). All the analysis was performed using SPSS 20 software.

Results

In this study, the age range from 40-69 years exhibited the highest number of GC patients (75%), and male patients (66.25%) were more prone to GC than females. A family history of any type of cancer in the first-degree relative was found in 32.5% of patients. The distal part of the stomach was reported to have highest tumor cases (73.75%). Out of the total 80 GC patients, 50% of the cases were found in stage III, well-differentiated cases were found in 8.75% of the patient, 46.25% were moderately differentiated and poorly differentiated cases were found in 32.5% of the patient.

In chi square distribution extra salt consumption was the highest significant risk factor (*p* value < 0.0001) followed by smoked food consumption (*p* value = 0.01), smoking (*p* value < 0.0001) and alcohol drinking (*p* value < 0.0001) and they are the high risk factors for developing GC.The univariate binary logistic regression analysis was performed for sex, BMI, dietary and lifestyle habits. Sex (*p*-value = 0.019) and BMI (*p*-value = 0.0001) were significant factors for the gastric cancer patients. Among the dietary factors, extra salt consumption (*p*-value = 0.007), smoked food consumption (*p*-value = 0.0001), Smokeless tobacco (tuibur) intake (*p*-value = 0.011), smoking (*p*-value = 0.0001) and alcohol consumption (*p*-value = 0.0001) are the major significant risk factors for the GC.Further, multivariate analysis was performed and five factors were predicted as significantly associated with GC risk with high OR and 95% CI in multivariate analysis. BMI (*p*-value = 0.0001), Extra salt consumers (*p*-value = 0.042), smoked food consumers (*p*-value = 0.001), smokers (*p*-value = 0.0007) and alcohol drinkers (*p*-value = 0.001) were the high-risk groups associated with GC development.

A risk score was estimated with the five factors using a logistic model and validated in the GC clinical cohort (Stage I, N = 20; Stage II, N = 14; Stage III, N = 44; Stage IV, N = 2) with the healthy controls. The exposer of five-panel epidemiological factors might be successful in predicting the GC risk with different early symptoms (area under the curve – AUC = 0.91; *p*-value < 0.0001). This five-panel epidemiological factor achieved high-risk score with significant-high positive probability values for GC patients with high sensitivity (79.45%) and specificity (91.72%). For predicting GC at early-stage, a risk score was estimated with the same 5 factors using a logistic model and was validated in the early stage (Stage I, N = 20 and II, N = 14) GC clinical cohort with the healthy control. The exposer of five-panel epidemiological factors (BMI, extra salt consumption, smoked food, alcohol drinking, and smoking) might be successful in predicting the GC risk during the premalignant stage with different early symptoms with higher AUC value (0.946; *p*-value < 0.0001). This 5-panel epidemiological factor achieved high-risk core with significant-high positive probability values for GC patients with high sensitivity (96.67%) and specificity (80.89%). This significant panel of epidemiological factors can be used to detect GC patient at early stage by counseling and proper public health practices.

Screening and Genotyping of *H. pylori* and EBV was done for 80 patients. Out of 80 samples, 71 (88.75%) cases were positive for the pathogens and 9 (11.25%) of them were negative for pathogens. EBV positive cases were 32 (40%), 50 (63%) were detected positive for *H. pylori* and, 11 (13.75%) were positive for both the pathogen. Out of 50 *H. pylori* positive cases (Figure 14), 46 cases were CagA, 17 were VacA, and 13 were both positive for both the genotypes (Figure 15 and 16). Out of 32 EBV cases, 29 were Type I, 7 were Type II positive and, 4 of them were having both the genotypes. In case of MSI analysis, PCR amplification was done for each marker and the representative gel images is given. After screening 80 patients for MSI detection, 32 (40%) of them were detected as MSI-H and 48 (60%) of cases were reported as Microsatellite stable.

In 32 MSI cases, 18 (56%) were positive for *H. pylori* infection and, 13 (41%) were EBV positive and one of them was negative for both pathogens. In the case of *H. pylori* infected cases, 17 (53%) were CagA positive cases, 7 (23%) were VacA positive cases, six were positive for both the genotype. In EBV infected cases, all the 13 were positive for the Type I genotype (41%) and 13% cases were Type II positive. Four (04) of them were positive for both Type I and Type II genotypes. EBV Type II cases were found more in MSI subgroup and other all the genotypes were evenly distributed in both the subgroup. EBV (+) cases was found more in MSS subgroup, though it was not significant. The proportion of *H. pylori* (+) cases was more in MSS subgroup than *H. pylori* (-) cases. CagA genotype associated cases were higher in MSS subgroup, but interestingly CagA (-) cases were not found in MSI subgroup. Proportion of VacA(+) cases was similar in both the subgroup, but VacA(-) cases were found more in MSS subgroup. Proportion of EBV type I (+) cases was more in MSS subgroup and EBV type II (+) cases were found in similar proportion in both the subgroups

Further, the GC samples were classified as *H. pylori* (+), *H. pylori* (-), *EBV* (+), *EBV* (-), MMR deficient and MMR proficient and a comparison was made between all the subgroups with demographic, and lifestyle habit data to find out significant factors with each subgroup of GC patients. The chi square distribution test was performed to find out significant risk factors with each subgroup. Among all the risk factors, only smoked food consumption was significantly associated with *H. pylori* positive patient group (*p*-value= 0.006) and EBV infected patient group (p-value = 0.002). Smoked food is the prime risk factor for developing pathogen associated GC. Two lifestyle factors, tobacco chewing and alcohol drinking were found as significant risk factor with high OR, 95% CI (p-value = 0.04) and (p-value= 0.03), respectively for MMR deficient patients group. For further verification, binary logistic regression was performed for determining the odd ratio and 95% CI. A significant association was found between *H. pylori*-infected GC patients with consumption of smoked food (*p*-value = 0.007). Smoked food consumption (*p*-value=0.003) and tuibur intake (*p*-value =

0.05) were significant factors for *EBV* infected GC patients and tuibur consumption. Significant association was observed with chewing tobacco (p-value = 0.04) and alcohol drinking (p-value = 0.03) for the MMR deficient (MSI) patient.

In this cohort, the follow-up data for 3 years was used to study the overall survival (OS) rate of patients with the subgroup [*H. pylori* (+), *H. pylori* (-), *EBV* (+), *EBV* (-), MMR deficient, and MMR proficient] by unadjusted analysis using the Kaplan Meier curve to find out prognostic factors. A univariate Cox proportional hazards model demonstrated that *H. pylori* infection does not have significant relation for GC patient's prognosis with stage I, II, and III (HR: 1.13, 95% CI: 0.86 - 1.73; *p*-value = 0.13;). *EBV* infection and MSI were independent prognostic predictors for GC patients with stages I, II, and III). The GC patients group with *EBV* infection showed poor prognosis (HR: 2.22, 95% CI: 0.92 - 2.97; *p*-value = 0.05) with stages I, II, and III and were observed as high-risk group. The comparison between MMR deficiency and proficiency exhibited significant prognostic predictor for stages I, II, and III GC patient groups (HR: 3.43; 95% CI: 0.95 - 4.08; *p*-value = 0.03). In this cohort, MSI/MMR deficient cases showed a good prognosis for GC patient, whereas MSS/MMR proficient cases exhibited poor prognosis for GC patients. Further, we performed comparison study by retrieving the data of gastric cancer patients with the *H. pylori*, *EBV,* and MMR gene status as independent prognostic factors for stages I, II, and III gastric cancer patients group from TCGA-STAD cohort. In this approach, Cox proportional-hazards regression model showed that *H. pylori* status have no significant log-rank value and *p*-value, whereas MMR gene status exhibited as an independent prognostic factor in TCGA-STAD cohort (HR: 1.60; 95% CI: 1.04 – 1.91; *p*-value = 0.03).

The top ten mutated genes were *TP53* (47%), followed by *MUC6, FAT4, RNF43, BCOR, PTPRC, ERBB2, CTNNB1, SOHLH2,* and *FBXW7.* The data was compared with the TCGA and Asian Cancer Research Group (ACRG) study of GC. *TP53* and *FAT4* were found to be mutated in all the studies. *MUC6* and *APC* were found to be mutated in the Mizo population study and in the ACRG study. The similarity between top ten mutated gene of ACRG group and our study was more than TCGA group. The frequently mutated genes were analyzed according to the subgroups and *APC* gene (32%) was significantly mutated with EBV (+) gastric cases (Table 12). Enrichment of RNF43, ARID1A and ERBB2 mutations were found in EBV (+) subtypes and mutation of these genes were absent in EBV (-) cases. The low frequency of *TP53* mutation was found in EBV (+) cases compared to EBV (-) gastric cancer subtypes. *BNC2* was found to be mutated only in MSI compared to MSS gastric cancer subtypes. *BCOR* and *PTPRC* were found to be mutated in MSS cases, but mutation of these genes were absent in MSI gastric cancer subtypes

The mutation data was presented as a heat map and two prominent patient clusters were obtained. TP53 was significantly mutated with the cluster 1 group compared to cluster 2. The EBV (+) group was dominant in cluster 2, while only one sample exhibited TP53 mutation in this cluster. There were not significant differences in *H. pylori*, MSI or MSS subgroups between the two clusters. We have compared the mutated genes in both the cluster on the basis of their frequency. There was an enrichment for *PTPRC* (25%) gene in cluster 1 while enrichment of *ERBB2* (25%) was observed in cluster 2. In case of clinical data, 58% of moderately differentiated cases were found in cluster 1 group. 57% of poorly differentiated cases were found in cluster 2, showing that patient samples with aggressive tumors were found in high EBV infected group.

In this study, 183 variants were obtained, out of them 11 variants were predicted as pathogenic in CLINVAR database. In case of these 11 variants, 8 (R306*, G245S & R175H of *TP53*, D769Y and V842I of *ERBB2*, E545K and H1047R of *PIK3CA* and R876* of *APC* gene) were reported as pathogenic stomach cancer mutations in other populations. Twenty one missense variants and one stopgain were reported as pathogenic stomach cancer variants in COSMIC database. Most frequently mutated gene was *TP53* with 8 variants (R273C, G266V, P250L. R175H, S215N, L194R, L137Q & E358V) followed by *ERBB2* with 2 variants (S310F &Y781C) and *FAT3* (Y4395C & R3784H). A1792V of *MTOR*, S575R of *BNC2*, S1747L of *KMT2B*, G1517R of *KMT2C*, R5Q of *RHOA*, H86R of *RNF43*, R98* of *CTNNA1*, R352C of SLIT2, R332Q of *APC* and I3602L of FAT4 gene were also found in this study. All the variants occurred with 2% frequency only one variant (R273C) of *TP53* occurred with 4% occurrence frequency

In germline case *MAP3K4* (92%) was the top mutated gene in this study followed by *KMT2C* (65%), *ATN1* (33%), *MACF1* (27%), *BRCA2* & *FAT4* with 21%, *FAT3, KMT2B* & *PLB1* with 17% and *APC* with 15% frequency. *MAP3K4* gene exhibited only one homozygous in-frame deletion (A1199del) with 92% occurrence frequency. Here, *KMT2C* (*MLL3*) gene is an important driver gene for developing GC at germline level in this population. *ATN1* and *KMT2C* genes were highly mutated compared to females. *BRCA2* were highly mutated in females compared to males.

Binary logistic regression analysis was performed to get the significantly mutated genes or gene family with clinical factors. The most frequently mutated gene and gene family like *FAT3* and *FAT4* under FAT family, *EGFR* and *ERBB3* under EGFR family, *BRCA1* and *BRCA2* under DNA repair gene family and *MACF1* & *ATN1* independently were selected. The genes of FAT family, *FAT3/4* were strongly significant (*p*-value = 0.003) with well and moderately differentiated cases. *MACF1* gene was significantly (*p*-value = 0.02) mutated with advanced stage and with poor survival status (*p*-value = 0.03). *MACF1* gene was showing more aggressive tumour with poor prognosis. *BRCA1/2* were showing good prognosis (*p*-value = 0.03).

CTNNA1, PMS2 and *KMT2C* identified significant mutations in familial GC cases. These three genes are the significant genes which might develop familial GC, besides *CDH1* in Mizo population. Survival analysis was done by selecting the familial patients having mutations in *CTNNA1, PMS2* and *KMT2C* to find out prognostic risk factors by unadjusted analysis of follow-up data using the Kaplan Meier curve. A univariate cox proportional hazards model demonstrated that *KMT2C* gene was independent prognostic predictor for familial GC patients as it was showing poor prognosis (HR: 1.57, 95% CI: 0.76 - 3.26; *p*-value = 0.02). The panel of three genes were strong prognostic predictor with a significant *p* value (HR: 1.82, 95% CI: 0.68 - 4.85; *p*-value = 0.04) as it was showing poor prognosis. The panel of three genes might be successful in predicting the familial GC risk in this population with higher AUC value (0.68; *p*-value = 0.03). RTK-RAS pathway, Hippo-signalling pathway, Wnt signaling, TP53 pathway, PI3K, TP53 and NOTCH signalling pathway alterations might be responsible for developing Gastric cancer in this population.

In this germline data, out of 78 variants 23 were novel variants. Pathogenicity prediction was done for all the non-synonymous variants by four prediction tools (SIFT, PROVEAN, Polyphen2 and Mutation Taster) and predicted as pathogenic if the variants found to be predicted as damaging or deleterious in all the tools. Among all the missense or non-synonymous variants, only 12 variants were predicted as pathogenic in all the tools and they are as follows: C3121Y, P4952L and R5357Q (*MACF1* with 8.33%, 4.17% and 2.08% frequency, respectively), P922R and W4352G (*KMT2C* with 4.17% and 2.08% frequency, respectively), A2066G (*FAT3* with 2.08% frequency), Y856H (*BRCA1* with 8.33% frequency), P587R (KMT2B with 10.42% frequency), A667T (*MSH2* with 2.08% frequency), Q965L (*ABCA10* with 2.08% frequency) G2608A (*FAT4* with 2.08% frequency) and L114F (*PMS2* with 2.08% frequency). Out of 12 pathogenic variants, 3 were novel variants (P4952 L - MACF1, Q965L - ABCA10, G2608A - FAT4). In this study, 9 indels were found. Among them, *ATNI* exhibited one novel In-frame deletion (L1740_S1741del) with 6.25% frequency. *MAP3K4* gene exhibited one in-frame deletion (A652del) in 92% frequency which was the highest occurrence frequency of variants in this study.

In whole exome sequencing the top ten mutated genes were *HLA-DRB1, HLAB, FLG, HLAC, RFPL4AL1, MAML3, MUC6, BAGE5, PRB1* and *KCNJ12*. Out of the top ten genes, *HLA-DRB1, HLAB* and *HLAC* plays a key role in the immune system. 34 genes were mutated frequently in more than 90% cases. In case of germline analysis variants were considered as polymorphism, if they occur in higher frequency. Most of them were polymorphism but six of them (*COL18A1, KCNJ18, CMYA5, FCGBP, HLA-DRB1* and *OR4M2)* exhibited 13 pathogenic mutations with high frequency. G1072R (*COL18A1* with 2.7% frequency), E430G (*KCNJ18* with 100% frequency), Y3957H, T3515N & F3628S (*CMYA5* with 2.7% frequency in each case), G3871R & C3904F (*FCGBP* with 5.4% and 2.7% frequency, respectively), T80R, D70N, Y152C, V188M & G197A (*HLA-DRB1* with 2.7%, 8.10%, 16.21%, 2,7% and 2.7% frequency, respectively) and S202C (*OR4M2* with 18.91% Frequency) were found as pathogenic variants in this study.

In this study, 40 novel mutated genes were not reported in other studies for association with Gastric Cancer. These genes might be responsible for developing Gastric Cancer in this population. The genes are as follows: *SUSD2, CNTNAP38, TTN, PDE4DIP, POLR2J3, SORBS1, DNAH1, ATIC, HSPA6, KRT6B, RASA4, LIMS1, PDE4D, SIRPB1, LAMA5, SLC66A2, SYNE1, TPTE, ZNF638, DNAH9, OBSCN, SEC16A, ZRANB3, CELSR1, FAI1, GNPTG, USP8, EYS, LOXHD1, NEB, SLCO2A1, SVIL, XIRP2, ARHGAP21, ARHGEF10, CEP295, CYP2C8, FAM43B* and *NRIP1.*

About 26 commonly mutated cancer related genes were derived in this study. Most of these genes were also found to be mutated in targeted germline data. The genes are as follows: *FAT4, ERBB3, FAT2, CREBBP, NOTCH3, ABCA10, FAT3, KMT2C, NOTCH1, PIK3C2A, APC, TP53, CTNNA3, FAT1, KMT2B, ALDH1A2, CDH19, EPCAM, MSH2, NOTCH2, BRCA1, BRCA2, EP300, PLB1, STK11* and *XRCC1*. RTK-RAS, Hippo, Wnt, PI3K, TP53 and NOTCH pathway gene alterations were obtained both in targeted re-sequencing data as well as whole exome sequence data.

Fifty three (53) pathogenic germline heterogeneous variants were identified in WES analysis and out of them 14 were novel mutations. R3053P (2.7%) of *FAT1*, E1617G (2.7%) & R2606T (2.7%) of *FAT3*, P4194H (2.7%) of *FAT2*, S1723I (2.7%) & D1987G (2.7%) of *NOTCH1*, G3961T (2.7%) of *NOTCH3*, C1876A (2.7%) of *CDH1*, C2765G (2.7%) of *KMT2C*, V66M (2.7%) of *ALDH1A2*, A339V (2.7%), A339T (10.81%) & D284Y (2.7%) of *FAT4*, and N111H (2.7%) of *EPCAM* were the novel variants. Six germline variants obtained from targeted data were also present in Whole exome data. These variants (*N2198Y & N2544S of MACF1, E319V of TP53, K253R of EGFR, P587R of KMT2B* and T1261I of *APC*) might play important roles for developing GC in this population.

The copy number analysis of *TP53* and *HER2* mutated samples (17 patients) were performed in ddPCR and 35.29% samples had a variation for *HER2* gene. There was gain for *HER2* copy number in five samples and in one sample it was a loss in copy number. *TP53* copy number was altered in 23.52% cases, among them there was a gain in copy number in 3 samples and one samples exhibited copy number loss. One missense (Y781C) mutation was responsible for *HER2* copy number gain in this study

Immunohistochemistry staining, there was a higher expression of BAX protein in tumor cases compared to adjacent normal tissues. The expression of BAX was higher in Stages I, II and III but unexpectedly the expression of BAX was low for Stage IV samples. Interestingly BAX expression was significantly (*p*-value = 0.05) associated with EBV (+) GC cases. BAX expression was not associated with *H. pylori* infected GC cases. In survival analysis the patient group with high BAX expression was at risk group (HR: 1.36; 95%CI: 0.26-6.87; p-value = 0.37) compared to low expression cases, though it was not significant. TP53 expression was higher in late stage compared to early stage GC cases. In gnomically stable cases, TP53 expression was higher compared to MSI associated GC cases. HER2 expression was higher in early stage GC cases, EBV (+) and MSI cases. ERCC1 expression was little higher in early stage GC. ERCC1 expression positive cases were higher in *H. pylori* (+) and MSI associated GC cases (Figure 41D).

The genes which were mutated in 90% cases and among novel genes 26 were identified as predicting, diagnostic, prognostic and therapeutic biomarker in different type of cancer (Table 24). This might be used as biomarker for gastric cancer also.

Discussion and Conclusion

This study to the best of our knowledge, this is the first case-control study, designed to assess the detailed epidemiological risk factors along with the potential role of *EBV / H. pylori* infections, MMR gene status and Genomics and the prognosis of GC patients in Northeast India. The GC patients from Mizo population exhibited higher pathogen associated GC cases. Smoked food, extra salt consumption, smoking and alcohol are the major risk factors for developing GC and obese persons are at risk for developing GC. EBV infection was significantly associated with the unique risk factor (tuibur). *EBV* infection is a strong risk factor for GC and poor prognosis in this Indian high-risk population. TP53 mutations were also a significant factor for GC risk. This study has found that this population might be genetically predisposed with *MAP3K4, HLA-DRB1*, *HLAB, HLAC* and *KCNJ12* pathogenic mutations and novel genes are also found associated with GC which may develop GC by following a combination of pathways. The panel of *KMT2C, PMS2* and *CTNNA1* genes may be useful in predicting familial GC in Mizo population.

Older aged people have more exposure to toxins and unhealthy food habits and some undesirable exposure like sun light over time. Precancerous cells can develop at any time during the lifespan, but as elderly people have weak immunity so it may not protect against the development of cancer cells. Male gastric cancer patients were found at significant risk group for developing GC than females in our study. H. pylori causes severe inflammation which can lead to Gastric cancer, male persons are found to be affected more with H. pylori infections (de Martel et al. 2006). The estrogen hormone can prevent the infection in women, studies has reported that increased level of estrogen is responsible for the decreasing risk of gastric cancer in females (Camargo et al. 2012)

In my findings, 73.75% of the tumor developed in distal site of the stomach. Studies have reported that most of the GC cases found to be in distal part and *H. pylori* present in gastric mucosa can develop a severe tissue injury in the distal stomach which may lead to Gastric Cancer (Hu et al. 2012; Piazuelo et al .2010). About 50% of the patients were in stage III indicating that most of the patients were diagnosed at advance stage only and 32.5% cases were found to be familial cases of cancer, with any type of cancer in the first degree relatives. Till date, Mizo population practice endogamy and this might be a cause of high risk for Gastric Cancer in this population.

In multivariate analysis, obese persons with excess BMI was found to be associated with an increased risk of GC development (OR = 0.69, 95% CI = 0.60 – 0.79; p-value = 0.0001). One meta-analysis showed that excess BMI is a significant risk factor with gastric cancer development in Asian population (Hirabayashia et al. 2019; Bae et al. 2020). Consumption of extra salt, a dietary habit was found as a risk factor for developing GC in this study. Extra salt provides the possible condition for colonization of *H. pylori* by increasing the mucin level of the surface mucus in the stomach and studies have reported that *H. pylori* is a significant risk factor of stomach cancer (Fox et al. 1999, Kato et al. 2006). Another dietary factor, smoked food was found as significant risk factor associated with GC in this population, as it was a common food habit in more than 60% of patients. Smoking in an oven or by burning of wood or charcoal and grilling method is used to cook smoked food (McDonald et al. 2015), and during this process antioxidants and antimicrobial properties along with carcinogenic chemicals like Polycyclic Aromatic Hydrocarbons (PAH) are produced (Varlet et al. 2006). Benzo[a]pyrene (BaP), a group I carcinogen, is a member of PAH family found in the smoked food, and plays an important role in GC disease progression along with other cancers. In Mizoram, it is common practice to make smoked foods rich in salt and in turn it can create a favorable condition for *H. pylori* infection which will ultimately lead the development of GC in this population. In the study, smoked food consumption was found to be a significant risk factor with EBV infected GC cases. Smoking cigarettes and consumption of smoked food are significant contributing factors, for the development of carcinogenesis in GC patient, which might be amplified by the presence of EBV. It has been reported that there is a strong association of smoking with the risk of developing EBV-positive Hodgkin's lymphoma (Kamper-Jorgensen et al. 2013) and that tobacco, which is a risk factor for GC, may contain EBV-activating substances (Jia et al. 2012).

Two lifestyle factors, smoking and alcohol, were found to be the associated risk factors with GC development in this study. Studies have reported that smoking is a strong significant risk factor for developing GC (Bersten et al. 2013, Bonequi et al. 2013). The association between alcohol drinking and GC development is always a matter of conflict. ALDH2 enzyme converts alcohol to acetate and any metabolic change in the enzyme activity will lead to accumulation of Acetaldehyde (class I carcinogen). In Asian populations, there is a prevalence of particular mutations which can inactivate ALDH2 enzyme (Ghosh et al. 2017). Studies have reported alcohol as an independent risk factor associated with GC in China (Moy et al. 2010). In this present study, two lifestyle factors, chewing tobacco and alcohol drinking were found as associated significant risk factor with MMR deficient GC patients. Several studies have reported that MSI-H colorectal cancer cases are strongly associated with tobacco and alcohol drinking (Diergaarde et al. 2003; Eaton et al. 2005; Poynter et al. 2009; Warneke et al. 2003; Ghatak et al. 2016).

This study achieved a panel of five epidemiology factors (BMI, Extra salt consumption, smoked food, drinking and smoking) with high AUC and sensitivity value for detecting Gastric Cancer patients in early-stage for therapy implementation and prognosis.

In this study, 88.75% cases were associated with pathogens indicating that EBV and *H. pylori* are playing major role in developing Gastric cancer in this population. EBV enters the body through saliva or oral contacts and in the Mizo people have a common practice to share water glasses or cigarettes with each other, while drinking tuibur (tobacco infused water) or alcohol and smoking. This might be a cause for this high prevalence of EBV associated GC cases in this population. Prevalence of *H. pylori* cases can be found in developing countries (Aziz et al. 2014). *H. pylori* is a class I carcinogen which can lead to produce proinflametary cytotoxins, oxidative stress and necrosis in the cells which in turn can develop chronic inflammation to lead GC cancer (Singh et al. 2017; Carlos et al. 2019). This study has shown that *EBV* infected GC patients are more aggressive with poor prognosis and the prognostic value of *EBV* infection was confirmed by multivariate analysis, even after adjustments for other clinical factors. In this study MMR proficient GC cases were showing poor prognosis and considered as a high-risk group with more aggressive tumors, while MMR deficient GC patients exhibited good prognosis. The result is consistent with other studies which reported that MSI shows a better prognosis than MSS cases in gastric cancer (Beghelli et al. 2006; Kim et al. 2020; Choi et al. 2014; Smyth et al. 2017).

In this study, the top somatically mutated gene was *TP53* (47%) like reported in other studies in Gastric cancer (Park et al. 2016; Busuttil et al. 2014). *TP53* mutations used to associate with late stage or advance stage of Gastric cancer, similar to this study. Studies have also reported that *TP53* mutations are associated with the risk of developing distal GC (Perez-Perez et al. 2005; Bellini et al. 2012). In this study most of the tumour occurred at the distal part of the stomach. Frequently mutated genes like *MUC6, FAT4 & APC* were also found to be mutated frequently in TCGA and ACRG studies, supporting our data. Other top genes like *RNF43, BCOR, PTPRC, ERBB2, CTNNB1, SOHLH2,* and *FBXW7* were also found to be associated with Gastric cancer in TCGA and ACRG studies (Cancer Genome Atlas Research Network et al. 2014; Cristescu et al. 2015). In this study, *TP53* was the top mutated gene in all the subgroups (EBV +, MSI and MSS) of the samples.

APC gene was significantly mutated with EBV associated gastric cases like other studies, though the percentage is much higher in this study (Shinozaki-Ushiku et al. 2015). A study reported hyper methylation of APC gene association for the development of EBV infected non-cardiac GC cases (Geddert et al. 2010). Enrichment of *ERBB2* mutation was found only in EBV associated cases. Studies have reported that the crosstalk between EBV and *HER-2* might play an important role to develop EBV associated GC through receptor kinase signaling pathway (Gulley et al. 2015; Cyprian et al. 2018).

In this study, two molecular subtypes were found: one with TP53 mutation dominant group and another group was found with EBV infected cases. Higher frequency of high grade tumor and enrichment of ERBB2 mutation in EBV associated cases indicates that they are more aggressive tumor having poor prognosis (Ming et al, 2000). *TP53* mutations were less in EBV (+) group like in another study (Kim et al. 2016). *TP53* somatic mutation and EBV infections are the two drivers for developing Gastric cancer in Mizo population.

Till date, there are few studies which gave us insights about germline mutated genes, except CDH1 in Gastric cancer. In this present study, frequent mutations in *MAP3K4, KMT2C, ATN1, MACF1, BRCA2, FAT4, FAT3, KMT2B, PLB1* and *APC* genes were obtained, except CDH1. Very few studies reported mutations in other gene besides CDH1 in case of hereditary GC (Gaston et al. 2014; Villacis et al. 2016). MAP3K4 gene is a member of mitogen-activated protein kinase (MAPK) pathway and plays an important role for Cancer development by activating the CSBP2, P38 and JNK MAPK pathways by phosphorylating MAP2K4 and MAP2K6 of MAP3K family.

MLL3, a member of the myeloid/lymphoid or mixed-lineage leukemia (MLL) family is a chromatin remodeling gene. The products of chromatin responsible can regulate the structure of chromatin for altering DNA accessibility and transcriptional efficiency and were observed as frequently mutated genes in GC (Cancer Genome Atlas Research Network. 2014).

Cell adhesion genes *CTNNA1, FAT3* and *FAT4* were mutated with 2%, 16.66% and 25% frequency, respectively. The genes of FAT family, FAT3/4 were strongly significant (p-value = 0.003) with well and moderately differentiated cases. One study has reported that tumor suppressor gene *FAT4* is a modulator of Wnt/β-catenin can be a novel therapeutic target for clinical development in GC (Cai et al. 2015). These Cadherin family genes besides *CDH1* might play important role for developing Gastric Cancer in this population. Another important gene *MACF1* which maintains Cell Motility and involved in metastatic invasion by regulating the cytoskeleton structure were reported as significantly mutated gene in Gastric cancer (Cancer Genome Atlas Research Network. 2014). This gene was also mutated in our study with 31% occurrence frequency and significantly mutated with the group of advanced stage patients. *MACF1* is showing poor prognosis as all the variants have aggressive effects and *BRCA1/2* were showing good prognosis (*p*-value = 0.03). BRCA2 germline mutation was found in 20% cases indicating the increasing risk of Gastric cancer relation with BRCA1/2 mutations (Hiroshi et al, 2020). The variants were reported as pathogenic for Hereditary Breast and ovarian cancer.

Three genes *KMT2C* (*MLL3*), *PMS2* and *CTNNA1* mutations were found significant with familial GC cases. Among them, *PMS2* and *CTNNA1* were already present in the hereditary Gastric Cancer panel made by Chicago university. *KMT2C* or *MLL3* is the new gene which was significantly associated with familial gastric cancer samples in this study. KMT2C gene were showing poor prognosis for familial gastric cancer patients. MLL3 gene mutation was associated with lynch syndrome (Villacis et al.

2016) which is associated with GC development. Besides *CDH1*, *CTNNA1* of cadherin family was also associated with hereditary diffuse GC development (Lauren et al. 1965).

Hippo, Wnt, PI3K, TP53 and NOTCH. RTK-RAS and Hippo pathway and these pathways were associated for developing Gastric Cancer in this study like reported in other studies also (Luciya et al, 2020 and Yiting et al, 2018).

In whole exome germline analysis, immune system related genes (*HLA-DRB1, HLAB* and *HLAC*) were the top mutated genes. One study has reported that EBV infects B lymphocytes to enter into the host body and HLA class II molecules used to act as a cofactor for initiation of this infection of B lymphocytes (Li et al. 1997). These results indicate that our immune related genes were mutated frequently in this population due to high prevalence of EBV infection, which is playing the prime role for developing GC in Mizo population.

Thirty four (34) genes were found to be mutated in more than 90% samples, in case of germline analysis though those variants might be polymorphic. But among them, six genes (*COL18A1, KCNJ18, CMYA5, FCGBP, HLA-DRB1* and *OR4M2*) exhibited pathogenic mutations. Fourty (40) novel genes were found in case of germline mutation, which were not reported in other studies for association with Gastric Cancer. These novel genes might be following some different pathway for developing Gastric Cancer in this population.

In copy number analysis, we found that in tumor tissue *ERBB2* is having copy number gain compared to adjacent normal. A negative correlation was found between copy number among TP53 and ERBB2 genes. This supports that Tumor oncogenes have a gain in copy number and tumor suppressor genes have deletion in copy number in cancer tissue samples (Lawrence et al, 2019).

BAX expression was significantly higher in EBV (+) group. It has reported in one study that BCL 2 expression was higher in EBV positive cases and BAX expression was comparative higher in EBV negative group. The present study reported a contradictory report, which suggests that EBV infection might contribute to apoptosis method (Lima et al. 2008). BAX expression was higher in Low grade tumors than high grade tumors while in advance stage the expression was lower (Gazzaniga et al, 1995). Cell death might play a role to develop cancer at early stage, as BAX is an apoptotic gene it is also following the same trend.

In this study we selected TP53 mutated samples for IHC and as a result they were showing more positive cases in late stage. During dysplasia, at the last stage of disease progression might be some stress used to drive TP53 mutations, which contributes to the progression of GC. TP53 expression was less in EBV (+) group, which supporting the sequence data of this study and other studies also reported that TP53 expression was more in EBV (-) subgroups (Kim et al. 2016). EBV infection might not alter the TP53 pathway for developing GC.

HER2 positive cases were found more in early stages, which indicating that HER2 can be targeted as a therapeutic marker for Gastric cancer in this population, which can develop the treatment strategy of GC. HER2 cases were showing positive expression on EBV positive and MSI cases, which supporting the sequence data of enrichment of ERBB2 mutation on EBV subgroup. EBV infection might effecting receptor kinase signalling pathway (Gulley et al. 2015; Cyprian et al. 2018) for developing GC in this population.

ERCC1 positive expression cases were more in *H. pylori* positive cases and MSI cases. This data is showing that *H. pylori* infection might play role in DNA damage repair pathway (Kim et al. 2008; Wang et al. 2014; Kwon et al. 2007).

Conclusion

The prospective of this present study is that high incidence of Gastric Cancer in Mizo population might be due to the effect of smoked food, tuibur, alcohol and smoking with EBV infection. Mizo population being a homogeneous population has unique set of driver genes with pathogenic alterations may play a role to initiate the progression of Gastric cancer in this population. The panel of five epidemiological risk factors which can predict early stage GC cases, which is very necessary in clinical field for making decision on patient treatment. The study will help clinicians to opt decision for the right therapy by applying the prognostic assessment of this study. Further study is necessary with large cohort which would be beneficial to support our data.

The present study reported novel genes which were not earlier related to gastric cancer and some genes were mutated in 90% of the patients, among them some of the gene were identified as biomarker for other cancer, like lungs, head and neck, colorectal, pancreatic etc. and chronic diseases. *HLADRB1, HLA-C, HLA-B, MAML3, MUC6, CMAY5, FCGBP, SUSD2, TTN, SORBS1, ATIC, HSPA6, KRT6B, LIMS1, PDE4D, SIRPB1, LAMA5, ZNF638, OBSCN, CELSR1, USP8, SLCO2A1, XIRP2, ARHGEF10* and *CYP2C8* might be identified as biomarkers for Gastric cancer in this population.

This study reported that unique food habits and lifestyle factors along with pathogen and microsatellite status might be driving the novel driver mutations for developing Gastric Cancer in Mizo population. Novel set of genes identified in this study might be the drivers for developing GC in this high risk population. This study reporting new epidemiological markers as well as gene markers for detecting early Gastric cancer and familial GC cases, respectively which will help the clinicians for taking correct diagnostic and therapeutic decisions.

Summary

The present study was accomplished to find out the significant risk factors associated with Gastric cancer in this population along with pathogen infection and MSI status. This study was also carried out to find out the novel driver alterations and genes associated with GC development. Statistical analysis was performed to find out significant risk factors. Screening of *H. pylori*, EBV and MSI were performed for molecular subtyping. Targeted re-sequencing was performed for paired tumor and blood samples, to find out driver genes associated with GC in this population. Sequencing was performed on Illumine Hi-seq machine by capturing hybrids of interested panel genes. Whole exome sequencing was also performed to find out novel set of genes which might play for developing GC in this population. In addition IHC was performed with tumor suppressor gene, oncogenes and apoptotic genes for studying there expression and prognosis on GC patient on the basis of clinical and mutation data.