Chapter III:

Materials and Methods

3.1.1. Data collection

We collected data for 24 different cancers, as outlined in Table 3.1, utilizing cBiportal (90-92) an online platform that facilitates the retrieval of extensive cancer data sourced from TCGA and ICGC projects. Additionally, we obtained a list of genes linked to immune responses from ImmPort(93). It is a comprehensive repository that stores and shares immunology-related data, encompassing clinical, experimental, and molecular information. The gene catalogue comprised of 1492 genes associated with different aspects of immune responses.

Sl. No	**Abbreviation**	**Full name**	**Dataset**	**Sample Number**		
				All	**Mutations**	**RNA seq**
1	ACC	Adrenocortical carcinoma	Adrenocortical Carcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	92	91	78
2	BLCA	Bladder Urothelial Carcinoma	Bladder Urothelial Carcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	411	410	407
3	BRCA	Breast invasive carcinoma	Breast Invasive Carcinoma (TCGA, Cell 2015) TCGA, Cell 2015	818	817	817
4	CHOL	Cholangiocarcinoma	Cholangiocarcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	36	36	36

5	COAD	Colorectal adenocarcinoma	Colorectal Adenocarcinoma (TCGA, Nature 2012) TCGA, Nature 2012	276	224	274
6	CESC	Cervical squamous cell carcinoma	Cervical Squamous Cell Carcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	297	291	294
7	HCC	Hepatocellular carcinoma	Harding et al. Clin Cancer Res 2018	127	127	
8	HNSC	Head and Neck Squamous cell Carcinoma	Head and Neck Squamous Cell Carcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	523	515	515
9	KICH	Kidney Chromophobe	Kidney Chromophobe (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	65	65	65
10	KIRC	Kidney renal clear cell carcinoma	Kidney Renal Clear Cell Carcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	512	402	510
11	KIRP	Kidney renal papillary cell carcinoma	Kidney Renal Papillary Cell Carcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	283	276	283
12	LUAD	Lung adenocarcinoma	Lung Adenocarcinoma (TCGA, Nature 2014) TCGA, Nature 2014	230	230	230

13	LUSC	Lung squamous cell carcinoma	Lung Squamous Cell Carcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	487	484	484
14	OV	Ovarian carcinoma	Ovarian Serous Cystadenocarcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	585	523	300
15	PRAD	Prostate adenocarcinoma	Prostate Adenocarcinoma (TCGA, Cell 2015) TCGA, Cell 2015	334	333	290
16	ESCA	Esophageal adenocarcinoma	Esophageal Adenocarcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	182	182	181
17	THCA	Thyroid carcinoma	Thyroid Carcinoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	500	498	490
18	UCEC	Uterine Corpus Endometrial Carcinoma	Uterine Corpus Endometrial Carcinoma (TCGA, Nature 2013) TCGA, Nature 2013	373	248	333
19	GBM	Glioblastoma multiforme	Glioblastoma Multiforme (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	592	397	160

20	LGG	Brain Lower Grade Glioma	Brain Lower Grade Glioma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	514	514	514
21	PCPG	Pheochromocytoma and Paraganglioma	Pheochromocytoma and Paraganglioma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	178	178	178
22	DLBC	Diffuse large B cell lymphoma	Diffuse Large B-Cell Lymphoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	48	41	48
23	LAML	Acute Myeloid Leukemia	Acute Myeloid Leukemia (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	200	200	173
24	SARCOMA	Sarcoma	Sarcoma (TCGA, PanCancer Atlas) TCGA PanCancer Atlas, Cell 2018	255	255	253

Table 3.1: Details of the datasets downloaded from cBioportal

3.1.2. Analysis of alterations in the DNA sequence of the immune related genes

We obtained mutation data for genes listed in the immune-related gene catalogue acquired from the ImmPort database. The mutation data was sourced from the dataset covering 24 different human cancers that we had previously obtained from cBioportal. Using the maftools software, we generated detailed summaries of mutations in each cancer (94). We conducted a screening for Copy Number Alterations (CNA) in the frequently mutated immune related genes of the particular cancer. This involved analysing available CNA data from the 24 different cancer dataset, using GISTIC 2.0(95). The analysed data (p-value-0.01)was visualised for CNA in frequently mutated immune related genes using maftools (94) in R.

3.1.3. Survival analysis considering mutations

Using mutation data and clinical data available in 9 datasets, Kaplan Meier plots in 5 years survival time were generated for cancers by maftools. Univariate and multivariate cox-proportional hazard analysis considering age for the mutated and non-mutated PIK3CA was performed using the “survminer” and “survival” packages.

3.1.4. Identification of differentially expressed genes (DEGs)

Data was divided into two dataset one mutated and another non-mutated, for malignancies illustrating the influence of gene mutations on survival. The edgeR(96), R package was used for detection of differentially expressed genes (DEGs) among the groups. The statistical cutoff for differential expression was p-value and false discovery rate (FDR) of less than 0.05 and a fold change greater than 1.5.

3.1.5. Identification of over-represented pathways

The DEGs from the analysis were used as input for pathway over-representation analysis using WebGestalt 2019 (97), database used was Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways database, pathways with FDR < 0.05 were considered in the study.

3.1.6. Estimating infiltration of immune cell to the tumor immune microenvironment

CIBERSORTx was used to estimate the number of immune cells infiltrating the tumor (98). The gene expression profile of both mutated and non-mutated dataset was used as input. The profile of gene expression from both the mutant and unmutated datasets was utilized as input. The fraction of twenty-two immune cell subtypes was assessed using an LM22 signature. (The LM22 signature is a gene expression profile utilized to deconvolve immune cell populations from bulk tumor gene expression data. It comprises 547 genes representing 22 immune cell subsets) *p* value calculation was done with 100 permutations. The difference in infiltration was compared.

3.1.7. Statistical analysis and data visualization

The statistical analyses for the study were conducted using the R 4.2.3 statistical framework and the ggplot2 R package was used for generating plots. KEGG pathway plots were generated using SRplots(99).

3.2.1. Expression of 4genes (HLA-A, HLA-B, HLA-DRB1 and CIITA) and survival analysis

The expression profile of 4genes for LGG was analysed by GEPIA (Gene Expression Profiling Interactive Analysis, http:// gepia.cancer-pku.cn/index.html) (100), a platform utilized to analyse RNA sequencing data from TCGA and GTEx projects and generate boxplots for differential expression and Kaplan Meier plots for survival by "Single Gene Analysis" module.

3.2.2. Relationship Between 4genes (HLA-A, HLA-B, HLA-DRB1 and CIITA) and immune infiltration in the tumor microenvironment

The quantity of immunological infiltrates(B cells, CD4+ T cells, CD8+ T cells, NK Cells, Tregs, T follicular helper, macrophages, and dendritic cells) was estimated using TIMER(101). The "Gene module" of the TIMER database was used in this study to assess the relationship between the expression of four genes under investigation and immune cell infiltration.

3.2.3. Correlation between 4genes (HLA-A, HLA-B, HLA-DRB1 and CIITA) and HLA-E

The correlation between 4genes (HLA-A, HLA-B, HLA-DRB1 and CIITA) and HLA-E was analysed using "correlation of expression analysis module" under expression analysis of OncoDB(102).

3.3.1. Identification of putative G-quadruplex in human genome

The study was initiated by acquiring FASTA sequences for the human genome from the National Center for Biotechnology Information-NCBI (GRCh38). These sequences were utilized as input for G4 Hunter (103) to predict locations for putative G-quadruplex (G4) structures along the sequence.

3.3.2. Visualization of overlap between G4 structures and immune-related genes

To visualize the overlap between G4 structures and immune-related genes, we utilized KaryoplotR(104), to generate a karyoplot that could illustrate the intersection.

3.3.3. Annotation of putative G4 locations onto human genome

Putative G4 locations from G4 Hunter were annotated using bedtools intersect (105) on reference human genome GRCh38 from Gencode (106). Thus, the precise location of the putative G4 along genes were determined. The list of proto-oncogenes was sourced from the cancer gene census (107) in the COmmon Software Measurement International Consortium (COSMIC) database. A catalogue of 803 proto-oncogenes from source http://ongene.bioinfo-minzhao.org/ongene_human.txt was downloaded. The list of housekeeping genes was downloaded from https://www.tau.ac.il/~elieis/HKG/ , the list had a catalogue of 2832 genes. The list of immune genes and proto-oncogenes had 101 genes in common, these 101 genes were removed from both the list before proceeding with further analysis. There were no such similar genes in list of housekeeping genes to either immune genes or cancer related genes.

3.3.4. Analysing the density of G4

The prevalence of G4 structures in cancer-related genes has been extensively established in prior research(108). G4 are involved in RNA regulation, hence it was expected that the frequency of G4 in housekeeping genes would be low. Therefore, comparative analysis of G4 abundance within immune genes and as opposed to proto-oncogenes and housekeeping genes was done. The density of G4 was considered as putative G4 per 1000 base pairs. Dunnett’s test was used for the comparison between the groups.

3.3.5. Examining the frequency of mutations in G4 locations

We overlaid mutations from all 24 different datasets included in our study onto the putative G4 locations along the human genome using bedtools intersect.