EXPLAINABLE ARTIFICIAL INTELLIGENCE TO STRATIFY PAN-CANCER PATIENTs FOR IMMUNE CHECKPOINT INHIBITOR DECISION MAKING

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Yuanyuan Shen

Dr. Jonathan B. Mitchem, Dissertation Supervisor

December 2022

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

EXPLAINABLE ARTIFICIAL INTELLIGENCE TO STRATIFY PAN-CANCER PATIENTS FOR IMMUNE CHECKPOINT INHIBITOR DECISION MAKING

presented by Yuanyuan Shen,

a candidate for the degree of Doctor of Philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Jonathan B. Mitchem

Dr. Chi-Ren Shyu

Dr. Trupti Joshi

Dr. Kevin F. Staveley-O’Carroll

Dr. Ayesha N. Shajahan-Haq

ACKNOWLEDGEMENTS

While I am finishing the journey of Ph.D. study in Mizzou, I have lots of appreciation to my supervisors, committees and University of Missouri-Columbia. First, I would like to appreciate Dr.Shyu who brought me into informatics world. In China, we have an old saying, swift horse is everywhere, however, and who has the good judgement and recognize the swift horse is the real master. Dr.Shyu is this kind of person to me. Next, I would like to appreciate Dr.Mitchem. In the past 5 years, Dr. Mitchem and I were working together and achieved a lot of accomplishment. Meanwhile, Dr.Mitchem is supporting me and helping me to dive into my passion by combining my medical knowledge with informatics. Dr.Mitchem is always responsive, responsible and always share my happiness and tough moments. And he is always encouraging me in any moment. Then I would like to thank all my Ph.D. committees, thank you all for the supportive and opportunities offered for me and all the supports, friendship from all the team members at iDAS lab, Mitchem lab, cancer omics research group and Rob and Tracy. In the very last, I would like to appreciate my families to support me and put up with me entire time, then I could totally focus on my Ph.D. study and career development.

Table of Contents

LIST OF ILLUSTRATIONS

ABSTRACT

Immune checkpoints are a normal part of the immune system. It engages when proteins on the surface of immune cells called T cells recognize and bind to partner proteins on other cells, such as some tumor cells. Immune based therapies such as ICIs work by blocking checkpoint proteins from binding with their partner proteins. This prevents the “off” signal from being sent, allowing the T cells to kill cancer cells. One such drug act against a checkpoint protein called PD-1 or its partner protein PD-L1. Some tumors turn down the T cell response by producing lots of PD-L1. Recent years FDA have granted accelerated approval for the immunotherapy drug on treating specific subgroups or advanced stage of solid tumors. This groundbreaking treatment has shown remarkable promise which could prevent tumors from growing and allowing some patients who receive the treatments to essentially be cured. The fact that some patients treated with immunotherapy have a durable response to cancer shows this treatment’s potential. But despite response rates between 20 and 50 percent in certain groups, such as microsatellite instability-high (MSI-H) colorectal cancer which are characterized by high mutational load, neoepitope formation, and an intense lymphocytic infiltrate when compared to microsatellite stable (MSS) tumors [1]. And some cancer, such as non-small cell lung cancer, rely on some indicators like PD-L1 protein, tumor mutation burden [2, 3]. But how accurate these indicators could predict patient’ responses are still in debating. Moreover, scientists still don’t know why the majority of people with cancer do not respond to immunotherapy drugs [4]. For those patients are not supposed to receive immunotherapy may cause unnecessary long term side effect, such as adrenal insufficiency, and financial burden[5]. Completion of the work we propose here would help us to identify characteristics of patients as subgroups who might benefit from ICIs from multi-omics. In this study, we applied an explainable AI approach for patient stratification. Exploratory mining is contrast pattern-based data mining process. We will integrate these two methods to explore clinically explainable subgroups with phenotypical and genotypical features. These features would be the labels for identify patients who really need ICIs. We applied this method in pan-cancer population and identified the patients’ subgroups based on distinctive features including demographic, phenotypical and genomic characteristics. We believe these distinctive features contribute together to the patients’ response sensitivity to the ICIs. Further wet lab experiments to validate these findings are required prior to initiating clinical trials using these identified features.

Chapter 1

Introduction

1. INTRODUCTION

Worldwide, an estimated 19.3 million new cancer cases (18.1 million excluding nonmelanoma skin cancer) and almost 10.0 million cancer deaths (9.9 million excluding nonmelanoma skin cancer) occurred in 2020 [6]. An estimated 1,806,590 new cases of cancer will be diagnosed and 606,520 people will die from the disease in the United States of year 2020. The goals of treatment are to "cure" the cancer, if possible; prolong survival; and provide the highest possible quality of life during and after treatment. Cancer treatment can include localized therapies, such as surgery, radiation therapy, cryotherapy, and heat or chemical ablation, and/or systemic therapies (e.g., chemotherapy, hormonal therapy, immune therapy, and targeted therapy) used alone or in combination. In recent years, there is four primary types of immunotherapies are currently applying on cancer treatment, which is monoclonal antibodies, immune checkpoint inhibitors, cancer vaccines, and some other non-specific immunotherapies. As one of the promise immunotherapies, ICIs have shown remarkable treatment effect which could prevent tumors from growing and allowing some people who receive the treatments to essentially be cured [7].

1.1 IMMUNE CHECKPOINT INHIBITOR AND EFFECTIVENESS

PD-1 as a critical biomarker of the immune checkpoint, which is a surface protein on immune cells called T cells. It normally acts as a type of "off switch" that helps keep the T cells from attacking other cells in the body. When PD-1 binds to PD-L1, it basically tells the T cell to leave the other cell alone. Some cancer cells have large amounts of PD-L1, which helps them evade immune attack. PD-1 inhibitors and PD-L1 inhibitors as monoclonal antibodies that target either PD-1 or PD-L1 can block this binding and boost the immune response against cancer cells. After several clinical trials, in 2017, FDA approved pembrolizumab (anti-PD-1 immune checkpoint inhibitors) could be used on patients with molecular identity, microsatellite instability-high (MSI-H), such as MSI-H colorectal cancer; and very recently in this year, 2019, FDA approved atezolizumab (anti-PD-L1 immune checkpoint inhibitors) could be used on triple-negative breast cancer [8, 9]. FDA have granted accelerated approval for the immunotherapy drug on treating specific subgroups or advanced stage of solid tumors. However, only colorectal cancer and breast cancer have been clearly indicated the subtypes which could get the benefit from immunotherapy despite response rates between 20 and 50 percent in these groups [10]. In the other solid tumors, FDA only could approve these immunotherapies could be used on advanced cancers with possible PD-L1 protein or/ and tumor mutation burden level testing [11]. In another word, these groundbreaking therapies have a substantial challenge: FDA only regard a small number of patients would benefit from the ICIs, but for the majority patients we haven't had a “gold standard” to indicate if there are portion of them would potentially qualified. Moreover, these incredible advances and the promise of cancer cures also come with eye-popping price tags that reach well past at least $100,000 per patient[12]. And when ICIs like nivolumab take the brakes off of cancer-killing immune cells, these activated immune cells can also harm healthy tissues, leading to side effects such as adrenal insufficiency. Therefore, for those patients who are not supposed to receive immunotherapy may cause unnecessary long term side effect and financial burden.

1.2 RELATED WORK

Therefore, with those questions in mind, it makes revealing the subgroups cancer patients who would benefit ICIs become a necessary and urgent issue. Recently some studies have shown that each single-tissue cancer type can be further divided into three to four molecular subtypes. These sub-classifications are more based on recurrent genetic and epigenetic alterations that converge on common pathways, such as p53 and/or Rb checkpoint loss, RTK/RAS/MEK or RTK/PI3K/AKT activation, etc. Moreover, these single-tissue tumor subtypes have shown meaningful differences in clinical behaviors, some of them could directly lead to therapies that target these subtype-specific molecular alterations [13]. Fortunately, more and more large-scale and multi omics-included genomics projects now producing detailed molecular characterizations for thousands of tumors datasets are available.

We brought to a unique combination of informatic approaches, a number of which we have recently developed, for integrating multi-omics and exploring the characteristics of ICI-sensitive subgroups[14-16]. Using our distinctive combination of pipeline and informatic expertise, we will determine that how would the population, phenotypical and transcriptome variables interact with each other; how multi-omics to modulate CTL activity in pan-cancer patients; the subgroups containing ICI-sensitive feature in large heterogeneous populations. In short, we contributed knowledge of the characteristics of ICI-sensitive subgroups in multi-omics and how the multi-omics variables interact and affect ICI-sensitivity in pan-cancer scale. This contribution would be significant because it will identify more ICI-sensitive subpopulations and avoid unnecessary treatment. We expect that our pan-cancer and multi omics-based taxonomy will lead to improved quality of cancer patients' life and survival outcomes, and reduced healthcare costs.

Since the Initial clinical trial results with IgG4 PD1 antibody Nivolumab were published in 2010 and was approved by the FDA in 2014 for inoperable or metastatic melanoma, it has been 8 years [17]. However, we haven't had a clear understanding of the difference of benefits were obtained from patients and cancer types. There are some previous studies have categorized the molecular subtypes by machine learning methods for solid tumors, such as colorectal cancer, breast cancer, etc. [18, 19]. However, in these methods, the molecular-based subtypes only limited in the one type of cancer, and it couldn't be applied for the other type of cancers.

1.3 PRELIMINARY RESULTS

In our previous study, we were using TCGA data to determine the heterogeneity and homogeneity by the mRNAseq data which based on well-accepted molecular subtype in COADREAD and breast cancer. COADREAD patient with MSI-h status and triple-negative breast cancer (TNBC) are the two subtypes which have been approved by FDA to treat by ICIs. We did a DEGs analysis for COADREAD microsatellite stable (MSS) vs. MSI-h patients and non-TNBC vs. TNBC in breast cancer patients. Then we applied venn-diagram overlap and pathway analysis by DEGs from two cancers. In BRCA, there were 3,499 DEGs and 676 DEGs in COADREAD. Three hundred eighteen DEGs are crosstalk genes from two cancers. Even it looks the overlap genes occupied in BRAC is close Figure 1 to 10%, however, in COADREAD, the percentage close to 47%. This result inspired us that even in two different tissue-of-origin cancers, it shows a significant common in the type of patients which potentially could acquire a benefit from immunotherapy.

Some of the overlap DEGs might be the “driver-gene” could be identified in pan-cancer scale study. In addition, we did a pathway enrichment analysis for these overlapped DEGs. Interestingly, in all the pathway events, there are 56 DEGs, which close to 1/5 of total overlapped DEGs enriched in the immune pathway event. We assumed these 56 immunes functional DEGs might be the signature genes of immunotherapy sensitive type.

In sum, the above preliminary data indicate that in two different tissue of-origin cancers, it shows a significant common in the type of patients which potentially could acquire a benefit from immunotherapy. Furthermore, the identified immune functional DEGs might be the signature genes of immunotherapy-sensitive type. From the primary results inspired us to determine a therapeutic-orientated, multi omics-based subtypes identification would support the clinical decision making for ICIs application. The following experiments of this proposal are designed to explore these questions.

Chapter 2
Methods

2.1 INTRODUCTION

This Patient Stratification and feature detection framework is composed of three major steps. (1) Build up a pipeline to integrate phenotypical and transcriptome data to identify immunotherapy targets in breast cancer (BRCA) and colorectal cancer (COADREAD). We collected COADREAD patient data from The Cancer Genome Atlas (TCGA) dataset. Then utilizing the Microenvironment Cell Population Counter (MCP-Counter) to create tumor cytotoxic lymphocyte (CTL) abundance scores. Grouping patients based on cytotoxic lymphocyte abundance score, stage, and tumor anatomic location. Last, running mRNA-seq differential gene expression analysis, pathway enrichment and survival analysis to identify phenotypic feature involved pathway and genomic targets. (2) Developed an approach to integration transcriptome data and Epigenetic data to determine CTL activity modulators in COADREAD. Again, we retrieved publicly available data repositories. And grouping patients based on consensus molecular subtypes (CMSs, signature based COADREAD classification); and exploring crosstalk gene in differentially expressed genes and differentially methylated genes from subtypes. Last, integrating crosstalk genes with MCP counter score, neoantigen, tumor mutation burden and cell type score to determine CTL activity modulators and possible mechanisms.

2.2 MATERIALS AND DATA PROCESSING

The input data for the patient stratification framework consists of genotypic and phenotypic variables for a disease population. The phenotypic, genotypic, and heterogeneous features are used to guide subgroup discovery. In this study, the genotypic and phenotypic data for patients were obtained from TCGA. As part of the human-in-the-loop process, a physician panel involved in the care of cancer patients selected the phenotypic and clinical variables to be included in the analysis. Additionally, many of these phenotypic variables were continuous, which required stratification into categories for inclusion in the data mining algorithm. The medical guidelines and the physician panel guided the categorization of all continuous variables. For example, the original data set of COADREAD contains the age of each patient. Patient age was categorized into four age groups by quartile. The genotypic data in this study are genes differentially expressed between normal and tumor tissues. The differential expression analysis using edgeR was implemented on the RNA-seq data of the patients [20]. The dimensionality reduction was made by deciding the p-value to be less than 0.05. In addition, our study contains 14-17 different types of biomedical variables. The differences of number of variables we included were based on the features of type of cancer, such as BRCA and COADREAD have subtypes information collected but not all the other cancers.

2.2.1 THE CANCER GENOME ATLAS DATA ACCESS AND PROCESSING

Data repositories such as TCGA allow for the in-depth study of patients on a molecular and clinical basis. Recently, a novel computational method for predicting the abundance of different cells within the tumor microenvironment using RNA-seq data was developed and validated with histologic specimens called the Microenvironment Cell Population Counter (MCP-Counter) [21]. This method allows for an effective comparison of the composition and pathways associated with cellular infiltration in the tumor microenvironment, improving over other methods primarily based on microarray data and gene set enrichment analysis. The mRNA-seq from TCGA data is an ideal input for starting group patients by MCP-counter score.

The data will be retrieved here is based upon data originally generated and organized by FireCloud from the Broad Institute. Full permission access transcriptomic data was obtained from dbGAP. We will download COADREAD and BRCA patients' open access, pre-processed mRNA expression data (level 3 data) from both platforms, IlluminaHiSeq and Illumina-GA; as well as mRNA-Seq by RSEM normalized data; and patients' clinical data from the cohorts TCGA_COAD_ControlledAccess, TCGA_READ_ControlledAccess, TCGA_BRCA_ ControlledAccess by gsutil Tool.

For the COADREAD, we integrated pertinent clinical data (age, gender, microsatellite status, anatomical location, pathologic stage, Tumor Node and Metastasis (TNM) classification, days to last follow up, and vital status), and mRNAseq by participant ID. Each COADREAD patient has pre-identified microsatellite status labeled as "microsatellite instability test results". Thirty-four patients with colon cancer and 18 patients with rectal cancer were excluded due to missing information including indeterminate microsatellite status, unclear anatomical location or pathologic stage, and unmatched RNA-seq data. Then for every patient, we implemented the R package Microenvironment Cell Populations-Counter on the RNA-Seq by Expectation Maximization (RSEM) normalized RNA-seq data to create cell type abundance scores [21]. There are 10 cell populations simultaneously quantified in the tumor microenvironment, including 8 immune cell populations (T cells, CD8 T cells, Cytotoxic lymphocytes, NK cells, B lineage, Monocytic lineage, Myeloid dendritic cells, Neutrophils), endothelial cells and fibroblasts. Specifically, the gene set for cytotoxic lymphocytes includes the genes: CD8A, EOMES, FGFBP2, GNLY, KLRC3, KLRC4, and KLRD1. We used the median value of the CTL score to group the patients into high (≥ median value, CL-High) and low (< median value, CL-Low) groups. We then grouped patients by anatomical location and stratified by cytotoxic lymphocyte score and pathologic stage. For consistency in clinical intervention while increasing group size, we assigned pathologic stage I-II to “early” stage, stage I-III as “localized”, and stage IV as “metastatic” (Figure 1). Patients were grouped by tumor subsite: tumors located in the cecum, ascending colon, hepatic flexure, and transverse colon were categorized as having right-sided colon cancer (Figure 1, RSC); tumors located in the splenic flexure, descending colon, sigmoid colon or rectosigmoid junction were categorized as having left-sided colon cancers (Figure 1, LSC); patients with tumors located in the rectum were kept in this group (Figure 1, REC).

As well as BC, we have identified 918 breast cancer tumor samples and compared RNAseq gene expression based on molecular subtypes and anatomic locations of biopsies (i.e., right, left, lower inner quadrant, lower outer quadrant, upper inner quadrant or upper outer quadrant). Genes with significantly different expression (p<0.01) were selected for survival analysis. R, Reactome Pathway Browser were used to retrieve and analyze data (Figure 2).

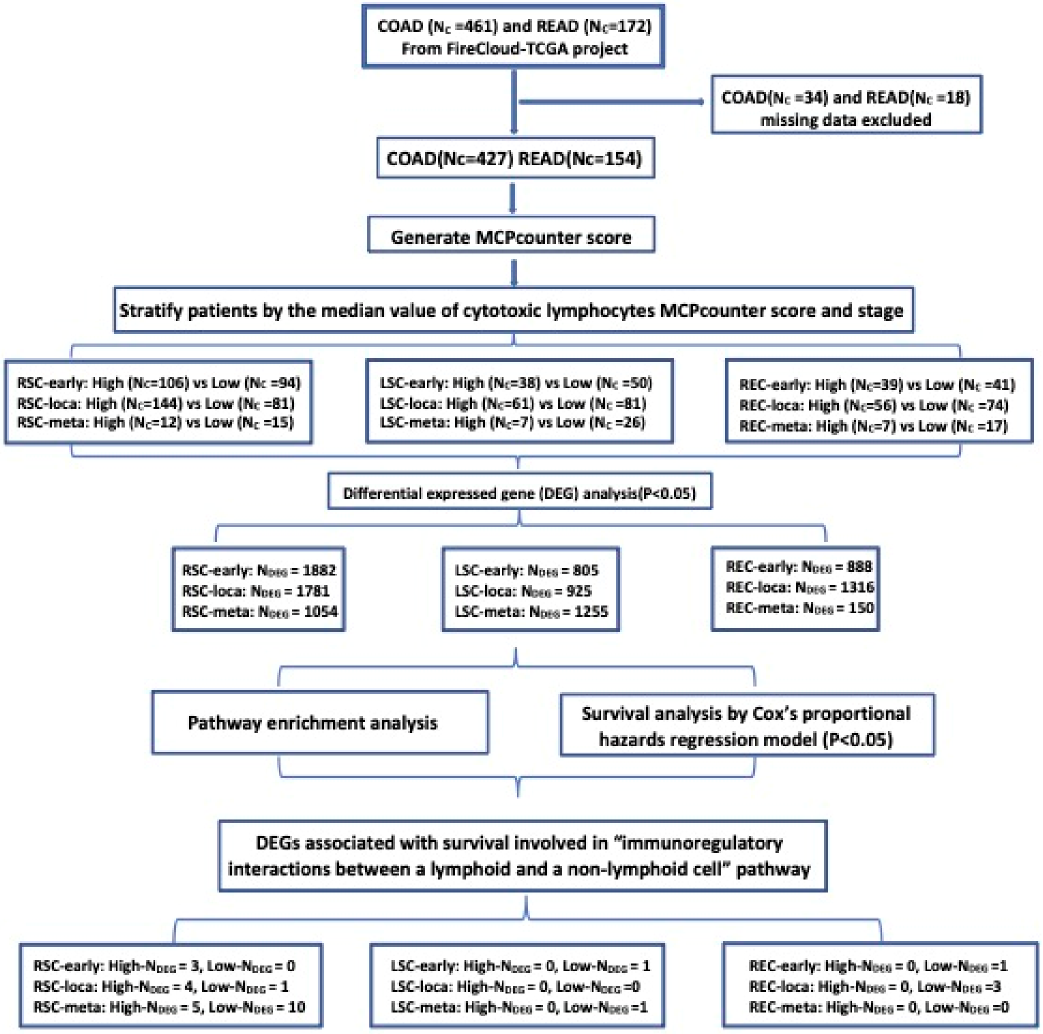

COAD (N_C =461) and READ (N_C=172)
From FireCloud-TCGA project
COAD(N_C =34) and READ(N_C =18)
missing data excluded
COAD(Nc=427) READ(Nc=154)
Generate MCPcounter score
Stratify patients by the median value of cytotoxic lymphocytes MCPcounter score and stage
RSC-early: High (Nc=106) vs Low (Nc =94)
RSC-loca: High (N_C=144) vs Low (N_C =81)
RSC-meta: High (N_C=12) vs Low (N_C =15)
LSC-early: High (N_C=38) vs Low (N_C =50)
LSC-loca: High (N_C=61) vs Low (N_C =81)
LSC-meta: High (N_C=7) vs Low (N_C =26)
REC-early: High (N_C=39) vs Low (N_C =41)
REC-loca: High (N_C=56) vs Low (N_C =74)
REC-meta: High (N_C=7) vs Low (N_C =17)
Differential expressed gene (DEG) analysis(P<0.05)
RSC-early: N_{DEG} = 1882
RSC-loca: N_{DEG} = 1781
RSC-meta: N_{DEG} = 1054
LSC-early: N_{DEG} = 805
LSC-loca: N_{DEG} = 925
LSC-meta: N_{DEG} = 1255
REC-early: N_{DEG} = 888
REC-loca: N_{DEG} = 1316
REC-meta: N_{DEG} = 150
Pathway enrichment analysis
Survival analysis by Cox's proportional hazards regression model (P<0.05)
DEGs associated with survival involved in "immunoregulatory interactions between a lymphoid and a non-lymphoid cell" pathway
RSC-early: High-N_{DEG} = 3, Low-N_{DEG} = 0
RSC-loca: High-N_{DEG} = 4, Low-N_{DEG} = 1
RSC-meta: High-N_{DEG} = 5, Low-N_{DEG} = 10
LSC-early: High-N_{DEG} = 0, Low-N_{DEG} = 1
LSC-loca: High-N_{DEG} = 0, Low-N_{DEG} =0
LSC-meta: High-N_{DEG} = 0, Low-N_{DEG} =1
REC-early: High-N_{DEG} = 0, Low-N_{DEG} =1
REC-loca: High-N_{DEG} = 0, Low-N_{DEG} =3
REC-meta: High-N_{DEG} = 0, Low-N_{DEG} =0

Figure 1. An outline of the methods and organization of this study.

2.2.2 DEMOGRAPHICS AND CLINICOPATHOLOGIC CHARACTERISTIC ANALYSIS

We group COADREAD patients by level of CTL score. It is necessary to determine how would the demographics and clinicopathologic characteristics are going to associate with CTL score level. And we group BC patients by subtypes, luminal A/B, HER2+ or Triple negative breast cancer (TNBC). Since FDA have been approved TNBC patients would benefit from ICIs. We assumed that there will be a significant difference in some of demographics and clinicopathologic characteristics between cytotoxic lymphocyte-high (CTL-High) and cytotoxic lymphocyte-low (CTL-Low) groups in COADREAD, likewise, TNBC compares with other subtypes.

Demographic, clinical, and pathologic characteristics were retrieved as stated above for COADREAD and BC. Statistical analyses were performed using Prism 7 (GraphPad Software). Patients' basic clinical features were summarized by descriptive statistics, including means and standard deviation, and an unpaired t-test was used for normally distributed continuous data. Categorical variables were compared using Fisher’s exact and chi-square tests. A p value < 0.05 was considered statistically significant.

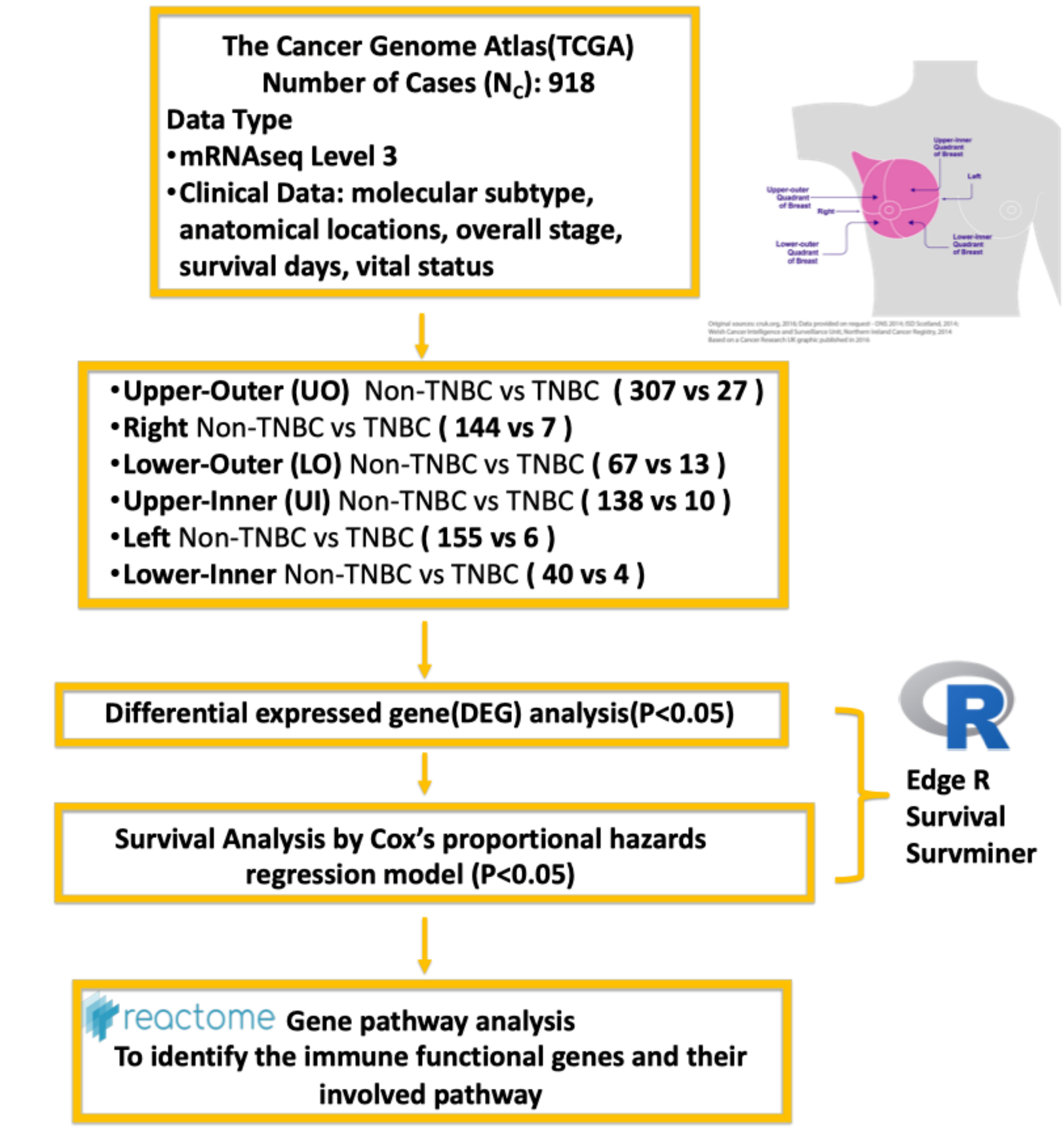

The Cancer Genome Atlas(TCGA)
Number of Cases (N_C): 918
Data Type
•mRNAseq Level 3
•Clinical Data: molecular subtype, anatomical locations, overall stage, survival days, vital status
Upper-Inner Quadrant of Breast
Left
Upper-outer Quadrant of Breast
Right
Lower-outer Quadrant of Breast
Lower-Inner Quadrant of Breast
Original sources: cruk.org, 2016; Data provided on request - ONS 2014; ISD Scotland, 2014; Welsh Cancer Intelligence and Surveillance Unit, Northern Ireland Cancer Registry, 2014
Based on a Cancer Research UK graphic published in 2016
•Upper-Outer (UO) Non-TNBC vs TNBC (307 vs 27)
•Right Non-TNBC vs TNBC (144 vs 7)
•Lower-Outer (LO) Non-TNBC vs TNBC (67 vs 13)
•Upper-Inner (UI) Non-TNBC vs TNBC (138 vs 10)
•Left Non-TNBC vs TNBC (155 vs 6)
•Lower-Inner Non-TNBC vs TNBC (40 vs 4)
Differential expressed gene(DEG) analysis(P<0.05)
Survival Analysis by Cox's proportional hazards regression model (P<0.05)
Edge R
Survival
Survminer
reactome Gene pathway analysis
To identify the immune functional genes and their involved pathway

Figure 2. An outline of the methods and organization of this study.

2.2.3 RNA-SEQ DIFFERENTIAL GENE EXPRESSION ANALYSIS

COADREAD and BC, however, is a heterogeneous disease made up of multiple subgroups. Additionally, recent studies have suggested that the use of other markers including lymphocyte infiltration may better predict survival and the potential for response to immune based therapy. Therefore, we assumed that in each heterogeneous subgroup would identify the significant transcriptome differences.

RNA-seq differential gene expression analysis was performed with the edgeR package using the raw data downloaded from the Illumina- HiSeq and Illumina-GA platforms [20]. Differentially expressed genes were defined as genes with an absolute fold change >1 between patients with high and low cytotoxic lymphocyte scores with a p value <0.05 (20). Genes with Benjamini-Hochberg adjusted False Discovery Rate (FDR) <0.05 were considered to be significantly differentially expressed for further steps. For each cohort, we identified 20,531 total genes by RNA-seq raw counts. The Reactome online browser was used to identify immune functional differentially expressed genes [22]. Figure 1 outlines this process.

2.2.4 PATHWAY ENRICHMENT AND SURVIVAL ANALYSIS

The immune functional differentially expressed genes identified from last aim would express similarity function from the heterogeneous subgroups. To further determine whether would have overlapping pathways associated with cytotoxic lymphocyte infiltration in COADREAD, we will perform the Reactome pathway enrichment analysis at each location based on stage. We hypothesize that among heterogeneity subjects, it should have conserved pathway to adjust the basic immune function across stage and location. Each individual gene in these pathways would positively and negatively affect patients' survival. Pathway enrichment analysis will be performed to evaluate the pathways associated with differentially expressed genes. The genes included in the MCP-counter CTL gene panel (CD8A, EOMES, FGFBP2, GNLY, KLRC3, KLRC4, and KLRD1) will be excluded from pathway enrichment analysis. Dotplot will be used to illustrate the comparison of enriched Reactome pathways among differentially expressed genes in each location and stage. These results will be analyzed by clusterProfiler, DOSE, and ReactomePA R packages. Next, we will perform survival analyses using the identified differentially expressed genes. Patients would be organized by stage and location as outlined above. The normalized mRNA-seq data of differentially expressed genes will be used for survival analysis and be processed using the Survival R package [22]. For each differentially expressed gene, if the normalized gene expression value is more than the median level, we labeled it as “high,” and otherwise as “low.” The Kaplan–Meier survival curves will be generated is going to be assessed by the Cox regression model for each immune functional differentially expressed gene using the

Survminer R package. The survival curves of patients with high gene expression and low gene expression are going to be compared by log-rank test. For each patient, overall survival (OS) will be used as the endpoint, either the days from diagnosis to death, or to the last follow-up.

2.3 MULTI-OMICS INTEGRATION

Develop an approach to integration transcriptome data, mutation, and epigenetic data to determine CTL activity and CD8+ T cell dysfunction modulators. We applied in COADREAD and BC as an example.

We first developed a data-driven approach to identify the key modulators by integrating multi-omics data. To address this, we were: (1) Retrieving publicly available data repositories. (2) Grouping patients based on consensus molecular subtypes (CMSs, signature based COADREAD classification); and exploring crosstalk gene in differentially expressed genes and differentially methylated genes from subtypes. (3) Integrating crosstalk genes with MCP counter score, neoantigen, tumor mutation burden and cell type score to determine CTL activity modulators and possible mechanisms. Figure 3 describe overall methods for this step.

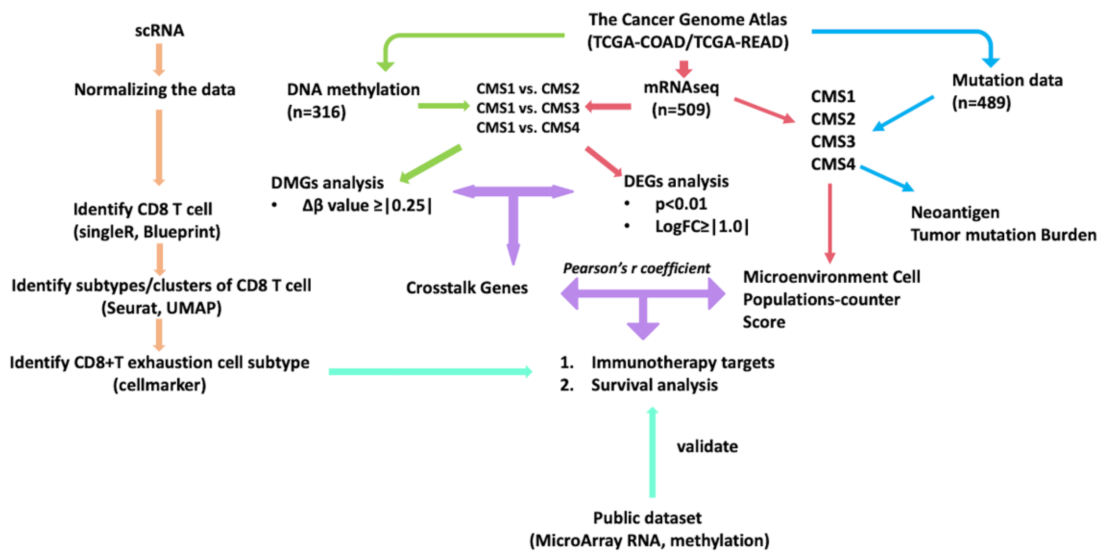


Figure 3. An outline of the methods and organization of this study.

2.3.1 RETRIEVING PUBLICLY AVAILABLE DATA REPOSITORIES

Most recently, large-scaled public data repositories such as TCGA, cbioportal, Gene Expression Omnibus, EMBL-EBI, as well as publicly available single cell sequencing data from each study are available for researchers. We hypothesized that comprehensively include multi-omics and large-scaled datasets would allow us to perform the in-depth study of cancer patients on a molecular and clinical basis from multi omics. We first will search and focus on publications and all public repositories for looking dataset which would contain scRNAseq, mRNAseq, microarray, mutation data and methylation data. In this study, we utilized 15 datasets, 2,391 COADREAD patients and 7 omics data and integrate comprehensive clinical, genomic data (epigenetic, scRNAseq, mRNAseq data, mutation), well-accepted immune infiltration indicator (TMB, neoantigen), and cell marker score, MCP-counter scores from multiple datasets to reveal that TBX21 is a critical regulator in CD8+T cell exhaustion (Figure 3). Meanwhile, we also included BC data from TCGA.

We searched and focused on publications and all public repositories for looking dataset which would contain scRNAseq, mRNAseq, microarray, mutation data and methylation data. From our previous systematic review, we have found a few public data repositories to match our requirement. Currently, there are two clinical trials ongoing to investigate COADREAD patients categorized by CMS. One of them is focusing on evaluating the clinical benefit of anti-PDL1/TGFbetaRII fusion protein M7824 in treating COADREAD patients that has metastases status or cannot be removed by surgery (NCT03436563); another study is aimed to develop a genome-based platform to predict patients who can achieve pathologic complete response after neoadjuvant concurrent chemoradiotherapy in locally advanced rectal cancer and CMS status of these patients would be identified (NCT04738214). We hypothesized that starting with a robust subtype classification would bring a more precise evaluation and improvement in the application of ICK inhibitor therapy in COADREAD. We found 10× scRNA-seq raw_UMI_count_matrix and cell_annotation data from Gene Expression Omnibus (GEO) with accession number GSE132465. Bulk RNA-seq, DNA methylation (450k) and mutation data of COADREAD will be collected from TCGA (https://tcgadata.nci.nih.gov/). Moreover, we will retrieve publicly available data deposited in the ArrayExpress database at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-7036 (methylation), EMTAB-8148 (microarray), and additional microarray data from Colorectal Cancer Subtyping Consortium (GSE13067, GSE13294, GSE17536, GSE20916, GSE33113, GSE37892, GSE39582, KFSYSCC) were stored under the synapse repository (https://www.synapse.org/#!Synapse:syn2623706/files/). Each COADREAD patient corresponded mutations (per Mb), neoantigens with its HLA alleles, and mutated expressed peptides can be retrieved from The Cancer Genome Atlas (http://tcia.at). Potential Problems and Alternatives: Every patient would be retrieved by us may not include the same omics data. The number of patients in each dataset may varies. For this case, we would perform the analysis separately for the patients cannot be integrated or use them for the evaluation of our results.

2.3.2 Demographics and clinicopathologic analysis

We integrated pertinent clinical data (age, gender, microsatellite status, pathologic stage, days to last follow up, and vital status), mRNAseq, and DNA methylation (450k) by participant ID. Each COADREAD patient has pre-identified CMS subtypes from the original study[23]. Due to missing information, including indeterminate microsatellite status,

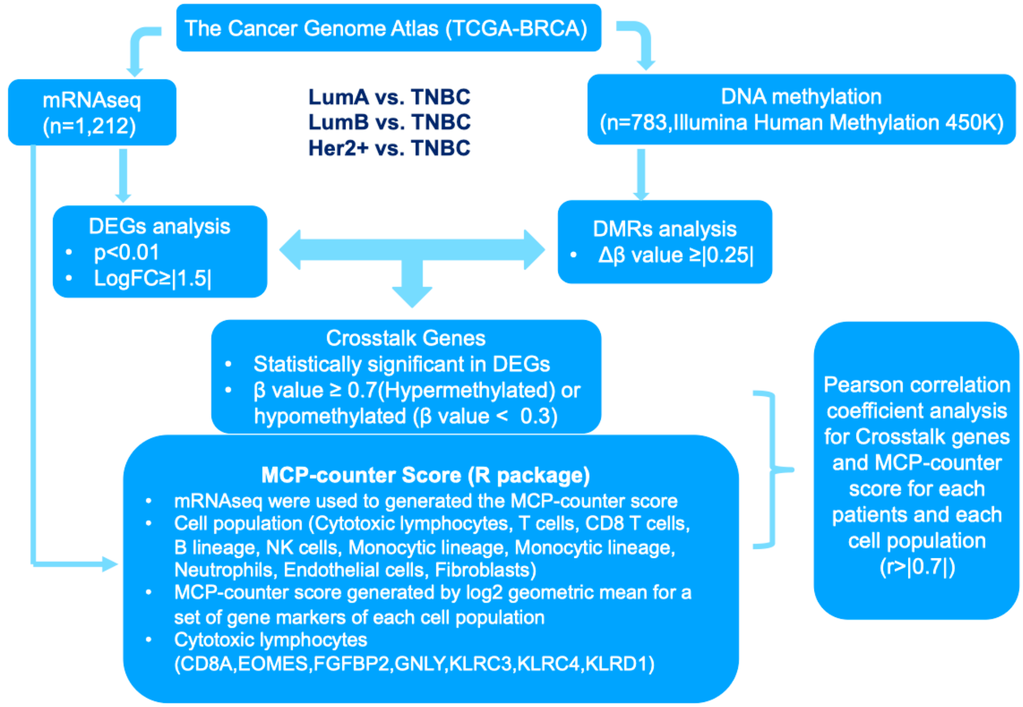

The Cancer Genome Atlas (TCGA-BRCA)
mRNAseq
(n=1,212)
LumA vs. TNBC
LumB vs. TNBC
Her2+ vs. TNBC
DNA methylation
(n=783,Illumina Human Methylation 450K)
DEGs analysis
• p<0.01
• LogFC≥|1.5|
DMRs analysis
• Δβ value ≥|0.25|
Crosstalk Genes
• Statistically significant in DEGs
• β value ≥ 0.7(Hypermethylated) or hypomethylated (β value < 0.3)
MCP-counter Score (R package)
• mRNAseq were used to generated the MCP-counter score
• Cell population (Cytotoxic lymphocytes, T cells, CD8 T cells, B lineage, NK cells, Monocytic lineage, Monocytic lineage, Neutrophils, Endothelial cells, Fibroblasts)
• MCP-counter score generated by log2 geometric mean for a set of gene markers of each cell population
• Cytotoxic lymphocytes (CD8A,EOMES,FGFBP2,GNLY,KLRC3,KLRC4,KLRD1)
Pearson correlation coefficient analysis for Crosstalk genes and MCP-counter score for each patients and each cell population (r>|0.7|)

Figure 4. An outline of the methods and organization of this study.

and unclear CMSs, eventually, we identified 316 patients with DNA methylation data (450k) and 509 patients with mRNAseq data with distinct CMS labels. We then grouped DNA methylation data and mRNAseq data by CMSs. Statistical analyses based on grouped patients were performed by Prism 7 (GraphPad Software). Patients' basic clinical features were summarized by descriptive statistics, including means and standard deviation, and an unpaired t-test and Mann-Whitney test were used for normally distributed continuous data. Categorical variables were compared using Fisher's exact and chi-square tests. A p-value < 0.05 was considered statistically significant. For BC patients, We retrieved RNA sequencing (n=1,212) and DNA methylation (Illumina Human Methylation 450K; n=783) databases to detect the gene expression and methylation profiles based on molecular subtypes: LumA vs TNBC, LumB vs TNBC, Her2+ vs TNBC. To be consistent with clinical diagnosis, our TNBC patients included all ER-, PR-, Her2 -/1+/2+ patients (n=192, Figure 4).

2.3.3 DIFFERENTIAL GENE EXPRESSION, AND DNA METHYLATION DIFFERENTIAL REGION ANALYSIS

The mRNAseq differential gene expression analysis was performed with the edgeR package using the raw data downloaded from the TCGA dataset Illumina- HiSeq and TCGA_Illumina-GA platforms [24]. Differentially expressed genes were defined as genes with an |logFC| >1 and p-Value <0.05 for comparisons of CMS1 vs. CMS2, CMS1 vs. CMS3, CMS1 vs. CMS4 and CMS1 patients with high-low vs. low-high for COADREAD patients. We applied the pipeline for BC subtypes comparison of LumA vs TNBC, LumB vs TNBC, Her2+ vs TNBC. Genes with Benjamini-Hochberg adjusted False Discovery Rate (FDR) <0.05 were considered to be significantly differentially expressed for further steps. For each cohort, we identified 20,531 total genes by mRNAseq raw counts.

The Microarray data obtained from EMBL-EBI and Colorectal Cancer Subtyping Consortium had CMSs defined from the original study. We grouped the patients by CMSs for DEG analysis. Probe identification numbers were converted into gene symbols. The Affy package normalized gene expression values [25]. The Limma R package was applied to identify the DEGs between CMS1 and other CMSs [26]. Genes with P values < 0.05 and adjusted P values <0.05 as the DEGs cutoff.

COHCAP R package was used to identify differentially methylated regions (DMRs) [27]. Δβ and p-value between every two groups were calculated based on β -values of CpG sites using COHCAP. DMRs were defined as methylated regions with $|\Delta\beta| > 0.25$ and P-value <0.05 (hypermethylated DMRs: $\Delta\beta > 0.25$ and P-value <0.05; hypomethylated DMRs: $\Delta\beta < 0.25$ and P-value <0.05) for both COADREAD and BC patients. Gene symbols were annotated based on DMRs using the annotation files of methylation profiling. DMRs corresponding to no gene symbol and multiple gene symbols were not obtained for further analysis. Genes with DMRs were defined as differentially methylated genes (DMGs).

2.3.4 INTEGRATING CROSSTALK GENES

Integrating crosstalk genes with MCP counter score, neoantigen, tumor mutation burden and cell type score to determine CL activity modulators and possible mechanisms. genomic instability, epigenetic abnormality, and gene expression dysregulation are primarily identified hallmarks of COADREAD. Tumorigenesis is never caused by a single factor. Genetic and epigenetic factor cooperate with other factors are responsible for COADREAD progress. The goal of this Aim is to test the hypothesis that key modulators would regulate CTL activity and CD8+ T cell dysfunction through multi-omics.

Then for every patient, we implemented the R package Microenvironment Cell Populations (MCP)-Counter on the mRNAseq by Expectation-Maximization (RSEM) normalized mRNAseq data to create cell type abundance scores [21]. Ten cell populations are simultaneously quantified in the tumor microenvironment, including eight immune cell populations (T cells, CD8 T cells, Cytotoxic lymphocytes, NK cells, B lineage, Monocytic lineage, and Myeloid dendritic cells, Neutrophils), endothelial cells, and fibroblasts. Finally, we used heatmap from ggplot2 R packages to visualize the percentile of MCP-Counter scores for each CMS, BC subtypes and cell population.

We identified the DEGs from mRNAseq DEG analysis and methylation DMG as crosstalk genes from each comparison. Pearson's r correlation coefficient performed correlations between crosstalk DEGs mRNAseq expression with cytotoxic lymphocytes scores and $|r| \geq 0.7$ used to determine the highly correlated genes.

We separate patients into four subgroups based on the median of TBX21 expression and mean of TBX21 methylation β value. The Kaplan–Meier survival curves generated were assessed by the Cox regression model for each immune functional differentially expressed gene using Prism 7. The survival curves between each subgroup of patients were compared by log-rank test. Overall survival was used as the endpoint for each patient, either the days from diagnosis to death or the last follow-up.

Pathway enrichment analysis was performed to evaluate the pathways associated with differentially expressed genes of high-low vs. low-high in CMS1 patients. We highlighted the DEGs on the bar chart. And, these DEGs are also markers CD8+ T_{EX}. Barplot was used to illustrate the comparison of enriched Reactome pathways among differentially expressed genes [21]. These results were analyzed by clusterProfiler, DOSE, and ReactomePA R packages [28-30].

Tumor mutation burden, neoantigens analysis and cell marker score calculation. TMB is a measure of the total number of mutations per megabyte of tumor tissue. The mutation density of tumor genes is also defined as the average number of mutations in the tumor genome, including the total number of gene coding errors, base substitution insertions, or deletions. The 38 Mb is routinely taken based on the length of the human exon, so the TMB estimate for each sample is equal to the total mutation frequency/38. TMB per megabase is calculated by dividing the total number of mutations by the size of the coding region of the target.

We retrieved neoantigens data from TCIA [31]. For each patient, we counted the number of genes, neopeptides generated by each gene and the uptake HLA alleles. And we applied ordinary one-way ANOVA for the statistical analysis.

We utilized mRNAseq expression of identified signature genes for each immune cell type. The modified geometric mean, jamGeomean R package, was applied to calculate the cell marker score, which is value of normalized each signature gene's mRNAseq expression of each cell population. Machine learning-based methods identified these cell marker genes based on all TCGA patients and other large datasets from the TCIA database [31].

2.3.5 SCRNASEQ DATA ANALYSIS

All R packages were used to process the 10× scRNAseq data analyses under R version 4.0.3. SingleCellExperiment and scatter R packages were used to integrate cell_annotation and raw_UMI_count_matrix data as SingleCellExperiment object was collected from GEO [32]. After removing genes not expressed in any cell, we normalized the SingleCellExperiment object by log2-transformed. Principal component analysis (PCA) was performed based on the 27,953 genes to analyze the normalized object. Based on the available annotation from the original study, we subset the "Tumor" from the "Class" annotation label, including 47,285 COADREAD cells. R package with Blueprint/EncodeData reference was applied to annotate immune cell types [32]. Next, we collected cell subtypes annotated by "CD8+ T-cells","CD8+ Tcm" and "CD8+ Tem" as an overall CD8+ T-cells. Seurat R package was used to convert SingleCellExperiment object to Seurat object, and the "FindVariableFeatures" function was used to select the top 2000 highly variable genes [32]. The "FindClusters" function with resolution 0.5 was applied to the Seurat object following CD8+ T-cells clustering. Uniform manifold approximation and projection (UMAP) were applied to explore the subclusters. After identifying the CD8+ T-cells subclusters, the "FindAllMarkers" function was used to define highly differentially expressed genes between clusters. Moreover, the "Idents" and "Simplot" functions were utilized to visualize the overlap between annotated CD8+ T-cell subtypes with CD8+ T-cell subclusters. CellMarker dataset was applied to recognize the CD8+ T-cell subclusters by highly differentially expressed genes [33].

2.4 PAN-CANCER ICI-SENSITIVE PATIENT' FEATURES DETECTION AND SUBGROUP MINING

Integrated phenotypic, bulk genomic data from TCGA dataset to explore subgroups, which would potentially benefit from ICIs. The subgroups would determine potential genomic targets and features in Pan-cancer scales. The goal of this Aim is to test the hypothesis that there are conserved modulators with certain phenotypical patterns would be responsible to regulate CL activity and CD8+ T cell dysfunction in pan-cancer. To address this hypothesis, we will: (1) Collecting solid tumor cancer which FDA approved to apply ICIs from TCGA data for the primary selection. Applying previous created pipeline to categorize each patient phenotypical and genotypical data. (2) Applying the pipeline developed previously for identifying targetable crosstalk gene across multi-omics. (3) Running exploratory subgroup contrast mining to determine subgroups and key modulators would possibly regulate CTL activity and CD8+ T cell dysfunction through multi-omics in pan-cancer.

2.4.1 COLLECTING SOLID TUMOR TYPES WHICH FDA APPROVED TO APPLY ICIS FROM TCGA DATA FOR THE PRIMARY SELECTION

Applying previous created pipeline to categorize each patient phenotypical based on the CL score. FDA has approved ICIs are approved to treat some people with a variety of cancer types such as triple-negative breast cancer, MSI-h colorectal cancer but not all cancer. The evidence for supporting approved solid tumor types and patients who would benefit from ICIs are not concrete. However, it has been well accepted that the effectiveness of ICIs is highly and positively correlated with CTL infiltration. We assumed that for each patient with the high level of CTL score would has varies status with tumor mutation burdens, subtypes, cell marker scores, etc.

	DATA TYPE	DATA SOURCE	ANALYSIS PLATFORM/TOOL
mRNAseq	raw counts, normalized counts, zscore	TCGA-Firehose Legacy	EdgeR, cBioPortal
Methylation	HM450 β-value	TCGA-Firehose Legacy	cBioPortal
Mutation	Processed mutated Gene	TCGA-Firehose Legacy	cBioPortal
CNV	copy number level	TCGA-Firehose Legacy	cBioPortal
RPPA	protein level	TCGA-Firehose Legacy	cBioPortal
Clinical	patient based clinical	TCGA-Firehose Legacy	cBioPortal

Table 1. Data Omics and Data resource

Currently, Immune checkpoint inhibitors have been approved to treat some people with a variety of cancer types, including BRCA, bladder cancer (BLCA), cervical cancer (CESC), COADREAD, head and neck cancer (HNSC), liver cancer (LIHC), lung cancer (LUAD,LUSC), renal cell cancer (KIRC,KIRP), skin cancer (SKCM), esophagus and stomach cancer (ESCA, STAD) or any solid tumor that is not able to repair errors in its DNA that occur when the DNA is copied. We next collected all the omics data from TCGA data from TCGA-Firehose Legacy, cBioPortal for 13 selected solid tumors mostly were FDA approved for receiving (Table 1). The thirteen cancers we included were breast cancer, bladder cancer, cervical cancer, colorectal cancer, head and neck cancer, liver cancer, lung cancer, renal cell cancer, skin cancer, stomach cancer. Again, we will integrate pertinent clinical data (age, race, BMI, gender, anatomical location, prelabeled subtypes, pathologic stage, TNM classification, vital status) and genotypical data (mRNA-seq, methylation, mutation, CNV, CTL score, TMB) by participant ID.

2.4.2 PATIENT STRATIFICATION AND VARIABLES CATEGORIZATION

Finding a homogeneous subgroup cohort is a crucial step in enabling precision medicine and clinical decision support. In this study, the focus was to systematically and strategically group patients into phenotypic subgroups based on their genotypic characteristics. The exploratory data mining method was extended by outcome-oriented analysis to guide the subpopulation discovery process. This exploratory data mining method provides an automatic subpopulation discovery tool that computationally investigates a large pool of subpopulations that have underlying factors differentiating each subpopulation within a given larger group (Figure 5).

For each type of selected cancers, we enrolled 14-17 phenotypical variables. The differences of the number of variables we enrolled for each cancer caused by the clinical data we collected. For example, BRAC patients doesn't have patient's weight and height recorded, therefore, BRAC patients doesn't have BMI as variables. Moreover, except BRCA, COADREAD and KIRP patients, the rest of other cancers don't have subtype were recorded in the TCGA clinical data.

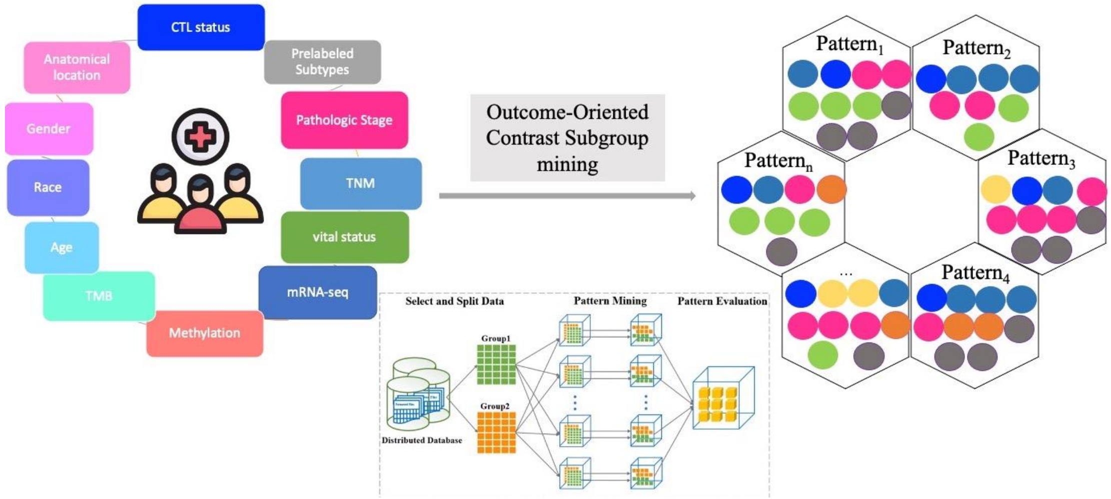


Figure 5. Each container indicates each variable.

And each variable would be categorized to multiple categorical sub-variables. After we categorized each patient each variable, subgroup mining will start to analyze and group the commonly occurred variables together. If a patient population all has the similar variables, then this population forms a pattern. After we have all patterns, outcome-oriented patterns would help us to identify how clinical meaningful of each pattern.

The results are presented as subgroups, each defined by a set of population criteria and underlying factors which differentiate each subgroup from the CTL-high vs. CTL-low from each cancer type. These criteria are the phenotypic variables, such as gender, age, and cancer stage. We then categorized each phenotypic variables based on pre-label or quartile (Table 2).

As table 2 showing, we totally enrolled 17 phenotypical variables for 13 cancers. The first 12 variables are general variables. The last 5 variables are specific for certain cancers. Anatomical location, gender, race, subtypes, tumor site, breslow depth, clark level at

Phenotypical Variable	Categorized variable
Age	Quartile
Anatomical Location	Labelled In TCGA
BMI	Quartile
Gender	Labelled In TCGA
Pathologic Categories M	M0, M1, MX
Pathologic Categories N	N0, N1, N2, N3, NX
Pathologic Categories T	T1, T2, T3, T4
Pathologic Stage	Stage I, Stage II, Stage III, Stage IV
Race	Labelled In TCGA
Subtypes	Labelled In TCGA
Tumor Site	Labelled In TCGA
Vital Status	Alive/Dead
Breslow Depth (SKCM)	Labelled In TCGA
Clark level at diagnosis (SKCM)	Labelled In TCGA
Primary Tumor Laterality (KIRP)	Labelled In TCGA
AFP At Procurement (LIHC)	Labelled In TCGA
Liver fibrosis ishak score category (LIHC)	Labelled In TCGA

Table 2. Phenotypical variables and categorization

diagnosis, primary tumor laterality, AFP at procurement, and liver fibrosis ishak score category were labelled by the recorded patient ID – based information. Age and BMI were applying quartile. And pathologic categories TNM and pathologic stage were labelled by the one higher categories, such as in LIHC, we labelled T3a, T3b patients by T3. As well as pathologic categories N, M and pathologic stage.

An example subgroup could be males, stage II cancer when got diagnosis at median high age across same type of cancer patients. When a phenotypic feature is added, a focus subpopulation is created and contrasted with the rest of the population. Adding additional population variables is desired, assuming there is statistical evidence to do so. The determination of the significance of a subgroup is based on underlying factors which are the genotypic patterns that are statistically unique to the subgroup in comparison to the rest of the population by utilizing Contrast Pattern Mining [34].

The patient subgroup stratification module takes a three-level approach. The top-level method, path expansion, includes a large number of second-level floating subgroup selection processes, each of which is supported by a series of third-level Inclusion and Exclusion procedures. This method is exploratory and differs from a decision tree approach in which samples are divided based on the decision for each node, and each leaf node contains a group of samples which are exclusively in a particular node. Unlike a traditional decision tree, the proposed method has a large number of dynamic fanouts for each node without dividing the samples during the expansion process, and each node represents a subgroup. As a result of the patient subgroup stratification process, a patient could be in multiple subgroups through branching expansion.

Other than the phenotypical variables, the number of genotypical variables we selected and enrolled are varies based on each cancer characteristics, for example, SKCM has 610 phenotypical and genotypical variables enrolled for subgroup mining; however, the LUAD has the fewest number of variables, which is 63 totally. The differences of the number of variables most caused by the number of DEGs from CTL- high vs. CTL-low comparison and if these DEGs were associated with survival.

After collecting thirteen cancers and eight omics, for every patient in each cancer, we implemented the R package MCP-Counter on the mRNAseq by Expectation-Maximization (RSEM) normalized mRNAseq data to create cell type abundance scores. Ten cell populations are simultaneously quantified in the tumor microenvironment, including eight immune cell populations (T cells, CD8 T cells, CTL), NK cells, B lineage, Monocytic lineage, and Myeloid dendritic cells, Neutrophils), endothelial cells, and fibroblasts. We then processed CTL score to quartile then label patients under lower quartile with “low”; lower quartile to median with “median-low”; median to upper quartile with “median-high” and patients were out of upper quartile with label “high”. We then group patients with median-high and high CTL score to CTL-high, and patients with median-low and low CTL score to CTL-low.

mRNAseq differential gene expression analysis was performed with the edgeR package using the raw data downloaded from the TCGA dataset Illumina- HiSeq and TCGA_Illumina-GA platforms [24]. Differentially expressed genes were defined as genes with an |logFC| >1 and p-Value <0.05 for comparisons of CTL-low and CTL-high. Genes with Benjamini-Hochberg adjusted False Discovery Rate (FDR) <0.05 were considered to be significantly differentially expressed for further steps. For each cohort, we identified 20,531 total genes by mRNAseq raw counts.

Next, we performed survival analyses using the identified DEGs with patient's vital status and last to follow up days were recorded from TCGA. The normalized RNA-seq data of differentially expressed genes used for survival analysis was processed using the Survival R package. For each differentially expressed gene, if the normalized gene expression value was more than the median level, we labeled it as "high," and otherwise as "low." The Kaplan–Meier survival curves generated were assessed by the Cox regression model for each immune functional differentially expressed gene using the Survminer R package. The survival curves of patients with high gene expression and low gene expression were compared by log-rank test. For each patient, overall survival was used as the endpoint, either the days from diagnosis to death, or to the last follow-up. The DEGs associated with survival were going to use for the next steps integration.

2.4.3 INTEGRATION OF IDENTIFIED TARGETABLE CROSSTALK GENE ACROSS MULTI-OMICS

We have known that tumorigenesis is the gain of malignant properties in normal cells, including primarily dedifferentiation, fast proliferation, metastasis, evasion of apoptosis and immunosurveillance, dysregulated metabolism and epigenetics, etc., which have been generalized as the hallmarks of cancer. In another word, any tumorigenesis is not caused by one progress, it must be contributed by complicated mechanisms and combination of these hallmarks. But each these hallmarks may share the common features. We hypothesized that different type of tumors may have its dominated tumorigenesis mechanisms, therefore, integrate multi-omics by key modulators will show the common features in different tumorigenesis mechanisms of pan-cancer scale.

The cBioPortal for Cancer Genomics (http://cbioportal.org) provides a Web resource for exploring, visualizing, and analyzing multidimensional cancer genomics data. The portal reduces molecular profiling data from cancer tissues and cell lines into readily understandable genetic, epigenetic, gene expression, and proteomic events. The query interface combined with customized data storage enables researchers to interactively explore genetic alterations across samples, genes, and pathways and, when available in the underlying data, to link these to clinical outcomes. The portal provides graphical summaries of gene-level data from multiple platforms, network visualization and analysis, survival analysis, patient-centric queries, and software programmatic access. The intuitive Web interface of the portal makes complex cancer genomics profiles accessible to researchers and clinicians without requiring bioinformatics expertise, thus facilitating biological discoveries.

We selected 13 cancers from more than 25 cancer studies. When selecting genomic profiles, mutations and CNVs are specified by default. When available, relative mRNA expression or relative protein and phosphoprotein abundance data can also be selected. Protein and phosphoprotein data are based on reverse phase protein array (RPPA) experiments. For mRNA data and RPPA, z scores are precomputed from the expression values, and users can specify the threshold or use the default setting (2 SDs from the mean). The z scores for mRNA expression are determined for each sample by comparing a gene's mRNA expression to the distribution in a reference population that represents typical expression for the gene. As we know, Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. We will use it to calculate for each patient by each row. If the Z-Score of a gene is greater than 2, then the gene is considered “over”; if the Z-Score of a gene is less than -2, then the gene is considered “under”; and if the Z-Score of a gene is less than 2 and greater than -2, then the gene is considered “normal”.

Since we customized comparison groups for each cancer type, then based on patient ID with pre-labelled CTL-high and low to defining CTL-high and CTL-low patients’ sets for each cancer analysis. The default option is set to match the selected genomic profiles and we can choose two groups to compare by cBioPortal. For example, cases with sequencing data will be selected if querying for mutations only. However, the user can change this selection by choosing from the drop-down list of case sets defined by the available data (for example, tumors with mutations, CNA data, gene expression, or RPPA data) or by known tumor subtypes. For example, from Data sets, we select Bladder Urothelial Carcinoma (TCGA, Firehose Legacy). In Groups module, we have grouped CTL-high and low patients. Selecting two groups' patients to compare and analyze, cBioportal gives us immediate analysis and results. There are overlap, Survival, Clinical, Genomic Alterations, mRNA, protein and DNA methylation results from the comparison of CTL-high and low. Since we are interested to integrate DEGs associated with survival and copy number variation/alterations (CNV), RPPA, DNA methylation, we then download the analysis results for CNV, RPPA and DNA methylation.

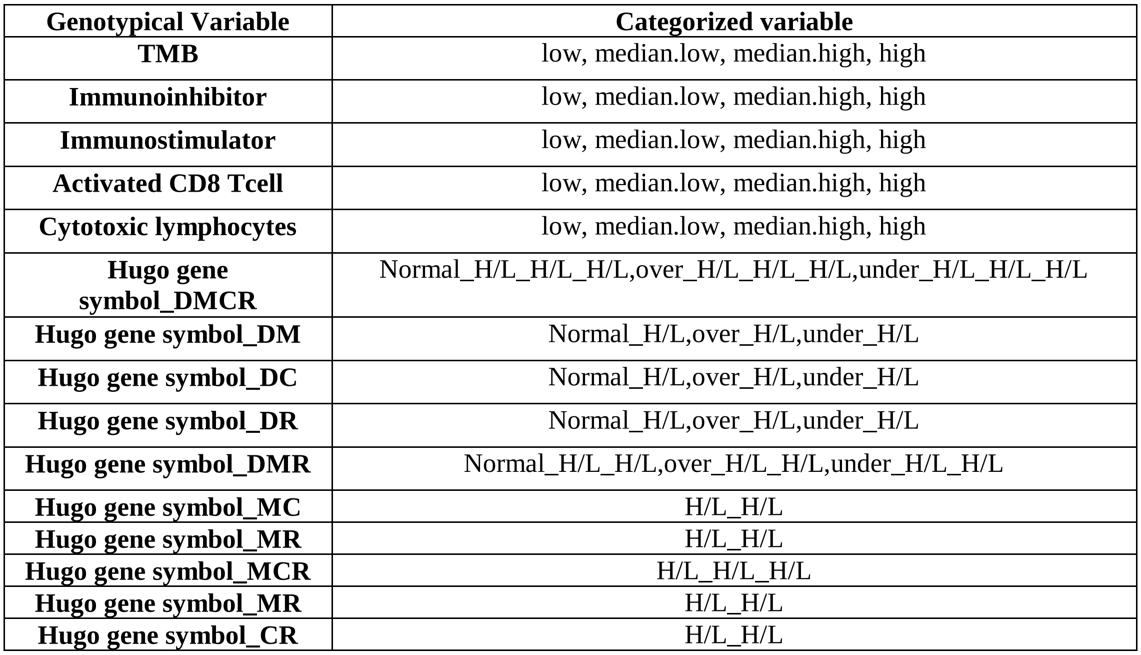
Thresholded copy number calls in the TCGA Firehouse Legacy datasets are generated by the GISTIC 2.0 algorithm and obtained from the Broad Firehose. The results for CNV contain Hugo gene symbol, cytoband, number patients and percentage in the CTL-high or low with this gene had amplification or deletion, co-occurrence pattern, log Ratio (Log2 based ratio of (percentage (pct) in (A) BLCA_high/pct in (B)BLCA_low)), p-value (Derived from one-sided Fisher Exact test), q-value (Derived from Benjamini-Hochberg procedure) and enriched in (Log ratio >0 Enriched in (A)BLCA_high Log ratio <=0:Enriched in (B) BLCA_low). For identifying if there are significances between CTL-high vs. low, we primally choose p-Value < 0.05 as the threshold to select the differential copy number alteration genes.

In the RPPA result, it included gene Hugo symbol, cytoband, μ in (A)BLCA_high (Mean log2 expression of the listed gene in samples in (A)BLCA_high), μ in (B)BLCA_low (Mean log2 expression of the listed gene in samples in (B)BLCA_low), σ in (A) BLCA_high (Standard deviation of log2 expression of the listed gene in samples in (A)BLCA_high), σ in (B) BLCA_low (Standard deviation of log2 expression of the listed gene in samples in (B)BLCA_low), Log Ratio (Log2 of ratio of (unlogged)mean in (A)BLCA_high to (unlogged)mean in (B)BLCA_low), p-Value(Derived from Student's t-test), q-Value (Derived from Benjamini-Hochberg procedure) and Higher expression in (Log ratio >0 Enriched in (A)BLCA_high Log ratio <=0:Enriched in (B) BLCA_low) . As same as CNV, for identifying if there are significant differences between CTL-high vs. low protein expression, we chose p-Value < 0.05 as the threshold to select the genes with differential protein expression.

For the differentially methylated gene (DMG) analysis, cBioPortal applied HM450 methylation data for analysis. The results listed gene Hugo symbol, cytoband, μ in (A)BLCA_high (Mean methylation of the listed gene in samples in (A)BLCA_high), μ in (B)BLCA_low (Mean methylation of the listed gene in samples in (B)BLCA_low), σ in (A) BLCA_high (Standard deviation of methylation of the listed gene in samples in (A)BLCA_high), σ in (B) BLCA_low (Standard deviation of methylation of the listed gene in samples in (B)BLCA_low), Log Ratio (Log2 of ratio of (unlogged)mean in (A)BLCA_high to (unlogged)mean in (B)BLCA_low), p-Value(Derived from Student's t-test), q-Value (Derived from Benjamini-Hochberg procedure) and Higher expression in (Log ratio >0 Enriched in (A)BLCA_high Log ratio <=0:Enriched in (B) BLCA_low) . As same as CNV, for identifying if there are significant differences between CTL-high vs. low protein expression, we chose p-Value < 0.05 as the threshold to select the genes with differential methylated gene.

With DEGs associated with survival, differential CNV, RPPN, and DMG, we did overlap analysis between these four omics. In overlap analysis, we labelled DEGs as D, Methylation by M, CNV by C and RPPA with R. We then identified each two, three and all four omics for overlap genes. To be more specific, we looked up DMCR, DM, DC, DR, DMR, MC, MR, MCR, MR, CR and DMR for overlap genes. For example, DM indicates the overlapped gene between DEGs and DMG. If any of omics identified overlapped gene with DEGs, it also indicated that the genes would associate with patient's survival. In the genotypical variables, the genes would be selected for contrast mining analysis and labelled with "Hugo gene symbol_DM". For examples, in BLCA, gene FBXW4 showed the significant differences in methylation and CNV between patients with CTL-high and low. If patient A with CTL-low score, patient A would be labelled Low_High (L_H). it indicated that FBXW4 gene has lower level of methylation and higher copy number variation in patients with CTL-low status. Any overlapped genes from each two, three or all omics were regarded as key modulators of certain type of cancers.

Besides the overlapped genes as the genotypical variables, TMB is a measure of the total number of mutations per megabyte of tumor tissue. The mutation density of tumor genes is also defined as the average number of mutations in the tumor genome, including the total number of gene coding errors, base substitution insertions, or deletions. The 38 Mb is routinely taken based on the length of the human exon, so the TMB estimate for each sample is equal to the total mutation frequency/38. TMB per megabase is calculated by dividing the total number of mutations by the size of the coding region of the target. We retrieved TMB from cBioPortal clinical data. We then utilized mRNAseq expression of identified signature genes for each immune cell type. The modified geometric mean formula, jamGeomean R package, was applied to calculate the cell marker score, which is value of normalized each signature gene's mRNAseq expression of each cell population. Machine learning-based methods identified these cell marker genes based on all TCGA patients and other large datasets from the TCIA database [31]. For the contrast subgroup mining, we were interested the cell marker score from Activated CD8 T cell, Immunoinhibitor and Immunostimulator. It has evidence to support that Activated CD8 T cell, Immunoinhibitor and Immunostimulator are closely associated with CTL function. Again, we applied quartile to categorize CTL score, TMB, Activated CD8 T cell, Immunoinhibitor and Immunostimulator into low, median.low, median.high and high. We summarized the enrolled genotypical variables to table 3.



Genotypical Variable	Categorized variable
TMB	low, median.low, median.high, high
Immunoinhibitor	low, median.low, median.high, high
Immunostimulator	low, median.low, median.high, high
Activated CD8 Tcell	low, median.low, median.high, high
Cytotoxic lymphocytes	low, median.low, median.high, high
Hugo gene symbol_DMCR	Normal_H/L_H/L_H/L,over_H/L_H/L_H/L,under_H/L_H/L_H/L
Hugo gene symbol_DM	Normal_H/L,over_H/L,under_H/L
Hugo gene symbol_DC	Normal_H/L,over_H/L,under_H/L
Hugo gene symbol_DR	Normal_H/L,over_H/L,under_H/L
Hugo gene symbol_DMR	Normal_H/L_H/L,over_H/L_H/L,under_H/L_H/L
Hugo gene symbol_MC	H/L_H/L
Hugo gene symbol_MR	H/L_H/L
Hugo gene symbol_MCR	H/L_H/L_H/L
Hugo gene symbol_MR	H/L_H/L
Hugo gene symbol_CR	H/L_H/L

Table 3. Genotypical variables and categorization

2.4.4 EXPLORATORY SUBGROUP MINING

Running exploratory subgroup mining to determine subgroups and key modulators would regulate CTL activity and CD8+ T cell dysfunction through multi-omics in pan-cancer. Rationale: ICIs mean to target a homogeneous subgroup which its CD8+ T cell cytotoxicity has been blocked from tumor. Regaining CTL abilities to kill cancer cells is the mission of ICIs. Deep Exploratory subgroup mining would help to identify these homogeneous subgroups. We hypothesized that ICIs sensitive homogeneous subgroups would be determined by deep Exploratory subgroup mining. These homogeneous subgroups would share the similar features of tumorigenesis mechanisms in pan-cancer. The features would be generated from previous categorized phenotypical and genotypical variables.

As discussed in the previous section, the evaluation of subgroup significance is performed by measuring the contrast between the subgroup and its outer population. For each candidate subgroup representing a set of phenotypic characteristics, the algorithm finds all genotypic patterns that are frequent within the subgroup but infrequent in the remaining population. Support is used to evaluate whether a given pattern is frequent in a subgroup and growth rate to evaluate the contrast of the pattern in the selected subgroup [34, 35].

Let *D* be the patient dataset in a subgroup, which includes n genotypic variables, $G = (g1, g2 , \ldots, gn)$. Pattern *p* that is commonly shared within patients in a given subgroup is defined as a set of genotypic variables, such as $p = (g1,e1 ,g2,e2 ,\ldots,gi,ei),$ where $g_{i,ei}$ is the expression level or mutation status of gene *i*. The expression level or the mutation status should be represented as a categorial value. This process is accomplished using the PATTERN_MINING() function in the pseudo-code.

The pattern is "frequent" if its support is greater than a user-defined threshold. The support of pattern *p* is the number of records (patients) that have that pattern ($|<D,p>|$) divided by the total number of records in the dataset *D* (|D|):

$$Support(p, D) = \frac{|<D,p>|}{|D|} \quad (1)$$

To find the contrast pattern (*cp*) between the focus subpopulation and the rest of the population, S_{G1} represents the focused subgroup and S_{G2} represents the remaining population, where $S_{G2} = D - S_{G1}$. The support of the contrast pattern should be significantly different between S_{G1} and S_{G2}. Let *s1* be the support of a contrast pattern in S_{G1} and *s2* the support of the same pattern in S_{G2}. The growth is used to measure the difference between the two groups. The growth of contrast pattern cp between subgroup S_{G1} and the remaining population S_{G2} is defined as follows:

$$Growth\ (cp, SG1, SG2) = \frac{\text{Max}\{s1,s2\}}{\text{Min}\{s1,s2\}} \quad (2)$$

The growth ratio is normalized to be between 0 and 1 using an extended version of the tanh function [36]. Let α be the threshold for the support and β the threshold of growth rate. To ensure that a cp is frequent and has significant differences between the two groups, the following condition should be held:

$$(Support\ (cp, SG1\ ) \geq \alpha\ OR\ Support\ (cp, SG2\ ) \geq \alpha)\ AND\ (Growth\ (cp, SG1\ , SG2\ ) \geq \beta)\ (3)$$

This condition identifies two sets of contrast patterns CP_1 and CP_2 for the target subgroup and the outer population, respectively. For each contrast pattern cp n with multiple genotype variables, the subset of the pattern $cp_i \subseteq cp_n$ will be kept when Growth $(cp_i, S_{G1}, S_{G2}\ ) -$ Growth$(cp_n, S_{G1}, S_{G2}) > 0$. These selected contrast patterns are utilized to evaluate each subgroup during the floating and path expansion procedure discussed in previous section.

2.4.5 OUTPUT AND SUBGROUPS PRIORITIZATION

The number of candidate subgroups selected by the floating and expansion process could be hundreds. The G score could be used to rank the subgroups. The higher the G score, there are larger differences of two populations. The G score is created by William I. Baskett, one of iDAS lab member. The basic idea G score is the rate of two support from each population with adding a small constant value to avoid the infinite value:

$$G\ score = \frac{Support\ (cp,SG1\ )}{Support\ (cp,SG2\ ) + 0.000001}\ (4)$$

Because this method was developed to improve patient care, all steps should be explainable and acceptable for practitioners. For clinically meaningful results, a physician-in-the-loop process was necessary to prioritize the subgroups further using a two-phase method. First, physician-in-the-loop provides a filtering mechanism where the focus will be on only a subset of the subgroups instead of going through the hundreds of subgroups resulting from our method. Second, the physicians may decide the most relevant subgroups by evaluating the top subgroups using the G score or using initial hypotheses formed by clinical observations and literature to prioritize all candidate subgroups. For example, in the COADREAD study, the physician investigators chose to focus on the subgroups with microsatellite (MS) status as one of the clinical variables in the COADREAD case study in which seven subgroups with Microsatellite Instability (MSI) test results were further examined as a phenotypic characteristic among the statistically significant subgroups (p-value<0.05). The rationale for the selection of groups based on MS status is related to therapeutic selection and tumor biology [36]. Microsatellite instable (MSI) tumors are associated with hypermutation due to the inactivation of mismatch repair genes via either germline mutation or methylation, accounting for 13-15% of COADREADs. The remaining 85% of colorectal cancers develop via the chromosomal instability pathway, referred to as microsatellite stable, following a well-described pathway acquiring mutations through the adenoma to carcinoma sequence as described in seminal work by Vogelstein, et al [37]. While these tumors appear to be biologically different, most critically, these tumors are also characterized by different prognosis, response to standard therapy, and response to novel therapy including both targeted and immune-based therapy [38, 39]. Therefore, this designation was felt to be highly clinically relevant. The subgroups that are selected through the physician-in-the-loop process are then chosen as the input for the next step in which drug candidates are evaluated and analyzed for each subgroup.

Chapter 3
Results

3.1 INTRODUCTION

Despite recent advances in detection and therapy, colorectal cancer (COADREAD) remains the second leading cause of cancer-related death in the US [40]. Immune based therapies such as immune checkpoint inhibition have recently made significant advances in a number of difficult to treat malignancies like non-small cell lung cancer, melanoma, and renal cell cancer [41]. However, these results have not yet extended to the majority of patients with COADREAD [39]. This is despite significant data that antitumor immunity is important for prognosis and treatment response in these patients [21, 42]. This suggests that there is significant progress to be made in the application of immune based therapy in COADREAD.

When considering immune based treatments, a critical factor is effective tumor infiltration of cytotoxic lymphocytes [43]. In colorectal cancer, this is evident as patients who demonstrate response to immune checkpoint inhibition and currently have an FDA approved indication for this therapy, are those with microsatellite instability-high (MSI-H) tumors [1]. These tumors are characterized by high mutational load, neoepitope formation, and an intense lymphocytic infiltrate when compared to microsatellite stable (MSS) tumors [44]. Microsatellite instability high tumors, however, are also associated with increased mutations in immune related genes and expression of negative regulatory genes, demonstrating that tumors try to dampen the immune response by multiple pathways [45]. Additionally, recent studies have suggested that the use of other markers including lymphocyte infiltration and tumor mutational burden may better predict survival and the potential for response to immune based therapy [7, 42]. It is therefore critical to develop a better understanding of immune resistance mechanisms to improve therapy in colorectal cancer patients.

In the current era of precision medicine, research is concentrated on providing more effective treatments by focusing on patient specific factors. This is particularly important in colorectal cancer, as subsets of patients responsive to targeted therapy, immune-based therapy, and chemotherapy have previously been identified [8, 9, 39]. Colorectal cancer, however, is a heterogeneous disease made up of multiple subgroups [23]. Even simple clinical characteristics often overlooked in molecular studies, such as anatomic location, are important for prognosis [18, 19]. Despite these differences in subtype and clinical characteristics, T lymphocyte infiltration has been demonstrated to be important for prognosis [42].

Data repositories such as The Cancer Genome Atlas (TCGA) allow for the in-depth study of patients on a molecular and clinical basis. Recently, a novel computational method for predicting the abundance of different cells within the tumor microenvironment using RNA-seq data was developed and validated with histologic specimens called the MCP-Counter [21]. This method allows for an effective comparison of the composition and pathways associated with cellular infiltration in the tumor microenvironment, improving over other methods primarily based on microarray data and gene set enrichment analysis. In this study, we use the MCP-counter program to create tumor CTL abundance scores. After grouping patients based on cytotoxic lymphocyte abundance score, stage, and tumor location, we found one immune pathway that was highly enriched at all tumor locations and stages, the "Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell" pathway, suggesting specific targets to improve immune based therapy in colorectal cancer patients.

3.2 PATIENT CHARACTERISTICS

In the most recently updated TCGA dataset (June 01, 2016), there are 461 colon cancer (COAD) and 172 rectal cancer (READ) cases (Figure 1). Thirty-four patients with colon cancer and 18 patients with rectal cancer were excluded due to missing information. The RNA-seq data with matched clinical data were integrated from COAD patients (Nc = 427) and READ patients (Nc = 154). Cytotoxic lymphocyte abundance scores were generated using the MCP-counter method. Patients were then separated based on the median cytotoxic lymphocyte score (26.72; 95% CI: 24.1–30.1). Patients with cytotoxic lymphocyte scores ≥ the median were

Characteristic	High	Low	p Value
Gender, n			
RSC	156	96	0.3637
Male	77	54	
Female	77	42	
LSC	68	106	0.1231
Male	41	51	
Female	27	55	
REC	63	91	0.5118
Male	37	48	
Female	26	43	
MS, n			
RSC	156	96	<0.0001*
MSS/MSI-L	92	89	
MSI-H	64	7	
LSC	68	106	0.0771
MSS/MSI-L	64	105	
MSI-H	4	1	
REC	63	91	0.0264*
MSS/MSI-L	59	91	
MSI-H	4	0	
Age,Mean ± SD			
RSC	70±13.69	67±12.95	0.2222
LSC	66±11.11	64±13.07	0.4267
REC	64±10.56	65±12.32	0.81
Pathologic stage, n (%)			
RSC	156	96	0.0278*
I	33(21.2%)	11(11.5%)	
II	73(46.8%)	38(39.6%)	
III	38(24.4%)	32(33.3%)	
IV	12(7.7%)	15(15.6%)	
LSC	68	106	0.1374
I	12(17.6%)	17(16.0%)	
II	26(38.2%)	33(31.1%)	
III	23(33.8%)	30(28.3%)	
IV	7(10.3%)	26(24.5%)	
REC	63	91	
I	10(15.9%)	20(22.0%)	0.0279*
II	29(46.0%)	21(23.1%)	
III	17(27.0%)	33(36.3%)	
IV	7(11.1%)	17(18.7%)	

Table 3. Patients’ characteristics

RSC Right-sided colon cancer, LSC Left-sided colon cancer, REC Rectal cancer, MSS Microsatellite stable, MSI-L Microsatellite instability-low, MSI-H Microsatellite instability-high *Indicates statistically significant difference (p < 0.05) classified as cytotoxic lymphocyte-high (CTL-high) and those with scores < the median were classified as cytotoxic lymphocyte-low (CTL-low). We then confirmed that there was a significant difference in cytotoxic lymphocyte scores between CTL-high and CTL-low groups (73.66 ± 64.2 v 14.07 ± 6.69, $p < 0.0001$). Colorectal cancer patients were then separated by anatomical location, cytotoxic lymphocyte score, and stage (Fig. 1). The demographic, clinical, and pathologic characteristics of each patient cohort is summarized in Table 1. Microsatellite status composition of patients with CTL-high and CTL-low tumors was significantly different in the right-sided colon cancer ($p < 0.0001$) and rectal cancer ($p = 0.0264$) groups with more MSI-H patients among the CL-High patients at both locations (Table 3). Additionally, cytotoxic lymphocyte scores correlated significantly with pathologic tumor stage in right-sided colon cancer ($p = 0.0278$) and rectal cancer ($p = 0.0279$) patients, but not in left-sided colon cancer patients (Table 3). This analysis demonstrates that there are significant differences based on tumor location, suggesting that this variable is an important consideration when analyzing patient data [18, 46].

3.3 DIFFERENTIAL GENE EXPRESSION ANALYSIS

For each cohort, we next performed RNA-seq differential gene expression analysis. Expression of 20,531 genes was determined for each tumor sample from the TCGA. Then gene expression was compared between patients with CL-High and CL-Low tumors in each cohort based on tumor location and stage using the edgeR. In right-sided early, localized, and metastatic colon cancer patients, 1882, 1781, and 1054 differentially expressed genes were observed, respectively. In left-sided patients, 805, 925, and 1255 genes were differentially expressed in each stage. And in rectal cancer patients, 888, 1316 and 150 genes were differentially expressed at each stage (Figure 7). In the left-sided

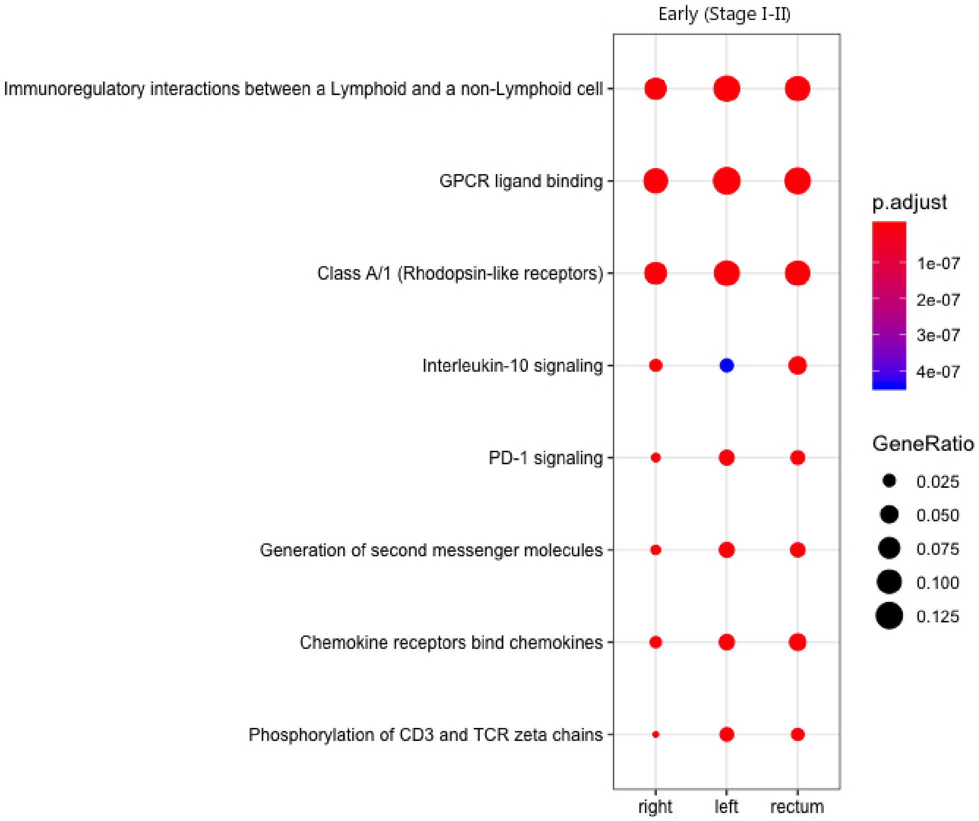


Figure 7a. Pathway Enrichment analysis based on tumor location and stage.

Pathway enrichment was ranked using a composite of the adjusted p value for false discovery rate (color) and gene ratio (size). a) pathway enrichment for early stage COADREAD (Stage I and II) based on location;

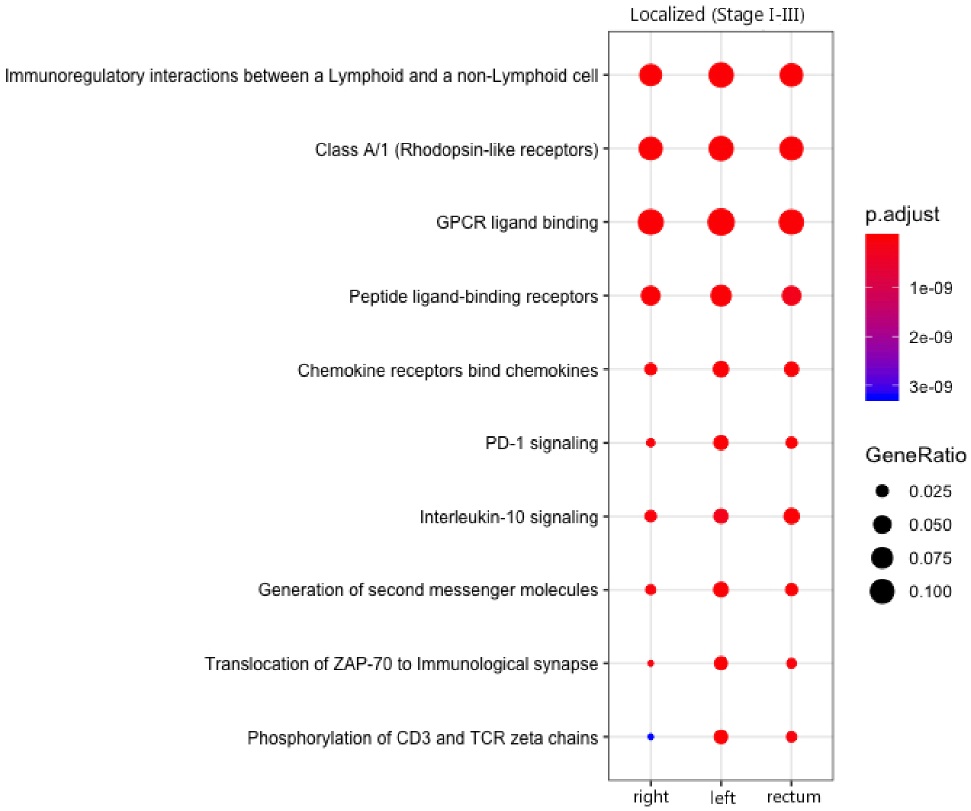


Figure 7b. pathway enrichment analysis for localized COADREAD (Stages I-III) based on location;

group, differentially expressed genes were highest in the metastatic cohort; however, in right-sided and rectal cancer patients, the metastatic cohort had the lowest number of differentially expressed genes. This again suggests that both tumor location and stage are important considerations when analyzing alterations in gene expression. Differentially expressed genes found in the above analysis were subsequently imported into the Reactome Pathway Browser to determine involvement in immune related functional pathways (Fig. 7). Interestingly, despite significant variation in the number of differentially expressed genes, the ratio of genes associated with immune function was similar in all sites and stages.

3.4 PATHWAY ENRICHMENT AND SURVIVAL ANALYSIS

To further determine whether there were overlapping pathways associated with cytotoxic lymphocyte infiltration in colorectal cancer, we then compared the Reactome pathway enrichment analysis at each location based on stage. Using the p value adjusted for false discovery rate and the ratio of differentially expressed genes in each pathway, we found that the “immunoregulatory interactions between a lymphoid and a non-lymphoid cell” was the most highly enriched pathway in early and local patients at all tumor locations. Additionally, this was the most highly enriched pathway in patients with metastatic right-sided cancer. This pathway was also among the top pathways enriched among patients with metastatic left-sided colon cancer and rectal cancer (Figure 8).This suggests that despite significant heterogeneity among subjects, redundant pathways of deregulation may be conserved across stage and location.

To further understand the potential for targetable genes within this pathway, we then took differentially expressed genes in the “immunoregulatory interactions between a lymphoid and non-lymphoid cell” pathway and performed a survival analysis using the KaplanMeier

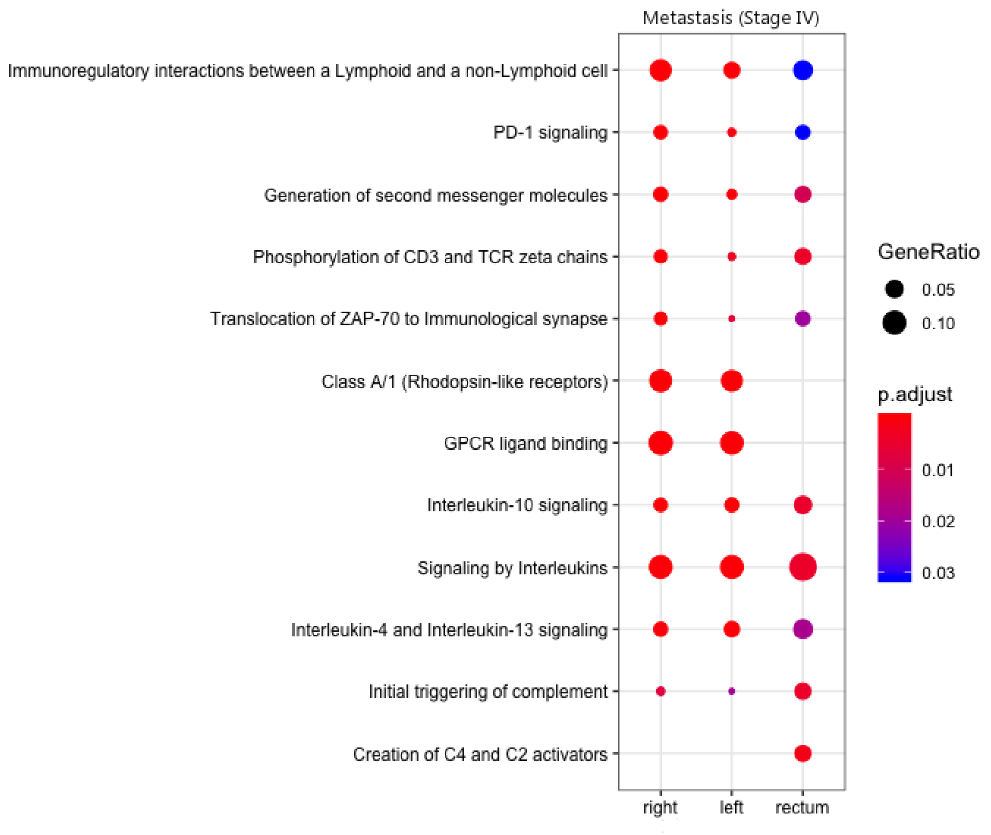


Figure 8. pathway enrichment analysis for metastatic stage COADREAD (Stage IV) based on location.

This analysis showed that the “immunoregulatory interactions between a lymphoid and a non-lymphoid cell” was the most highly enriched pathway in early and local patients at all sites. Additionally, this was the most highly enriched pathway in patients with metastatic right-sided cancer. This pathway was also among the top pathways enriched among patients with metastatic left-sided colon cancer and rectal cancer. (Input data included in Additional file 2)

method and Cox-Proportional Hazards Model based on differentially expressed gene, location, cytotoxic lymphocyte score, and pathologic stage (Figure 9, Table 4). There are a total of 297 genes included in this pathway, and we found 21

(7.1%) unique genes associated with survival in this pathway. As figure 1 demonstrates, the number of differentially expressed genes were variable with most genes associated with survival in the right-sided colon cancer group. Additionally, the positive and negative impact of differentially expressed genes on survival depended

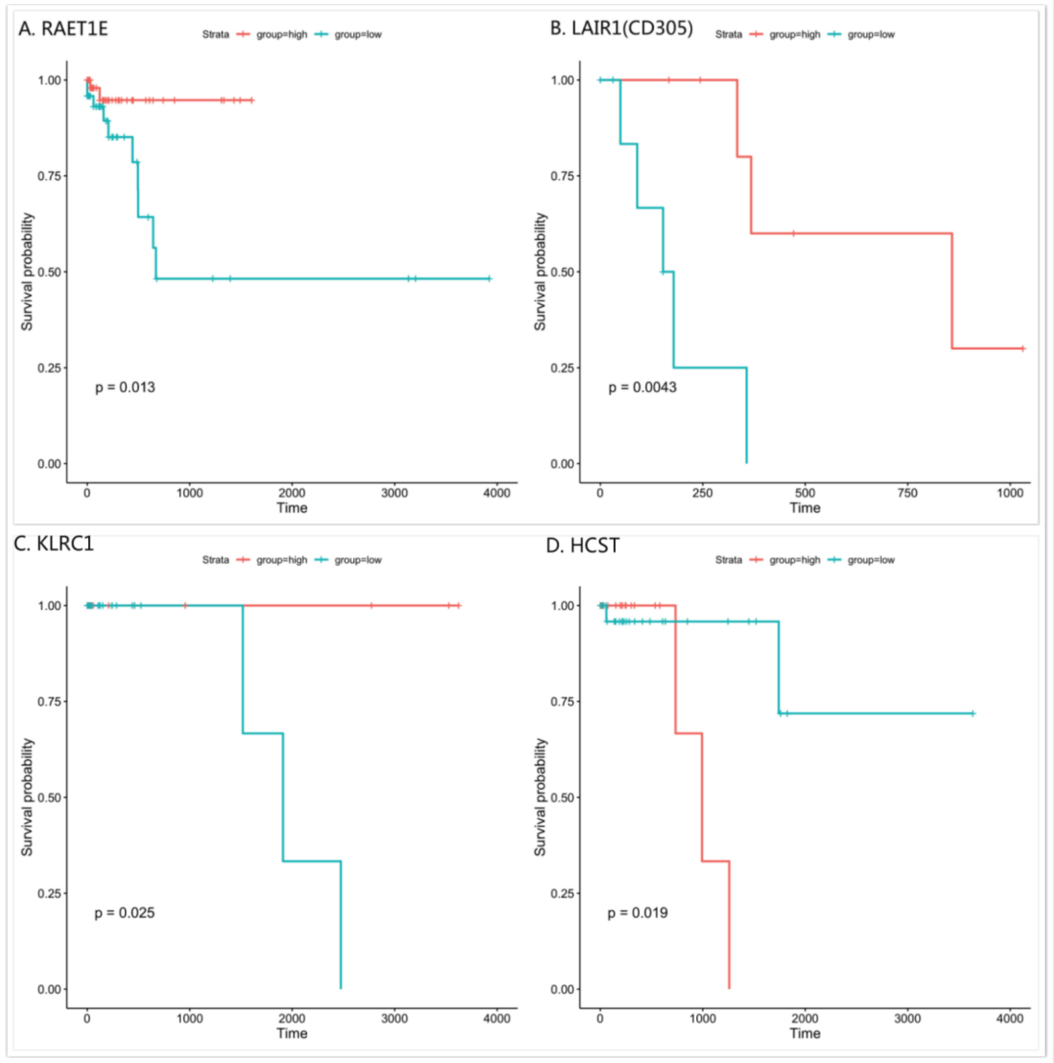


Figure 9. Representative survival curves based on tumor location and stage.

Survival curves with the p values derived from Kaplan-Meier analysis. a RAET1E was positively associated with survival in right-sided colon cancer patients with high cytotoxic lymphocyte scores in early and localized stages; b LAIR1(CD305) was positively associated with survival in right-sided colon cancer patients with low cytotoxic lymphocyte scores in the metastatic stage; c KLRC1 was positively associated with survival in left-sided colon cancer patients with low cytotoxic lymphocyte scores in the early stage; d HCST was negatively associated with survival in rectal cancer patients with low cytotoxic lymphocyte scores in the localized group.

on cytotoxic lymphocyte abundance scores. The majority of genes associated with a positive impact on survival (bold in Table 4) were in the CTL-low group whereas the majority of genes with a negative impact on survival (italicized in Table 4) were

Table 4a. Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell

Location	Stages	CLs Status	Pathways	Gene Symbol	HR	95% CI
Right	Early (I,II)	High	NKG2D homodimer interacting with ligands	**RAET1E**	0.88	0.78, 1.00
			LILRs interact with MHC Class I	*LILRA1*	1.04	0.98, 1.09
			Sialic acid binds SIGLEC	CD33	1	0.98, 1.02
		Low	NA			
	Localized (I,II,III)	High	NKG2D homodimer interacting with ligands	**RAET1E**	0.9	0.81, 0.99
			LILRs interact with MHC Class I	*LILRA1*	1.04	0.99, 1.09
				LILRA4	1.02	0.98, 1.06
			Sialic acid binds SIGLEC	CD33	1	0.98, 1.02
		Low	KLRC1:KLRD1 heterodimer interacts with HLA-E	*KLRC1*	1.23	0.93, 1.63
	Metastasis (IV)	High	MADCAM1-1 binds Integrin alpha4beta7	*MADCAM1*	1.02	0.94, 1.11
			Ligands bind L-selectin	*MADCAM1*		
			Fc gamma receptors interact with antigen-bound IgG	*FCGR2B(CD32)*	1.01	0.98, 1.03
			Sialic acid binds SIGLEC	*SIGLEC8(CD329)*	1.06	0.94, 1.21
			PILRA binds PIANP, COLLEC12 trimer, NPDC1, CLEC4G	*CLEC4G*	42.7	0.00, 469970
		Low	CD96 binds PVR	**CD96**	0.98	0.93, 1.03
			LILRs interact with MHC Class I	**LILRB1**	0.97	0.94, 1.00
			ICAM1-5 bind Integrin alphaLbeta2 (LFA-1)	ITGB2(CD18)	1	1.00, 1.00
			CD40L binds CD40	**CD40LG**	0.97	0.89, 1.06
			C3d-complexed antigen binds to complement receptor	C3	1	1.00, 1.00
			Sialic acid binds SIGLEC	**SIGLEC9**	0.96	0.89, 1.04
			SAP and EAT2 binds SLAMF6	**SH2D1A**	0.95	0.85, 1.05
			LAIR1 binds collagen	**LAIR1(CD305)**	0.99	0.98, 1.00
			TREM,CD300 binds lipids	**CD300A**	0.99	0.97, 1.00

Table 4b. Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell

Left	Early	High	NA			
		Low	KLRC1:KLRD1 heterodimer interacts with HLA-E	**KLRC1**	0.21	0.02, 1.78
	Localized	High	NA			
		Low	NA			
	Metastasis	High	NA			
		Low	CLEC2B binds KLRF1 dimer	**CLEC2B**	0.91	0.79, 1.05
Rectum	Early	High	NA			
		Low	PILRA binds PIANP, COLLEC12 trimer, NPDC1, CLEC4G	*CLEC4G*	1.55	0.55, 4.37
	Localized	High	NA			
		Low	NKG2D homodimer interacting with ligands	*HCST*	1.09	0.99, 1.19
			viral HA binds NCRI	*FCGR1A(CD64)*	1.05	0.97, 1.14
			NCR3LG1 binds NCR3	*FCGR1A(CD64)*		
			CMVPP65 binds NCR3	*FCGR1A(CD64)*		
			PILRA binds PIANP, COLLEC12 trimer, NPDC1, CLEC4G	*CLEC4G*	1.19	0.40, 3.52
	Metastasis	High	NA			
		Low	NA			

found in the CTL-High group. In CTL-High right-sided colon cancer patients with metastatic disease, all differentially expressed genes had a negative impact on survival; however, all immune functional differentially expressed genes in the CTL-Low group had a positive impact on survival. Often, patients with rectal cancer and left-sided cancer are considered to have similar disease biologically. While we found few differentially expressed genes in this pathway associated with survival in the left sided and rectal cancer groups, differentially expressed genes in the left-sided colon cancer group primarily had a positive impact on survival with the converse being true in patients in the rectal cancer group. Together this data demonstrates that even within conserved immune pathways, there is significant heterogeneity in the impact on patient survival. This further suggests the importance of a patient centered approach for the application of immune based therapy in colorectal cancer.

3.5 IMMUNOGENOMIC PATHWAY AND SURVIVAL ANALYSIS IN BREAST CANCERS BASED ON TUMOR LOCATION AND MOLECULAR SUBTYPES

3.5.1 DIFFERENTIAL GENE EXPRESSION ANALYSIS

Using the TCGA dataset, we have identified 918 breast cancer tumor samples and compared mRNAseq gene expression based on molecular subtypes and anatomic locations of biopsies (i.e., right , left , lower inner quadrant (LI), lower outer quadrant (LO), upper inner quadrant (UI) or upper outer quadrant (UO). After group 918 into six groups, for each cohort, we next performed RNA-seq differential gene expression analysis. Expression of 20,531 genes was determined for each tumor sample from the TCGA. Then gene expression was compared between patients with CTL-High and CTL-Low tumors in each cohort based on tumor location and stage using the edgeR package. In the UO group, there were 307 non-TNBC patients and 27 TNBC patients; in the patients who labelled with Right, there were 144 non-TNBC patients and 7 TNBC patients; in the LO group, there were 67 non-TNBC patients and 13 TNBC patients; in the UI group, there were 138 non-TNBC patients and 10 TNBC patients; in the left group, there were 155 non-TNBC patients and 6 TNBC patients; in the LI group, there were 40 non-TNBC patients and 4 TNBC patients (Table 5).

For each cohort, we next performed RNA-seq differential gene expression analysis. Expression of 20,531 genes was determined for each tumor sample from the TCGA. Then gene expression was compared between patients with non-TNBC and TNBC tumors in each cohort based on tumor location and stage using the edgeR package. In UO, right, LO, UI, left, LI, BRAC patients, 3794, 2622,2447,3233,2379, and 1847 differentially expressed genes were observed, respectively. After applying Reactome pathway enrichment analysis, in UO, right, LO, UI, left, LI, BRAC patients, we identified 320, 291, 318, 301,208 and 168 immune associated DEGs. The percentage of these immune associated DEGs were 8.43%, 11.10%, 9.23%, 9.31%, 8.74% and 9.10% (Table 5, Figure 10). Interestingly, despite significant variation in the number of differentially expressed genes, the ratio of genes associated with immune function was similar in all sites.

The percentage of Immune DEGs vary by each location

	DEGs	Immune DEGs	Percentage of Immune DEGs
UO	3794	320	8.43%
Right	2622	291	11.10%
LO	3447	318	9.23%
UI	3233	301	9.31%
Left	2379	208	8.74%
LI	1847	168	9.10%

Table 5. DEGs by each location

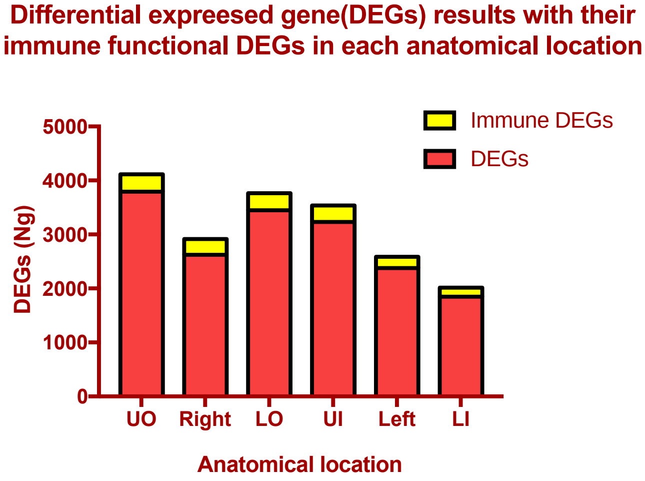
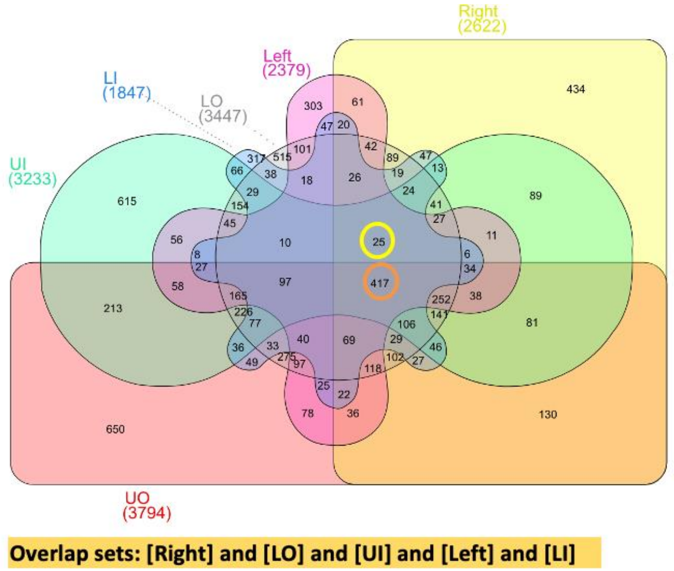


Figure 10. The number of DEGs and immune associated DEGs

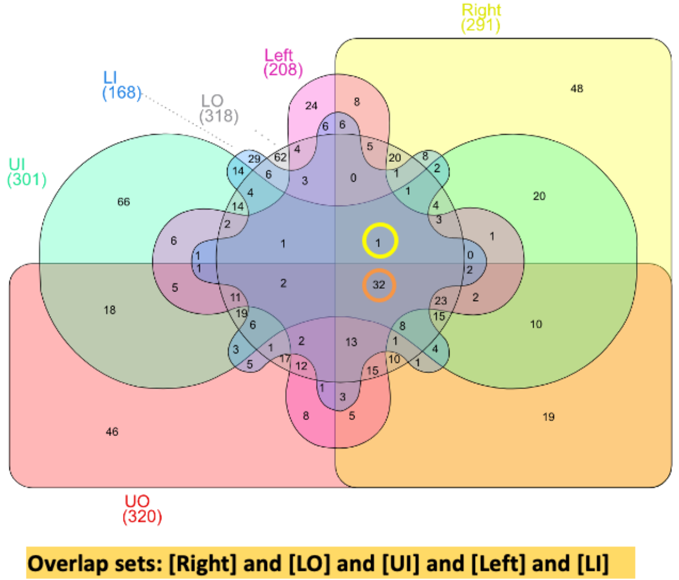
We then did the Venn diagram overlap analysis for DEGs and immune associated DEGs (Figure 11). We observed that there were 25 DEGs from overlapping set - [Right] and [LO] and [UI] and [Left] and [LI]. And gene GBP1 were the only immune associated DEGs. We assume that GBP1 might be a homogeneous modulator in these sites (Figure). It has been reported that in breast, colorectal, and skin cancers, transcriptional and immunohistochemical profiling of patient samples has revealed that high GBP1 signatures are favorable prognostic indicators (15, 51–53, 55, 56) associated with decreased disease progression and greater overall survival (10.3389/fimmu.2019.03139)



Overlapping Genes:

LOC100190939,CCDC77,GBP1,USP6NL,SUV39H2,FOLH1,CLDN16,C1orf135,TAS1R1,LRFN2,IQCG,MCM7,LDLRAD1,RAET1K,FAT1,SEMA3F,SLC26A3,MREG,ACOT4,C14orf182,DIAPH3,LALBA,CHST2,CROT,RCAN1

Figure 11a. The number of overlapped DEGs



Overlapping Gene:

GBP1

Figure 11b. The number of overlapped immune associated DEGs

3.5.2 Pathway enrichment and survival analysis

To further determine whether there were DEGs/immune DEGs associated survival in different site, we then compared the Reactome pathway enrichment analysis at each location based on non-TNBC and TNBC patients. As table 6 shows, mostly DEGs/immune DEGs associated with survival were enriched in UO (5/59) and right site (26/272).

From immune pathway analysis, genes involved in the antigen activates B cell receptor (BCR) pathway (p<0.05) were associated with overall survival (OS) in right and left sided Luminal A/B and HER2 tumors and right sided TNBC tumors. Genes from the antigen processing (ubiquitination and proteasome degradation) pathway (p<0.05) was associated with OS in left and right sided lower outer quadrant

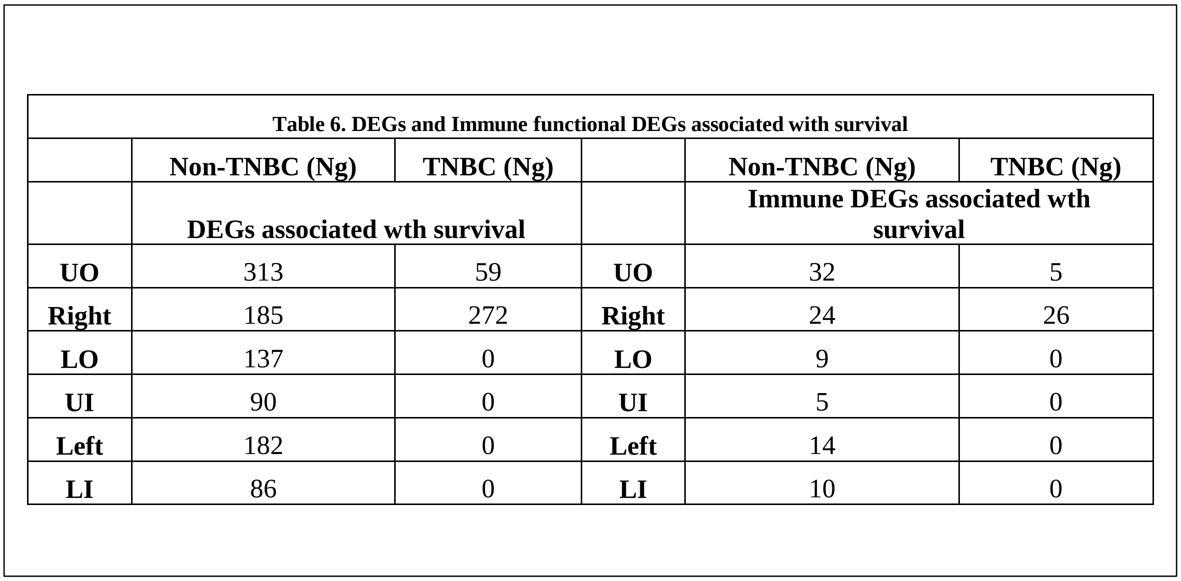


Table 6. DEGs and Immune functional DEGs associated with survival					
	Non-TNBC (Ng)	TNBC (Ng)		Non-TNBC (Ng)	TNBC (Ng)
	DEGs associated wth survival			Immune DEGs associated wth survival	
UO	313	59	UO	32	5
Right	185	272	Right	24	26
LO	137	0	LO	9	0
UI	90	0	UI	5	0
Left	182	0	Left	14	0
LI	86	0	LI	10	0

in luminal A/B and HER2 tumors and all right TNBC tumors. Finally, genes from pathway involved in immune-regulatory interactions between a lymphoid and a non-lymphoid cells were associated with OS in lower outer quadrant, upper outer quadrant tumors in luminal A/B and HER2 cases and right sided tumors in TNBC ($p<0.05$).

3.6 DISCUSSION

Cytotoxic lymphocyte infiltration is critical for response to immune based therapy [43] and has been shown to predict survival and treatment response in colorectal cancer. A better understanding of potential targets is critical for the improvement of immune based therapy in colorectal cancer as currently utilized therapy is not effective in the majority of patients. Therefore, in this study we have combined publicly available data resources with computational methods to focus on genes that may have an impact both on tumor associated cytotoxic lymphocytes and survival. Comparing patients with high and low cytotoxic lymphocyte abundance scores, we found many differentially expressed genes at all tumor locations and stages. Unsurprisingly, the group with the highest number of immune related differentially expressed genes was the right-sided colon cancer group. This may be a reflection of the higher number of MSI-H patients in this group, which is expected to have a higher mutation rate, and therefore, potentially more genes with altered expression.

To further define potential therapeutic targets, we then performed pathway enrichment analysis. In this analysis, we found the pathway, “immunoregulatory

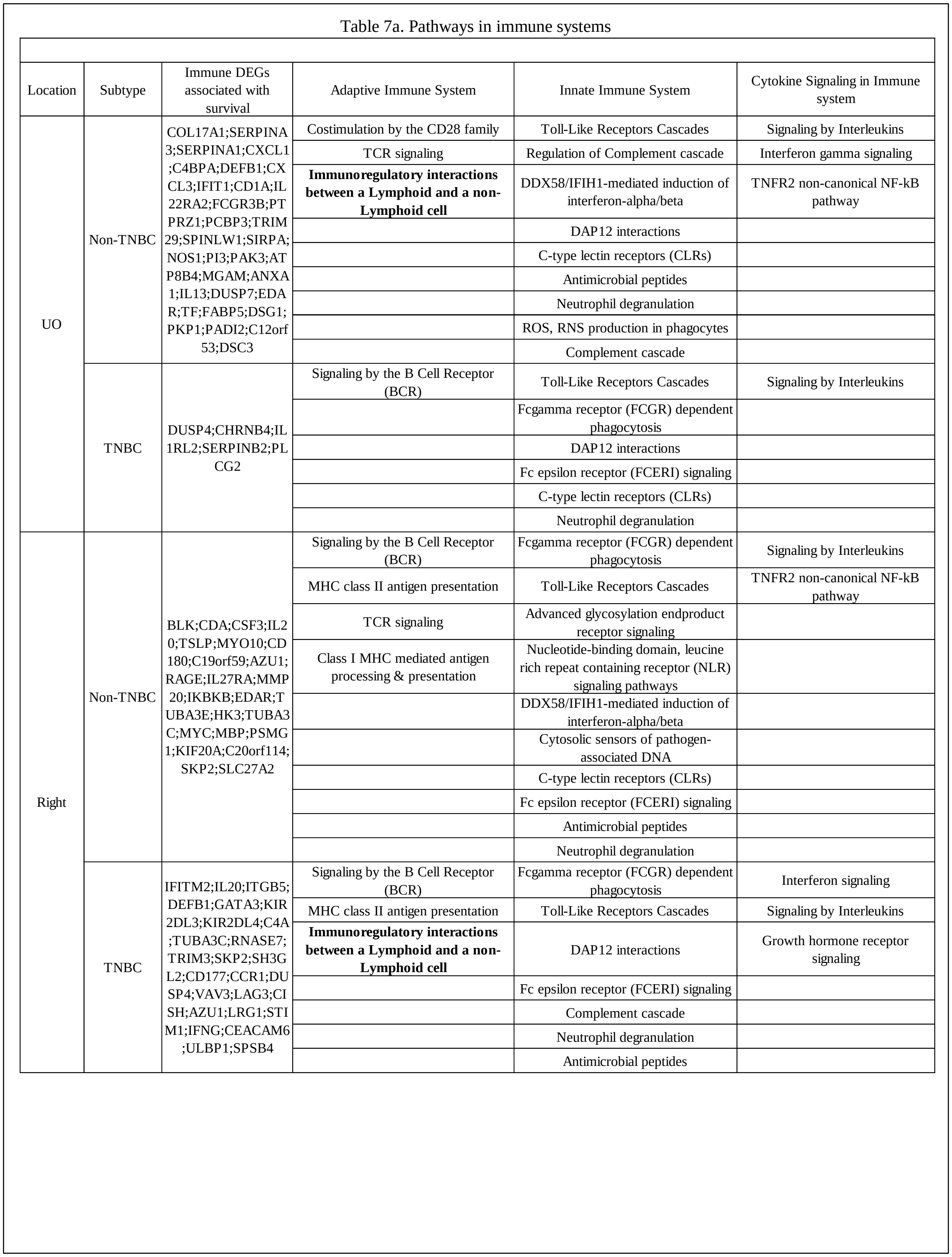


Table 7a. Pathways in immune systems

Location	Subtype	Immune DEGs associated with survival	Adaptive Immune System	Innate Immune System	Cytokine Signaling in Immune system
UO	Non-TNBC	COL17A1;SERPINA3;SERPINA1;CXCL1;C4BPA;DEFB1;CXCL3;IFIT1;CD1A;IL22RA2;FCGR3B;PTPRZ1;PCBP3;TRIM29;SPINLW1;SIRPA;NOS1;PI3;PAK3;ATP8B4;MGAM;ANXA1;IL13;DUSP7;EDAR;TF;FABP5;DSG1;PKP1;PADI2;C12orf53;DSC3	Costimulation by the CD28 family	Toll-Like Receptors Cascades	Signaling by Interleukins
			TCR signaling	Regulation of Complement cascade	Interferon gamma signaling
			Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	DDX58/IFIH1-mediated induction of interferon-alpha/beta	TNFR2 non-canonical NF-kB pathway
				DAP12 interactions	
				C-type lectin receptors (CLRs)	
				Antimicrobial peptides	
				Neutrophil degranulation	
				ROS, RNS production in phagocytes	
				Complement cascade	
	TNBC	DUSP4;CHRNB4;IL1RL2;SERPINB2;PLCG2	Signaling by the B Cell Receptor (BCR)	Toll-Like Receptors Cascades	Signaling by Interleukins
				Fcgamma receptor (FCGR) dependent phagocytosis	
				DAP12 interactions	
				Fc epsilon receptor (FCERI) signaling	
				C-type lectin receptors (CLRs)	
				Neutrophil degranulation	
Right	Non-TNBC	BLK;CDA;CSF3;IL20;TSLP;MYO10;CD180;C19orf59;AZU1;RAGE;IL27RA;MMP20;IKBKB;EDAR;TUBA3E;HK3;TUBA3C;MYC;MBP;PSMG1;KIF20A;C20orf114;SKP2;SLC27A2	Signaling by the B Cell Receptor (BCR)	Fcgamma receptor (FCGR) dependent phagocytosis	Signaling by Interleukins
			MHC class II antigen presentation	Toll-Like Receptors Cascades	TNFR2 non-canonical NF-kB pathway
			TCR signaling	Advanced glycosylation endproduct receptor signaling	
			Class I MHC mediated antigen processing & presentation	Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways	
				DDX58/IFIH1-mediated induction of interferon-alpha/beta	
				Cytosolic sensors of pathogen-associated DNA	
				C-type lectin receptors (CLRs)	
				Fc epsilon receptor (FCERI) signaling	
				Antimicrobial peptides	
				Neutrophil degranulation	
	TNBC	IFITM2;IL20;ITGB5;DEFB1;GATA3;KIR2DL3;KIR2DL4;C4A;TUBA3C;RNASE7;TRIM3;SKP2;SH3GL2;CD177;CCR1;DUSP4;VAV3;LAG3;CISH;AZU1;LRG1;STIM1;IFNG;CEACAM6;ULBP1;SPSB4	Signaling by the B Cell Receptor (BCR)	Fcgamma receptor (FCGR) dependent phagocytosis	Interferon signaling
			MHC class II antigen presentation	Toll-Like Receptors Cascades	Signaling by Interleukins
			Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	DAP12 interactions	Growth hormone receptor signaling
				Fc epsilon receptor (FCERI) signaling	
				Complement cascade	
				Neutrophil degranulation	
				Antimicrobial peptides	

interactions between a lymphoid and non-lymphoid cell", was among the most highly enriched and altered in all sites and stages. And it was noy only enriched in the CTL-high of COADREAD patients, but TNBC BRCA patients in multiple sites. A few pathways were occasionally more highly enriched, however were not affected at all sites or stages, therefore, we chose to focus on this pathway. The "immunoregulatory interactions between a lymphoid and non-lymphoid cell" pathway involves a number of cell surface signaling pathways that are involved in the regulation of anti-tumor immunity [22]. After performing survival analyses using the differentially expressed genes from this pathway, we found the majority of genes affecting survival were in the right-sided patient group consistent with the differential gene expression analysis. Patients with right-sided colon cancer have significantly worse survival than other tumor locations at all stages, and right-sided colon cancer patients with metastatic disease demonstrate poorer survival with current chemotherapy regimens. This group, therefore, likely represents the group with the most important need for new therapeutic options [18, 46]. There was, however, a clear dichotomy between patients with high and low cytotoxic lymphocyte abundance scores. Nearly all genes affecting survival found in the CTL-low patients had a positive impact, whereas nearly all genes affecting survival in the CTL-high patients had a negative impact. This is not entirely unsuspected given we know that tumors attempt to evade anti-tumor immunity

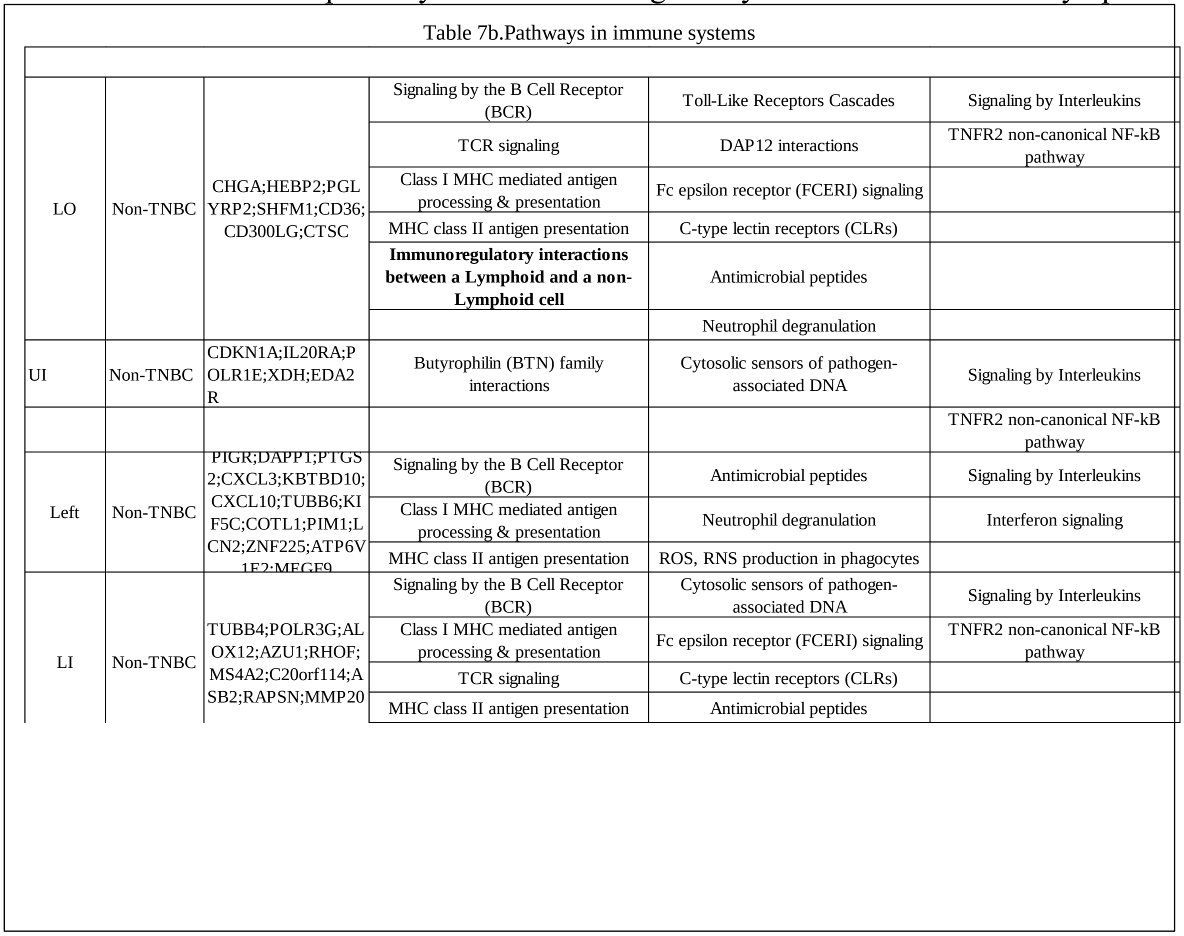


Table 7b.Pathways in immune systems

LO	Non-TNBC	CHGA;HEBP2;PGLYRP2;SHFM1;CD36;CD300LG;CTSC	Signaling by the B Cell Receptor (BCR)	Toll-Like Receptors Cascades	Signaling by Interleukins
			TCR signaling	DAP12 interactions	TNFR2 non-canonical NF-kB pathway
			Class I MHC mediated antigen processing & presentation	Fc epsilon receptor (FCERI) signaling	
			MHC class II antigen presentation	C-type lectin receptors (CLRs)	
			Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	Antimicrobial peptides	
				Neutrophil degranulation	
UI	Non-TNBC	CDKN1A;IL20RA;POLR1E;XDH;EDA2R	Butyrophilin (BTN) family interactions	Cytosolic sensors of pathogen-associated DNA	Signaling by Interleukins
					TNFR2 non-canonical NF-kB pathway
Left	Non-TNBC	PIGR;DAPP1;PTGS2;CXCL3;KBTBD10;CXCL10;TUBB6;KIF5C;COTL1;PIM1;LCN2;ZNF225;ATP6V1E2;MEGF9	Signaling by the B Cell Receptor (BCR)	Antimicrobial peptides	Signaling by Interleukins
			Class I MHC mediated antigen processing & presentation	Neutrophil degranulation	Interferon signaling
			MHC class II antigen presentation	ROS, RNS production in phagocytes	
LI	Non-TNBC	TUBB4;POLR3G;ALOX12;AZU1;RHOF;MS4A2;C20orf114;ASB2;RAPSN;MMP20	Signaling by the B Cell Receptor (BCR)	Cytosolic sensors of pathogen-associated DNA	Signaling by Interleukins
			Class I MHC mediated antigen processing & presentation	Fc epsilon receptor (FCERI) signaling	TNFR2 non-canonical NF-kB pathway
			TCR signaling	C-type lectin receptors (CLRs)	
			MHC class II antigen presentation	Antimicrobial peptides	

through various mechanisms [19]. Other groups have also demonstrated this in the context of MSI-H colorectal cancer, noting a significant upregulation of multiple negative regulators of immunity in these patients. This data further underscores the need to develop and test new immune based therapy in patients with colorectal cancer tailored to patient specific factors.

In our survival analysis, we identified several potential targets for combination therapy. CD40L is a cell surface marker expressed on activated T cells that promotes maturation of antigen presenting cells, upregulating costimulatory molecules and activating antigen presentation machinery, and may represent the most attractive target identified in this work. In our analysis, this gene demonstrated a positive impact on survival in metastatic patients with low cytotoxic lymphocyte abundance scores. In preclinical models, CD40 agonists have demonstrated a significant ability to activate anti-cancer immunity, overcome immune checkpoint inhibition resistance, and work in concert with other immune based treatments [47]. Currently, a number of clinical trials are open studying these drugs in combination with other immune based treatments; however, none are specifically directed at colorectal cancer patients. The fact that there are drugs available targeting this interaction may lend itself to rapid translation in these patients. Additionally, we found potentially attractive targets in CD96 and CD18 (ITGB2), each of which has demonstrated some significant impact on anti-tumor immunity in preclinical studies with the potential for translation in the future [48, 49].

One limitation of this study is related to patient numbers and clinical data available, as with many database studies. Due to patient numbers, we included patients in Stage I and II in the analysis for both “early” and “local” disease. This was done to increase patient numbers assigned to each group and improve our analysis. Based on our results, we felt this helped to support findings in the “early” stage patients as the Stage III patients contributed 40–60% of “local” patients depending on disease location. Another important potential confounding factor, however, is significant heterogeneity in therapy, most notably in patients with metastatic disease (Stage IV). These were real world patients not treated on specific study protocols, so this heterogeneity in treatment may represent an impactful difference. Additionally, the patients included in this study did not receive immune based therapy, so the impact of cytotoxic lymphocyte infiltration on response to immune based treatments cannot be directly assessed. However, a number of studies have previously shown that cytotoxic lymphocyte infiltration in colorectal cancer predicts survival and response to therapy, therefore augmenting anti-tumor immunity is likely to be impactful when considering combination with conventional treatments such as chemotherapy, or immune based therapy alone. Recent studies in cancer therapy have also begun to understand that the immune response is critical to the efficacy of chemotherapy and radiotherapy, further highlighting the need to understand altered immune pathways in cancer [50]. Despite these limitations, resources such as the TCGA, when combined with informatics-based analysis, yield highly impactful results that can be used to develop future human studies and inform translational pre-clinical studies. The goals of precision-based oncology will be best met by combining studies of all types to select both the best therapy for each patient, as well as the best patient for each therapy.

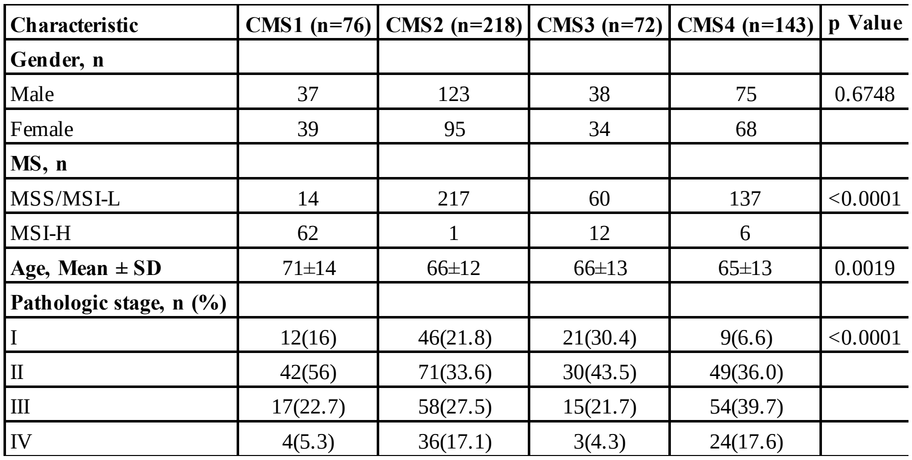
Conclusion In this study, we integrate comprehensive RNA-seq data, clinical and pathologic data, and cytotoxic lymphocyte scores to determine pathways associated with immune response and survival in patients with colorectal cancer. We identified one pathway, "immunoregulatory interactions between a lymphoid and non-lymphoid cell", that was highly enriched and included in all tumor locations and stages. We then found specific genes associated with survival, primarily in patients with the worst survival, those with metastatic right-sided colon cancer, that may be targeted to improve therapy. Future studies will focus on further exploration of immune pathway interactions using multi-omics analysis in humans, and mechanistic studies of T lymphocyte recruitment and activation in murine models of colorectal cancer.

3.6.1 TBX21 METHYLATION AS A POTENTIAL REGULATOR OF IMMUNE SUPPRESSION IN CMS1 SUBTYPE COLORECTAL CANCER

COADREAD is a heterogeneous disease characterized by distinct genome-wide changes, with the third-highest incidence rate and the second-highest rate of cancer-related deaths worldwide [10]. To better design treatments for COADREAD, it is crucial to understand tumor heterogeneity and how this contributes to therapeutic resistance and disease progression. Genomic instability and epigenetic abnormalities with resultant dysregulation of gene expression are hallmarks of COADREAD. The high frequency of DNA somatic copy number alterations (SCNA) and APC tumor suppressor gene loss of function closely links with CIN-caused deletions, gains, translocations, and other chromosomal rearrangements, one of the primary pathways of COADREAD development [51]. Additionally, ~15% of COADREAD demonstrate alterations in DNA mismatch repair (MMR) proteins which lead to hypermethylation and cancer development [52-54]. Importantly, a single factor does not lead to tumorigenesis and underlines the importance of understanding the molecular features of each individual tumor so that may improve therapy using a precision based approach.

Immune-based therapies such as ICI have recently made significant advances in difficult-to-treat malignancies like non-small cell lung cancer, melanoma, and renal cell cancer [55, 56]. However, these treatments are limited to patients with microsatellite instability- high (MSI-H) COADREAD as they have not yet demonstrated efficacy in other patient groups [57]. Additionally, even in patients with an indication for use of ICI therapy, response is limited [58]. This is despite data demonstrating that the number of CTLs within the tumor microenvironment (TME) is a critical prognostic marker for COADREAD [59, 60]. This again highlights the importance for improved methods of assessing the TME and understanding therapeutic resistance.

Dysregulated methylation impacts signaling pathways associated with apoptosis avoidance, metastasis, and therapeutic resistance, including immunotherapy and represents an important



Characteristic	**CMS1 (n=76)**	**CMS2 (n=218)**	**CMS3 (n=72)**	**CMS4 (n=143)**	**p Value**
Gender, n					
Male	37	123	38	75	0.6748
Female	39	95	34	68	
MS, n					
MSS/MSI-L	14	217	60	137	<0.0001
MSI-H	62	1	12	6	
Age, Mean ± SD	71±14	66±12	66±13	65±13	0.0019
Pathologic stage, n (%)					
I	12(16)	46(21.8)	21(30.4)	9(6.6)	<0.0001
II	42(56)	71(33.6)	30(43.5)	49(36.0)	
III	17(22.7)	58(27.5)	15(21.7)	54(39.7)	
IV	4(5.3)	36(17.1)	3(4.3)	24(17.6)	

Table 8. Patient Characteristics. CMS, consensus molecular subtypes; MSS, Microsatellite stable; MSI-L, Microsatellite instability-low; MSI-H, Microsatellite instability-high.

process in COADREAD [61]. Additionally, combined treatments with drugs targeting epigenetic modification exploit the dynamic nature of epigenetic changes to potentially modulate responses to immunotherapy [62]. However, current drugs targeting epigenetic modification are globally hypomethylating agents, such as Azacitidine, Decitabine, which are non-selective and may cause have unexpected effects [63]. We found gene TBX21, MX1, and SP140 may play a crucial role impacting function through the TP53/P53 pathway in modification of TBX21 methylation level and then further upregulate TBX21 expression [64].Therefore, identifying more specific candidate epigenetic biomarkers and targets may provide a rationale for patient stratification and targeted therapy, maximizing the chances of treatment success while minimizing unwanted effects. Recently, large-scale public data repositories such as The Cancer Genome Atlas (TCGA) and cBioPortal, as well as publicly available single-cell sequencing data, have allowed us to perform further in-depth study of cancer patients on a molecular and clinical basis using multi-omics [65-67]. To better classify COADREAD patients, the consensus molecular subtypes (CMS) were developed by an expert panel using eighteen different COADREAD mRNAseq and microarray datasets, settling on four different subtypes based on molecular features [23]. Patients classified as CMS1 are characterized by microsatellite instability (MSI), hypermutation, and are considered the "immune" subtype. CMS2 is referred to as "canonical", characterized by marked Wnt/β-catenin/ TCF7L2 pathway activation, and APC mutation. Characterized by metabolic deregulation, KRAS mutations, and mixed MSI patients, CMS3 is the least common. Tumors classified as CMS4 are "mesenchymal", demonstrating prominent TGF-β activation, stromal infiltration, angiogenesis, and epithelial-mesenchymal transition (EMT) . Recognizing the importance of molecular classification, there are ongoing pre-clinical trials utilizing this system for patient stratification and selection of immune based treatments

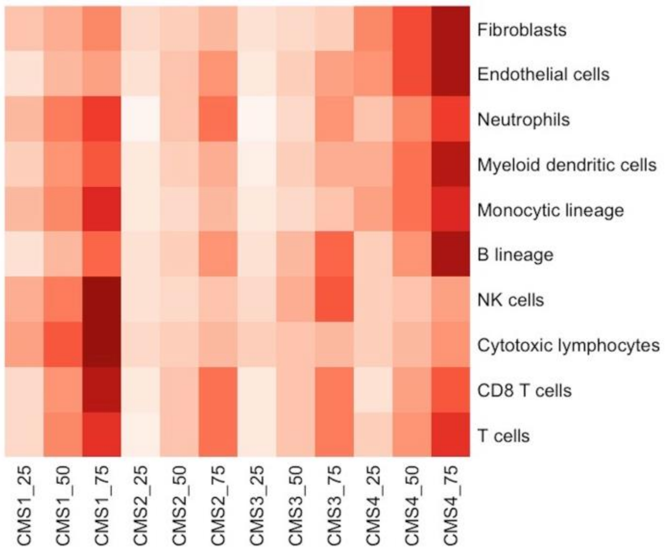


Figure 12. Ten tumor microenvironment cell populations of MCP-counter score. CMS1 patients have significantly higher cytotoxic lymphocytes scores than the others. Each three columns represent one CMS subtype, each row represents the cell population.

(NCT03436563, NCT04738214).To address limitations regarding the current knowledge in resistance to immune-based therapy and the role of epigenetic modification of gene expression, in this study we sought to combine multi-omics data with cutting-edge data analytics and comprehensive molecular classification using CMS (Figure 2). Utilizing publicly available data from four data repositories, including 15 datasets, 2,391 COADREAD patients, and seven -omics datasets, we found a difference in survival based on differential expression and methylation of the critical T cell regulatory factor T-bet (TBX21). Additional exploration of this data pointed to alterations in CD8 T cell exhaustion, suggesting that T-bet methylation and expression is a critical factor for T cell dysfunction impacting survival in patients with COADREAD.

3.6.2 CMS1 SUBTYPE PATIENTS DEMONSTRATE FEATURES CONSISTENT WITH IMMUNE ACTIVATION

To address the question of the clinical impact of DNA methylation on gene expression and resistance to immune based therapy, we chose to use the TCGA, the largest publicly available multi-omics dataset with substantial clinical annotation. First, we downloaded and integrated the 461 colon cancer (COAD) and 171 rectal cancer (READ) patients from this dataset. We then classified patients in CMS subtypes to use the most comprehensive current molecular classification. Using this system, there were 316 patients with DNA methylation

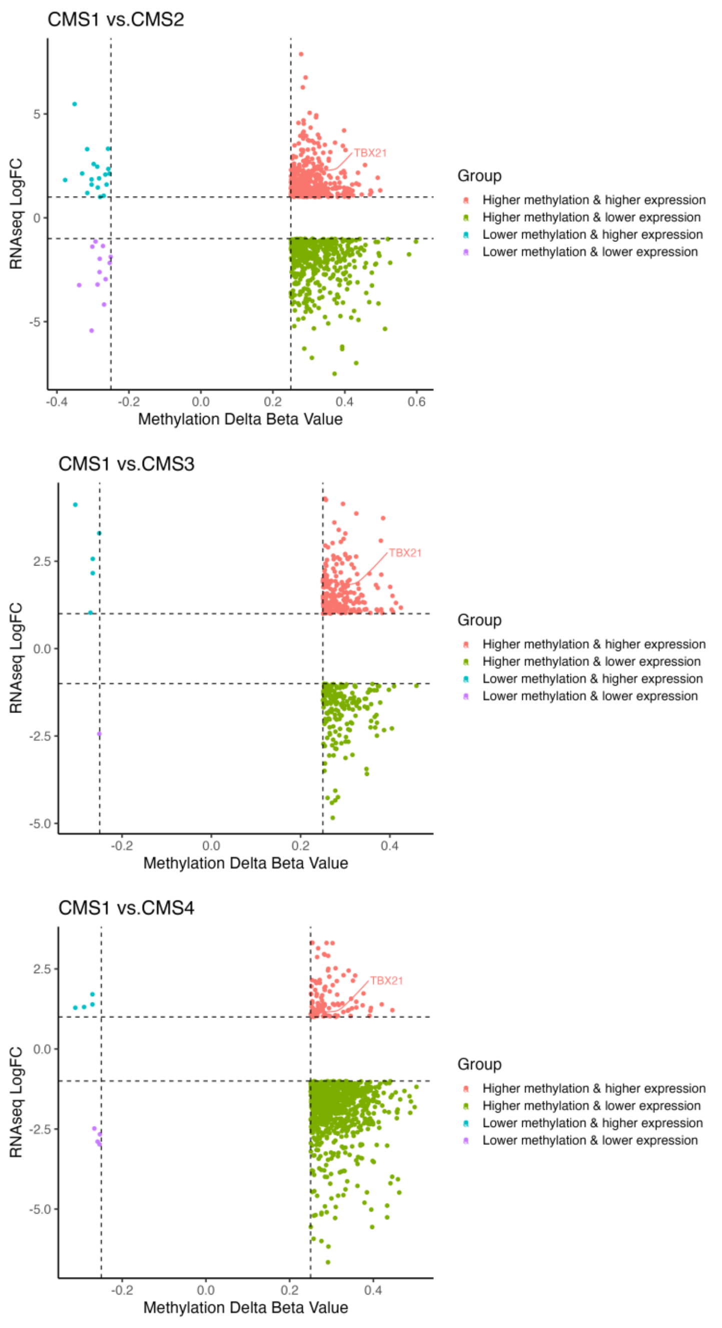


Figure 13. The distribution of crosstalk genes.

With identified crosstalk genes for each comparison, most crosstalk genes enriched in Quadrant 2 and 3. TBX21 is the only gene differentially expressed, differentially methylated, across three comparisons X axis is the delta beta value of methylation, Y axis is the RNAseq LogFC.(a) Distribution of crosstalk genes from CMS2 vs. CMS1 DEG and DMG analysis. (b) Distribution of crosstalk genes from CMS3 vs. CMS1 DEG and DMG analysis. (c) Distribution of crosstalk genes from CMS4 vs. CMS1 DEG and DMG analysis.

data (450k) and 509 patients with mRNAseq data labelled with CMS subtypes.The demographic, clinical, and pathologic characteristics of each patient group by CMS subtype are summarized in Table 8 .As expected, the microsatellite status composition of patients between CMS subtypes was significantly different ($p < 0.0001$) with the CMS1 subtype containing more MSI-H patients and the CMS3 subtype containing more MSI-L/MSS patients. The median age of CMS1 patients' diagnosis was 71, ranging from 58 to 85, significantly higher than patients in the other CMS subtypes ($p = 0.0019$). Additionally, in the CMS2 and CMS4 subtypes, more patients were diagnosed at stage III and IV than the other subtypes ($p < 0.0001$).

Next, we sought to combine CMS classification with stratification by cytotoxic lymphocyte (CTL) infiltration as a marker for anti-tumor immunity [68]. MCP-counter scores were derived for each patient, and patients were then stratified by MCP-counter score quartile and CMS classification (Figure 12) [16, 21]. Patients in the CMS1 subtype had the highest median CTL abundance score, and the majority of patients in this subgroup were in the highest quartile of CTL infiltration. Additionally, we also found that when considering tumor mutational burden (TMB) and neoantigen predicted peptides, CMS1 subtype patients were again much higher than the other subtypes. This data combined supports the idea that CMS1 subtype patients represent an "inflamed" phenotype and are the most immune active.

3.6.3 TBX21 IS THE ONLY GENE DIFFERENTIALLY EXPRESSED, DIFFERENTIALLY METHYLATED, AND HIGHLY CORRELATED WITH CTL INFILTRATION ACROSS ALL CMS SUBTYPES

CTL infiltration is critical for anti-tumor immunity and response to immune-based therapy, therefore, we next sought to understand how gene expression and DNA methylation were associated with CTL infiltration and molecular alteration. Using CMS1 as the reference subtype given its high level of CTL infiltration and predicted response to ICI, we performed differential gene expression and differential methylated region analysis (Figure 2). Expression of 20,531 genes and 293,276 methylation loci were determined for each tumor sample from the TCGA. There were 2,977 differentially expressed genes and 11,541 differentially methylated regions (3,500 differentially methylated genes) from the comparison of CMS1 and CMS2. In the comparison of CMS1 and CMS3, there were 2,385 differentially expressed genes and 4,231 differentially methylated regions (1,583 differentially methylated genes). And in the comparison of CMS1 and CMS4, there were 3,246 differentially expressed genes and 9,393 differentially methylated regions (2,359 differentially methylated genes). To identify "crosstalk genes", genes both differentially expressed and differentially methylated genes, we then performed a correlative analysis plotting differentially expressed genes (|logFC| >1 and p-Value <0.05 adjusted for FDR) against differentially methylated genes (|Δβ| > 0.25 and P-value <0.05, adjusted for FDR, Figure 13). Identified crosstalk genes were predominantly enriched in quadrants “higher methylation & higher expression” and “higher methylation % lower expression (Figure 14), demonstrating increased methylation (Δβ >0.25) associated with either increased or decreased differentially expressed gene expression. To validate these findings, we analyzed publicly available gene expression and DNA methylation datasets [69].Using this data, we confirmed that TBX21 was consistently differentially methylated when comparing CMS1 vs. other CMS subtypes. Specifically, we found that one specific loci in the coding region of TBX21 (SiteID: cg26281453, Loc: 45810610) was differentially methylated in all datasets.

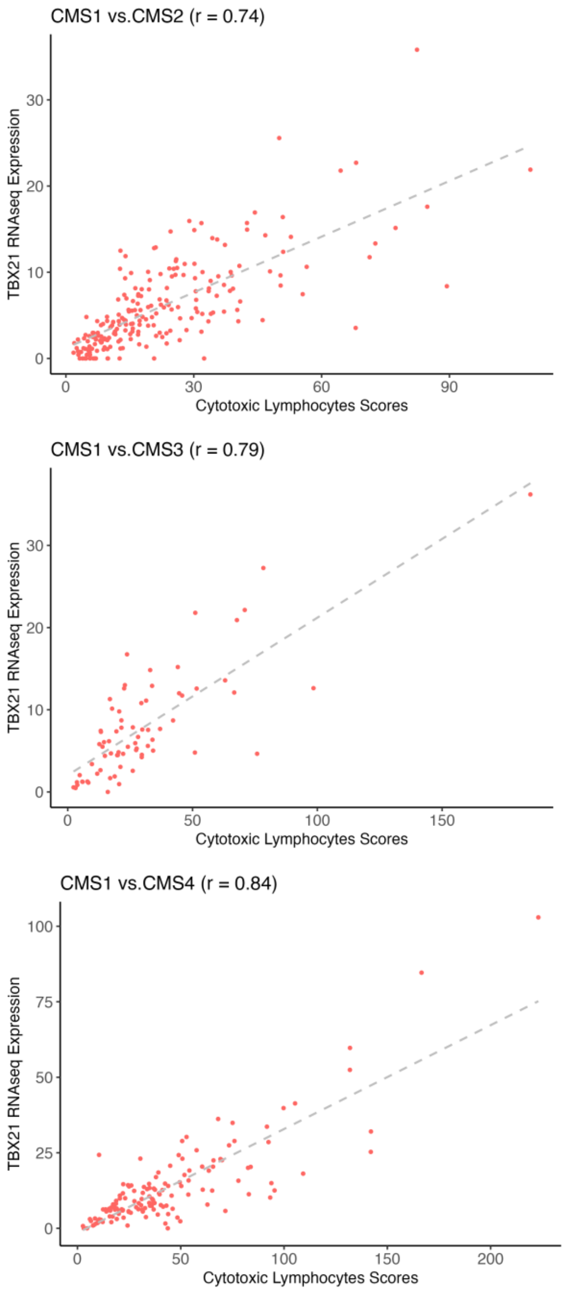


Figure 14. TBX21 is the highly correlated with CTL.X axis is the CTL score,Y axis is the normalized TBX21 mRNAseq expression.(a)Normalized TBX21 expression highly correlated with CTL in CMS2.(b)Normalized TBX21 expression highly correlated with CTL in CMS3.(c)Normalized TBX21 expression highly correlated with CTL in CMS4.

To explore the relationship of crosstalk genes with CTL infiltration, we correlated the crosstalk gene expression with MCP-counter CTL abundance scores. Interestingly, TBX21 was the only gene differentially expressed, differentially methylated, and highly correlated with CTL from each CMS1 comparison (r > |0.7|, Figure 13). Interestingly, we observed that TBX21 was more highly expressed in CMS1 patients, but was also more highly methylated (Figure 13). However, the typical expectation is that methylation is an epigenetic modification that leads to decreased gene expression [54]. Moreover, previous studies have suggested that TBX21 is an important transcriptional regulator of tumor-reactive CD8+ T-cells which are critical for response and survival [70]. Given these results, we hypothesized that the importance of alterations in the expression of TBX21 via methylation may be most important within CMS1 patients.

3.6.4 CMS1 PATIENTS WITH HIGH TBX21 EXPRESSION AND LOW TBX21 METHYLATION HAVE THE BEST SURVIVAL

To better understand the relationship between TBX21 expression and methylation and CMS1 patient outcomes, we next performed a survival analysis. When using TBX21 expression or methylation values as independent variables, there was no difference in survival (data not shown). Therefore, we further classified CMS1 patients using the median value of TBX21 expression and mean methylation β value to separate these patients into four groups (Figure 15). Patients were grouped as: high TBX21 expression with low methylation (high-low); high TBX21 expression with high methylation (high-high); low TBX21 expression with high methylation (low-high); and low TBX21 expression with low methylation (low-low). We then repeated the survival analysis with CMS1 patients stratified by these four subgroups (Figure 15). Patients classified as high-low (high expression, low methylation) demonstrated the best survival, followed by low-high patients. Interestingly, the group of patients classified as high-high, appeared to represent a potential intermediate subgroup (Figure 15) with worse survival (Figure 15) than both high-low and low-high subgroups. This data suggests that the interaction between TBX21 expression and methylation plays an important role in patient prognosis.

3.6.5 THERE WERE NO SIGNIFICANT CLINICAL DIFFERENCES BETWEEN HIGH-LOW AND LOW-HIGH CMS1 PATIENT SUBGROUPS

This result inspired us to investigate the potential mechanisms of the observed difference in survival between CMS1 patient subgroups. Therefore, we retrieved the patient clinical attributes from cBioPortal including 41 separate attributes. We found no clinical attributes

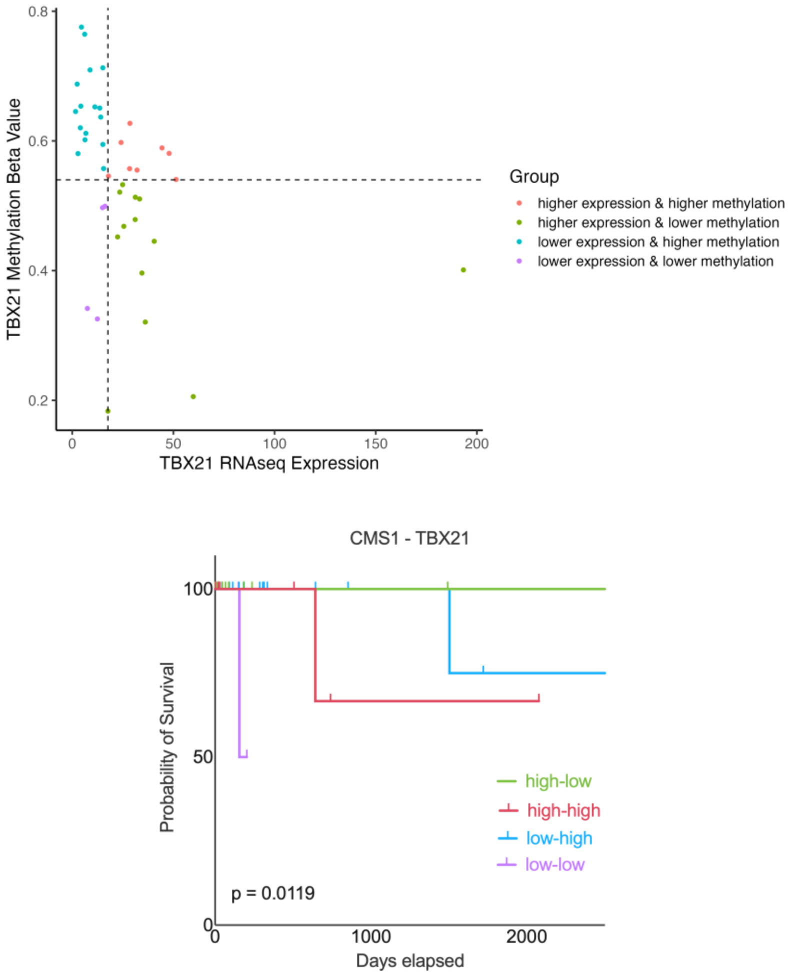


Figure 15 TBX21 in CMS1 patients. (a) Subtype of CMS1 patients based the expression and methylation of TBX21. There were 24 patients with higher TBX21 methylation and 17 patients with lower methylation. X axis is TBX21 mRNAseq expression, Y axis is TBX21 methylation beta value. high-low: high TBX21 mRNA expression, low TBX21 methylation (number of patients(n) =13); highhigh: high TBX21 mRNA expression, high TBX21 methylation (n=8); low-high: low TBX21 mRNA expression, high TBX21 methylation (n=16);low-low: low TBX21 mRNA expression, low TBX21 methylation (n=4). (b) Patients with higher expression and low methylation of TBX21 has the best survival, and the patients with lower expression and higher methylation has the worse survival than high-low group and it also has the largest number of patients with this status in this group. X axis is patients' days elapsed, Y axis is percent of patients were still survive.

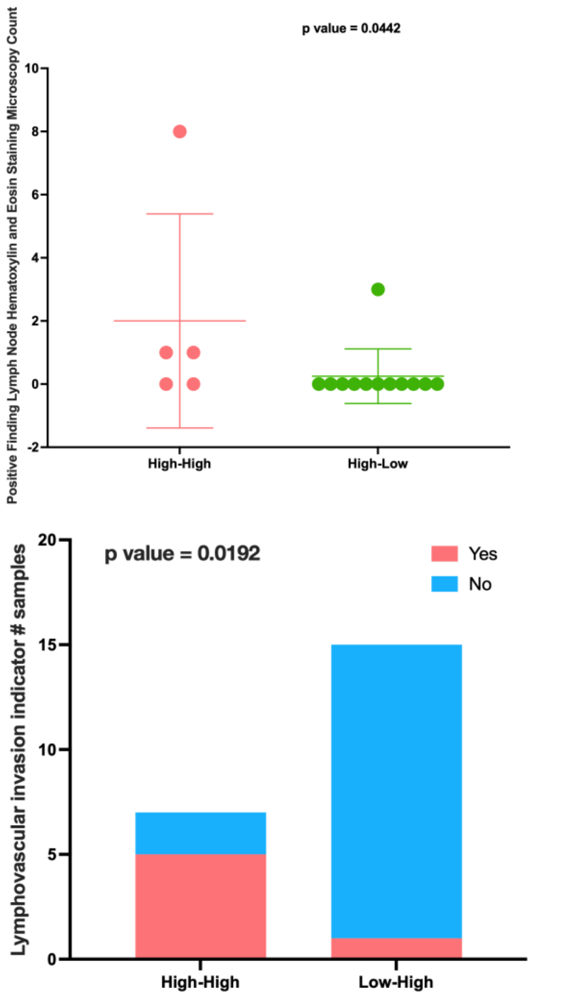


Figure 16 Clinical status in CMS1 patients. (a) high-high patients had more positive lymph nodes than high-low patients. (b) high-high patients have more patients had lymphovascular invasion than low-high patients.

with significant differences when comparing high-low and low- high patients. However, when comparing the high-high group to other groups, we found multiple significant factors. There were eight clinical attributes with significant differences between high-low and high-high patient groups. Most specifically, we noted that high-high patients had a significantly higher number of Positive Lymph Nodes than high-low patients (p-Value: 0.0442, Figure 15), a known marker of poorer survival. Additionally, high-high patients had significantly higher rates of lymphovascular invasion (LVI) than low-high patients (p-Value: 0.0192, Fig 6b), another important clinical risk factor for survival. These results suggest that the survival difference demonstrated by the high-high group may be driven by clinical factors, however, the survival difference between the high-low and low-high patient groups is more likely driven by molecular factors.

3 3.6.6 Patients with high TBX21 expression and low methylation are the most highly immune infiltrated

To further evaluate the impact of TBX21 on survival and anti-tumor immunity, we looked at other indicators of immune resistance. Tumor mutational burden (TMB) has been shown to predict survival and response to immune checkpoint, so we first derived these scores for each patient [25]. However, when comparing the subgroups of CMS1 patients, we observed no significant differences in TMB (Table S4). Next, we derived neopeptides a potentially better marker of immune reactive antigen in the tumor microenvironment. But again, we did not observe any differences between subgroups of CMS1 patients (Table S4). Given the lack of overt differences in these measures, we sought to further investigate other aspects of anti-tumor immunity.

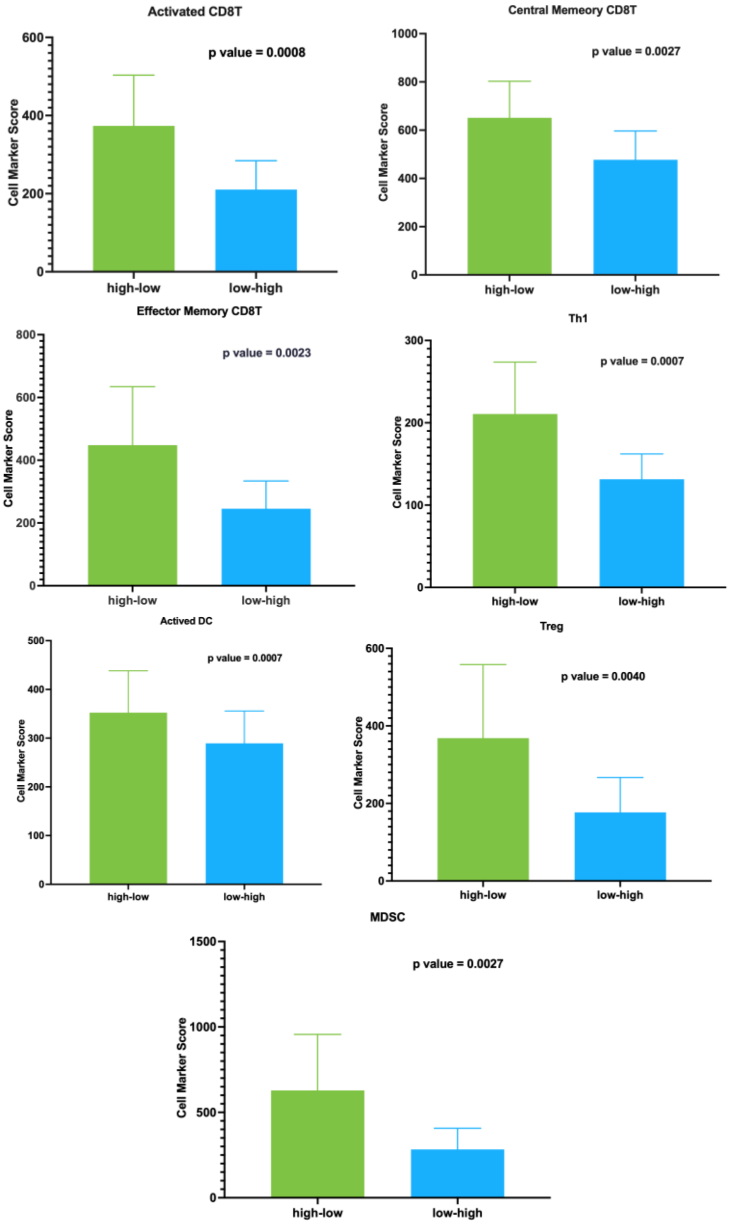
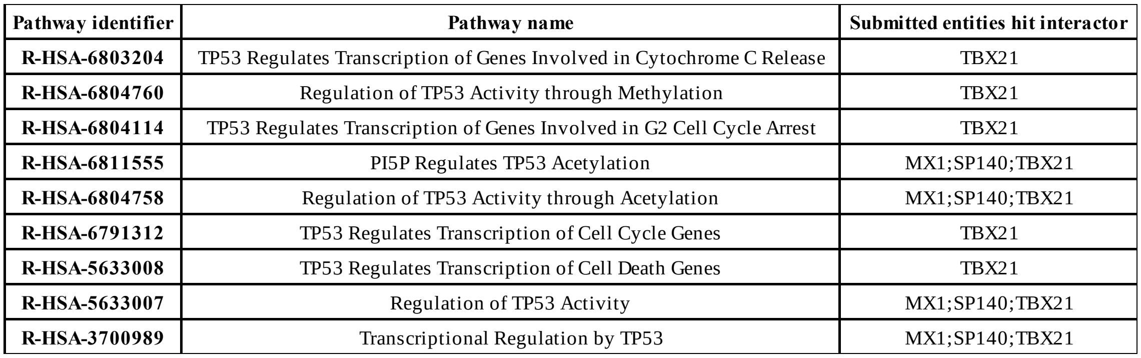


Figure 17. Cell marker score analysis for immune cell populations. High-low patients had the significant higher infiltration of CD8+T cell subtypes including (a)activated CD8+T cell,(b)central memory CD8+T cell,and (c)effector memory CD8+T cell;We also found increased infiltration of (d)T helper cells(Th1)and (e)activated dendritic cells (DC);we also saw increased infiltration of cell subtypes suggest to be immune suppressive,(f)Treg and (g)myeloid derived suppressor cells.

To obtain a more in-depth look at immune alterations in CMS1 patient subgroups, we next looked more specifically at different cell populations in the TME using cell marker score analysis [26]. Signature genes for calculating the cell marker score of each cell population were obtained from the TCIA dataset (Table S5)[26]. Using this analysis, we found that high-low patients had significantly higher infiltration of CD8+ T cell subtypes including activated CD8+ T cell (p value = 0.0008), central memory CD8+ T cell (p value = 0.0027), and effector memory CD8+ T cell (p value = 0.0023) (Fig 7a-c). We also found increased infiltration of T helper cells (Th1, p value = 0.0007) and activated dendritic cells (DC, p value = 0.0007) (Fig 7d, e). However, in addition to the significantly higher infiltration of cell subtypes that are consistent with anti-tumor immune profiles, we also saw increased infiltration of immune suppressive cell subtypes, such as Treg (p value = 0.0040) and myeloid derived suppressor cells (p value = 0.0027) (Fig 7f, g). TH17 and monocyte subsets were not significantly different (data not shown). Given the evidence of increased immune cell infiltration, both pro-and anti-inflammatory, we sought to further understand the molecular differences that may suggest a mechanism for the observed difference in survival associated with TBX21 in these patients.

3.6.7 Epigenetic modification of MX1, SP140, and TBX21 caused their expression to be upregulated in high expression-low methylation patients

To look at the question of molecular differences impacting survival in CMS1 patient subgroups, we completed DEG analysis, focusing on the differences between high-low and low-high patients (Table 9). Notably, we found that there were many more genes differentially expressed when comparing high-low and low-high patients than when comparing the other subgroups (1,482 DEGs, Table S6), further supporting an important molecular difference in these patient groups. Next, DEGs obtained from comparing high-low and low-high subgroups were analyzed using Reactome gene enrichment analysis. We found 741 DEGs that were enriched in 41 pathways. Eighteen CD8+ TEX marker genes were enriched in 13 pathways, all of which were upregulated in high-low patients (Figure



Pathway identifier	Pathway name	Submitted entities hit interactor
R-HSA-6803204	TP53 Regulates Transcription of Genes Involved in Cytochrome C Release	TBX21
R-HSA-6804760	Regulation of TP53 Activity through Methylation	TBX21
R-HSA-6804114	TP53 Regulates Transcription of Genes Involved in G2 Cell Cycle Arrest	TBX21
R-HSA-6811555	PI5P Regulates TP53 Acetylation	MX1;SP140;TBX21
R-HSA-6804758	Regulation of TP53 Activity through Acetylation	MX1;SP140;TBX21
R-HSA-6791312	TP53 Regulates Transcription of Cell Cycle Genes	TBX21
R-HSA-5633008	TP53 Regulates Transcription of Cell Death Genes	TBX21
R-HSA-5633007	Regulation of TP53 Activity	MX1;SP140;TBX21
R-HSA-3700989	Transcriptional Regulation by TP53	MX1;SP140;TBX21

Table 9. Reactome pathway analysis for crosstalk gene with CD8Tex feature from the DEG and DMR analysis of high-low and low-high CMS1 patients.

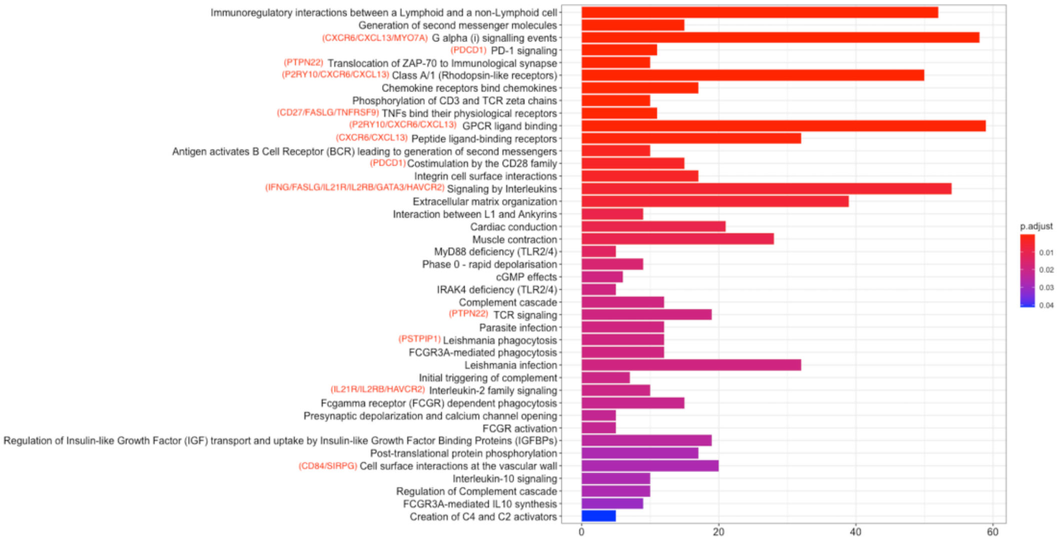


Figure 18. Eighteen CD8+ TEX marker genes were enriched in 13 pathways, all of which were upregulated in high-low patients. Longer bars indicate that had more DEGs enriched in the pathway. Bar color from blue to red indicate that the DEGs enriched pathways had higher p adjust value.

18). After processing DMG analysis between low-high and high-low patients, five crosstalk genes associated with the CD8+ TEX signature were identified .We next applied Reactome analysis specifically for crosstalk genes. These five crosstalk genes participated in 303 pathways. Notably, we observed that genes MX1, SP140, and TBX21 were frequently enriched in the "Regulation of TP53 Activity" and its cascade pathways. Moreover, we identified the function of SP140 from the EpiFactors database as a Zinc finger structure that mainly targets Histone modification read and transcription factor (TF) regions to participate in epigenetic modification [71]. Together this data suggests a critical role for epigenetic modification of TP53 pathway genes impacting TBX21 and patient outcome in these subgroups.

3.6.8 TBX21 IS A KEY MODULATOR IN CD8+ T EXHAUSTED CELLS

To validate our hypothesis regarding the importance of TBX21 in TEX in COADREAD, we explored publicly available scRNAseq data from 23 COADREAD patients with 65,362 matched normal and tumor single cells. Data was first normalized and then utilizing cell subtypes identified by the original study, we retrieved 47,285 tumor cells labeled as TP. Blueprint/ENCODE reference from the SingleR subtype identifier was then used to re-annotate tumor cells [72]. We identified 36 pure stroma and immune cell types, including 2,268 central memory CD8 positive alpha-beta T cell (CD8+ TCM); 4,214 effector memory CD8 positive alpha-beta T cell (CD8+ TEM); and 84 CD8 positive alpha-beta T cell (CD8+ T-cells). To focus on exhausted CD8+ T cells, we retrieved all CD8+ T cell subsets (CD8+ T-cells) and performed an independent cluster analysis. In this analysis, we found eight distinct CD8+ T cell subclusters, each exhibiting a distribution of clusters. To annotate these clusters, we then used CellMarker, a marker-based annotation database, and identified clusters 1 and 7 as CD8+ T exhausted (CD8+ TEX) cells [29]. Compared with the pre-identified annotation from the Blueprint/ENCODE reference, CD8+ TEX predominantly overlapped with CD8+ TEM cells and a small proportion of CD8+ TCM.Additionally, we selectively looked at the expression of transcription factors, checkpoint receptors, and effector molecules, noting that in subclusters of CD8+ TEX, cluster 7 was enriched with cells that have a significantly higher average

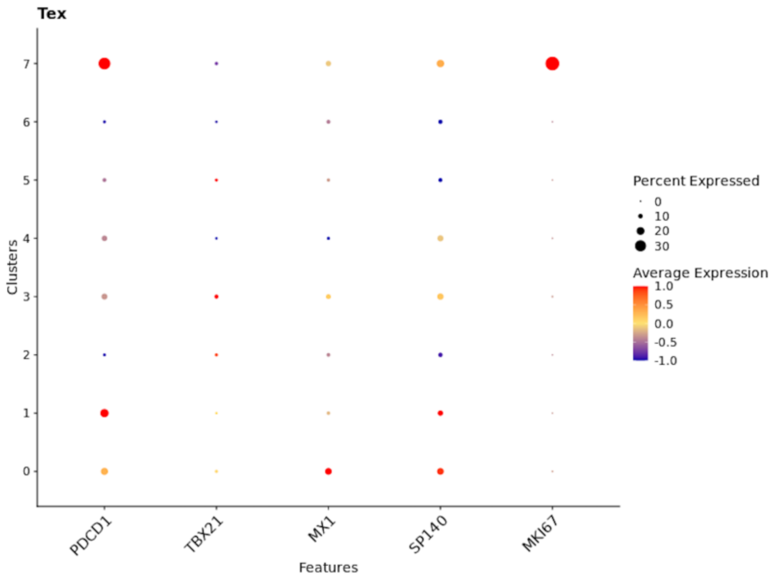


Figure 19. Cluster 1 and cluster 7 are identified CD8+ T cell sub-clusters with CD8+TEX features. X axis is the interested gene with scaled average expression (low to high: -1.0 to 1.0) and bigger dot indicate higher percentage of expression of interested genes in totally analyzed cells; Y axis is the identified CD8+ T cell subclusters.

Figure

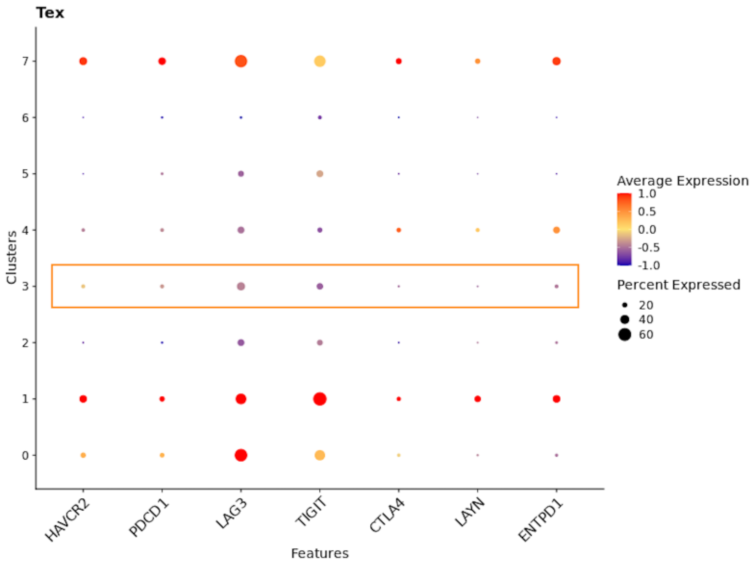


Figure 20 Cluster 3 of CD8+T cell sub-clusters has highest TBX21 expression associate with low levels of checkpoint receptors (PDCD1,LAG3,and TIGIT).X axis is the interested gene with scaled average expression (low to high:-1.0 to 1.0)and bigger dot indicate higher percentage of expression of interested genes in totally analyzed cells;Y axis is the identified CD8+T cell sub- clusters.

expression of MKI67, PDCD1, and lower expression of TBX21 than cluster 1 (Figure 19) [33].Cluster 1 demonstrated expression of cells that are consistent with the identified low-proliferative TEX cluster in the previous research consistent with the idea that TBX21 expression is associated with more highly functional cells. In contrast, cluster 3 contained cells with the highest TBX21 expression and low levels of expression of checkpoint receptors, such as PDCD1, LAG3, and TIGIT, when compared with CD8+ TEX (clusters 1 and 7) (Fig 19, 20). We then performed featureplots to show the distribution of TBX21, PDCD1, and EOMES. Specifically, PDCD1 showed the highest density in CD8+ TEX (clusters 1 and 7) and was associated with CD8+ TEM (Figure 21) and there was minimal overlap in cells expressing TBX21 and PDCD1. To further confirm our findings in the bulk mRNAseq and DNA methylation data we then looked to see if the expression of MX1 and SP140 were associated with TBX21 expression and found that the expression of these genes was associated with TBX21 expression, supporting our findings. Together, this data suggests strongly that MX1, SP140 utilize epigenetic modification and cooperate with TBX21 through TP53 cascade pathways to decrease expression of TEX cell markers and improve function in CD8+ T cells.

In addition, beside the COADREAD, as we know, low CTLs infiltration and poor immunogenicity in the BC microenvironment is a challenge to treatment with ICIs. And ICIs were recently approved by the FDA in TNBC, which are characterized by high levels of tumor infiltrating CTL. However, DNA methylation is a component of epigenetic modification involved in gene expression programming that can promote the progression of cancers, including BC. Therefore, the association between transcriptional and methylation changes that modulate CTL infiltration in the tumor microenvironment in specific subtypes of breast cancer is still unclear.

We then next compare transcriptional and methylation profile data in patients with different BC subtypes to identify targetable genes to improve CTL infiltration and ICI efficacy in BC patients. We enrolled 1,212 BC patients with mRNAseq and 783 patients with HM450 methylation data. Again, we applied MCP-counter to generate the abundance of ten cell populations for further analysis. As expected, Patients in the TNBC subtype had the highest median CTL abundance score, and the majority of patients in this subgroup were in the highest quartile of CTL infiltration (Figure 22).

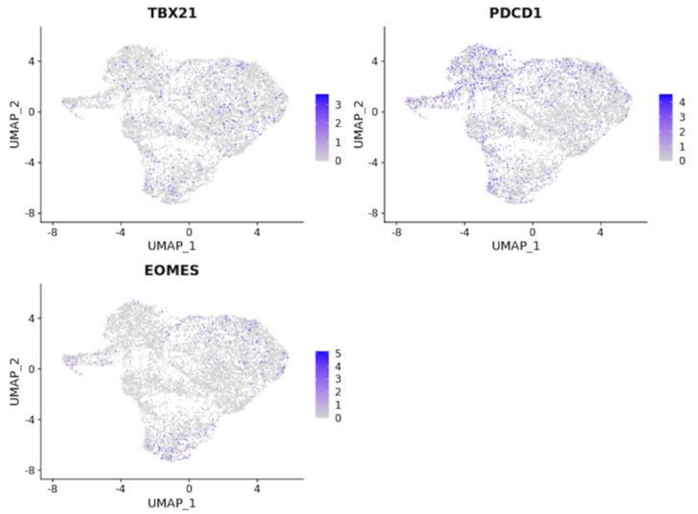


Figure 21. Feature plot to show the distribution of TBX21, PDCD1, and EOMES in each cluster. PDCD1 showed the highest density in CD8+TEX (clusters 1 and 7) and was associated with CD8+TEM. X axis and Y axis are two dimensions of UMAP.

As previously described in the methods, using TNBC as the reference subtype given its high level of CTL infiltration and predicted response to ICI, we performed differential gene

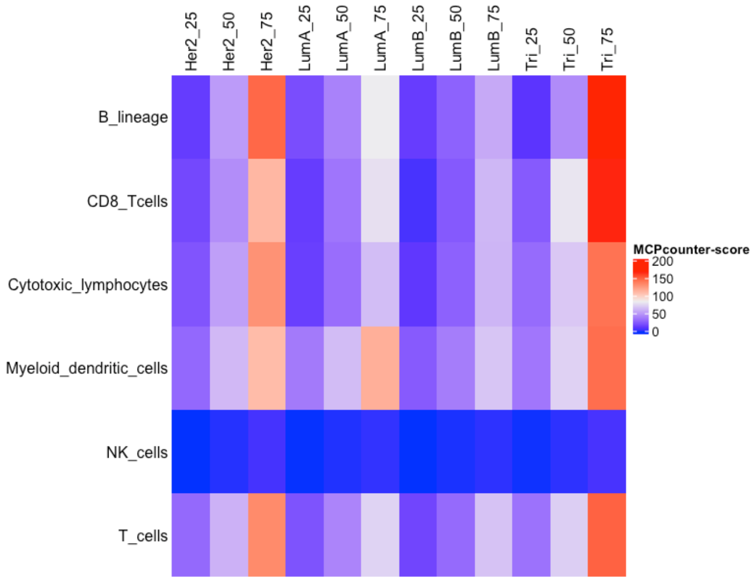


Figure 22. Quartiles of MCP-counter scores in molecular subtypes: 1) Selected microenvironment cell populations scores were presented by Quartiles for each molecular subtype of BC; 2) Red box highlighted the differences of CTL cell population MCP – counter score between subtypes, and the CTL score in Triple negative BC(the 75th percentile) is significantly higher than the others.

expression and differential methylated region analysis. Expression of 20,531 genes and 293,276 methylation loci were determined for each tumor sample from the TCGA. We then did Pearson correlation coefficient analysis for crosstalk genes and MCP-counter score for each patients and each cell population (r>|0.7|). We observed that CD38, HLA.DOB, LAMP3, RUNX3 were hypomethylated in LumA and highly correlated with CTL scores; AIM2, SEL1L3, TOX were hypermethylated in LumA and highly correlated with CTL scores; AIM2 , GPR55 , UBD were hypermethylated in LumB and highly correlated with CTL scores; CHST2 were hypermethylated in LumB and highly correlated with CTL scores (Figure 23). Specifically, interferon-inducible protein AIM2 is a gene that was altered in both LumA and LumB, is associated with response to ICK in murine models and may be a potential target to improve response to immune based therapy in these subtypes of BC patients.

3.6.9 DISCUSSION

DNA methylation plays an important role in the development of COADREAD; however, its potential role in immune dysfunction is less well characterized [54]. This may be most important in the subgroup of patients that demonstrate hypermethylation, MSI-H patients.

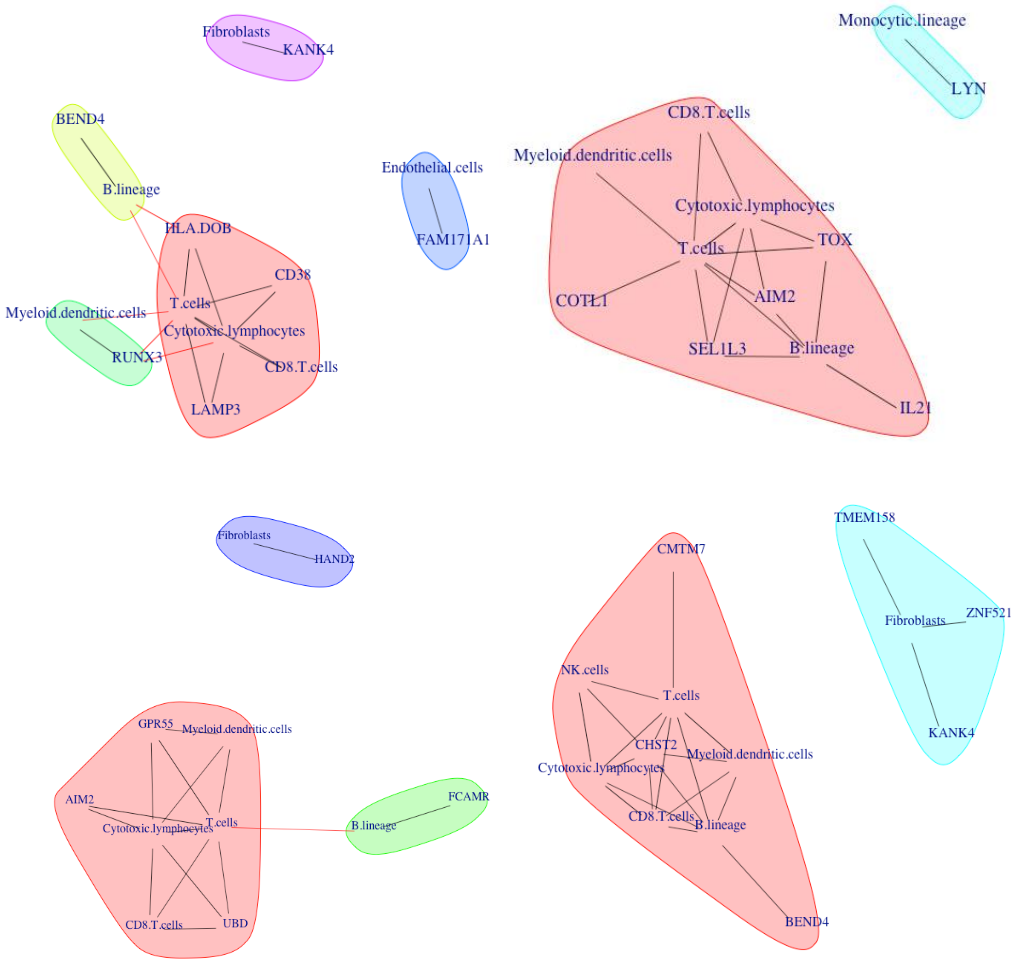


Figure 23. Correlations between microenvironment subpopulations and crosstalk genes: a, CD38， HLA.DOB， LAMP3， RUNX3 were hypomethylated in LumA and highly correlated with CTL scores; b, AIM2， SEL1L3， TOX， were hypermethylated in LumA and highly correlated with CTL scores; c, AIM2， GPR55， UBD， were hypermethylated in LumB and highly correlated with CTL scores; Fig 2d, CHST2 were hypermethylated in LumB and highly correlated with CTL scores;

These are the patients most likely to respond to immune checkpoint inhibition and are the only COADREAD patients with a current FDA approved indication for ICI therapy. However, MSI status has been shown to be an incomplete marker, leading to the establishment of CMS subtypes to better characterize this heterogeneous disease on a molecular level [73]. To better understand the impact of DNA methylation on immune dysfunction and potential resistance to immune-based therapy in COADREAD patients, we integrated data from the TCGA with cutting edge bioinformatic techniques demonstrating that in CMS1 patients, the “inflamed” subtype of COADREAD, TBX21 methylation and expression stratified patients into groups with significantly different survival. A known critical transcriptional factor in T cell function, this suggested an important role for methylation of TBX21 in COADREAD [70]. Using further in-depth analysis validated in publicly available scRNAseq data, we found evidence that TP53 pathway genes MX1 and SP140 may participate in the epigenetic modification of TBX21, impacting patient survival.

Unlike the canonical pattern of epigenetic modification, TBX21 demonstrated both increased methylation and expression (high-high) in CMS1 patients. This may primarily be related to the majority of CMS1 patients being MSI-H. However, this implied that methylation of TBX21 was most important in CMS1 patients. Previous research suggests that genes in the high-high quadrant have a more complex and dynamic manner of regulation of gene expression by DNA methylation, especially during carcinogenesis and metastasis [74]. Our results identified that hypermethylated loci in TBX21 are mainly enriched in CTCF, a promoter, and some coding areas in CMS1 patients. In this case, it is reasonable to suggest that these patients potentially developed a protective mechanism that hypermethylated selective promoters within CpG islands in the TBX21 gene under tumor associated stress that blocks the binding of transcriptional tumor-induced repressor proteins to facilitate active TBX21 transcription, contributing to our observations in this study.

There is significant research linking TBX21 with T cell exhaustion, therefore it fits that this gene would impact survival in the subset of COADREAD patients characterized by an active anti-tumor immune response. Early studies first demonstrated that a gradient of TBX21 expression led to direction of CD8 T cells towards short-lived, highly active effectors versus long-term “slow burn” effectors in response to viral illness [75]. Further work has then developed the story into showing clearly important roles for TBX21 in the regulation of interferon-γ production by regulating the accessibility of the IFNG gene by chromatin remodeling in the context of infection[76]. In another murine infection model, others further demonstrated that TBX21 appeared to be a critical regulator of PD-1 expression and was susceptible to epigenetic disruption impacting CD8 T cell exhaustion[77]. Most recently, Beltra, et al published an elegant description of the impact of transcriptional alteration of TBX21 and TOX on CD8 T cell exhaustion where they also demonstrate that the CTL exhaustion depicted in chronic viral illness correlates with similar makers in CTL from a small group of patients with melanoma [78]. In this study, we further build on the data exhibited by Beltra, et al by demonstrating that TBX21 expression and epigenetic modification have an impact on patient outcome in COADREAD which has not previously been shown. Additionally, we connect the work of Barili, et al by showing that epigenetic alteration of TP53 pathway genes is dysregulated in conjunction with TBX21, suggesting novel therapeutic combinations that may work to improve outcomes in "inflamed" COADREAD [79]. Despite the substantial improvement in outcome that has been seen with the application of ICI therapy in these patients, most recent data suggests that only 40% of patients demonstrate therapeutic response [58]."

Although we have attempted to control for weaknesses, our study suffers from some significant limitations. First, as with any retrospective analysis of clinical data, this study is subjected to bias based on clinical factors. Additionally, while the TCGA is the most robust publicly available dataset including multi-omics and clinical data, our analysis suffers from limitations due to patient number in the specific subgroups [80]. Additionally, due to the nature of the data we are unable to review patient records for accuracy and to further explore potential impact of confounding variables on outcomes of interest. We attempted to compensate for this by comparative analysis of clinical factors associated with survival, which we noted no differences between the high-low and low-high CMS1 patients on univariate analysis. To validate our findings, we used other publicly available methylation datasets; however, there were no datasets in which to confirm all our findings, particularly in the context of CMS subtyping. To test our hypothesis regarding the importance of TBX21 in CTL function, we utilized a large publicly available scRNAseq dataset, but we are unable to directly explore TBX21 expression and methylation at the single cell level in COADREAD using existing available data. This represents an exciting area of future exploration utilizing advanced single cell techniques [81].

In this study, we integrate comprehensive scRNAseq, mRNAseq, methylation, TMB, neoantigen, clinical and MCP-counter scores from multiple datasets to explore the role of TBX21 in CD8+ T cell exhaustion and patient outcome in COADREAD. We demonstrate that in CMS1 subtype COADREAD patients, those with high TBX21 expression and low methylation have improved survival suggesting a critical role for epigenetic regulation of TBX21 in the outcome of these patients. Moreover, epigenetic modification of TBX21 along with MX1 and SP140 may provide this impact via the TP53/P53 pathway. Therefore, DNA methyltransferase inhibitors combined with immune checkpoint blockade may further support CTL function in CMS1 patients via protection of TBX21 expression. Future work focusing on enrolling sufficient patients with strict quality control and application cutting edge biomedical informatics techniques at the single cell level is required for further understanding of DNA methylation and gene expression in CD8+T cell function and patient outcome[82].

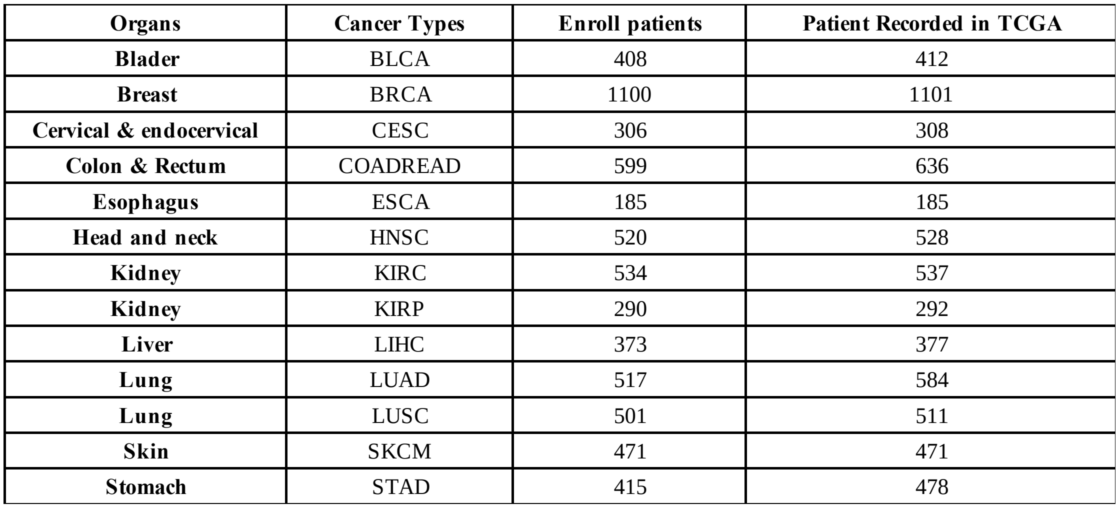
3.7 PAN-CANCER ANALYSIS

Cancer is a group of diseases that is the second leading cause of death in the United States. Each cancer type has a different mortality rate, and patients of each type have heterogeneous responses to treatment. However, all these types of cancer are characterized by uncontrolled growth and the spread of abnormal cells. This suggests the existence of common mechanisms among these types in addition to the unique mechanisms for each type. Studying the inter and intra heterogeneity of cancer is crucial to understanding the mechanisms of action, identifying biomarkers, finding drug targets, and developing or repurposing therapies. To study the heterogeneity of cancer, pan-cancer analyses need to be conducted over a wide range of cancer types. The purpose of these analyses is to find homogeneous subgroups of patients across different cancer types. Finding these subgroups enables targeting the cancer mechanism over different cancer types and this can reduce the cost of treating patients because one drug can be used as a regimen for more than one cancer type. Additionally, finding subgroups across cancer types will improve patient survival by finding a better treatment through targeting common mechanisms instead of only targeting a mechanism unique to a specific cancer type. Also, to reduce the cost associated with treating patients, old drugs can be investigated for new uses by developing and implementing drug repositioning methods over pan-cancer data after stratifying patients into subgroups.

Pan-cancer represents a comprehensive heterogeneity analysis required to solve the intra-heterogeneity problem which is the major barrier to classifying patients into potential benefited groups [83]. This type of analysis has been used for a variety of research questions including studying genes' effect on cancer in general instead of studying their effect on each cancer type [84, 85]. For the subgrouping, pan-cancer analysis has been done using different methods to stratify cancer patients into subgroups. It was used to stratify patients based on the expression status of one gene and its upstream, downstream, and correlated genes to study cancer prognosis in each subgroup. This type of study does not consider the wide range of the genetic variation because it focuses on a narrow set of genes. A wider set of molecular features were considered for stratifying patients into groups where patients in each subgroup share the same molecular features. These molecular features were represented by RNA signatures, Tumor Mutational Burden (TMB), Copy-Number Alteration (CNA), and genes expression using network analysis [83, 86, 87]. These methods represent an improvement on previous methods that depend on a limited set of genes, but they only focus on genotypic features without taking into account heterogeneity on the phenotypic level. Other methods addressed the importance phenotypic heterogeneity by using the phenotypic features to stratify patients into subgroups. For the stratification purposes, the surgery and radiotherapy status, socioeconomic status, mortality after surgery, race, age, and metastasis site were used [79]. While these methods do use the phenotypic features, they miss the importance of genotypic features to stratify patients into subgroups.

3.7.1 DATA DESCRIPTION AND PROCESSING

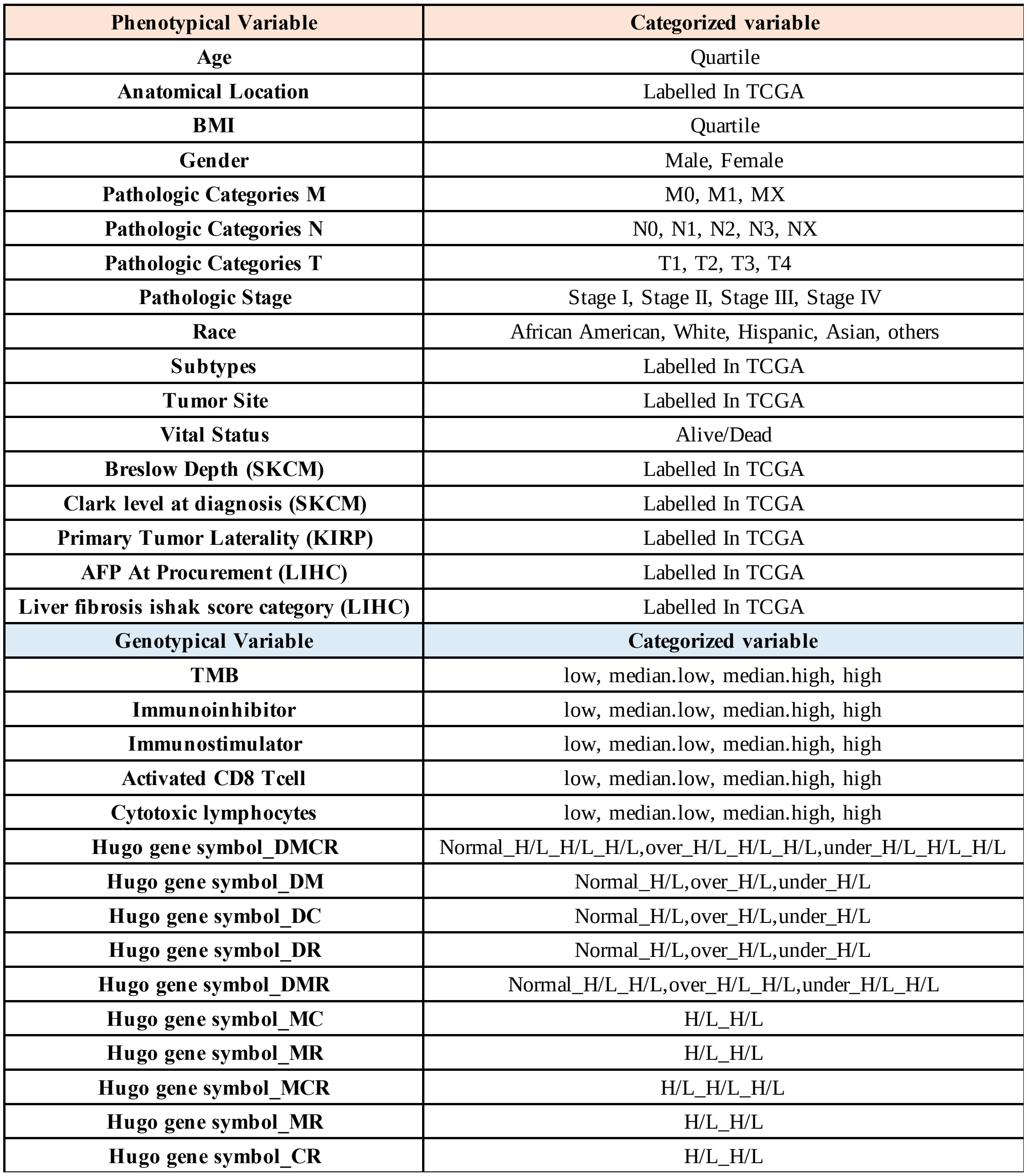
Next, we overviewed and did crosswise comparison to phenotypical and genotypical variables. From TCGA, we totally selected 13 cancers types for analysis, and patients got recorded in the TCGA dataset is 6420 and the patients were collected for analysis was 6129.



Organs	Cancer Types	Enroll patients	Patient Recorded in TCGA
Blader	BLCA	408	412
Breast	BRCA	1100	1101
Cervical & endocervical	CESC	306	308
Colon & Rectum	COADREAD	599	636
Esophagus	ESCA	185	185
Head and neck	HNSC	520	528
Kidney	KIRC	534	537
Kidney	KIRP	290	292
Liver	LIHC	373	377
Lung	LUAD	517	584
Lung	LUSC	501	511
Skin	SKCM	471	471
Stomach	STAD	415	478

Table 10. Types of cancers with enrolled number of patients

Fourteen caner types and one subtype in 11 organs were included (Table 10). The dataset for these patients consists of genotypic and phenotypic variables. It has 17 phenotypic and 15 categories genotypic variables (Table 11). The genotypic data consists of two parts. TMB, Immunoinhibitor, Immunostimulator, Activated CD8 T cell and Cytotoxic lymphocytes were general genotypical variables. In another word, these 5 genotypical variables were all enrolled for subgroup mining analysis. The rest of specific 12 genotypic variables were partially enrolled in different cancer. The different number of specific genotypic variables enrolled were mainly caused by the number DEGs, DEGs associated with survival and overlapped genes across different omics. The continuous variables in the phenotypic dataset were categorized by quartile. The genotypic variables were categorized based on the z-score of each gene in each cancer.



Phenotypical Variable	Categorized variable
Age	Quartile
Anatomical Location	Labelled In TCGA
BMI	Quartile
Gender	Male, Female
Pathologic Categories M	M0, M1, MX
Pathologic Categories N	N0, N1, N2, N3, NX
Pathologic Categories T	T1, T2, T3, T4
Pathologic Stage	Stage I, Stage II, Stage III, Stage IV
Race	African American, White, Hispanic, Asian, others
Subtypes	Labelled In TCGA
Tumor Site	Labelled In TCGA
Vital Status	Alive/Dead
Breslow Depth (SKCM)	Labelled In TCGA
Clark level at diagnosis (SKCM)	Labelled In TCGA
Primary Tumor Laterality (KIRP)	Labelled In TCGA
AFP At Procurement (LIHC)	Labelled In TCGA
Liver fibrosis ishak score category (LIHC)	Labelled In TCGA
Genotypical Variable	**Categorized variable**
TMB	low, median.low, median.high, high
Immunoinhibitor	low, median.low, median.high, high
Immunostimulator	low, median.low, median.high, high
Activated CD8 Tcell	low, median.low, median.high, high
Cytotoxic lymphocytes	low, median.low, median.high, high
Hugo gene symbol_DMCR	Normal_H/L_H/L_H/L,over_H/L_H/L_H/L,under_H/L_H/L_H/L
Hugo gene symbol_DM	Normal_H/L,over_H/L,under_H/L
Hugo gene symbol_DC	Normal_H/L,over_H/L,under_H/L
Hugo gene symbol_DR	Normal_H/L,over_H/L,under_H/L
Hugo gene symbol_DMR	Normal_H/L_H/L,over_H/L_H/L,under_H/L_H/L
Hugo gene symbol_MC	H/L_H/L
Hugo gene symbol_MR	H/L_H/L
Hugo gene symbol_MCR	H/L_H/L_H/L
Hugo gene symbol_MR	H/L_H/L
Hugo gene symbol_CR	H/L_H/L

Table 11. Variables enrolled and categorization

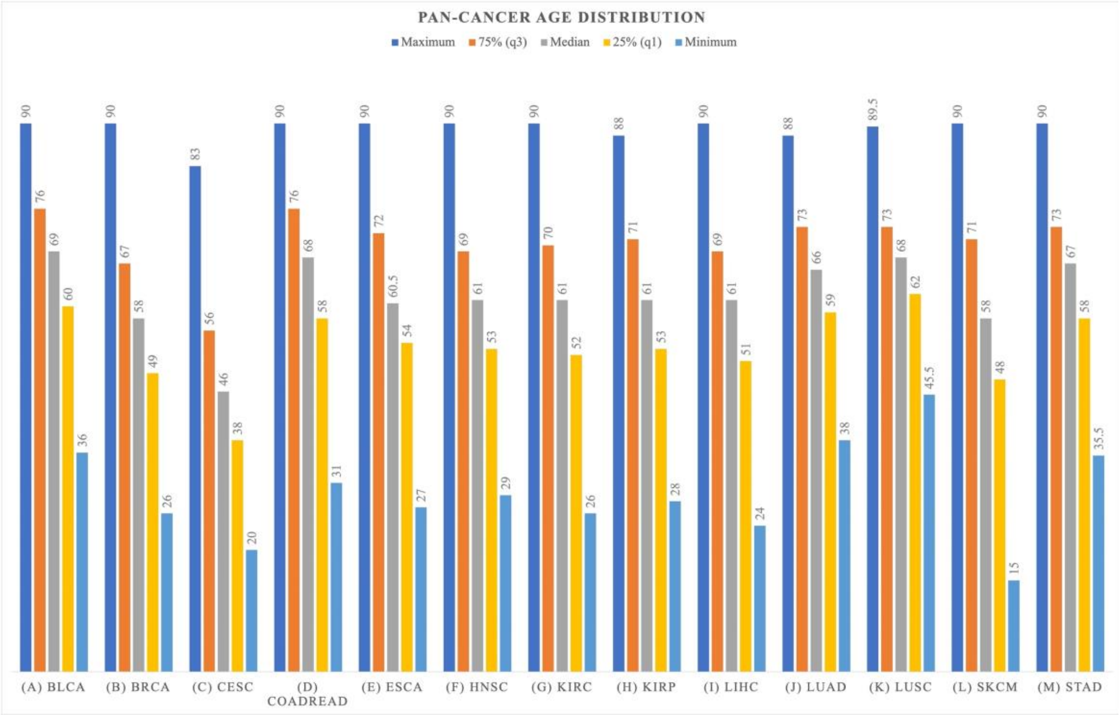


Figure 24 The distribution of number of patients

3.7.2 PATIENT CHARACTERISTICS

In totally enrolled 6,420 patients, there were 3,200 males (49.8%), 3,102 females (48.3%), 118 patients without records (1.8%). The lowest median age with CESC diagnosis is 46 years old, and highest median age with BLCA diagnosis is 69 years old (Figure 24). However, the SKCM has the earliest minimum diagnosis age. This distribution would help to make the decision to initiate the early screening for patients with high predispose or family histories. Overview patients' survival, 4,306 (67.1%) were living, and 1,966 (31.1%) were deceased. As figure 25 and table 12 shows that SKCM has the lowest survival rate and BRCA has the highest overall survival percentage.

Study ID	Living	Deceased	NA	Total	Deceased Percent
blca	231	182	0	413	44%
brca	949	155	4	1108	14%
cesc	236	73	1	310	24%
coadread	502	131	7	640	20%
esca	108	78	0	186	42%
hnsc	305	225	0	530	42%
kirc	361	177	0	538	33%
kirp	248	44	1	293	15%
lihc	247	132		379	35%
luad	336	288	62	686	42%
lusc	284	220	7	511	43%
skcm	251	228	1	480	48%
stad	268	175	35	478	37%

Table 12 The number of patients' overall survival

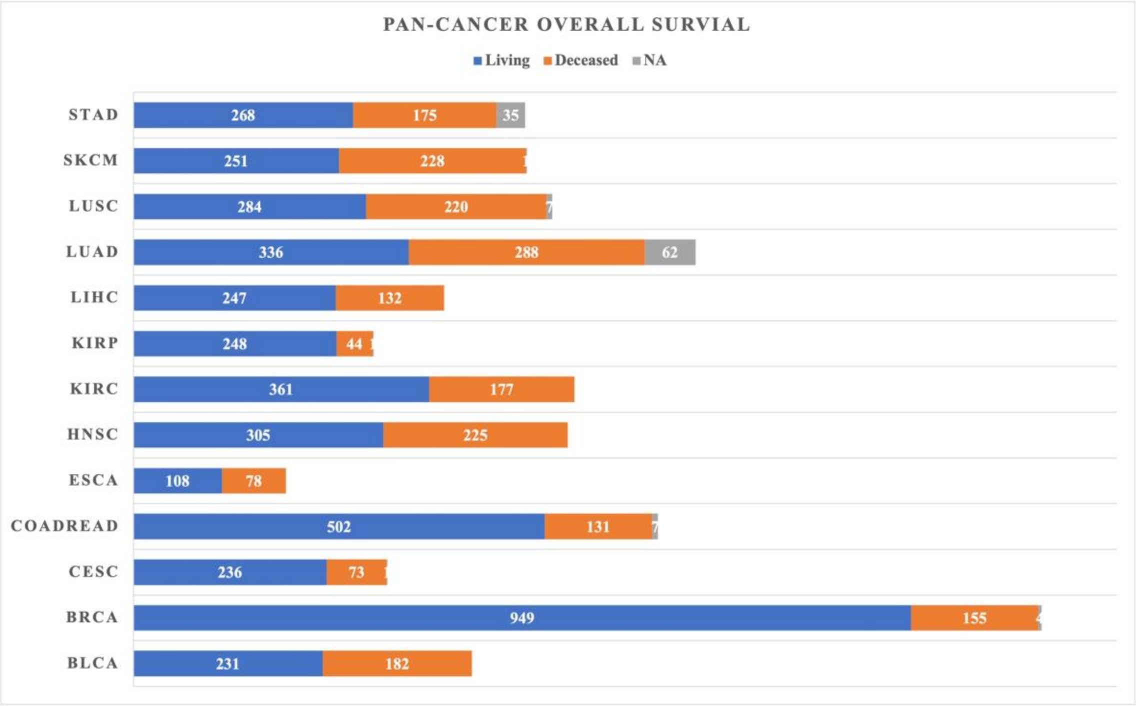


Figure 25. patients' overall survival

We next observed that BLCA all the patients recorded were white; KIRP has the highest black or African American than the other cancer; and Asian population has the highest portion in LIHC than any other cancers (Table 13, Figure 26).

Study ID	blca	brca	cesc	coadread	esca	hnsc	kirc	kirp	lihc	luad	lusc	skcm	stad
WHITE	100%	76%	78%	79%	60%	88%	88%	75%	51%	86%	90%	97%	73%
BLACK OR AFRICAN AMERICAN	0%	18%	11%	18%	16%	9%	11%	22%	5%	12%	8%	0%	3%
ASIAN	0%	6%	7%	3%	24%	2%	2%	2%	44%	2%	2%	3%	23%
AMERICAN INDIAN OR ALASKA NATIVE	0%	0%	3%	0%	0%	0%	0%	1%	1%	0%	0%	0%	0%
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 13 Patients' Race

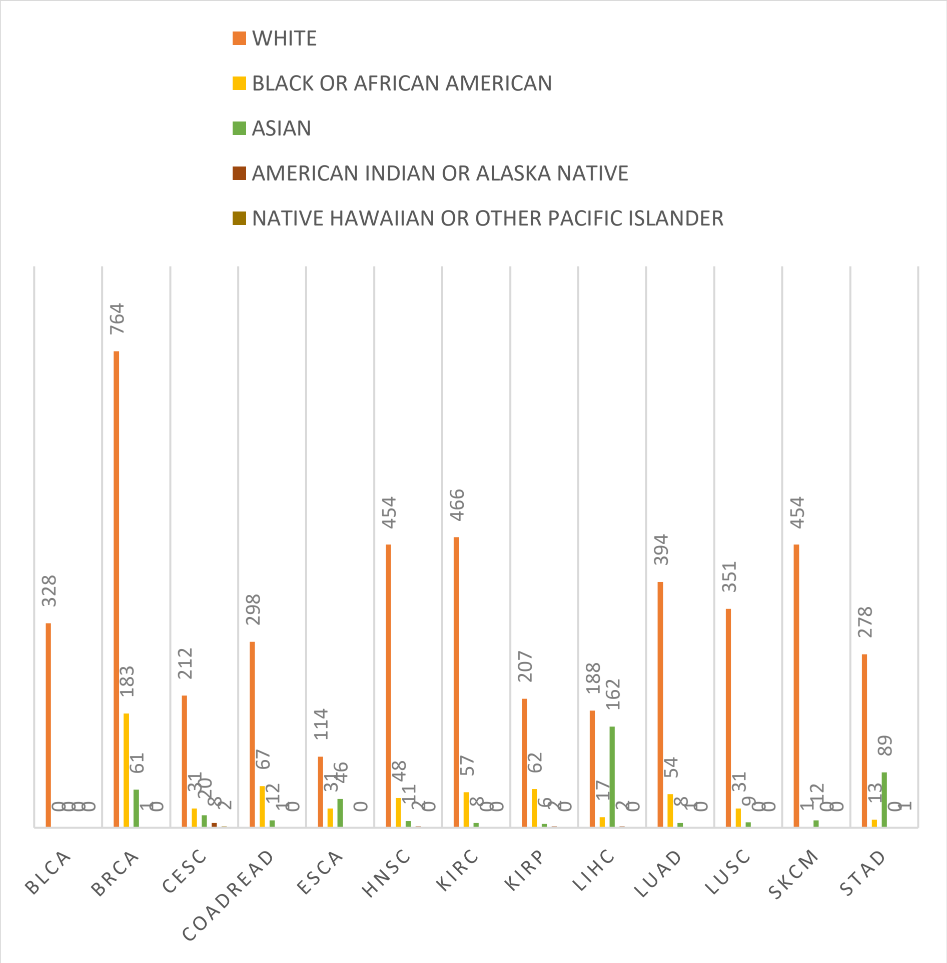


Figure 26. The distribution of race

Additionally, in the comparison of pathologic stages, KIRP has the highest ratio of patients were diagnosed at Stage I; BRCA patients were mostly diagnosed at Stage II; STAD and SKCM patients were most likely diagnosed at Stage III; additionally, HNSC patients close to 50% of patients were diagnosed at Stage IV. In the TNM comparisons, not surprisingly, HNSC has the highest rates of patients got diagnosis at T4; ESCA has the earliest lymph node metastasis; and KIRC has the earliest distal metastasis (Figure 27,28).

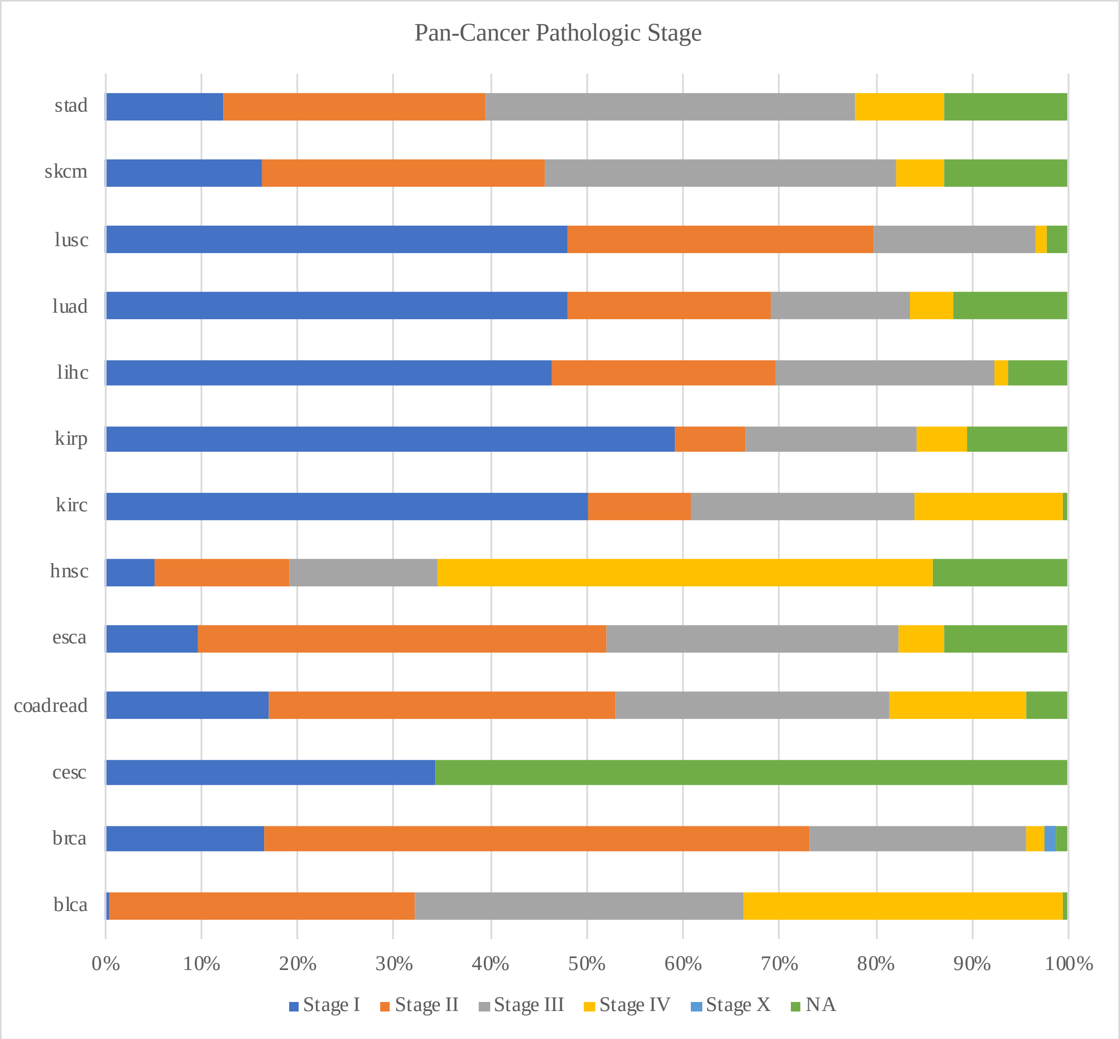


Figure 27. Percentage of Pathologic stage cross the cancers

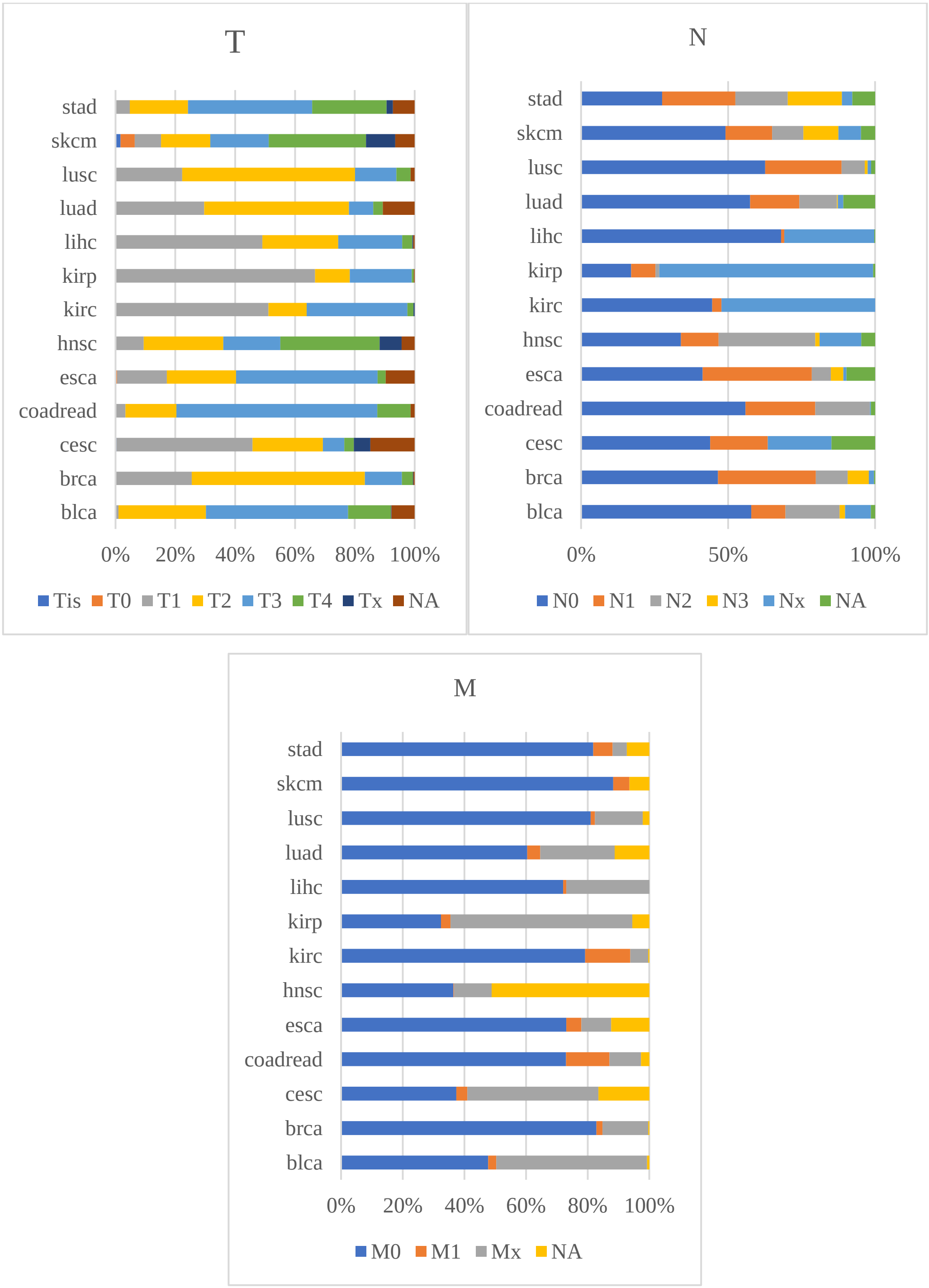
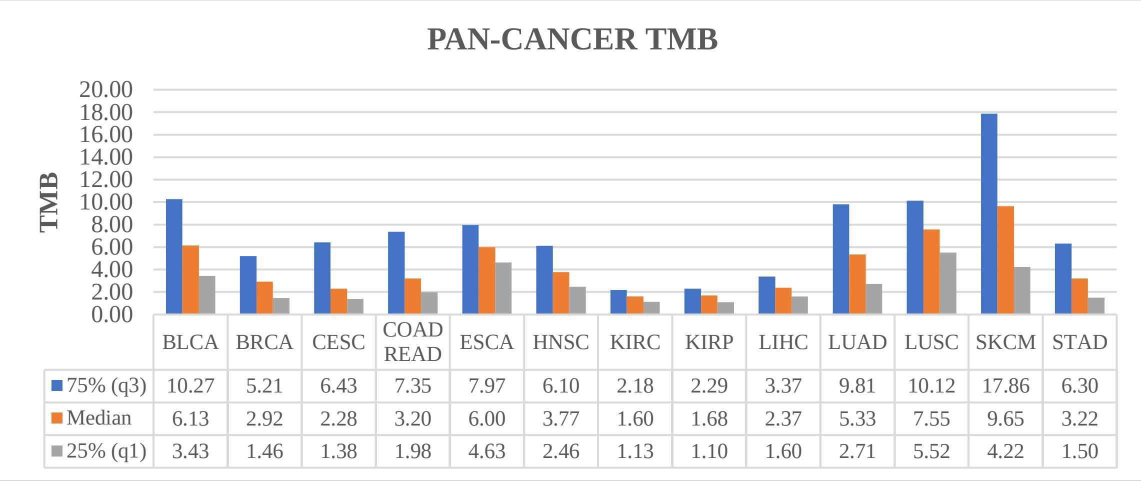


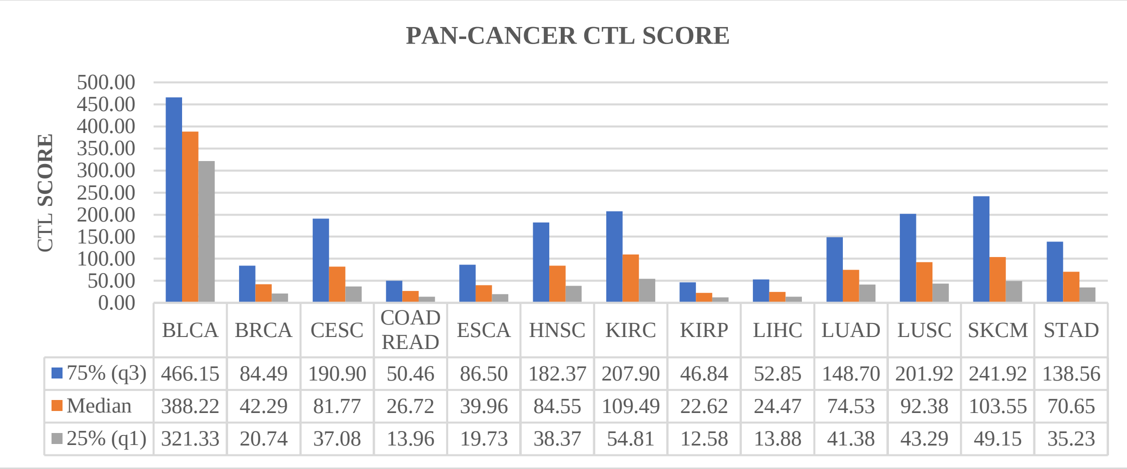
Figure 28. Percentage of TNM Pathologic stage cross the cancers

For each cancer, we applied the quartile methods to categorize the general genotypical variables, such CTL score. However, the range of each genotypical variables are widely different (Figure 29). For example, the maximum of KIRP TMB value is close to the median value of HNSC. Therefore, for better comparison and align across pan-cancer, we compared the quartile values for general variables as references for the further subgroup results interpretation. As figure 29 shows that SKCM has the highest median and 75 percentile TMB



	BLCA	BRCA	CESC	COAD READ	ESCA	HNSC	KIRC	KIRP	LIHC	LUAD	LUSC	SKCM	STAD
🟦 75% (q3)	10.27	5.21	6.43	7.35	7.97	6.10	2.18	2.29	3.37	9.81	10.12	17.86	6.30
🟧 Median	6.13	2.92	2.28	3.20	6.00	3.77	1.60	1.68	2.37	5.33	7.55	9.65	3.22
■ 25% (q1)	3.43	1.46	1.38	1.98	4.63	2.46	1.13	1.10	1.60	2.71	5.52	4.22	1.50

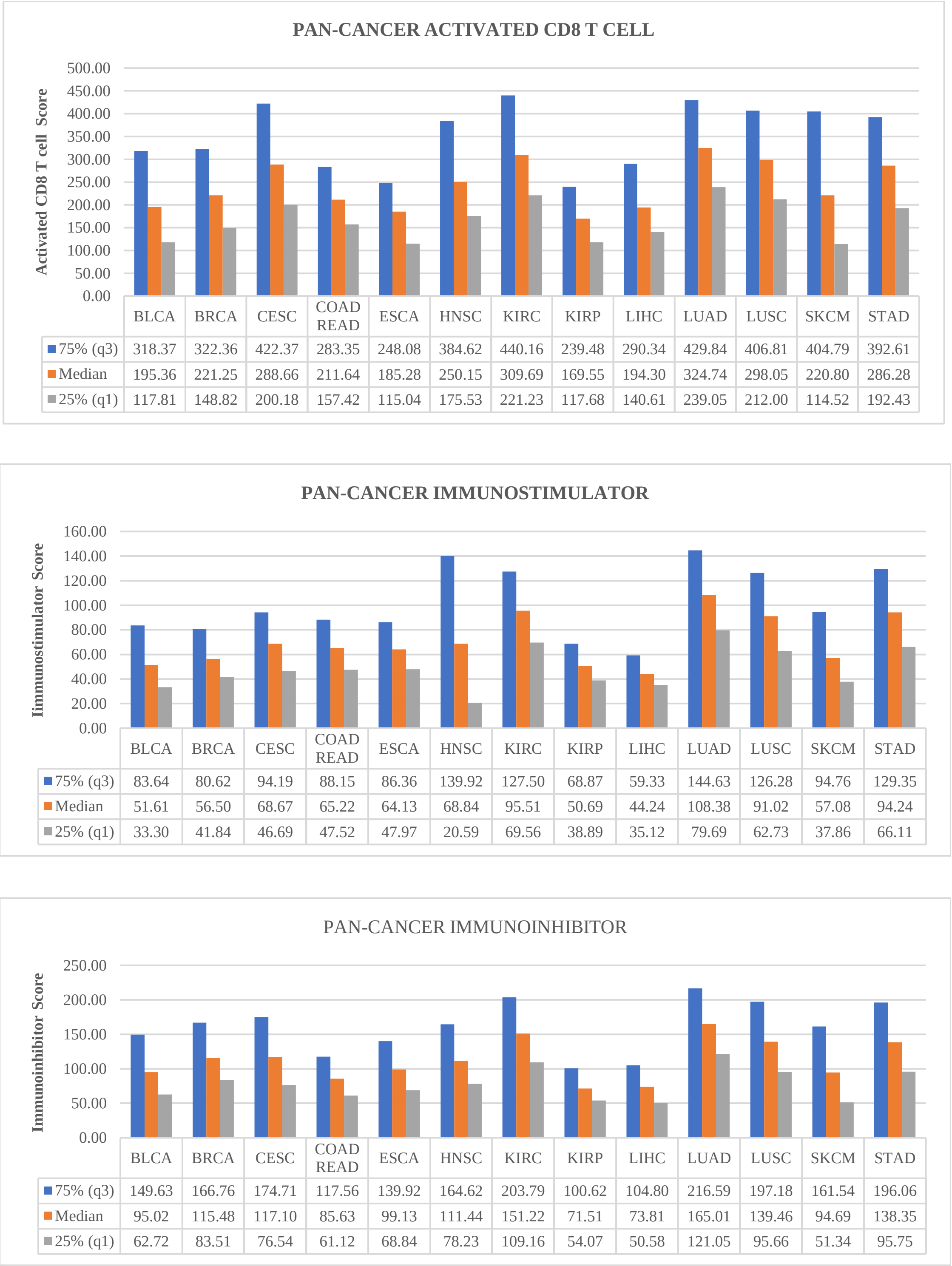
Figure 29. TMB level across pan-cancers



	BLCA	BRCA	CESC	COAD READ	ESCA	HNSC	KIRC	KIRP	LIHC	LUAD	LUSC	SKCM	STAD
🟦 75% (q3)	466.15	84.49	190.90	50.46	86.50	182.37	207.90	46.84	52.85	148.70	201.92	241.92	138.56
🟧 Median	388.22	42.29	81.77	26.72	39.96	84.55	109.49	22.62	24.47	74.53	92.38	103.55	70.65
■ 25% (q1)	321.33	20.74	37.08	13.96	19.73	38.37	54.81	12.58	13.88	41.38	43.29	49.15	35.23

Figure 30. CTL level across pan-cancers

level, whereas, KIRC has the lowest average TMB level. Next, we evaluated the CTL score


PAN-CANCER ACTIVATED CD8 T CELL

	BLCA	BRCA	CESC	COAD READ	ESCA	HNSC	KIRC	KIRP	LIHC	LUAD	LUSC	SKCM	STAD
🟦 75% (q3)	318.37	322.36	422.37	283.35	248.08	384.62	440.16	239.48	290.34	429.84	406.81	404.79	392.61
🟧 Median	195.36	221.25	288.66	211.64	185.28	250.15	309.69	169.55	194.30	324.74	298.05	220.80	286.28
⬜ 25% (q1)	117.81	148.82	200.18	157.42	115.04	175.53	221.23	117.68	140.61	239.05	212.00	114.52	192.43

PAN-CANCER IMMUNOSTIMULATOR

	BLCA	BRCA	CESC	COAD READ	ESCA	HNSC	KIRC	KIRP	LIHC	LUAD	LUSC	SKCM	STAD
🟦 75% (q3)	83.64	80.62	94.19	88.15	86.36	139.92	127.50	68.87	59.33	144.63	126.28	94.76	129.35
🟧 Median	51.61	56.50	68.67	65.22	64.13	68.84	95.51	50.69	44.24	108.38	91.02	57.08	94.24
⬜ 25% (q1)	33.30	41.84	46.69	47.52	47.97	20.59	69.56	38.89	35.12	79.69	62.73	37.86	66.11

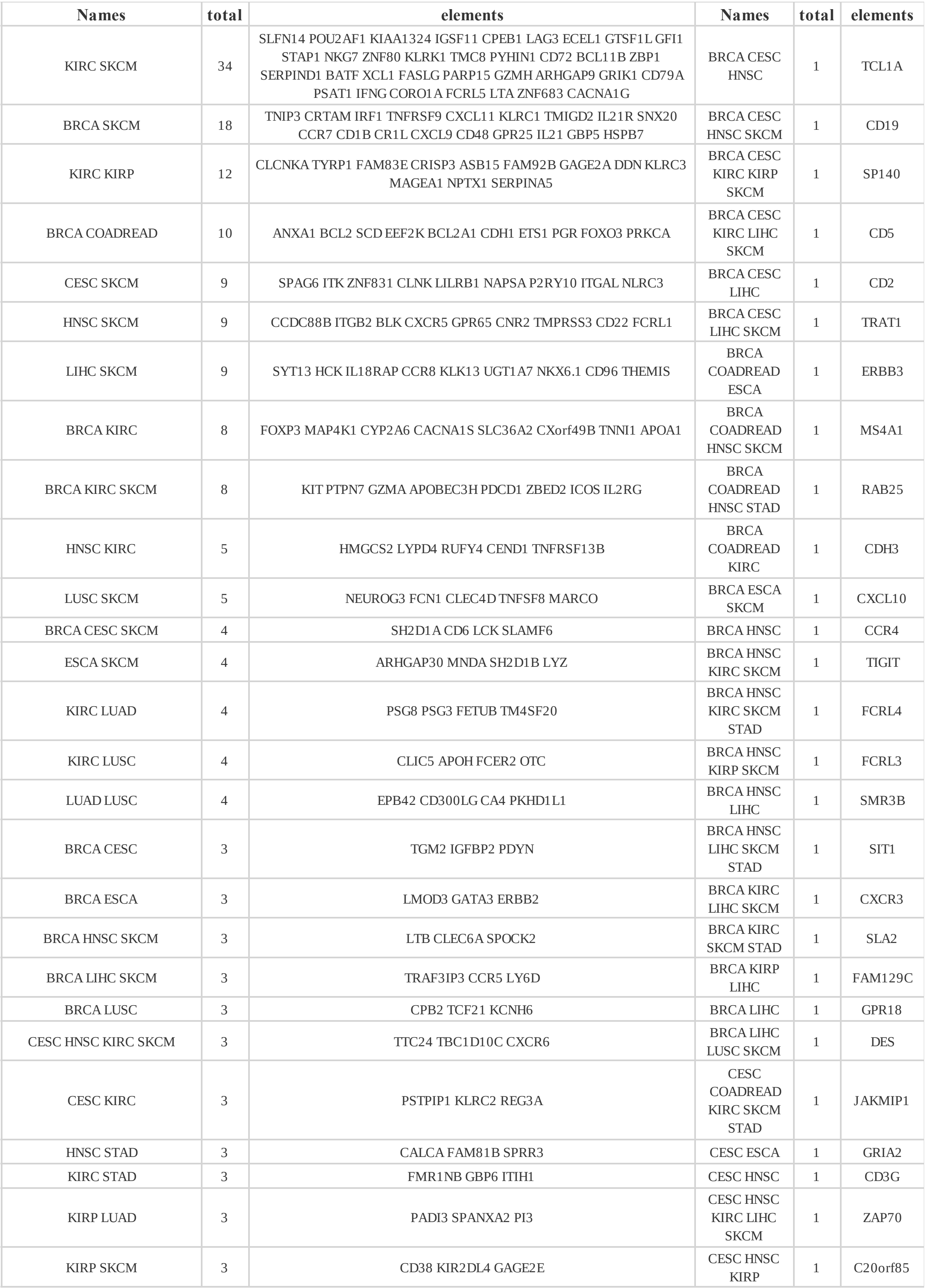
PAN-CANCER IMMUNOINHIBITOR

	BLCA	BRCA	CESC	COAD READ	ESCA	HNSC	KIRC	KIRP	LIHC	LUAD	LUSC	SKCM	STAD
🟦 75% (q3)	149.63	166.76	174.71	117.56	139.92	164.62	203.79	100.62	104.80	216.59	197.18	161.54	196.06
🟧 Median	95.02	115.48	117.10	85.63	99.13	111.44	151.22	71.51	73.81	165.01	139.46	94.69	138.35
⬜ 25% (q1)	62.72	83.51	76.54	61.12	68.84	78.23	109.16	54.07	50.58	121.05	95.66	51.34	95.75

Figure 31. cell marker score across pan-cancers.

and cell marker score (Figure 30). Interestingly, LUAD has the highest cell marker scores in all three subsets and KIRP has the lowest over all cell marker scores (Figure 31).

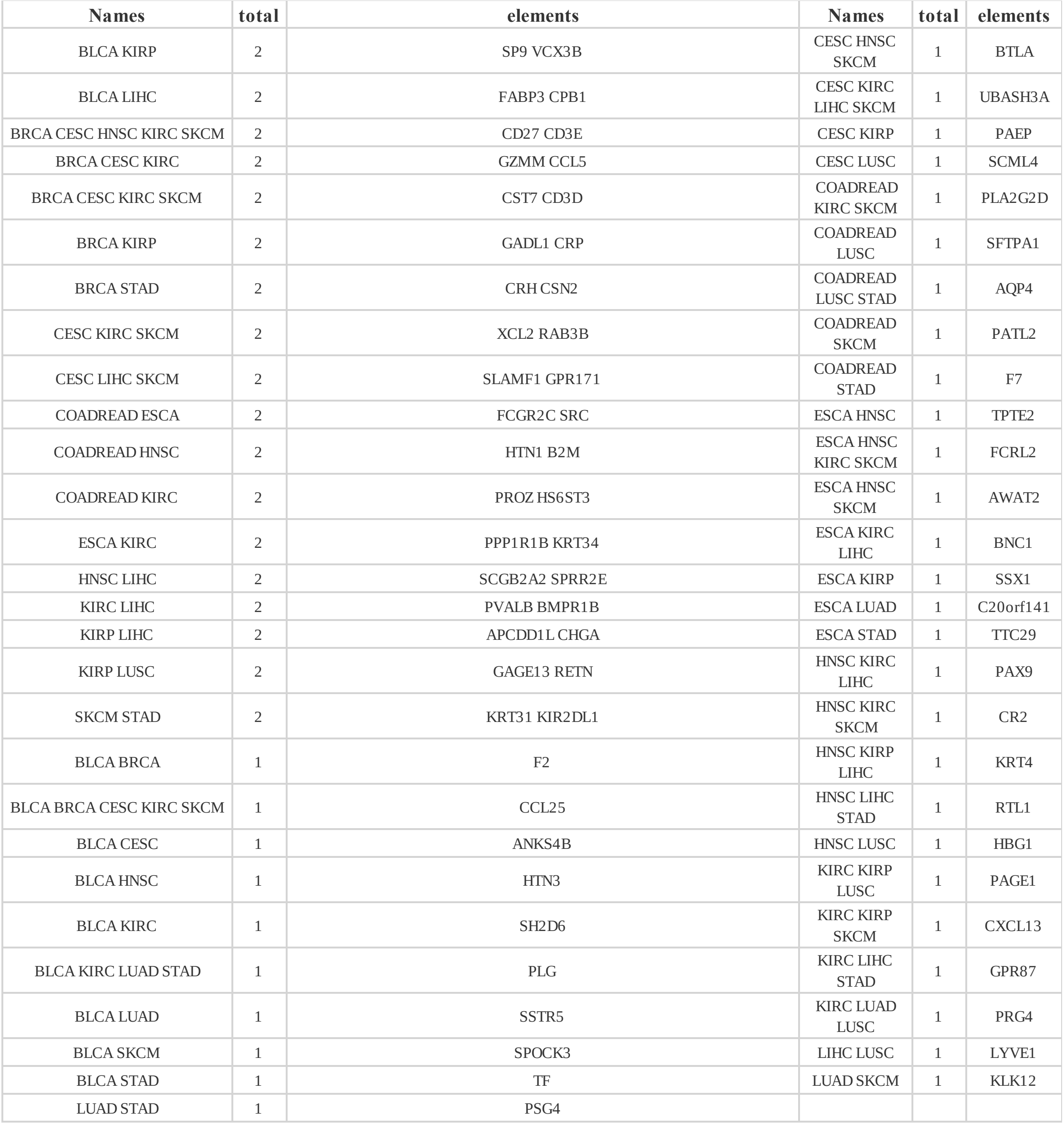
We further did the overlap analysis between each cancer to determinate if there are



Names	total	elements	Names	total	elements
KIRC SKCM	34	SLFN14 POU2AF1 KIAA1324 IGSF11 CPEB1 LAG3 ECEL1 GTSF1L GFI1 STAP1 NKG7 ZNF80 KLRK1 TMC8 PYHIN1 CD72 BCL11B ZBP1 SERPIND1 BATF XCL1 FASLG PARP15 GZMH ARHGAP9 GRIK1 CD79A PSAT1 IFNG CORO1A FCRL5 LTA ZNF683 CACNA1G	BRCA CESC HNSC	1	TCL1A
BRCA SKCM	18	TNIP3 CRTAM IRF1 TNFRSF9 CXCL11 KLRC1 TMIGD2 IL21R SNX20 CCR7 CD1B CR1L CXCL9 CD48 GPR25 IL21 GBP5 HSPB7	BRCA CESC HNSC SKCM	1	CD19
KIRC KIRP	12	CLCNKA TYRP1 FAM83E CRISP3 ASB15 FAM92B GAGE2A DDN KLRC3 MAGEA1 NPTX1 SERPINA5	BRCA CESC KIRC KIRP SKCM	1	SP140
BRCA COADREAD	10	ANXA1 BCL2 SCD EEF2K BCL2A1 CDH1 ETS1 PGR FOXO3 PRKCA	BRCA CESC KIRC LIHC SKCM	1	CD5
CESC SKCM	9	SPAG6 ITK ZNF831 CLNK LILRB1 NAPSA P2RY10 ITGAL NLRC3	BRCA CESC LIHC	1	CD2
HNSC SKCM	9	CCDC88B ITGB2 BLK CXCR5 GPR65 CNR2 TMPRSS3 CD22 FCRL1	BRCA CESC LIHC SKCM	1	TRAT1
LIHC SKCM	9	SYT13 HCK IL18RAP CCR8 KLK13 UGT1A7 NKX6.1 CD96 THEMIS	BRCA COADREAD ESCA	1	ERBB3
BRCA KIRC	8	FOXP3 MAP4K1 CYP2A6 CACNA1S SLC36A2 CXorf49B TNNI1 APOA1	BRCA COADREAD HNSC SKCM	1	MS4A1
BRCA KIRC SKCM	8	KIT PTPN7 GZMA APOBEC3H PDCD1 ZBED2 ICOS IL2RG	BRCA COADREAD HNSC STAD	1	RAB25
HNSC KIRC	5	HMGCS2 LYPD4 RUFY4 CEND1 TNFRSF13B	BRCA COADREAD KIRC	1	CDH3
LUSC SKCM	5	NEUROG3 FCN1 CLEC4D TNFSF8 MARCO	BRCA ESCA SKCM	1	CXCL10
BRCA CESC SKCM	4	SH2D1A CD6 LCK SLAMF6	BRCA HNSC	1	CCR4
ESCA SKCM	4	ARHGAP30 MNDA SH2D1B LYZ	BRCA HNSC KIRC SKCM	1	TIGIT
KIRC LUAD	4	PSG8 PSG3 FETUB TM4SF20	BRCA HNSC KIRC SKCM STAD	1	FCRL4
KIRC LUSC	4	CLIC5 APOH FCER2 OTC	BRCA HNSC KIRP SKCM	1	FCRL3
LUAD LUSC	4	EPB42 CD300LG CA4 PKHD1L1	BRCA HNSC LIHC	1	SMR3B
BRCA CESC	3	TGM2 IGFBP2 PDYN	BRCA HNSC LIHC SKCM STAD	1	SIT1
BRCA ESCA	3	LMOD3 GATA3 ERBB2	BRCA KIRC LIHC SKCM	1	CXCR3
BRCA HNSC SKCM	3	LTB CLEC6A SPOCK2	BRCA KIRC SKCM STAD	1	SLA2
BRCA LIHC SKCM	3	TRAF3IP3 CCR5 LY6D	BRCA KIRP LIHC	1	FAM129C
BRCA LUSC	3	CPB2 TCF21 KCNH6	BRCA LIHC	1	GPR18
CESC HNSC KIRC SKCM	3	TTC24 TBC1D10C CXCR6	BRCA LIHC LUSC SKCM	1	DES
CESC KIRC	3	PSTPIP1 KLRC2 REG3A	CESC COADREAD KIRC SKCM STAD	1	JAKMIP1
HNSC STAD	3	CALCA FAM81B SPRR3	CESC ESCA	1	GRIA2
KIRC STAD	3	FMR1NB GBP6 ITIH1	CESC HNSC	1	CD3G
KIRP LUAD	3	PADI3 SPANXA2 PI3	CESC HNSC KIRC LIHC SKCM	1	ZAP70
KIRP SKCM	3	CD38 KIR2DL4 GAGE2E	CESC HNSC KIRP	1	C20orf85

Table 14a. Overlapped specfic genotypical variables

homogeneous genes caused by the significant differences between CTL-high and CTL-low



Names	total	elements	Names	total	elements
BLCA KIRP	2	SP9 VCX3B	CESC HNSC SKCM	1	BTLA
BLCA LIHC	2	FABP3 CPB1	CESC KIRC LIHC SKCM	1	UBASH3A
BRCA CESC HNSC KIRC SKCM	2	CD27 CD3E	CESC KIRP	1	PAEP
BRCA CESC KIRC	2	GZMM CCL5	CESC LUSC	1	SCML4
BRCA CESC KIRC SKCM	2	CST7 CD3D	COADREAD KIRC SKCM	1	PLA2G2D
BRCA KIRP	2	GADL1 CRP	COADREAD LUSC	1	SFTPA1
BRCA STAD	2	CRH CSN2	COADREAD LUSC STAD	1	AQP4
CESC KIRC SKCM	2	XCL2 RAB3B	COADREAD SKCM	1	PATL2
CESC LIHC SKCM	2	SLAMF1 GPR171	COADREAD STAD	1	F7
COADREAD ESCA	2	FCGR2C SRC	ESCA HNSC	1	TPTE2
COADREAD HNSC	2	HTN1 B2M	ESCA HNSC KIRC SKCM	1	FCRL2
COADREAD KIRC	2	PROZ HS6ST3	ESCA HNSC SKCM	1	AWAT2
ESCA KIRC	2	PPP1R1B KRT34	ESCA KIRC LIHC	1	BNC1
HNSC LIHC	2	SCGB2A2 SPRR2E	ESCA KIRP	1	SSX1
KIRC LIHC	2	PVALB BMPR1B	ESCA LUAD	1	C20orf141
KIRP LIHC	2	APCDD1L CHGA	ESCA STAD	1	TTC29
KIRP LUSC	2	GAGE13 RETN	HNSC KIRC LIHC	1	PAX9
SKCM STAD	2	KRT31 KIR2DL1	HNSC KIRC SKCM	1	CR2
BLCA BRCA	1	F2	HNSC KIRP LIHC	1	KRT4
BLCA BRCA CESC KIRC SKCM	1	CCL25	HNSC LIHC STAD	1	RTL1
BLCA CESC	1	ANKS4B	HNSC LUSC	1	HBG1
BLCA HNSC	1	HTN3	KIRC KIRP LUSC	1	PAGE1
BLCA KIRC	1	SH2D6	KIRC KIRP SKCM	1	CXCL13
BLCA KIRC LUAD STAD	1	PLG	KIRC LIHC STAD	1	GPR87
BLCA LUAD	1	SSTR5	KIRC LUAD LUSC	1	PRG4
BLCA SKCM	1	SPOCK3	LIHC LUSC	1	LYVE1
BLCA STAD	1	TF	LUAD SKCM	1	KLK12
LUAD STAD	1	PSG4			

Table 14b. Overlapped specific genotypic variables.

(Table 14). Each part of table xx contains two *names*, *total* and *elements* columns. Each *elements* cell has the overlapped genes according to the cancer types under *names* cell. The *total* between *names* and *elements* were the number of overlapped genes. For example, between cancer KIRC and SKCM, there were 34 overlapped genes. These 34 genes were both DEGs from CTL-high and CTL-low comparison from each cancer. However, we couldn't determine if these genes were modulated by multiple omics. Since we labelled

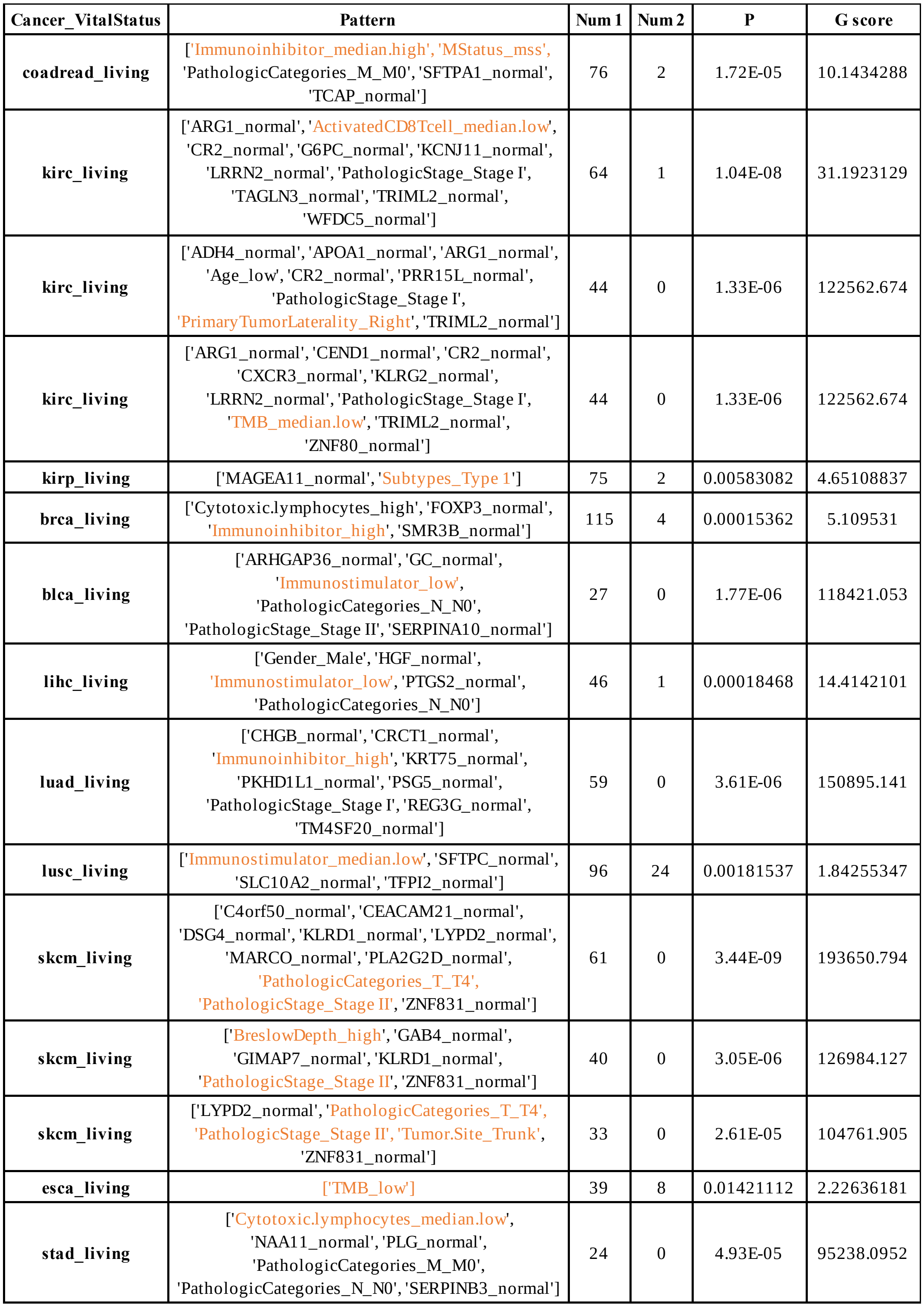
DEGs during categorization, we would identify if any DEG was modulated by multiple omics and associate with survival in subgroup mining analysis. To have an overview to the overlapped genes between different cancers, it would help to give us a clue for further integration of subgroups from different cancers.

3.7.3 PATIENT SUBGROUP MINING RESULTS

The implementation of the subgroup discovery process resulted in a set of subgroups with a contrast score (G score). Comparing with unsupervised contrast subgroup mining, our modified outcome-oriented contrast subgroup mining would be more comparative when integrate with other cancer types. Therefore, we predefined and selected five general genotypical variables and one phenotypical varible, which were CTL score, TMB, activated CD8 T cell score, immunoinhibitory, immunostimulatory and vital status as outcomes for processing for every cancer. After defining the outcome, each categorized under outcome would be contrasted, and all the other variables were not defined as outcome would be applied for the mining themselves. Every time, we only defined one outcome for contrast mining processing. The reason was that we tried to define multiple outcomes at same time, it turned out that the subgroups lost the tracking of co-occurrence between defined outcomes. The subgroups then were initially clustered based on the phenotypic variables' distribution in cancer types. After implementing subgroup mining, we used patients' overall survival as references as primary subgroups crosswise comparison.

We highlighted subgroups from each cancer with living and deceased overall survival status as an example. There were couple of steps of how we selected these subgroups: 1) Each outcome with its categorized value generates large amount of patterns, in here, we also called subgroups. 2) From each cancer with living or deceased, we removed the reduplicated variables from all the subgroups. 3) We next reviews by general phenotypical and genotypical variables. 4) We then pick out the patterns with interesting co-occurrence general phenotypical and genotypical variables. 5) Some contrast patterns were contained in their superior patterns; we then applied the G score as references to rank and analysis them.

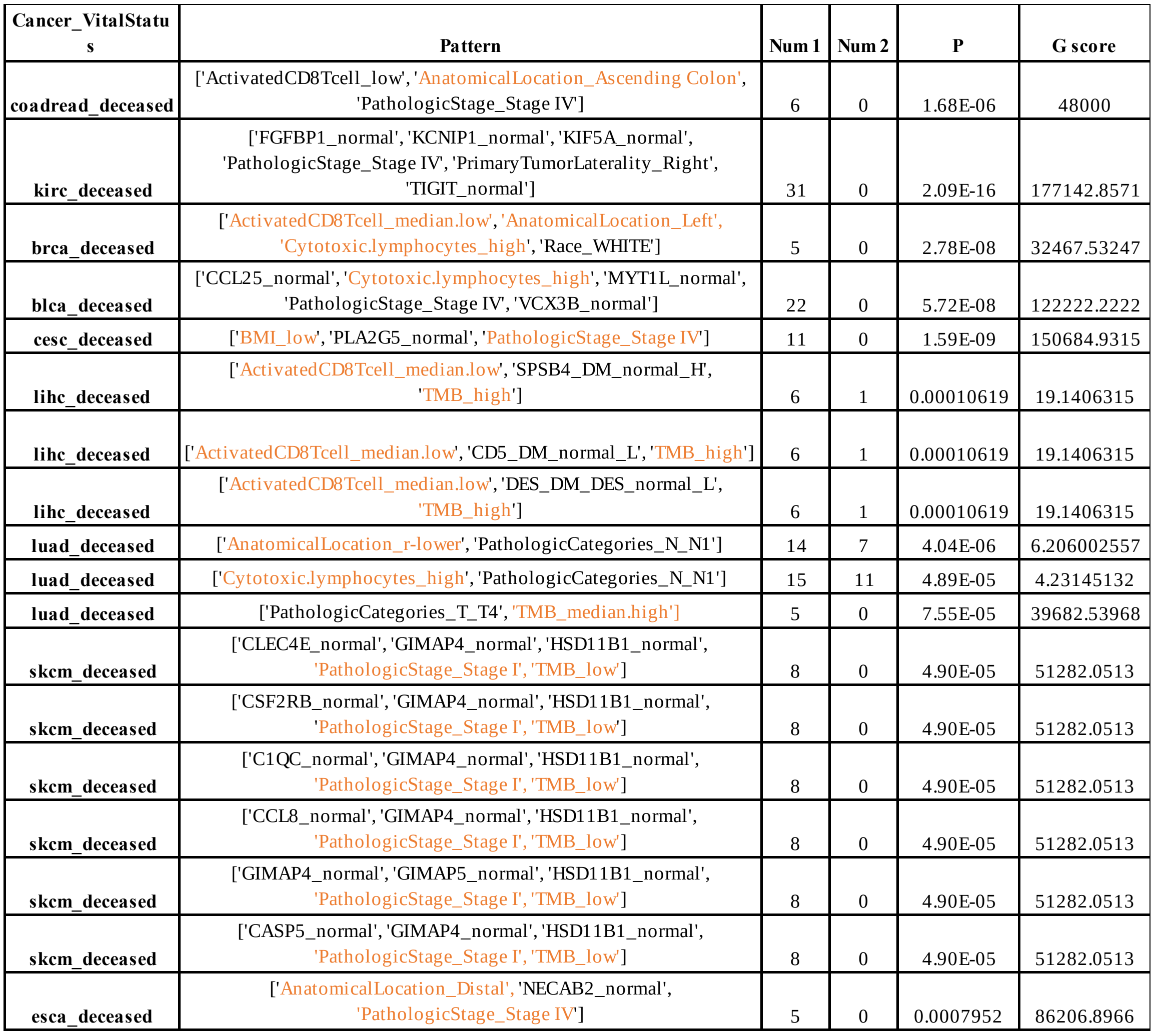
As table 15 shows that COADREAD, KIRC, BRCA, LIHC, LUAD, LUSC, and STAD patients with living status, we noticed that these patients with immunoinhibitor_median.high,



Cancer_VitalStatus	Pattern	Num 1	Num 2	P	G score
coadread_living	['Immunoinhibitor_median.high', 'MStatus_mss', 'PathologicCategories_M_M0', 'SFTPA1_normal', 'TCAP_normal']	76	2	1.72E-05	10.1434288
kirc_living	['ARG1_normal', 'ActivatedCD8Tcell_median.low', 'CR2_normal', 'G6PC_normal', 'KCNJ11_normal', 'LRRN2_normal', 'PathologicStage_Stage I', 'TAGLN3_normal', 'TRIML2_normal', 'WFDC5_normal']	64	1	1.04E-08	31.1923129
kirc_living	['ADH4_normal', 'APOA1_normal', 'ARG1_normal', 'Age_low', 'CR2_normal', 'PRR15L_normal', 'PathologicStage_Stage I', 'PrimaryTumorLaterality_Right', 'TRIML2_normal']	44	0	1.33E-06	122562.674
kirc_living	['ARG1_normal', 'CEND1_normal', 'CR2_normal', 'CXCR3_normal', 'KLRG2_normal', 'LRRN2_normal', 'PathologicStage_Stage I', 'TMB_median.low', 'TRIML2_normal', 'ZNF80_normal']	44	0	1.33E-06	122562.674
kirp_living	['MAGEA11_normal', 'Subtypes_Type 1']	75	2	0.00583082	4.65108837
brca_living	['Cytotoxic.lymphocytes_high', 'FOXP3_normal', 'Immunoinhibitor_high', 'SMR3B_normal']	115	4	0.00015362	5.109531
blca_living	['ARHGAP36_normal', 'GC_normal', 'Immunostimulator_low', 'PathologicCategories_N_N0', 'PathologicStage_Stage II', 'SERPINA10_normal']	27	0	1.77E-06	118421.053
lihc_living	['Gender_Male', 'HGF_normal', 'Immunostimulator_low', 'PTGS2_normal', 'PathologicCategories_N_N0']	46	1	0.00018468	14.4142101
luad_living	['CHGB_normal', 'CRCT1_normal', 'Immunoinhibitor_high', 'KRT75_normal', 'PKHD1L1_normal', 'PSG5_normal', 'PathologicStage_Stage I', 'REG3G_normal', 'TM4SF20_normal']	59	0	3.61E-06	150895.141
lusc_living	['Immunostimulator_median.low', 'SFTPC_normal', 'SLC10A2_normal', 'TFPI2_normal']	96	24	0.00181537	1.84255347
skcm_living	['C4orf50_normal', 'CEACAM21_normal', 'DSG4_normal', 'KLRD1_normal', 'LYPD2_normal', 'MARCO_normal', 'PLA2G2D_normal', 'PathologicCategories_T_T4', 'PathologicStage_Stage II', 'ZNF831_normal']	61	0	3.44E-09	193650.794
skcm_living	['BreslowDepth_high', 'GAB4_normal', 'GIMAP7_normal', 'KLRD1_normal', 'PathologicStage_Stage II', 'ZNF831_normal']	40	0	3.05E-06	126984.127
skcm_living	['LYPD2_normal', 'PathologicCategories_T_T4', 'PathologicStage_Stage II', 'Tumor.Site_Trunk', 'ZNF831_normal']	33	0	2.61E-05	104761.905
esca_living	['TMB_low']	39	8	0.01421112	2.22636181
stad_living	['Cytotoxic.lymphocytes_median.low', 'NAA11_normal', 'PLG_normal', 'PathologicCategories_M_M0', 'PathologicCategories_N_N0', 'SERPINB3_normal']	24	0	4.93E-05	95238.0952

Table 15.Patients with living overall status highlighted subgroups in pan-cancer

ActivatedCD8Tcell_median.low, Immunoinhibitor_high, Immunostimulator_low, and



Cancer_VitalStatus	Pattern	Num 1	Num 2	P	G score
coadread_deceased	['ActivatedCD8Tcell_low', 'AnatomicalLocation_Ascending Colon', 'PathologicStage_Stage IV']	6	0	1.68E-06	48000
kirc_deceased	['FGFBP1_normal', 'KCNIP1_normal', 'KIF5A_normal', 'PathologicStage_Stage IV', 'PrimaryTumorLaterality_Right', 'TIGIT_normal']	31	0	2.09E-16	177142.8571
brca_deceased	['ActivatedCD8Tcell_median.low', 'AnatomicalLocation_Left', 'Cytotoxic.lymphocytes_high', 'Race_WHITE']	5	0	2.78E-08	32467.53247
blca_deceased	['CCL25_normal', 'Cytotoxic.lymphocytes_high', 'MYT1L_normal', 'PathologicStage_Stage IV', 'VCX3B_normal']	22	0	5.72E-08	122222.2222
cesc_deceased	['BMI_low', 'PLA2G5_normal', 'PathologicStage_Stage IV']	11	0	1.59E-09	150684.9315
lihc_deceased	['ActivatedCD8Tcell_median.low', 'SPSB4_DM_normal_H', 'TMB_high']	6	1	0.00010619	19.1406315
lihc_deceased	['ActivatedCD8Tcell_median.low', 'CD5_DM_normal_L', 'TMB_high']	6	1	0.00010619	19.1406315
lihc_deceased	['ActivatedCD8Tcell_median.low', 'DES_DM_DES_normal_L', 'TMB_high']	6	1	0.00010619	19.1406315
luad_deceased	['AnatomicalLocation_r-lower', 'PathologicCategories_N_N1']	14	7	4.04E-06	6.206002557
luad_deceased	['Cytotoxic.lymphocytes_high', 'PathologicCategories_N_N1']	15	11	4.89E-05	4.23145132
luad_deceased	['PathologicCategories_T_T4', 'TMB_median.high']	5	0	7.55E-05	39682.53968
skcm_deceased	['CLEC4E_normal', 'GIMAP4_normal', 'HSD11B1_normal', 'PathologicStage_Stage I', 'TMB_low']	8	0	4.90E-05	51282.0513
skcm_deceased	['CSF2RB_normal', 'GIMAP4_normal', 'HSD11B1_normal', 'PathologicStage_Stage I', 'TMB_low']	8	0	4.90E-05	51282.0513
skcm_deceased	['C1QC_normal', 'GIMAP4_normal', 'HSD11B1_normal', 'PathologicStage_Stage I', 'TMB_low']	8	0	4.90E-05	51282.0513
skcm_deceased	['CCL8_normal', 'GIMAP4_normal', 'HSD11B1_normal', 'PathologicStage_Stage I', 'TMB_low']	8	0	4.90E-05	51282.0513
skcm_deceased	['GIMAP4_normal', 'GIMAP5_normal', 'HSD11B1_normal', 'PathologicStage_Stage I', 'TMB_low']	8	0	4.90E-05	51282.0513
skcm_deceased	['CASP5_normal', 'GIMAP4_normal', 'HSD11B1_normal', 'PathologicStage_Stage I', 'TMB_low']	8	0	4.90E-05	51282.0513
esca_deceased	['AnatomicalLocation_Distal', 'NECAB2_normal', 'PathologicStage_Stage IV']	5	0	0.0007952	86206.8966

Table 16. Patients with deceased overall status highlighted subgroups in pan-cancer

Cytotoxic.lymphocytes_median.low scores. In KIRC, ESCA patients, there some populations with TMB_low. Additionally, there was a subgroup of COADREAD patients with MSS status, meanwhile the immunoinhibitory score was high. And KIRC patients with tumor collected at right side and type 1 KIRP patients were associated with better survival. In SKCM, the subgroups had more varieties than the other cancer types. For example, one subgroup had BreslowDepth_high associated with PathologicStage_Stage II still may have better survival. As well as another subgroup, patients identified tumor at trunk and the tumor stage was T4.

We then analyzed the deceased patients’ subgroups (Table 16). In the COADREAD patients with AnatomicalLocation_Ascending Colon and ActivatedCD8Tcell_low were more likely associated with worse survival. And one BRCA subgroup patients were associated with worse survival even with Cytotoxic.lymphocytes_high score. However, this population were also with ActivatedCD8Tcell_median.low status, and the tumor location was at left breast. In addition, we observed that one LIHC population with TMB high status associated with ActivatedCD8Tcell_median.low status. In LUAD and ESCA subgroups, algorithms identified tumor were at right-lower side of lung and distal of esophagus were associated with worse survival.

3.8 DISCUSSION

Cancer is a group of diseases characterized by the abnormal growth of cells. The existence of common characteristics among these different types of cancer indicates the importance of pan-cancer analysis to study the inter heterogeneity in cancer. This part of the study aims to find homogeneous subpopulations that share genotypic and phenotypic characteristics across the cancer type. Exploring homogeneous subgroups will help identify "True eligible" ICI-sensitive populations that can apply on more cancer types and provide the healthcare more comprehensive clinical decision support for different subgroups of patients in each cancer type. In this pan-cancer analysis, the stratification algorithm was applied to identify homogeneous and heterogeneous features of patients across cancer types by multi-omics from each subgroup.

From TCGA, the genotypic and phenotypic data for 6,420 patients across 13 cancer types were obtained. The patient stratification framework was implemented to find the subgroups. Then, using the genotypic features of these homogeneous and heterogeneous subgroups, "True-eligible" patients were identified based on the phenotypic and genotypic features and their connection to each other. After filtering the resulting subgroup, vital status and general genotypical variables as outcome were selected to explore the subgroups. This resulted in 358 possibilities of outcomes were identified. Because the goal is to find the homogeneous and heterogeneous feature from each subgroup that should be identified in cancer types. The homogeneous feature may help to identify ICI-sensitive populations, and the heterogeneous features may provide more evidence to improve the ICI sensitivity in the rest of not sensitive populations.

The subgroups were identified based on the genotypical and phenotypical features with pre-selected outcomes were interested. In the section 3.7.3, we mainly introduced vital status as outcome-oriented subgroups. The highlighted subgroups were selected based on the variables included general genotypical, phenotypical and partially specific genotypical variables. In the table 15, each cancer as an example we selected interesting subgroups to show from living or deceased outcomes. From the living populations across pan-cancer, we revealed that single omic is not always representative. Any well accepted biomarkers are varies based on the differences of characteristics of cancers. For example, TMB is a measure of the total number of mutations per megabyte of tumor tissue. A TMB is generally predictive of response to ICI therapy across multiple tumor types[88]. But the universal cutoff of 10 mutations/Mb to signify a high burden may not be applicable for all tumors. For example, a LIHC population with deceased status were TMB_high but at meantime, the ActivatedCD8Tcell of this population was median.low. And in a SKCM population, the TMB status associated with low status and this population of patients were deceased in a very early pathologic stage. However, all the subgroups were determined by the data exploration, further analysis is needed in the context of wet lab experiments and clinical trials to validate these results before recommending them for clinical use.

Chapter 4

Conclusion and Future Work

4.1 CONCLUSION

This dissertation aims to stratify a disease population based on the genotypic and phenotypic features into phenotypic and genotypic features for each subgroup. To achieve that, an explainable artificial intelligence (XAI) framework was implemented. The design, implementation, and validation results demonstrate the potential of the XAI framework to stratify patients into subgroups based on their genotypic features. Stratifying patients using genotypic and phenotypic features, the ability of explaining the results, and the ability to provide clinical decision for each subgroup to overcome the limitation of identifying “True eligible” ICI receivers.

In Chapter Two, the development of the patient stratification and the outcome-oriented framework was introduced. The stratification process consists of a three-layer system where data mining was used to find co-occurrence phenotypic and genotypic features within a heterogeneous disease population.

In Chapter Three, the implementation of patient stratification, subgroup identification and interpretation were presented. The implementation was performed on different datasets from the TCGA database. First, the pipeline was developed on colorectal cancer and breast cancer data to integrate the phenotypical and genotypical data. MCP-counter were introduced by applying mRNAseq data and anatomic location. Then, we developed the pipeline for integrating of multi-omics by colorectal and breast cancer data. In the second step of building pipeline, the cut-in-edge single cell RNAseq data and 10 publicly available datasets were applied to validate the results. Last, we introduced the subgroup contrast mining results by vital status as an example to summarize and interpret.

From the identified subgroups, we confirmed that any individual of biomarkers is not accurate enough as a “gold” standard to guide the clinical decisions. For a comprehensive evaluation to a patient characteristic has to weight and consider from multi-omics.

4.2 LIMITATIONS

This is a data driven approach and the availability of the data represents a crucial need to do this kind of analysis. One of the limitations of this study is the availability of open access datasets. This could be a limitation for any other data driven approach, but what makes it more challenging in this study is the requirement to have both phenotypical and genotypical data for a large number of patients because this is a data mining-based analysis.

The other significant challenge that this study faced in all the implementations is the validation in a wet lab setting or ICI clinical trial data. To overcome this limitation, a literature review was used to find the biomedical merit for the results. Still, wet lab experimentation and clinical trial validation is needed before implementing any of our findings on patients.

4.3 CONTRIBUTION TO INFORMATICS AND CANCER RESEARCH

This work aims to find homogeneous subgroups of patients within a disease population to implement precision medicine in our healthcare system to improve patient survival and reduce treatment costs. This was addressed in this study as follows:

1 Developing a pipeline to determine the possible mechanism by genotypical, phenotypical feature.

Methods uses heterogeneous data types to represent the biological system. The genotypic data, phenotypic data, and biomedical entities were used to stratify patients and find groups of patients that share phenotypic and genotypic similarities. At the same time, they have significant contrast from the rest of the disease population. Finding these subgroups will improve drug efficiency, where the drugs will be used only for the patients who can benefit from them because the drug targets the common mechanism among the patients in that subgroup.

2 Presenting explainable results for the medical practitioners: The explainability of this method demonstrates the ability to explain the reasoning behind the selection of the

subgroups and drugs. Its explainability offers the possibility to understand the mechanism of action to address drug resistance and to find combined therapy. This represents an important factor in ensuring the implementation of the method in the clinical setting because applying black-box methods is challenging due to the lack of the explainability of the results. This explainability is crucial for medical practitioners to decide if a patient or a group of patients can be treated with any recommended drug when referring to a computational method.

4.4 FUTURE WORK

The future work will continue to identify the outcome-oriented subgroups and provide clinical meaning information. The "gold standard" of ICIs therapy for cancer patients is an urgent mission. However, from our multi-omics based subgroup mining analysis, we revealed that single omic based "gold standard" is not realistic. Integrating multi-omics on pan-cancer scale-based study is necessary and responsible to determine the dominate and common tumorigenesis in different cancer. Cancer type based homogeneous ICI-sensitive subpopulation identification is possible. Identifying these subgroups may provide more evidence and prediction if a cancer patient with specific feature would benefit from ICIs. The evaluation of the results would be needed to be done by using data from ICIs neoadjuvant clinical trials, the experts' knowledge, and the biomedical knowledge from literature. To increase the scope of this framework implementation, a web-based tool will be developed to ensure easy access to this framework by other researchers in the scientific community. In the next phase of this research, the plan is to validate these promising results at the bench in tumor cell lines in vitro and in vivo.

Reference

1. Le, D.T., et al., *Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade.* Science, 2017. **357**(6349): p. 409-413.

2. Strickler, J.H., B.A. Hanks, and M. Khasraw, *Tumor Mutational Burden as a Predictor of Immunotherapy Response: Is More Always Better?* Clin Cancer Res, 2021. **27**(5): p. 1236-1241.

3. Huang, R.S.P., et al., *A pan-cancer analysis of PD-L1 immunohistochemistry and gene amplification, tumor mutation burden and microsatellite instability in 48,782 cases.* Mod Pathol, 2021. **34**(2): p. 252-263.

4. Haslam, A. and V. Prasad, *Estimation of the Percentage of US Patients With Cancer Who Are Eligible for and Respond to Checkpoint Inhibitor Immunotherapy Drugs.* JAMA Netw Open, 2019. **2**(5): p. e192535.

5. Brahmer, J.R., et al., *Management of Immune-Related Adverse Events in Patients Treated With Immune Checkpoint Inhibitor Therapy: American Society of Clinical Oncology Clinical Practice Guideline.* J Clin Oncol, 2018. **36**(17): p. 1714-1768.

6. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. **71**(3): p. 209-249.

7. Fabrizio, D.A., et al., *Beyond microsatellite testing: assessment of tumor mutational burden identifies subsets of colorectal cancer who may respond to immune checkpoint inhibition.* J Gastrointest Oncol, 2018. **9**(4): p. 610-617.

8. Douillard, J.Y., et al., *Randomized, phase III trial of panitumumab with infusional fluorouracil, leucovorin, and oxaliplatin (FOLFOX4) versus FOLFOX4 alone as first-line treatment in patients with previously untreated metastatic colorectal cancer: the PRIME study.* J Clin Oncol, 2010. **28**(31): p. 4697-705.

9. Sinicrope, F.A., et al., *Prognostic impact of deficient DNA mismatch repair in patients with stage III colon cancer from a randomized trial of FOLFOX-based adjuvant chemotherapy.* J Clin Oncol, 2013. **31**(29): p. 3664-72.

10. Xie, Y.H., Y.X. Chen, and J.Y. Fang, *Comprehensive review of targeted therapy for colorectal cancer.* Signal Transduct Target Ther, 2020. **5**(1): p. 22.

11. Wang, Y., et al., *FDA-Approved and Emerging Next Generation Predictive Biomarkers for Immune Checkpoint Inhibitors in Cancer Patients.* Front Oncol, 2021. **11**: p. 683419.

12. Dolgin, E., *Bringing down the cost of cancer treatment.* Nature, 2018. **555**(7695): p. S26-S29.

13. Hoadley, K.A., et al., *Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin.* Cell, 2014. **158**(4): p. 929-944.

14. Liu, D., et al., *Exploratory Data Mining for Subgroup Cohort Discoveries and Prioritization.* IEEE J Biomed Health Inform, 2020. **24**(5): p. 1456-1468.

15. Al-Taie, Z., et al., *Explainable artificial intelligence in high-throughput drug repositioning for subgroup stratifications with interventionable potential.* J Biomed Inform, 2021. **118**: p. 103792.

16. Shen, Y., et al., *Immunogenomic pathways associated with cytotoxic lymphocyte infiltration and survival in colorectal cancer.* BMC Cancer, 2020. **20**(1): p. 124.

17. Guo, L., H. Zhang, and B. Chen, *Nivolumab as Programmed Death-1 (PD-1) Inhibitor for Targeted Immunotherapy in Tumor.* J Cancer, 2017. **8**(3): p. 410-416.

18. Loupakis, F., et al., *Primary tumor location as a prognostic factor in metastatic colorectal cancer.* J Natl Cancer Inst, 2015. **107**(3).

19. Schreiber, R.D., L.J. Old, and M.J. Smyth, *Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion.* Science, 2011. **331**(6024): p. 1565-70.

20. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

21. Becht, E., et al., *Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression.* Genome Biol, 2016. **17**(1): p. 218.

22. Gillespie, M., et al., *The reactome pathway knowledgebase 2022.* Nucleic Acids Res, 2022. **50**(D1): p. D687-D692.

23. Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer.* Nat Med, 2015. **21**(11): p. 1350-6.

24. Chen, Y., A.T. Lun, and G.K. Smyth, *From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline.* F1000Res, 2016. **5**: p. 1438.

25. Gautier, L., et al., *affy--analysis of Affymetrix GeneChip data at the probe level.* Bioinformatics, 2004. **20**(3): p. 307-15.

26. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

27. Warden, C.D., et al., *COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis.* Nucleic Acids Res, 2019. **47**(15): p. 8335-8336.

28. Wu, T., et al., *clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.* Innovation (N Y), 2021. **2**(3): p. 100141.

29. Yu, G., et al., *DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis.* Bioinformatics, 2015. **31**(4): p. 608-9.

30. Yu, G. and Q.Y. He, *ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization.* Mol Biosyst, 2016. **12**(2): p. 477-9.

31. Charoentong, P., et al., *Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade.* Cell Rep, 2017. **18**(1): p. 248-262.

32. Amezquita, R.A., et al., *Orchestrating single-cell analysis with Bioconductor.* Nat Methods, 2020. **17**(2): p. 137-145.

33. Zhang, L., et al., *Lineage tracking reveals dynamic relationships of T cells in colorectal cancer.* Nature, 2018. **564**(7735): p. 268-272.

34. Guozhu Dong, J.L., *Efficient Mining of Emerging Patterns: Discovering Trends and Differences.* KNO.E.SIS PUBLICATIONS, 8-1999.

35. Agarwal, R., R. Srikant, *Fast algorithms for mining association rules.* in Proc. of the 20th VLDB Conference, 1994.

36. Jain, A., K. Nandakumar, and A. Ross, *Score normalization in multimodal biometric systems.* Pattern recognition, 2005. **38(12): p. 2270-2285.**

37. Vogelstein, B., et al., *Genetic alterations during colorectal-tumor development.* N Engl J Med, 1988. **319**(9): p. 525-32.

38. Hasan, S., et al., *Microsatellite Instability (MSI) as an Independent Predictor of Pathologic Complete Response (PCR) in Locally Advanced Rectal Cancer: A National Cancer Database (NCDB) Analysis.* Ann Surg, 2020. **271**(4): p. 716-723.

39. Le, D.T., et al., *PD-1 Blockade in Tumors with Mismatch-Repair Deficiency.* N Engl J Med, 2015. **372**(26): p. 2509-20.

40. Siegel, R.L., et al., *Cancer statistics, 2022.* CA Cancer J Clin, 2022. **72**(1): p. 7-33.

41. Drake, C.G., E.J. Lipson, and J.R. Brahmer, *Breathing new life into immunotherapy: review of melanoma, lung and kidney cancer.* Nat Rev Clin Oncol, 2014. **11**(1): p. 24-37.

42. Mlecnik, B., et al., *Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability.* Immunity, 2016. **44**(3): p. 698-711.

43. Wei, S.C., C.R. Duffy, and J.P. Allison, *Fundamental Mechanisms of Immune Checkpoint Blockade Therapy.* Cancer Discov, 2018. **8**(9): p. 1069-1086.

44. Dudley, J.C., et al., *Microsatellite Instability as a Biomarker for PD-1 Blockade.* Clin Cancer Res, 2016. **22**(4): p. 813-20.

45. Grasso, C.S., et al., *Genetic Mechanisms of Immune Evasion in Colorectal Cancer.* Cancer Discov, 2018. **8**(6): p. 730-749.

46. Petrelli, F., et al., *Prognostic Survival Associated With Left-Sided vs Right-Sided Colon Cancer: A Systematic Review and Meta-analysis.* JAMA Oncol, 2017. **3**(2): p. 211-219.

47. Beatty, G.L., Y. Li, and K.B. Long, *Cancer immunotherapy: activating innate and adaptive immunity through CD40 agonists.* Expert Rev Anticancer Ther, 2017. **17**(2): p. 175-186.

48. Mittal, D., et al., *CD96 Is an Immune Checkpoint That Regulates CD8(+) T-cell Antitumor Function.* Cancer Immunol Res, 2019. **7**(4): p. 559-571.

49. Panni, R.Z., et al., *Agonism of CD11b reprograms innate immunity to sensitize pancreatic cancer to immunotherapies.* Sci Transl Med, 2019. **11**(499).

50. Galluzzi, L., et al., *Immunological Effects of Conventional Chemotherapy and Targeted Anticancer Agents.* Cancer Cell, 2015. **28**(6): p. 690-714.

51. Bach, D.H., W. Zhang, and A.K. Sood, *Chromosomal Instability in Tumor Initiation and Development.* Cancer Res, 2019. **79**(16): p. 3995-4002.

52. Muller, M.F., A.E. Ibrahim, and M.J. Arends, *Molecular pathological classification of colorectal cancer.* Virchows Arch, 2016. **469**(2): p. 125-34.

53. Nazemalhosseini Mojarad, E., et al., *The CpG island methylator phenotype (CIMP) in colorectal cancer.* Gastroenterol Hepatol Bed Bench, 2013. **6**(3): p. 120-8.

54. Moore, L.D., T. Le, and G. Fan, *DNA methylation and its basic function.* Neuropsychopharmacology, 2013. **38**(1): p. 23-38.

55. Emran, A.A., et al., *Targeting DNA Methylation and EZH2 Activity to Overcome Melanoma Resistance to Immunotherapy.* Trends Immunol, 2019. **40**(4): p. 328-344.

56. Healey Bird, B., et al., *Cancer Immunotherapy with Immune Checkpoint Inhibitors-Biomarkers of Response and Toxicity; Current Limitations and Future Promise.* Diagnostics (Basel), 2022. **12**(1).

57. Oliveira, A.F., L. Bretes, and I. Furtado, *Review of PD-1/PD-L1 Inhibitors in Metastatic dMMR/MSI-H Colorectal Cancer.* Front Oncol, 2019. **9**: p. 396.

58. Andre, T., et al., *Pembrolizumab in Microsatellite-Instability-High Advanced Colorectal Cancer.* N Engl J Med, 2020. **383**(23): p. 2207-2218.

59. Labani-Motlagh, A., M. Ashja-Mahdavi, and A. Loskog, *The Tumor Microenvironment: A Milieu Hindering and Obstructing Antitumor Immune Responses.* Front Immunol, 2020. **11**: p. 940.

60. Otegbeye, E.E., et al., *Immunity, immunotherapy, and rectal cancer: A clinical and translational science review.* Transl Res, 2021. **231**: p. 124-138.

61. Romero-Garcia, S., H. Prado-Garcia, and A. Carlos-Reyes, *Role of DNA Methylation in the Resistance to Therapy in Solid Tumors.* Front Oncol, 2020. **10**: p. 1152.

62. Villanueva, L., D. Alvarez-Errico, and M. Esteller, *The Contribution of Epigenetics to Cancer Immunotherapy.* Trends Immunol, 2020. **41**(8): p. 676-691.

63. Fenaux, P., et al., *Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase III study.* Lancet Oncol, 2009. **10**(3): p. 223-32.

64. Kuipers, E.J., et al., *Colorectal cancer.* Nat Rev Dis Primers, 2015. **1**: p. 15065.

65. Cerami, E., et al., *The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.* Cancer Discov, 2012. **2**(5): p. 401-4.

66. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.

67. Madeira, F., et al., *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic Acids Res, 2019. **47**(W1): p. W636-W641.

68. Farhood, B., M. Najafi, and K. Mortezaee, *CD8(+) cytotoxic T lymphocytes in cancer immunotherapy: A review.* J Cell Physiol, 2019. **234**(6): p. 8509-8521.

69. Fennell, L., et al., *Integrative Genome-Scale DNA Methylation Analysis of a Large and Unselected Cohort Reveals 5 Distinct Subtypes of Colorectal Adenocarcinomas.* Cell Mol Gastroenterol Hepatol, 2019. **8**(2): p. 269-290.

70. Lazarevic, V., L.H. Glimcher, and G.M. Lord, *T-bet: a bridge between innate and adaptive immunity.* Nat Rev Immunol, 2013. **13**(11): p. 777-89.

71. Medvedeva, Y.A., et al., *EpiFactors: a comprehensive database of human epigenetic factors and complexes.* Database (Oxford), 2015. **2015**: p. bav067.

72. Aran, D., et al., *Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage.* Nat Immunol, 2019. **20**(2): p. 163-172.

73. Zhao, L. and Y. Pan, *SSCS: A Stage Supervised Subtyping System for Colorectal Cancer.* Biomedicines, 2021. **9**(12).

74. Smith, J., et al., *Promoter DNA Hypermethylation and Paradoxical Gene Activation.* Trends Cancer, 2020. **6**(5): p. 392-406.

75. Joshi, N.S., et al., *Inflammation directs memory precursor and short-lived effector CD8(+) T cell fates via the graded expression of T-bet transcription factor.* Immunity, 2007. **27**(2): p. 281-95.

76. Yang, R., et al., *Human T-bet Governs Innate and Innate-like Adaptive IFN-gamma Immunity against Mycobacteria.* Cell, 2020. **183**(7): p. 1826-1847 e31.

77. Sen, D.R., et al., *The epigenetic landscape of T cell exhaustion.* Science, 2016. **354**(6316): p. 1165-1169.

78. Beltra, J.C., et al., *Developmental Relationships of Four Exhausted CD8(+) T Cell Subsets Reveals Underlying Transcriptional and Epigenetic Landscape Control Mechanisms.* Immunity, 2020. **52**(5): p. 825-841 e8.

79. Sun, W., et al., *Association between Socioeconomic Status and One-Month Mortality after Surgery in 20 Primary Solid Tumors: a Pan-Cancer Analysis.* J Cancer, 2020. **11**(18): p. 5449-5455.

80. Kaur, P., et al., *Comparison of TCGA and GENIE genomic datasets for the detection of clinically actionable alterations in breast cancer.* Sci Rep, 2019. **9**(1): p. 1482.

81. Wang, C., et al., *Integrative analyses of single-cell transcriptome and regulome using MAESTRO.* Genome Biol, 2020. **21**(1): p. 198.

82. Han, Y., et al., *Comparison of EM-seq and PBAT methylome library methods for low-input DNA.* Epigenetics, 2021: p. 1-10.

83. Huang, K., et al., *Multi-Omics Perspective Reveals the Different Patterns of Tumor Immune Microenvironment Based on Programmed Death Ligand 1 (PD-L1) Expression and Predictor of Responses to Immune Checkpoint Blockade across Pan-Cancer.* Int J Mol Sci, 2021. **22**(10).

84. Wang, H., et al., *A pan-cancer perspective analysis reveals the opposite prognostic significance of CD133 in lower grade glioma and papillary renal cell carcinoma.* Sci Prog, 2021. **104**(2): p. 368504211010938.

85. Chiu, Y.C., et al., *Deep learning of pharmacogenomics resources: moving towards precision oncology.* Brief Bioinform, 2020. **21**(6): p. 2066-2083.

86. Liu, Z. and S. Zhang, *Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features.* BMC Genomics, 2015. **16**: p. 503.

87. Liu, L., et al., *Combination of TMB and CNA Stratifies Prognostic and Predictive Responses to Immunotherapy Across Metastatic Cancer.* Clin Cancer Res, 2019. **25**(24): p. 7413-7423.

88. Lee, M., et al., *Tumor mutational burden as a predictive biomarker for checkpoint inhibitor immunotherapy.* Hum Vaccin Immunother, 2020. **16**(1): p. 112-115.

VITA

Yuanyuan Shen was a graduate student of the MU Data Science and Informatics Institute. An ultimate goal of her proposed research is to integrate data-driven biomedical informatics research with traditional research to form an efficiently scientific mode and improve cancer patients' life qualities and overall survivals. She has the expertise, experience, and motivation necessary to successfully carry out the proposed work.

Yuanyuan had started my traditional research since when she was in medical school in China. She spent all my summer and winter break to train my experiment skills and scientific thinking in a top, leading tumor biology lab of Chongqing Medical University. After Yuanyuan graduated from medical school, she chose to pursue a residency program in Surgery training and Breast cancer - reconstruction, plastics as her clinical and research area. Even she was buried by the busy surgical residency schedule, she still published three publications. There were two of them highly related our proposed research, and another one is surgical skills meta-analysis. Within the invitation from Yingqun Lab, Yale School of Medicine, she officially started her full-time tumor biology/glucose metabolism research work after she finished her residency and clinical fellow work in China. During the two years of training in an Ivy League lab, she participated in three projects, and there are two of them have been published. As a coauthor, the paper "The Steroidogenic Acute Regulatory Protein (StAR) Is Regulated by the H19/let-7 Axis." was published on Endocrinology nominated as one of "Top Endocrine Discoveries of 2017" by the Endocrine News.

However, just focusing on limited research samples in a project isn't her ultimate goal. Therefore, she started biomedical informatics Ph.D. study at the University of Missouri-Columbia from August, 2017. She conducted data mining skills on hundreds of thousands to millions of patients’ genomics and clinical information which gave a full picture and research clues on her projects. Her research work during the Ph.D. continued to serve the purpose of implementing informatics methods in exploring the solution and mechanisms to improve the sensitivities of immune checkpoint inhibitors treatment for pan-cancer patients. Yuanyuan recently accepted an offer and join Washington University in St.Louis as a postdoctoral research associate.