UNIVERSITY OF CALIFORNIA SAN DIEGO

# Tumor Associated Macrophages and Colorectal Cancer: AI-Assisted Predictive Modeling of Macrophage Polarization in Colorectal Cancer

A Thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Bioengineering

by

Ekta Dadlani

Committee in charge:

Professor Debashis Sahoo, Chair
Professor Shankar Subramaniam, Co-Chair
Professor Andrew McCulloch

2023

The Thesis of Ekta Dadlani is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| TAMs | Tumor Associated Macrophages |
| CRC | Colorectal Cancer/Colorectal Cancer Tissue |
| NT | Normal Colon Tissue |
| p-val/p | p-value |
| SMaRT | Signature of Macrophage Reactivity and Tolerance |
| ROC-AUC | Receiver Operating Characteristic Area Under the Curve |
| TPR | True Positive Rate |
| FPR | False Positive Rate |
| ISGs | Interferon-stimulated Genes |
| scRNAseq | Single Cell RNA sequencing |
| TPM | Transcripts Per Million |
| RPKM | Reads Per Kilobase of Transcript per Million |
| FPKM | Fragments Per Kilobase of Transcript per Million |
| CPM | Counts Per Million |
| BoNE | Boolean Network Explorer |
| BIR | Boolean Implication Relationship |
| BIN | Boolean Implication Network |
| CBIN | Clustered Boolean Implication Network |
| MMR | Mismatch Repair System Deficiency |
| MSI | Microsatellite Instability |
| MSS | Microsatellite Stable |

| | |
|---|---|
| NCBI GEO | National Center for Biotechnology Information Gene Expression Omnibus |
| BECC | Boolean Equivalent Correlated Clusters |
| FAP | Familial Adenomatous Polyposis |
| NSAIDS | Non-Steroidal Anti-Inflammatory Drugs |
| LAMs | Lipid-Associated Macrophages |
| SAMs | Scar-Associated Macrophages |
| DAMs | Disease-Associated Microglia |
| TME | Tumor Microenvironment |
| M0 | Unstimulated Macrophage |
| M1 | Reactive Macrophage (pro-inflammatory) |
| M2 | Tolerant Macrophage (anti-inflammatory) |
| CIMP+ | CpG Island Methylator Phenotype (positive status) |
| CIMP- | CpG Island Methylator Phenotype (negative status) |
| RSCC | Right-Side Colon Cancers (proximal tumors) |
| LSCC | Left-Side Colon Cancers (distal tumors) |
| DEG | Differentially Expressed Gene |

ACKNOWLEDGEMENTS

I am profoundly grateful for my professor and chair of my committee, Dr. Debashis Sahoo, for his continued mentorship, patience, gratitude, and feedback. His unwavering support has been extremely instrumental in my research endeavors. I would like to extend my appreciation to my defense committee, who generously provided support and constructive insights during this undertaking. I express my gratitude to the National Institutes of Health (NIH) for financially supporting my research; their investment not only played a pivotal role in my research but also deemed a potential impact in the scientific community.

I could not have undertaken this journey without the guidance, encouragement, and inspiration from my colleagues, especially Dr. Tirtharaj Dash and Daniella Vo.

Lastly, I would like to acknowledge and extend a special thank you to my family, friends, and dog (Buddy) for motivating me during this process and keeping my spirits high during the challenges encountered along the way.

Chapters 2, in part, is currently being prepared for submission for publication of the material. Dadlani, Ekta; Dash, Tirtharaj; Sahoo, Debashis. The thesis author is the primary researcher and author of this material.

Chapters 3, in full, is currently being prepared for submission for publication of the material. Dadlani, Ekta; Dash, Tirtharaj; Sahoo, Debashis. The thesis author is the primary researcher and author of this material.

Chapter 4, in full, is currently being prepared for submission for publication of the material. Dadlani, Ekta; Dash, Tirtharaj; Sahoo, Debashis. The thesis author is the primary researcher and author of this material.

Chapter 5, in full, is currently being prepared for submission for publication of the material. Dadlani, Ekta; Dash, Tirtharaj; Sahoo, Debashis. The thesis author is the primary researcher and author of this material.

ABSTRACT OF THE THESIS

Tumor Associated Macrophages and Colorectal Cancer: AI-assisted Predictive Modeling of Macrophage Polarization in Colorectal Cancer

by

Ekta Dadlani

Master of Science in Bioengineering

University of California San Diego, 2023

Professor Debashis Sahoo, Chair
Professor Shankar Subramaniam, Co-Chair

Colorectal cancer (CRC) is one of the leading malignant diseases in the United States, predominantly due to its poor prognosis and high metastasis. Tumor-associated macrophages (TAMs) are amongst the most common cells that play a significant role in cancer survival and

progression in the tumor microenvironment. By using single-cell CRC-specific RNAseq datasets and computational approaches developed in-house, I aim to answer two specific scientific questions: (Q1) Do TAMs show distinctive signature in CRC samples in contrast with healthy samples?; and (Q2) Can I relate the pro-inflammatory and anti-inflammatory polarization of TAMs to the prognosis of colorectal cancer? I filter macrophage cells from eight publicly-available single-cell CRC-specific RNAseq datasets, obtained from both human (Homo Sapiens) and mouse (Mus Musculus) samples and refine a computational model, called SMaRT, to identify a distinctive signature for accurately predicting macrophage-polarization states in the specialized context of CRC. The computational analysis suggests: (a) TAMs are consistently more reactive in tumorous cells as compared to the healthy cells and that the separation between their source samples is statistically significant; (b) A TAM-specific composite gene signature can be reliably used to separate samples that are cancerous versus samples that are healthy. Specifically, these findings provide sufficient and statistically significance evidence that TAMs have a distinctively different signature in CRC samples, majorly falling in the spectrum of the immuno-reactive polarization state.

# INTRODUCTION

Colorectal cancer (CRC), the third most common cancer worldwide and representing 10% of all cancers, consists of abnormal proliferation of cells within the colon or rectum of the large intestine [1, 2]. There were an estimated 1.9 million cases of newly diagnosed colorectal cancer and 0.9 million deaths globally in the year 2020 [3]. Factors that induce the initiation of the cancer include changes in age, environment, and lifestyle, all of which cause genetic and epigenetic alterations of the cells in the gut, including mutations that affect tumor suppressor genes, oncogenes, and DNA repair mechanisms [4]. The dysregulation of signaling pathways resulting from these alterations contribute to tumor onset and progression. Risk factors for colorectal cancer include obesity, smoking, and low physical activity [1]. The first line of treatments for colorectal cancer include surgical treatments (i.e. endoscopic treatment), chemotherapy, immunotherapy, and radiotherapy. For advanced-stage CRC, surgical resection involving lymph node dissection is often necessary [5]. Despite these treatment advances, colorectal cancer is the second leading cause of cancer-related deaths worldwide [1]. As a result, researchers all around the world are interested in further investigating the mechanisms and factors that underlie the cancer's progression.

The initiation and progression of colorectal cancer involves genetic alterations in both cancer cells and cells in the surrounding tumor microenvironment. The tumor microenvironment comprises of immune and stromal cells such as macrophages, dendritic cells, T lymphocytes, and natural killer cells, all of which can contribute to chronic inflammation [4]. Clinical studies have demonstrated that the long-term use of non-steroidal anti-inflammatory drugs (NSAIDS) can

reduce the risk of CRC by counteracting the inflammation that results from the tumor initiation [6].

The gastrointestinal tract houses the largest population of macrophages; tumor associated macrophages are the predominant immune cells in the tumor microenvironment in CRC [7]. Tumor associated macrophages (TAMs), derived from blood monocytes, are attracted to the tumor site by the activity of growth factors and chemokines within the tumor microenvironment. It is implicated that tumor associated macrophages in the tumor microenvironment support the progression of normal colonic epithelium to adenomatous polyps, or the gland-like growths on the large intestine's membrane, to invasive colon carcinoma [8].

Macrophages exhibit distinctive functions in response to the stimuli in their environment; they are traditionally classified as M0 macrophages (unstimulated, undifferentiated), M1 macrophages (reactive, anti-tumor), and M2 macrophages (tolerant, pro-tumor). M0 macrophages arise from circulating white blood cells (monocytes) that use chemokines and adhesive molecules that reside on the endothelium of blood vessels to migrate to a site of inflammation or infection [9, 10]. Unstimulated macrophages possess the ability to adopt a wide range of functions in response to specific stimuli in their microenvironment. M1 macrophages exhibit pro-inflammatory properties in response to the release of inflammatory cytokines and the presence of oxygen species within the microenvironment; M2 macrophages are characterized by their anti-inflammatory properties which emerge in response to cytokines produced by immune cells, growth factors released during wound healing, and metabolic changes [11, 12].

It is conventionally understood that the stimuli in a microenvironment can activate M0 macrophages to adopt a tolerant or reactive-like state based on the extreme functions of macrophages at these polarization states; however, this simplistic nomenclature overlooks the full extent of macrophage plasticity and the continuum of polarization states that macrophages adopt under steady-state conditions and during disease. While specialized macrophage sub-types, such as lipid-associated macrophages (LAMs) in atherosclerosis, scar-associated macrophages (SAMs) in liver fibrosis, and tumor-associated macrophages (TAMs), have distinct spatial localization, origin, and functional pathways, the traditional definition of reactive (M1) and tolerant (M2) macrophages do not possess reliable predictive or prognostic abilities [13]. This indicates that the current definitions of macrophage polarization states is insufficient to fully capture the complex dynamics and clinical implications of macrophage phenotype in various contexts.

Consequently, I employ the comprehensive Signature of Macrophage Reactivity and Tolerance (SMaRT) model, which offers a general framework and standardized definitions for the continuum of macrophage polarization states, to capture the immunophenotype of macrophages in colorectal cancer and provide valuable insights into their functional characteristics [13]. Built using a Boolean implication network trained on a pooled human macrophage transcriptome dataset, the SMaRT model contains a 338 gene-signature of universal, unbiased biomarkers for the spectra of macrophage polarization states that is represented across relevant tissues, organs, species, and immune cells [13]. Despite the model's ability to precisely capture the evolving macrophage cellular states in a diverse context, I seek to create a TAMs specific gene-signature that specifically captures the physiologic and pathologic spectra of

“immuno-reactivity” and “immuno-tolerance” in colorectal cancer. A refined signature prevents over fitting of the SMaRT gene signature to the specialized context of colorectal cancer; specific feature selection for TAMs allows for the classifications of normal colon tissue and colorectal cancer tissue samples while reducing the dimensionality of the dataset and countering the over fitting of the universal macrophage signature. The refined signature might inspire formal definitions for macrophage polarization states that maintain relevance and rationalize diagnostics and therapeutics to target deranged macrophage states in colorectal cancer.

# CHAPTER 1: LITERATURE REVIEW

In this chapter, I provide relevant background information for the premise of the project. Topics include: colorectal cancer, macrophages and macrophage polarization states, tumor microenvironment and tumor associated macrophages, tumor associated macrophages as potential prognostic biomarkers for colorectal cancer, and refinement of gene signatures in the Signature of Macrophage Reactivity and Tolerance (SMaRT) model.

## *Colorectal Cancer*

Colorectal cancer (CRC) is a type of cancer that primarily impacts the colon or the rectum of the large intestine. Globally, it is the second leading cause of cancer-related deaths and the third most common type of cancer amongst both men and women across various populations [14]. In 2020, there were an estimated of 1.9 million cases of newly diagnosed colorectal cancer and approximately 935,000 reported deaths [15]. As a result, colorectal cancer has been mandated as a critical public health concern with high mortality and incidence rates. Despite significant progress in the screening, diagnosis, and treatment options for colorectal cancer, there is a necessity for ongoing research and the development of therapies which can better predict treatment response and enhance patient outcomes. This highlights the interest in identifying biomarkers which can guide personalized treatment strategies.

Risk factors of colorectal cancer include increasing age, a high body mass index resulting from a poor diet (particularly a diet high in red and processed meats) and physical inactivity, inflammatory bowel disease, excessive alcohol consumption, smoking, and a genetic family history of colorectal cancer, the Lynch syndrome, or familial adenomatous polyposis (FAP) [15–

18]. About 91% of the CRC cases occur in individuals over the age of 50; there is also disparity in the incidence of CRC based on ethnicity, with African Americans and Non-Hispanic American Indians having the highest incidence rate compared to all other ethnic groups in the United States [15]. Screening tests, including colonoscopies, fecal occult blood tests, flexible sigmoidoscopy, and stool DNA tests, play a vital in the early detection and treatment of colorectal cancer; they can be used to identify and remove the precancerous growths, or adenomatous polyps. Early stage detection is associated with a significant improvement in the 5-year survival rate, with a survival rate of 90%; however, once the cancer progresses and metastasizes, the prognosis significantly worsens [19].

In most cases, genetic and epigenetic mutations of the colonic epithelial cells transform benign growths in the colon, called adenomatous polyps, to the malignant invasive adenocarcinoma form. During the early stages of adenomatous polyp development, mutations in the APC (adenomatous polyposis coil) tumor suppressor gene are known to be key events that initiate tumor onset and growth. These mutations result in the activation of the Wnt signaling pathway and the nuclear accumulation of β-catenin in malignant colon cells. The overexpression of specific genes within the Wnt-signaling pathway, including *c-Myc* and *Cyclin D1*, promote colorectal cancer cell proliferation; additional mutations in critical genes involved in the Wnt signaling pathway, such as *CTNNB1* and *AXIN1*, further contribute to the formation of the invasive cancer during the growth of the polyps [18]. Mutations of genes, such as *KRAS* and *TP53*, emerge later in the progression of colorectal adenoma and contribute to the development of invasive carcinoma. Mutations of *KRAS* lead to the activation of the MAPK pathway, which promotes tumor cell proliferation and survival. Mutations of *TP53* causes the loss of function of

the tumor suppressor protein p53, which is involved in apoptosis, DNA repair, and regulation of the cell cycle [20]. In addition to genetic mutations, environmental factors, methylation changes of the CpG islands, and chronic inflammation implicate the acquisition of colorectal cancer from adenomatous polyps.

Treatment options for colorectal cancer typically encompasses surgery, chemotherapy, and radiation therapy; these treatments modalities can be administered concurrently based on the stage and location of the cancer and the age and the overall health of the patient. Colectomy, the primary treatment choice for early stage CRC, involves the surgical removal of the affected colon and its surrounding lymph nodes. For rectal cancer, standard surgical treatments include low anterior resection, in which the affected rectum tissue is excised and the remaining rectum is reconnected to the colon, and abdominoperineal resection, in which the entire rectum, anus, and surrounding lymph nodes are removed [5]. Chemotherapy treatments, such as adjuvant therapy (i.e: cetuximab and bevacizumab) and palliative therapy, can be employed to reduce the risk of cancer recurrence, diminish tumor size prior to surgical procedures, or alleviate symptoms in advanced stages of CRC. Chemotherapy drugs, such as 5-fluorouracil (5-FU) and irinotecan, can be used in combination with other treatment modalities to improve the outcome of the procedure [21–23]. Under radiation therapy, an external or internal beam of high-energy radiation is directed to kill targeted cancer cells. The treatment has the potential to reduce the local recurrence of the cancer and improve the overall health of the patient; however, potential side effects of this treatment option include fatigue, skin irritation, diarrhea, fibrosis, scarring, and the emergence of secondary cancers in the bladder and sarcoma [24, 25]. Despite the existing treatment options and their abilities to specifically target cancer cells whilst causing minimal

damage to the surrounding healthy cells, heterogeneity of colorectal cancer across patients has hindered a standard clinical implementation of universal targeted immunotherapy treatment. Therefore, there is an urgency for the discovery of better biomarkers that can identify patients who would likely benefit from specific types of treatment options through continued research of the underlying genetic and molecular mechanisms of the cancer's development.

*Macrophages And Macrophage Polarization States*

Macrophages, discovered in 1883 by Élie Metchnikoff, are a specialized type of white blood cell responsible for detecting and engulfing cellular waste, foreign substances, and pathogens [12]. By employing phagocytic receptors that recognize conserved motifs, macrophages exhibit the capacity to discern pathogens and facilitate their engulfment, triggering an adaptive immune response. They originate from monocytes in the blood circulation that travel to and differentiate in different tissues located in the bone marrow, blood, lymphoid, and other non-hematopoietic tissues [9, 10]. Upon reaching local sites of injury or infections, the immune cells can contribute to acute or chronic inflammation, antimicrobial inhibitory activities, and tissue repair [26]. Resident macrophages reside in organs in the absence of inflammation and are involved in the homeostatic role of removing apoptotic cells at sites of injury or infection. Recent advances in the study of macrophage origin studies have revealed that resident tissue macrophages can also be established during embryonic development and maintained throughout adulthood without blood monocyte input under steady state conditions [27].

Unstimulated macrophages (M0) can adopt a wide range of specialized functions as a result of the stimuli in the extracellular environment; the states that identify these functions are

classically defined as the M1 and M2 continuum of macrophage polarization. M1-polarized macrophages exhibit pro-inflammatory and reactive properties that contribute to an anti-tumor response through their involvement in phagocytic and cytotoxic functions. They can recruit and sustain the activation of other immune cells, such as B cells and T cells, at a site of infection. M2-polarized macrophages are characterized as anti-inflammatory, tolerant, and pro-tumor. The balance between anti-tumor M1 macrophages and pro-tumor M2 macrophages is crucial for regulating a tumor's growth and survival [11, 12]. Despite this classic nomenclature of polarization states, macrophages exist as a spectrum of states; macrophages are able to adopt intermediate phenotype and perform diverse physiological roles. Defining the mechanism by which macrophage phenotype are defined along this gradient of polarization states continues to pose a challenge in the field. The conventional nomenclature used across macrophage analysis classifies the M1 and M2 phenotype as the endpoints of the continuum; however, this approach poses experimental challenges when assessing in vitro-generated macrophages [12].

The morphology of the macrophages, resulting from the monocyte precursors and their differentiation process, enables the immune cell to recognize various pathogens and secrete different levels of inflammatory cytokines under specialized contexts. The capacity of macrophages to induce endocytosis, phagocytosis, and the secretion of cytokines, growth factors, and metabolites within different contexts allow the immune cells to carry out trophic and toxic functions (i.e: tissue remodeling and adaptive immunity) during development and adulthood [28]. The differentiation and activation of macrophages are influence by the presence of growth factors, signaling pathways, and transcription factors within their specialized micro environments. Macrophage sub-types include lipid-associated macrophages in atherosclerosis

(LAMs), disease-associated microglia in neurodegenerative disorders (DAMs), scar-associated macrophages in liver fibrosis (SAMs), and tumor-associated macrophages (TAMs) [13].

*Tumor Microenvironment (TME) and Tumor Associated Macrophages (TAMs)*

The tumor microenvironment (TME), consisting of tumor cells, stromal cells, immune cells, and extracellular matrix (ECM) components, is a vital ecosystem that supports tumor growth, survival, and proliferation. The components that make up the TME engage in intercellular interactions to influence the behavior of a tumor. The tumor microenvironment can facilitate tumor growth and invasion by promoting the remodeling of the ECM via degradation caused by enzymes secreted by cancer and stromal cells [29]. Angiogenesis, or a cancer cell's ability to form new blood vessels, in the tumor microenvironment, enables a tumor to receive the nutrients and oxygen required for growth and supports the potential for metastasis at distant sites [29]. Paracrine signaling within the tumor microenvironment facilitates the communication between cancer cells, stromal cells, and immune cells, leading to the secretion of growth factors and cytokines that promote tumor growth and survival [29]. Immune cells are influenced to promote or inhibit tumor growth based on the local cytokine environment; cancer cells in the TME can release cytokines to suppress and evade the immune cells at the site of a tumor, thereby supporting tumor proliferation [30].

Tumor-associated macrophages (TAMs) are one of the key immune cell types in the TME that can contribute to tumor growth and progression. These specialized type of white blood cells are recruited by tumor cells into the TME through the release of tumor-derived chemokines and growth factors. Upon recruitment, TAMs have the potential to support tumor cell invasion,

migration, and proliferation [12]. The local cytokine environment influences the macrophages to polarize towards the M1 or M2 phenotype. TAMs can secrete growth factors and cytokines, such as the vascular endothelial growth factor (VEGF), interleukin-10 (IL-10), and transforming growth factor-beta (TGF-β), which promote angiogenesis, suppress the immune response, remodel the ECM, and induce an epithelial-mesenchymal transition [28, 29, 31]. Inhibiting the tumor-supporting TAMs to balance the presence of M1-like and M2-like TAMs in the TME can be used to maintain homeostasis; inhibition can disrupt angiogenesis, reduce the severity of ECM remodeling, and propagate the immune system to fight the tumor growth. As a result, TAMs in the TME have emerged as a compelling target for cancer therapy.

*Tumor Associated Macrophages as Potential Prognostic Biomarkers for Colorectal Cancer*

Despite the recent advances in treatment options, the heterogeneity in the survival rates of patients with advanced colorectal cancer marks an urgency for the discovery of prognostic markers that could tailor therapeutic options for CRC stratification. Tumor-associated macrophages (TAMs), which have been associated with different outcomes for a diverse range of cancers, have the potential to serve as prognostic biomarkers for diagnosed colorectal cancer patients despite the differences in the tumor onset and growth. It is known that the abundance of TAMs in the TME of CRC patients is associated with an increase in the tumor progression and metastasis, leading to a reduced survival rates [32, 33]. TAMs and other inflammatory cells are responsible for producing cytokines, growth factors, and chemokines that promote tumor growth and angiogenesis; early detection of CRC includes the risk of chronic inflammation [34]. TAMs also have a crucial role in regulating immune evasion. Therefore, the quantification of annotated

TAMs in CRC tumors and analyzing the expression of gene markers specific to the M1-type macrophages would provide useful information in predicting the disease prognosis.

Receptors involved in the interaction between TAMs and CRC cells include the colony-stimulating factor 1 (CSF-1) receptor, the epidermal growth factor receptor (EGFR), and the transforming growth factor-beta (TGF-β) pathway [32, 34]. In the CSF-1 receptor pathway, the CSF-1 receptor expressed on the surface of TAMs binds to its ligand, CSF-1, which is released by CRC cells or other cells in the TME; this interaction recruits TAMs to the tumor site and contributes to maintaining the pro-inflammatory properties associated with M1 macrophage function [11, 31, 32]. Activation of EGFR signaling promotes the production of the pro-inflammatory cytokines that support the M1 polarization state; overexpression of EGFR in CRC cells can lead to the secretion of factors that recruit and activate TAMs. Macrophages can also secrete EGFR ligands; the secretion leads to a positive feedback loop that further activates the EGFR signaling in CRC cells and promote tumor growth and metastasis [35]. TAMs can also secrete transforming growth factor-β (TGF-β), which promotes the migration and invasion of CRC cells and facilitates the induction of a M2-like phenotype [36]. Inhibition of the TGF-β pathway can enhance the pro-inflammatory responses the contribute to a M1 phenotype [11]. Targeting genes involved with these receptor pathways can block the activity of the receptor or disrupt the pathway.

*Tumor Associated Macrophages as Potential Prognostic Biomarkers for Colorectal Cancer*

The Signature of Macrophage Reactivity and Tolerance (SMaRT) model is a general computational framework for analyzing relationships among genes and different macrophage

polarization states [13]. A concern with the current model is on specializing this general framework to be well-suited for tumor-associated macrophages in colorectal cancers. One way to do this is to independently specialize each cluster in the Boolean network associated with SMaRT, especially, the most prominent transitive path (Path C13-C14-C3; refer to section *Methods: Computational Approaches: Signature of Macrophage Reactivity and Tolerance* for more details) to obtain a TAM-specific signature for predicting colorectal cancer diagnosis. I term this approach of specializing a general model as "refinement" and "purification". Evidently, refinement allows the number of genes in the Path C13-C14-C3 to reduce drastically and construct a context-specific gene-signature for colorectal cancer. This approach is akin to regularization in machine learning (reducing the number of free parameters in a model to control over fitting) and transfer learning (fine-tuning a pre-trained machine learning model towards learning on a related-yet-different problem) [37, 38].

# CHAPTER 2: PRELIMINARY

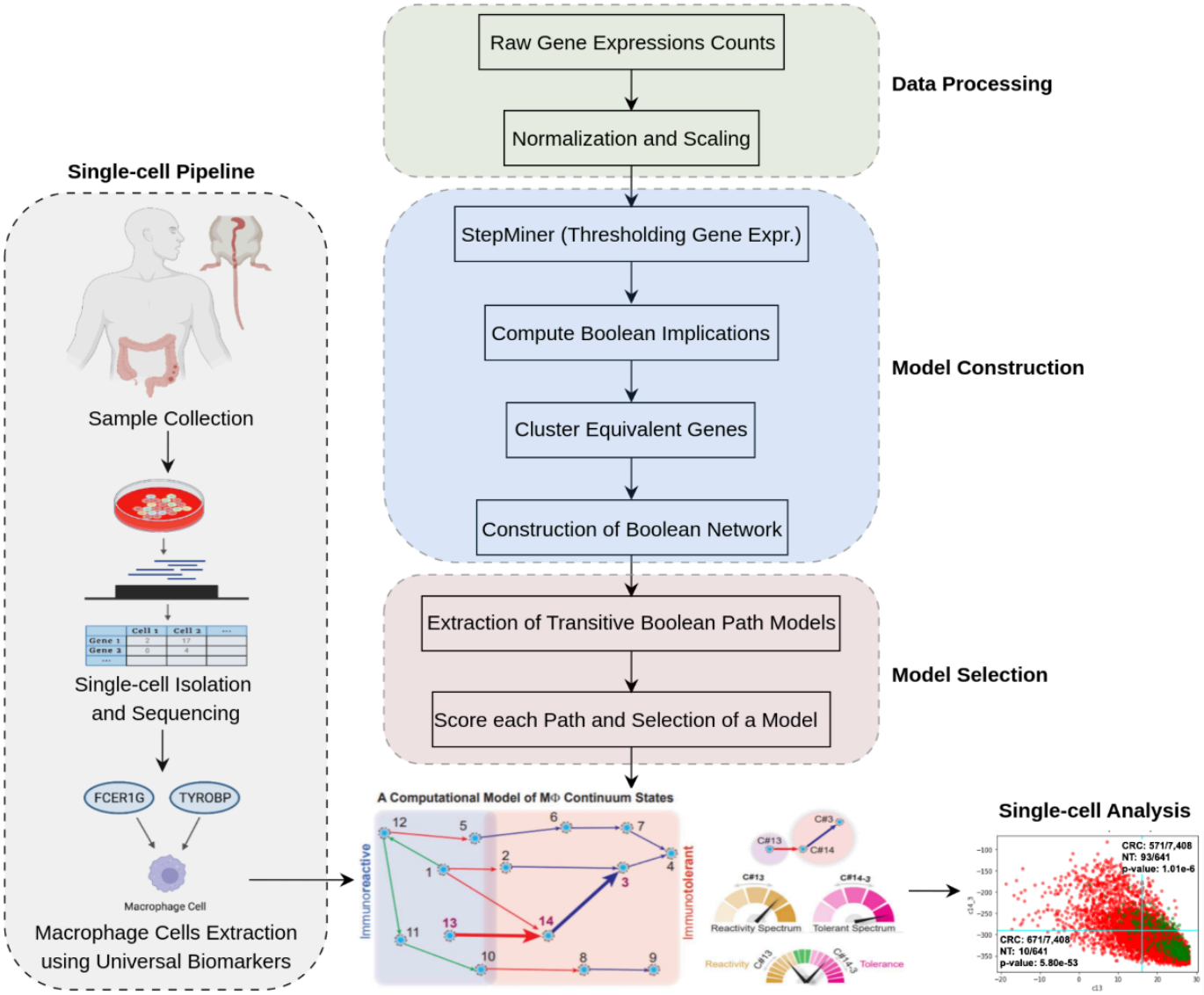
Boolean patterns in high-dimensional datasets pave the way for leveraging a diverse range of datasets with different experimental origins and conditions. In combination with machine learning networks, they allow us to extract biological principles and impactful features (correlations across gene expression patterns) that remain invariant across datasets with different variables. This allows for the identification of clinically relevant biomarkers that underlie the complex processes that differentiate normal and cancerous tissues. In order to capture the dynamics of macrophage polarization using gene expression datasets, I follow a step-wise computational pipeline. Minimally, these steps involved are summarized as follows, and described in detail in subsequent subsections (see Figure 2.1).

1. Data Processing: The raw gene expression counts are normalized and scaled, if not done already; however, this step is not required if the available gene expression counts are normalized and scaled.

2. Network construction: In this step, a global computational map representing various relationships among genes is constructed using the expression values across samples.

3. Model selection: This is the step where the underlying dynamics of the macrophage polarization is captured using the global computational map constructed in step (2) and using an evaluation metric.

Figure 2.1: Overview of Computational Approach to Differentiate Normal Colon Tissue Samples and Colorectal Cancer Tissue Samples at the Single-Cell Level. *Single Cell Pipeline*: The single-cell RNA sequencing datasets used for the analysis originate from human or mouse colon tissues annotated as colorectal cancer tissue or adjacent normal colon tissue. In general, viable, single cells from the targeted tissue are first isolated and lysed using poly[T]-primers to capture the mRNA molecules (usage of a primer prevents capturing ribosomal RNA). Reverse transcriptase alongside the addition of unique molecular identifiers (UMIs) is used to convert the poly[T]-primed mRNA to complimentary DNA (cDNA). Each tagged cDNA molecule is amplified by PCR or in vitro transcription, pooled, and sequenced using next generation sequencing (NGS) library preparation techniques. Preliminary bioinformatics tools are used to perform quality control, resulting in a single cell gene expression count matrix. The gene expression count matrices used for the computational analysis are taken from the NCBI GEO database. I normalize (CPM) and scale (log2) the gene expression count matrix and extract the macrophage cells by applying an expression threshold for the universal macrophage biomarkers *TYROBP* and *FCER1G*. Applying the SMaRT C13-C14-C3 signature on the macrophage-purified single cell RNA sequencing datasets allows for preliminary analysis of the difference in polarization states in colorectal cancer and normal colon tissue macrophage cells. Differential gene expression analysis is used for refinement, where the refined signature aims to capture the pattern of macrophage polarization observed at the single-cell level using the SMaRT C13-C14-C3 signature. *Pipeline of Developing the Underlying Signature of Macrophage Reactivity and Tolerance (SMaRT) Model*: The SMaRT model used to underlie this research is built using a pooled human macrophage annotated dataset (GSE134312, n = 197) from the NCBI GEO database. The gene expression is normalized (TPM) and scaled (log2). The StepMiner algorithm is applied on each gene to generate a statistically significant threshold separating low/high expression. The six Boolean implications are used to generate the initial Boolean Implication Network (BIN); equivalent relationships and annotations of macrophage polarization are used to generate the compact Clustered Boolean Implication Network (CBIN). Transitive paths within the CBIN are extracted using Boolean paths and depth-first traversal (DFS); a score is applied to each specified path. Validation of separation of polarization states using the composite scores of the clusters from specific paths across an invariant collection of macrophage annotated datasets (using the ROC-AUC metric to quantify the predictive power of gene signatures derived from the transitive Boolean paths) supports the accuracy and consistency of the C13-C14-C3 gene signature. As a result, the genes in the C13-C14-C3 transitive Boolean path make up an informative, universal macrophage gene signature.


Single-cell Pipeline
Sample Collection
Single-cell Isolation and Sequencing
FCER1G
TYROBP
Macrophage Cell
Macrophage Cells Extraction using Universal Biomarkers
Raw Gene Expressions Counts
Normalization and Scaling
Data Processing
StepMiner (Thresholding Gene Expr.)
Compute Boolean Implications
Cluster Equivalent Genes
Construction of Boolean Network
Model Construction
Extraction of Transitive Boolean Path Models
Score each Path and Selection of a Model
Model Selection
A Computational Model of MΦ Continuum States
Immunoreactive
Immunotolerant
Reactivity Spectrum
Tolerant Spectrum
Single-cell Analysis
CRC: 571/7,408
NT: 93/641
p-value: 1.01e-6
CRC: 671/7,408
NT: 10/641
p-value: 5.80e-53

*Data Processing: Normalization And Scaling*

Publicly available high throughput sequence data in the format of RNA mRNA profiling (RNA sequencing, microarray, scRNA sequencing), small RNA profiling (miRNA sequencing), ChIP sequencing, HiC-sequencing, methyl-sequencing, and bisulfite-sequencing is available on the National Center of Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. This model relies on mRNA profiling datasets, where prepared mRNA samples undergo fragmentation, conversion to cDNA using reverse transcription, and library creation for sequencing. The reads of cDNA fragments are mapped into a reference genome to determine the composition of the library. Normalization and scaling of the gene expression data enables the assessment of the relative abundance of a genomic feature without the bias of individual read counts that are influenced by factors such as the sequencing depth, feature length, and transcriptome composition [39]. Inter-sample normalization and scaling allows for comparable read counts between samples and experiments [40]. Normalization of the gene expression values is performed using Counts Per Million reads mapped (CPM); however, some datasets provide pre-processed gene expression matrices that are normalized using Reads Per Kilobase of Transcript per Million reads mapped (RPKM), Fragments Per Kilobase of Transcript per Million reads mapped (FPKM), or Transcripts Per Million reads mapped (TPM). Scaling is performed to model the proportional change in gene expression across a sample. I apply log2 scaling on the normalized gene expression values. The final gene expression values are log2(CPM) if CPM > 1 or log2(CPM - 1) if CPM $< 1$.

*Network Construction: StepMiner Analysis*

StepMiner is a tool used primarily to identify step-wise transitions in a time-series data. It directly addresses the questions: “What is the impact of a stimulus on the regulation of gene expression?” and “At what timestamp does a gene undergo a transition into an up regulated or down-regulated state?” [41]. StepMiner employs a regression approach to detect the most distinct transitions between low and high gene expression values, utilizing the sum-of-squares error as a quantitative measure; this provides insight into the temporal dynamics associated with gene expression-switching events. The step function is evaluated at all possible step positions with the objective to fit a one- or two-step function that optimally fits a set of n time points. An average of the gene expression values is calculated at each step position and linear regression is used to determine the values of the constant segments. The square error is computed for each fitted values corresponding to each step; the regression scheme selects the step positions that minimize this square error. A regression test statistic and its corresponding p-value is computed to find the best fit for the curve of the data.

Extracting networks of relationships from public-domain gene expression data has traditionally been rooted in pairwise relationships between genes (symmetric co-expression of genes). A wider range of relationships between gene pairs can be taken into consideration using Boolean implications [42]. The expression values for each gene are sorted in ascending order. A StepMiner threshold is used to fit a rising step function to the data, aiming to minimize the discrepancy between the fitted values and the measured values. The step function is applied to identify the most significant jump from low gene expression values to high gene expression

values; the threshold is placed where the step intersects the original data. In cases where the gene expression levels are evenly distributed from low to high, the threshold tends to be located near the mean expression level. When the gene expression levels are unevenly distributed, relying on the mean expression level can introduce a bias that fails to account for variance parameters. If the assigned StepMiner threshold for a gene is t, the expression levels above t + 0.5 are classified as “high” and the expression levels below t - 0.5 are classified as “low”. The “intermediate” expression values between t - 0.5 and t + 0.5 are ignored, because they are more likely to appear on the wrong side of the threshold due to noise. The interval width of 0.5 is based on the estimated minimum noise in gene expression of a gene whose standard deviation is at the 5th percentile from the bottom. When a minimum of 2/3 of the expression values of a gene are classified as “intermediate”, the gene is removed from consideration for further analysis due to the lack of dynamic range in expression values. After all significant gene expression values are extracted by using the StepMiner algorithm, all pairs of features can be analyzed using Boolean implications.

*Network Construction: Boolean Logic and Implications*

Traditional gene expression networks rely on pairwise relationships, indicating symmetrical co-expression between the genes. A larger set of relationships between gene pairs can be extracted using Boolean implications. Boolean logic encapsulates the simplistic, yet fundamental, mathematical relationship between two values which can be represented as binary states, such as 0/1, negative/positive, or low expression/high expression. After the normalized and log-scaled gene expression values are sorted from low to high, the StepMiner algorithm is

used to fit a rising step function to the series and identify a threshold for each gene; it characterizes the expression values of each gene as “low” and “high”. The thresholds for gene A and gene B are used to separate a scatter plot into four quadrants based on the Boolean values (low, low), (low, high), (high, low), and (high, high). Boolean analysis, or a statistical approach to create binary logical inferences, can be used to determine the relationship between the expression values of pairs of genes.

Implications are derived from an “if-then” rule, where a statement like “if the expression of gene A is high, then the expression of gene B is almost always low” establishes a logical relationship. In a concise format, these relationships are written as “A high implies B low”, or “A high => B low”. There are six possible Boolean implications between genes A and B: four asymmetrical implications (A low => B low, A low => B high, A high => B low, B high => A high) and two symmetrical implications (equivalent, opposite). Asymmetrical Boolean implications suggest that the statement “A high => B high” may be valid without the reverse statement “B high => A high” holding true. When both of the relationships “A high => B high” and “B high => A high” are observed, the symmetrical equivalent Boolean relationship, corresponding to positively correlated genes, is indicated. The symmetrical opposite Boolean relationship is indicated when both of the relationships “A high => B low” and “B high => A low” is observed, corresponding to highly negatively correlated genes.

The significance of a Boolean implication relationship is determined using BooleanNet statistics, which quantifies the sparsity of each quadrant after applying a StepMiner thresholds for gene A and gene B [43]. Boolean implications between gene expression values are observed

when there is sparsity in any of the four possible quadrants or in two diagonally opposite quadrants. The statistical metric to determine significance utilizes the number of samples in each quadrant, represented as $a_{00}$, $a_{01}$, $a_{10}$ and $a_{11}$.

The total number of samples is computed using the equation:

$$\text{total} = a_{00}+a_{01}+a_{10}+a_{11}$$

The total number of samples and the corresponding probabilities that are considered as A low and B low are computed using the equations:

$$nA_{low} = (a_{00}+a_{01}) \text{ and } p(A_{low}) = nA_{low} / \text{total}$$

$$nB_{low} = (a_{00}+a_{10}) \text{ and } p(B_{low}) = nB_{low} / \text{total}$$

The expected number of samples in each quadrant is computed by assuming independence for genes A and B. For example, for the quadrant A low and B low, the expected number of samples is the probability of A low multiplied by the probability of B low: $p(A_{low}) * p(B_{low})$.

If we let n denote the number of samples in the quadrant $a_{ij}$, and $n' = p(A_{low})*p(B_{low})*\text{total}$, a statistical test to determine whether a quadrant is sparse can be computed based on the expectation that in a sparse quadrant, $n' > n$. The quadrant A low and B low (that is, $a_{00}$) is considered sparse if $S_{00}$ is high (representing $n' > n$) and the error rate, $p_{00}$, is low:

$$S_{ij} = \frac{n'-n}{\sqrt{n'}}$$

$$p_{00} = \frac{1}{2}\left(\frac{a_{00}}{a_{00}+a_{01}} + \frac{a_{00}}{a_{00}+a_{10}}\right)$$

A threshold for $S_{ij} > S_{Thr}$ and $p_{ij} < p_{Thr}$ is chosen to check for quadrant sparsity. The thresholds $S_{Thr} = 3$ and $p_{Thr} = 0.1$, based on the previously used thresholds for the Boolean analysis in the test dataset GSE134213, are used as the standards for BooleanNet [44].

An equivalent Boolean relationship is found if the top-left ($a_{01}$) and bottom-right ($a_{10}$) quadrants are significantly sparse. An opposite Boolean relationship is found if the top right ($a_{11}$) and the bottom-left ($a_{00}$) quadrants are significantly sparse. For the asymmetrical Boolean implications, one quadrant is significantly sparse: A low => B low (top-left, $a_{01}$), A low => B high (bottom-left, $a_{00}$), A high => B high (bottom-right, $a_{10}$), A high => B low (top-right, $a_{11}$) (see Figure 2.2).

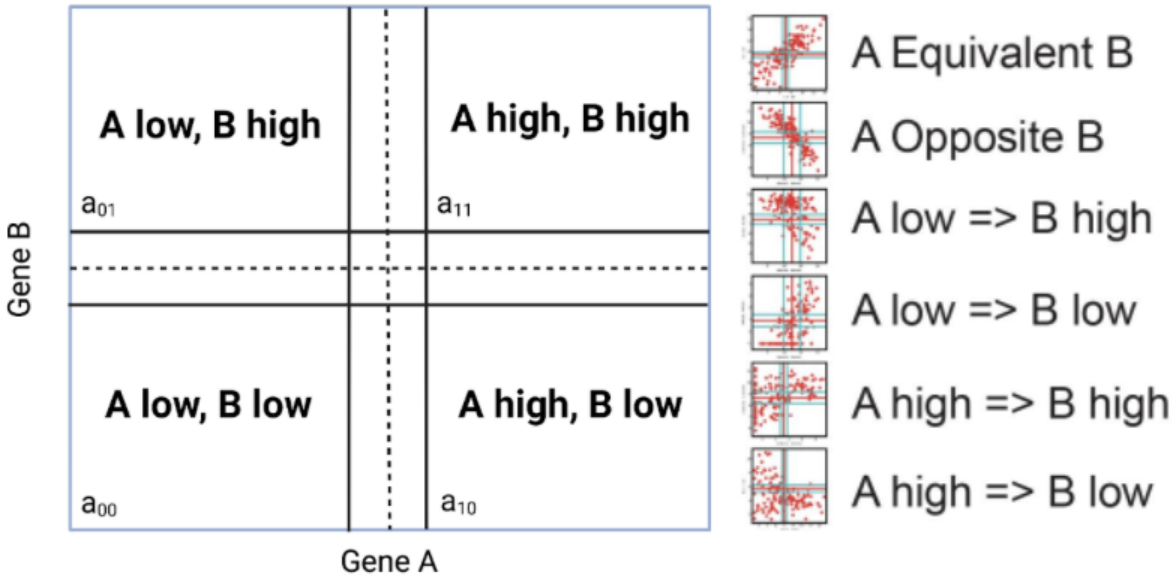



Figure 2.2: Scatter-plot separated into four quadrants based on the StepMiner thresholds and the noise error margin for arbitrary genes A and B and all 6 possible Boolean Implication relationships and their corresponding scatter plots to show for quadrant sparsity.

Boolean implication relationships remain robust even in the presence of noise and sample heterogeneity. The relationships can be invariantly observed in the expression levels of two genes; however, they do not imply causality. The biological significance of Boolean implications can be analyzed when a set of arrays of gene expression values exhibit a relationship that is likely to result from regulatory relationships, specificity to a cell types, or mutual exclusivity between the genes. Boolean implications enable investigators to uncover overlooked relationships that would otherwise be considered as weakly correlated in methods that solely rely on symmetrical relationships.

*Network Construction: Boolean Network Explorer (BoNE)*

The Boolean Network Explorer provides a visual representation of a network that captures the underlying, sequential changes in a biological process, such as the onset of a disease [45]. In order to construct such network using Boolean implications, three steps are computed: (1) The StepMiner algorithm is applied to convert all of the genes in a dataset into the binary values high expression and low expression (2) The relationship between the expression values for pairs of genes are classified as one of the six Boolean Implication Relationships (BIRs) and (3) Genes with similar expression relationship patterns, sharing at least half of the equivalences among the gene pairs, are clustered and organized into a network. In the resulting Clustered Boolean Implication Network, the cluster of equivalent genes make the nodes and the BIR between clusters make up the directed edges.

*Network Construction: Clustered Boolean Implication Network (CBIN)*

A basic Boolean Implication Network (BIN) is constructed by identifying all of the significant pairwise Boolean Implication Relationships (BIRs) within a dataset. It takes the form of a directed graph, in which genes represent the nodes and the edges correspond to the BIRs. Symmetrical Boolean implication relationships (equivalent and opposite) are represented as undirected edges and asymmetrical Boolean implication relationships (low => low, high => low, low => high, high => high) are indicated as directed edges. A BIN demonstrates robustness when the sample size is larger than 200; however, it can be applied for a smaller dataset in which genes with a dynamic range of expression values are filtered using the StepMiner threshold and analyzing the fraction of low and high values.

The complexity of the BIN can be simplified by clustering nodes based on equivalent BIRs, generating a clustered Boolean implication network (CBIN). Ideally, all genes in the same cluster should share as many BIRs to genes of other clusters as possible. The weak links in each component are eliminated to prevent noisy instances where two genes with an opposite Boolean implication relationship are included in the same cluster. A minimum spanning tree is built for the graph and the Jaccard similarity coefficient for every edge in the tree is computed to identify the weakest links. If the Jaccard similarity coefficient is $< 0.5$ for two members of the same cluster, the edges are dropped from further analysis. Eliminating the weak links ensures consistency within the clusters. A new graph is built by linking the individual clusters to each other using the four asymmetric Boolean relationships. The link between two clusters (A and B) is established by using the top node in cluster A that is connected to most of the members of A

and sampling 6 nodes from cluster B to identify the majority of BIRs between the nodes in each cluster.

*Model Selection: Boolean Paths*

The asymmetric Boolean Implication Relationships (BIRs) that form the edges in a directed Clustered Boolean Implication Network (CBIN) allow for a traversal of nodes to generate a Boolean path. A fundamental Boolean path comprises of two nodes connected by a directed edge, while a more intricate Boolean path involves multiple Boolean implication relationship. The order of clusters in a Boolean path is based on the hypothetical biological path defined by the sample order. To initiate the discovery of paths, a node that represents the biggest cluster in the CBIN is used in a greedy algorithm that traverses the nodes to choose the next biggest cluster connected to the nodes visited in sequence. In each subsequent step, the biggest cluster among the remaining nodes is chosen. Equivalence relationships from each cluster are utilized to expand the gene set within the cluster; the whole path is clustered based on these equivalence relationships. Depth-first traversal (DFS) is used to follow a path of big clusters with the specification of a Boolean implication that can be used to order samples. This process is repeated to find all the paths that connect the big clusters in the CBIN.

*Model Selection: Composite Score*

Given a set of genes that are clustered together in a specific Boolean path on a Boolean Implication network, a score can be computed to order the samples in a logical order. The genes presented in each cluster are normalized and averaged based on a modified Z-score approach

centered around the StepMiner threshold. Hence, the StepMiner-normalized gene expression value for a particular gene, denoted by $expr_{SM}$, is computed as follows:

$$expr_{SM} = \frac{expr - \theta_{SM}}{3\sigma}$$

Where *expr* is the gene expression value for a given sample, $\theta_{SM}$ is the threshold separating low/ high expression values for that gene, and is the standard deviation of the expression value from the $\theta_{SM}$.

A weighted linear combination of the averages from the clusters in the Boolean path is used to create a total score for each sample. The weights on the path are monotonically increasing or decreasing to make the sample order consistent with the logical order of the network.

$$\text{Composite Score: } a \cdot \sum A_{SM\ Norm} + b \cdot \sum B_{SM\ Norm} + c \cdot \sum C_{SM\ Norm} + d \cdot \sum D_{SM\ Norm}$$

Where $\Sigma\ X_{SM\ Norm}$ is the total sum of the StepMiner normalized gene expression values for all genes belonging to cluster X (where X is cluster A, B, C, or D), and a, b, c, and d are the assigned weights for that particular cluster. In theory, the weight is a model parameter that can be learned using training data; however, a more logical approach to determine its value for each cluster is by considering how the transitive Boolean path encodes the state transition from one pole of the map to the other. For example, in the C13-C14-C3 path, C13 refers to the immuno-reactive state and C14 and C3 are immuno-tolerant states; therefore, it is logical to put

contrasting values for the corresponding weights. In Ghosh et. al, these values are -1, +1 and +2 for C13, C14 and C3, respectively [13].

*Statistical Analysis for Model Selection: ROC-AUC Metric*

The performance of the predictive capacity to classify a set of binary annotations (i.e: healthy and diseased states) of the Boolean paths in a Clustered Boolean Implication Network (CBIN) is measured using the logistic regression model Receiver Operating Characteristic Area Under the Curve (ROC-AUC) metric [46]. The ROC curve considers the True Positive Rate (TPR) and False Positive Rate (FPR) at different classification thresholds. The given binary classification annotations from the dataset can be used to compute the number of true positives, false negatives, false positives, and true negatives. Lowering the classification threshold classifies more annotations as positive and increases the number of false positives and true positives. The equations for TPR and FPR are:

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

Where TP is the number of observed True Positives, FN is the number of observed False Negatives, FP is the number of observed False Positives, and TN is the number of observed True Negatives given a classification threshold.

The area under the ROC curve (AUC) provides a comprehensive evaluation of performance across all possible classification thresholds; it quantifies the probability of a logistic regression model ranking a random positive sample higher than a random negative sample. The ROC-AUC metric ranges from 0 to 1, where 0 reflects that all predictions are wrong, 0.5 reflects

random predictions, and 1 reflects that all predictions are correct; a higher ROC-AUC value indicates a better predictive power. This metric is scale invariant, assesses the quality of how well the predictions are ranked, and measures the predictive performance irrespective of the specific classification threshold chosen.

**Acknowledgements**

Chapter 2, in part, is currently being prepared for submission for publication of the material. Dadlani, Ekta; Dash, Tirtharaj; Sahoo, Debashis. The thesis author is the primary researcher and author of this material.

# CHAPTER 3 METHODS

In this chapter, I outline the detailed procedure used to derive the TAM gene signature specified for a colorectal cancer context. Refer to Figure 2.1 for an overview of the pipeline.

*Data Collection and Annotation*

Publicly available single-cell RNA sequencing (scRNA-seq) databases are downloaded from the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) database [47–49]. scRNA-seq captures the heterogeneity of RNA transcripts across individual cells; it allows for the understanding of a disease with higher resolution compared to gene expression pseudo-bulk datasets. Gene expression normalization is performed by computing CPM (Counts Per Millions) values; log2-normalized CPM counts are used as the final gene expression values [50, 51]. The datasets are annotated with the classifications of originating from tumor tissue samples or normal colon tissue samples, as noted in the metadata derived from the corresponding experiment.

*Macrophage Datasets Used for Network Analysis*

Macrophage network analysis for colorectal cancer specific analysis is performed using the datasets: GSE161277 human colorectal carcinoma tissue and normal colorectal tissue samples (n = 13, n cells = 54,782); GSE200997 human naive colorectal cancer patients and adjacent normal colonic tissue samples (n = 23, n cells = 49,859); GSE132257 human primary colorectal cancers and matched normal mucosa samples (n = 10, n cells = 18,409); GSE132465 human primary colorectal cancer and normal mucosa samples (n = 33, n cells = 63,689); GSE139555 human T cells in colon tumors and normal adjacent tissue samples (n = 4, n cells =

12,495); GSE222300 human rectal/sigmoid cancer and rectal adjacent normal tissue (n = 3 , n cells = 14,852); and GSE110009 human primary and metastatic tumors in colon cancer and adjacent normal tissues (n = 29, n cells = 8,160). To show consistency of the polarization patterns in mouse datasets, the macrophage network analysis is applied to GSE198758 APKS and AOM/DSS mouse models of colorectal cancer and normal colon tissue samples (n = 4, n cells = 20,849); and GSE224679 AOM/DSS mouse models of colorectal cancer and normal colon tissue samples (n = 4, n cells = 27,656) (see Table 3.1). For simple nomenclature of annotations, the colorectal cancer tumor samples are referred to as “colorectal cancer” (CRC) samples and the normal colorectal tissue, adjacent normal colonic tissue, and matched normal mucosa samples are referred as “normal tissue” (NT) samples.

Table 3.1: General information about scRNA sequencing datasets taken from the NCBI GEO Database for analysis. n Samples represents the number of targeted types of cells annotated (i.e: group of cells that undergo the same treatment/experimental procedure); n Cells represents the total number of cells provided in the dataset's gene expression matrix.

| GEO ID | Species | n Samples | n Cells |
|---|---|---|---|
| GSE161277 | Homo Sapien | 13 | 54,782 |
| GSE200997 | Homo Sapien | 23 | 49,859 |
| GSE132257 | Homo Sapien | 10 | 18,409 |
| GSE132465 | Homo Sapien | 33 | 63,689 |
| GSE139555 | Homo Sapien | 4 | 12,495 |
| GSE222300 | Homo Sapien | 3 | 14,852 |
| GSE110009 | Homo Sapien | 29 | 8,160 |
| GSE198758 | Mus musculus | 4 | 20,849 |
| GSE224679 | Mus musculus | 4 | 27,656 |

*Computational Approaches: Signature of Macrophage Reactivity and Tolerance (SMaRT)*

For the Signature of Macrophage Reactivity and Tolerance (SMaRT) model, the gene expression summarization for publicly available microarray and RNASeq databases are performed by normalizing Affymetrix platforms by RMA (Robust Multichip Average) and RNASeq platforms by computing TPM values when normalized data is not available in the GEO database [13]. Final gene expression values for analyses are computed as log2(TPM) if TPM > 1 and (TPM-1) if TPM < 1. Publicly available datasets normalized using RPKM, FPKM, and CPM are also used for validation purposes.

The SMaRT model is built using the published pooled macrophage dataset from GEO (GSE134312, n = 197) assayed on the Human U133 Plus 2.0 (GPL570), Human U133A 2.0 (GPL571), and Human U133A (GPL96) platforms. The M0, M1, and M2 phenotype are manually annotated for this dataset. This dataset consists of primary tissue-derived macrophages obtained from both healthy and diseased tissues and cultured macrophage cell lines that are either untreated or treated with specific ligands that induce either the M1 polarized state (n = 13) or the M2 polarized state (n = 8). Testing of the macrophage gene signature is performed using the validation datasets: GSE35449 (7 M0, 7 M1, 7 M2), GSE46903 (64 M0, 29 M1, 40 M2), GSE61298 (6 M0, 6 M1, 6 M2), GSE55536 human peripheral blood mononuclear cell-derived macrophage (6 M0, 6 M1, 6 M2), and GSE55536 iPSC derived macrophages (3 M0, 3 M1, 3 M2).

To create the SMaRT model, a Boolean Implication Network (BIN) is created using all of the significant Boolean implication relationships (BIRs) for the dataset GSE134312 (n = 197) (refer to Figure 2.1). Since the macrophage dataset has less than 200 samples, the fraction of high and low values after applying the StepMiner algorithm is used to filter all genes that have a reasonable dynamic range of expression values. Probe sets that contained less than 5% of the high or low values are dropped. A Clustered Boolean Implication Network (CBIN) is created in which every cluster of genes is associated with healthy or diseased samples based on where the cluster is highly expressed. Locating the ‘M1’ labeled samples and ‘M2’ labeled samples on the resulting network shows a segregation of the two polarization states, indicating a continuum of cellular states in macrophages within the immunologic spectrum rather than the discrete categories. The paths of high => high, high => low, and low => low are used to order the

samples from the healthy to diseased states. As a result, Boolean paths that intersect these relationships and large clusters in the CBIN is used to show the continuum of polarization from the reactive pole to the tolerant pole of the network. Depth-first traversal (DFS) is used to follow the longest possible paths of low => low where the biggest clusters are visited first. The direction of the paths is derived from the connection of a reactive cluster to a tolerant cluster. The top continuum paths are analyzed using reactome pathway analysis, resulting in a list of enriched pathways. From these enriched pathways, specific paths are selected for testing the classification of samples into the immuno-reactive (M1-like) and immuno-tolerant (M2-like) states. The effectiveness of various Boolean paths in their ability to utilizing the composite scores for clusters in the path to classify the samples is assessed for sample classification.

Multivariate analysis of the top five Boolean paths shows that the path that connects clusters C13 => C14 => C3 is performs the best ($p > 0.001$) at discriminating the M1-like (ROC-AUC: 0.98) and M2-like (ROC-AUC: 0.99) polarization states. Independently, C13 accurately predicts the reactivity (M1-like) state with a ROC-AUC of 1.0 and the path C14 => C3 demonstrates close to perfect prediction of the tolerant (M2-like) state, with an ROC-AUC ranging between 0.8 - 1.0. The C13-14-3 signature, consisting of 48 genes in C13 and 290 genes in the path C14 => C3), successfully identifies the M1/M2 polarization states under a diverse range of tissue-resident macrophages, in both human and mice, and in other immune cells.

*Computational Approaches: Macrophage Extraction*

Traditional biomarkers for macrophages, such as *CD14*, *ITGAM*, *CD68*, and *EMR1*, exhibit variable expression patterns in different tissues due to the intricate nature of macrophage

biology and variability in experimental techniques and purification methods. In Dang et. al, a computational approach called BECC (Boolean Equivalent Correlated Clusters) analysis is used to identify and validate *FCER1G* (Fc fragment of IgE receptor Ig) and *TYROBP* (TYRO protein tyrosine kinase-binding protein) as universal biomarkers for macrophages in human and mouse tissues [44]. The BECC model uses Boolean methodologies to differentiate the asymmetric and symmetric relationships that identify the genes that mirror each other's gene expression patterns. It compares the normalized expression levels of two genes across a set of provided datasets, searching for two sparsely populated quadrants diagonally opposite to each other amongst the four possible quadrants (high-low and low-high). Boolean Equivalent relationships are used to identify functionally related genes. The macrophage BECC model is constructed using the seed gene *CD14*, which is expressed in a majority of macrophage populations; Boolean equivalent relationships, pairwise correlation, and linear regression analysis are used to rank a list of 33 probe sets that serve as training data the model. The probe sets correspond to 21 unique genes with similar expression patterns as *CD14*. A StepMiner threshold is applied to identify high-confident macrophage genes. The threshold results in 18 significant probe sets and 13 unique genes. *FCER1G* is the top candidate gene and *TYROBP* is the fourth top candidate based on the BECC ranking; out of the 13 gene candidates *TYROBP* and *FCER1G* has the strongest correlation patterns across human and mouse datasets. The consistent equivalent Boolean expression pattern of *TYROBP* and *FCER1G* suggests that the tight correlation is expressed in a similar context in all tissues coming from pure macrophage samples. *TYROBP* is an adapter protein that form non-covalent associations with activating receptors present on the surface of immune cells. Its primary function is to mediate signaling and cell activation upon ligand

binding by receptors. In the case of *FCER1G*, interactions with an allergen, triggers cell activation and the induction the release of mediators involved in allergic responses.

For each colorectal cancer dataset, the expression of the two universal biomarkers, *TYROBP* and *FCER1G*, are used to filter cells that are potential macrophages. A threshold of > 2.0 or a threshold of > 0.0 for the expression of the macrophage biomarkers at the raw count gene level (GSE132465, GSE224679) is applied. This significantly reduces the number of cells for each dataset (see Table 3.2). On average, approximately 8.6% of all cells in each dataset are classified as a macrophage cell using this filtering method.

Table 3.2: General information about scRNA sequencing datasets after the macrophage cells are extracted using thresholds for the normalized and scaled gene expressions of *TYROBP* and *FCER1G* (*TYROBP* > 2.0 & *FCER1G* > 2.0 or *TYROBP* > 0.0 & *FCER1G* > 0.0). For some datasets, the number of samples reduces due to the lack of macrophage cells in that particular cell group.

| GEO ID | Species | n Samples | n Samples (after macrophage cell extraction & with annotations) | n Macrophage Cells |
|---|---|---|---|---|
| GSE161277 | Homo Sapien | 13 | 7 | 2,704 |
| GSE200997 | Homo Sapien | 23 | 23 | 969 |
| GSE132257 | Homo Sapien | 10 | 10 | 1,144 |
| GSE132465 | Homo Sapien | 33 | 33 | 8,049 |
| GSE139555 | Homo Sapien | 4 | 4 | 672 |
| GSE222300 | Homo Sapien | 3 | 3 | 441 |
| GSE110009 | Homo Sapien | 29 | 29 | 1,916 |
| GSE198758 | Mus musculus | 4 | 3 | 7,379 |
| GSE224679 | Mus musculus | 4 | 4 | 1,478 |

*Computational Approaches: Applying the SMaRT Model to Characterize Polarization Dynamics in Single Cell RNA-seq Colorectal Cancer Datasets*

The 338 gene universal macrophage SMaRT signature (48 genes measuring the immuno-reactive gradient and 290 genes measuring the immuno-tolerant gradient in humans; 71 genes measuring the immuno-reactive gradient and 227 genes measuring the immuno-tolerant gradient

in mouse) is applied to each macrophage filtered scRNA sequencing dataset to compare the immuno-reactive samples to the immuno-tolerant samples (see Figure 3.1).

Figure 3.1: Workflow of Application of SMaRT Model on Single Cell Dataset. Publicly available single cell RNA sequencing (scRNAseq) colorectal cancer (CRC) centered datasets annotated with colorectal cancer tissue (CRC) and normal colon tissue (NT) samples are used for analysis. A threshold for the expression of the two universal macrophage biomarkers, *TYROBP* and *FCER1G*, at the raw count gene expression level is applied; this reduces the number of cells under further scrutiny by approximately 92%. Normalization/scaling (CPM and log2) of the raw gene expression count matrix is performed. The SMaRT Model's C13-C14- C3 signature is used to generate an immuno-reactive and immuno-tolerant composite score for each macrophage cell in each dataset. The StepMiner algorithm is used to threshold the C13 (immuno-reactive) scores and the C14-C3 (immuno-tolerant) scores as a high/low composite score. A negative weight is applied to C13; a composite score below the StepMiner threshold for C13 is considered highly reactive. Positive weights are applied for C14 and C3; a composite score above the StepMiner threshold for C14-C3 is considered highly tolerant. Therefore, the bottom left quadrant (low C13 and low C14-C3) consists of highly reactive macrophage cells and the top right quadrant (high C13 and high C14-C3) consist of highly tolerant macrophage cells. A normal (z) test is applied to compare the number of CRC macrophage cells to the number of NT macrophage cells; the p-value for the normal test computed in the highly reactive and the highly tolerant quadrants is used to determine the significance in the number of macrophage cells annotated as CRC and NT. *A.* For the human dataset GSE137465, the threshold of > 0.0 is used for the raw gene expression values of *TYROBP* and *FCER1G*. The gene expression values are normalized and scaled prior to implementation of the SMaRT model's C13-C14-C3 signature. *B.* For the mouse dataset GSE198758, a threshold of > 2.0 is used for the raw gene expression values of *Tyrobp* and *Fcer1g*. The gene expression values are normalized and scaled prior to implementation of the SMaRT model's C13-C14-C3 signature.

**A** **Macrophage Polarization in single cell dataset GSE132465**

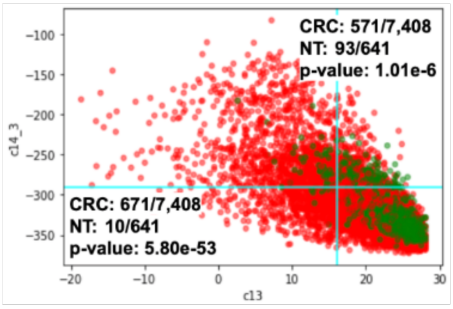
GSE132465 (n = 33, ncells = 63,689)
Colorectal cancer (n = 23)
Normal Colon Tissue (n = 10)

↓

Macrophage selection (ncells = 8,049):
ENSG00000011600 (*TYROBP*) > 0.0
ENSG00000158869 (*FCER1G*) > 0.0

↓

Colorectal Cancer (ncells = 7,408)
Normal Colon Tissue (ncells = 641)




**B** **Macrophage Polarization in single cell dataset GSE198758**

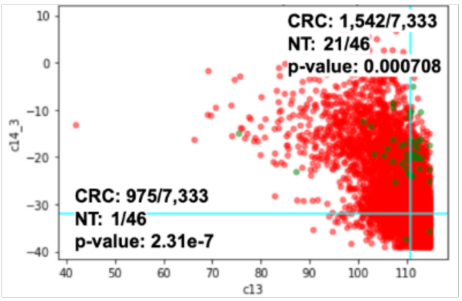
GSE198758 (n = 4, ncells = 20,849)
Colorectal Cancer (n = 3)
Normal Colon Tissue (n = 1)

↓

Macrophage selection (ncells = 7,379):
ENSMUSG00000030579 (*Tyrobp*) > 2.0
ENSMUSG00000058715 (*Fcer1g*) > 2.0

↓

Colorectal Cancer (ncells = 7,333)
Normal Colon Tissue (ncells = 46)



For each dataset, all macrophage cells are annotated as tumor tissue (CRC) or normal colon/ rectum tissue (NT). For each dataset, the composite score for the reactive cluster (C13) is computed using a negative weight (-1); the composite scores for the tolerant clusters (C14 and C3) are computed using positive weights (+1 for C14, +2 for C3). The StepMiner algorithm is used to create a statistically-significant threshold for separating low C13 composite scores from high C13 scores and for separating low C14-C3 composite scores from high C14-C3 composite scores.

For each dataset, samples that are confined to the low C13/low C14-C3 quadrant are classified as highly reactive macrophages. Samples that are confined to the high C13/high C14-C3 quadrant are classified as highly tolerant macrophages. To determine whether the number of tumor specific macrophage cells and normal colon specific macrophage cells are significantly comparable to each other, a test for proportions based on normal (z) test is performed:

$$p_1 = \frac{successes}{n} \qquad Z = \frac{p_1 - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Where successes is the number of observed normal colon macrophage cells confined to the specified quadrant (reactive or tolerant), n is the total number of observed normal colon macrophage cells confined to both the reactive and tolerant quadrants, and $p_0$ is the null hypothesis of number of observed tumor macrophage cells confined to the specified quadrant (reactive or tolerant)/number of observed tumor macrophage cells confined to both the reactive and tolerant quadrants.

The p-value is used to indicate the statistical significance of the two proportions z-test. The null hypothesis for p-value is that there is no difference in the proportion of NT samples to CRC samples in the given quadrant. The threshold of $< 0.05$ indicates that we can reject the null hypothesis in favor of the alternative hypothesis which states that there is a difference in the proportion of NT samples to CRC samples likely not due to result of random chance; a p-value of $< 0.05$ for the two proportion z-test is considered statistically significant.

*Computational Approaches: Reactivity Gene Signature Refinement*

In order to remove noise from the existing C13-C14-C3 macrophage signature due to its relevance across different tissues, organs, species, and immune cells, refinement is performed. The ultimate goal is to create a tumor-associated macrophage (TAMs) specific gene signature that captures the dynamics of immuno-reactivity for the specialized context of colorectal cancer. Refinement leads to a sensitive and specific gene set for the specific context of TAMs in colorectal cancer patients. Using the 338-gene signature from the SMaRT model, macrophages from the colorectal cancer tumor samples are classified as more reactive compared to macrophages from the normal colon samples. As a result, refinement of C13 (which captures immuno-reactivity), consisting of 48 human genes and 71 equivalent mouse genes, would alone capture the polarization difference of macrophage data comparing colorectal cancer and normal colon tissue samples at the pseudo-bulk level. To indicate that refinement is necessary, I identify a subset of genes within C13 that produces a great deal of noise for the difference between NT and CRC samples, including flipping or scrambling the expected pattern of CRC samples being more reactive.

Refinement is performed using differential gene expression analysis on the single cell dataset, GSE132465 (human, n = 33, n cells = 63,689). After selecting for macrophages using the thresholds for normalized and log-scaled expression of *TYROBP* and *FCER1G* to > 0, the dataset consists of 8,049 macrophage cells. To specify the macrophage polarization state, the SMaRT model is applied to the tumor and normal colon macrophage cells. The StepMiner algorithm is used to annotate the threshold of low/high composite scores for C13 and C14-C3 and separate the samples to C13 low/high and C14-C3 low/high classifications. The highly reactive (C13 low and C14-C3 low) samples and the highly tolerant (C13 high and C14-C3 high) samples are extracted and used with the annotation of "Reactive" (highly reactive) and "Tolerant" (highly tolerant) to perform differential gene analysis.

In order to determine whether a change in the gene expression is observed across a population, I cannot consider each cell (technical replicates) as an independent variable and must consider the difference across biological replicates, or samples. Therefore, I artificially generated the pseudo-bulk representation of the dataset at the level of all macrophages, macrophages confined to the highly tolerant quadrant, all tumor-specific macrophages, and purified epithelial cells. The equivalent pseudo-bulk representation of a scRNA-seq dataset is created by summing the gene expression values from all cells that contribute to a specific sample; normalization and scaling is performed after the transformation. The tumor microenvironment in colorectal cancer consists of various cell types, including immune cells (macrophages) and tumor cells, which originate from the epithelium. The expression of epithelial genes can introduce heterogeneity in the gene expression patterns of macrophages as a result of potential contamination of housekeeping genes. To retrieve significant epithelial cells, a threshold of <= 0 for *TYROBP* and

*FCER1G* (to remove potential macrophages) and a threshold of > 2 for the epithelial biomarker *EPCAM*, which is involved in making the epithelial cellular adhesion molecule (EpCAM), is applied. Four differential gene lists are considered for refinement: separating the CRC and NT samples from only the highly tolerant macrophages, separating reactive and tolerant tumor macrophages, separating the CRC and NT samples in all purified macrophages, and separating the CRC and NT samples in the purified epithelium. Since the single cell dataset used for training has far more tumor cells compared to normal colon cells, the pseudo-bulk representation of the dataset can be biased, leading to an incorrect ranking of highly expressed genes. As a result, differential gene expression is performed across samples that are derived from a comparable number of cells (Tumor: GSM3868434, GSM3868436, GSM3868427, GSM3868430, and GSM3868431; Normal Colon: GSM3868448, GSM3868451, GSM3868452, GSM3868456, and GSM3868457) for the cases of separating the CRC and NT samples from only the highly tolerant macrophages and separating reactive and tolerant tumor macrophages.

Differential gene expression analysis aims to identify genes that have a distinctive expression between across a set of conditions solely based on a biological phenomenon. It is performed using a t-test, in which the average expression of gene is compared across two conditions. The null hypothesis of a tests is that a gene has the same average expression across the two binary groups. The differential value (fold change) used to compute the change in expression is calculated by dividing the difference in the mean expression value by the variation of the gene expression values from the mean values. A positive value indicates an increase in the expression of the gene in a binary group and a negative value indicates a decrease in the expression of the gene in the respective binary group. A two fold increase (equivalent to a Log

Fold Change of 1) for gene A in the “Reactive” state compared to the “Tolerant” state indicates that gene A is expressed twice as much in the “Reactive” state.

The differential expression analysis produces a ranking of genes; I extract all of the genes that exist in the C13 cluster in descending order of differential values. The top ranking of C13 genes from the results of the differential gene expression analysis using the different conditions are used as an ensemble to generate two signatures: a “noisy” signature composed of a subset of 6 genes from C13 that contains artifacts that negatively impact the predictive potential of the universal macrophage signature for diagnosing colorectal cancer and a “refined” signature composed of a subset of 15 genes from C13 (equivalent to 40 homologous mouse genes) that are highly expressed in the reactive tumor macrophages. The noisy signature is generated by overlapping C13 genes that are up regulated in tolerant tumor macrophages and up regulated in tolerant normal colon macrophages. The refined signature is generated by overlapping C13 genes that are down regulated in normal epithelial cells and up regulated in reactive macrophages.

*Computational Approaches: Predictive Modeling*

In order to assess the predictive capability of the refined signature for classifying CRC samples and NT samples, a logistic regression model is employed. The model utilizes the composite score of the signature along with the sample annotations to make binary classifications. In order to produce comparable results to C13, I assign a weight of -1 to the refined signature. In order to compare the expression pattern of genes within the signature, the expression value of each gene is converted into its equivalent StepMiner normalized expression value; the composite score is calculated using the weighted linear combination of the assigned

weight and sum of the StepMiner normalized gene expression values. The logistic regression model, trained on dataset GSE132465, makes predictions of tissue type based on the probability of being classified as CRC or NT samples; it is a supervised learning classification model that uses the logistic (or sigmoid) function to convert the composite score that is used for the binary classifications to a value between 0 to 1. It predicts the probability of the tissue type (CRC and NT) based on the independent variable of composite score of the refined signature (15 genes). The performance of the model on the training dataset is compared to that of C13 and a subset of 6 genes from C13.

*Computational Approaches: Validating The Refined Reactivity Signature*

Primary validation of the refined, tumor-associated macrophage specific signature is performed on artificially generated human pseudo-bulk datasets with the annotations of CRC and NT and the artificially generated macrophage purified pseudo-bulk datasets with the annotations of CRC and NT. The single-cell RNAseq datasets with a statistically significant difference for the proportion of colorectal cancer macrophage cells and normal colon macrophage cells in the highly reactive quadrant, GSE161277, GSE200997, GSE222300, GSE139555, and GSE132257, are used as the initial validation datasets. After applying the thresholds for *TYROBP* and *FCER1G* expression at the raw count level, the expression matrices with single cell data are converted into normalized and log-scaled pseudo-bulk matrices. Generation of the composite score for the refined signature and the annotations of “Normal Colon Tissue” (Normal) samples and “Colorectal Cancer Tumor” (Tumor) samples allow me to analyze whether the TAMs specific signature could prognostically predict the difference between normal and cancerous

samples solely based on the composite scores of the signature. Secondary validation is performed on large CRC microarray and RNA-seq datasets without the specification of macrophages. These datasets include GPL570 (microarray; 170 NT and 1,662 CRC), TCGA 2017 mRNA cohort (microarray; 51 NT and 644 CRC), GSE20916 (microarray; 34 NT and 36 CRC), GSE146009 (RNA-seq; 9 NT and 9 CRC), GSE44076 (microarray; 98 NT and 98 CRC), GSE62932 (microarray; 4 NT and 64 CRC).

**Acknowledgements**

Chapter 3, in full, is currently being prepared for submission for publication of the material. Dadlani, Ekta; Dash, Tirtharaj; Sahoo, Debashis. The thesis author is the primary researcher and author of this material.

# CHAPTER 4: RESULTS

In this chapter, I provide the findings of my research using the integrative methodology described in the previous chapter. In general, I provide evidence to support the conclusions: (1) colorectal cancer tumor tissue derived macrophage cells are more reactive compared to normal colon tissue derived macrophage cells and (2) I can apply a refined macrophage signature of 15 human genes (40 equivalent mouse genes) on a single cell macrophage specific colorectal cancer dataset in order to capture the difference in the immuno-reactivity polarization state for colorectal cancer and normal colon tissues.

## *Polarization in Colorectal Cancer and Normal Colon Macrophage Specific Cells*

After extracting the macrophage specific cells from each single cell dataset, the SMaRT C13-C14-C3 universal macrophage signature, which provides a set of standardized definitions for macrophage polarization, is utilized to segregate the M1 (reactive) and M2 (tolerant) macrophages from normal colon tissue (NT) samples and colorectal cancer tissue (CRC) samples at the single cell level [13]. A scatter plot representing the composite scores for C13 and C14-C3 for the macrophage specific cells, annotated as tumor samples and normal colon samples, and applying a StepMiner threshold to separate the C13 and C14-C3 into low/high scores, showcases that at the macrophage cell level, tumor samples are more immuno-reactive compared to normal colon samples; there are more tumor samples compared to normal colon samples confined to the highly reactive quadrant. This pattern is consistent with a statistically significant difference across multiple human single cell datasets (GSE161277, GSE200997, GSE132257, GSE222300,

GSE132465, GSE139555) and mouse single cell datasets (GSE198758 and GSE224679) (see Figure 4.1).

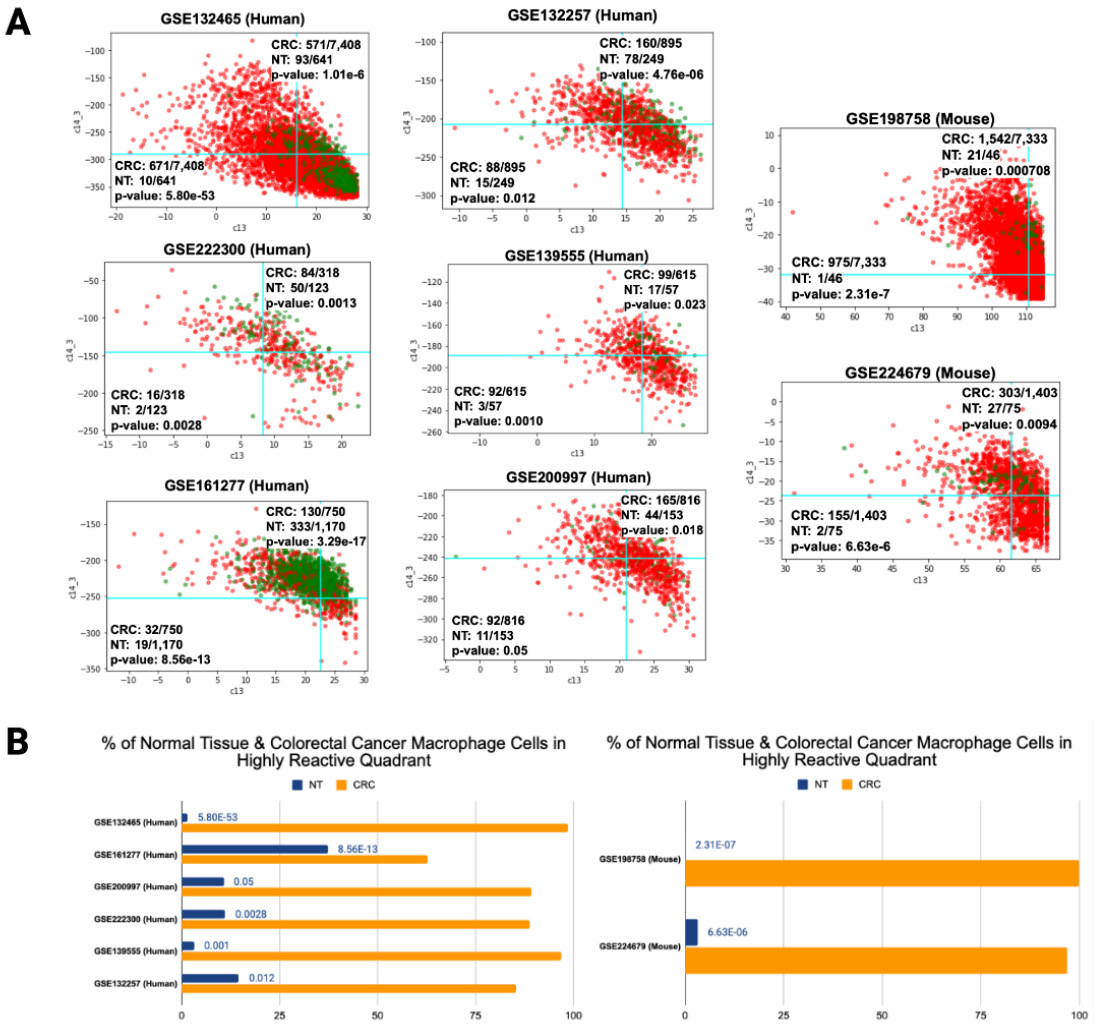



Figure 4.1: Difference in Polarization States in Macrophage Specific Human and Mouse Single Cell Colorectal Cancer Datasets. *A*. By applying the SMaRT model's C13-C14-C3 signature on the macrophage purified single cell datasets and a threshold the composite scores for C13 and C14-C3 using the StepMiner algorithm allows me to qualitatively and quantitatively analyze the difference in the tumor (CRC) samples from the normal tissue (NT) samples. On the scatterplot(s), the bottom left quadrant (low C13 and low C14-C3) consists of highly reactive macrophage cells and the top right quadrant (high C13 and high C14-C3) consist of highly tolerant macrophage cells. Comparing the proportions of CRC samples and NT samples in the lower left quadrant showcases a pattern in which there are more reactive CRC samples compared to NT samples. This trend is consistent and statistically significant (computed using the p-value of a normal (z) test) across multiple human and mouse datasets. *B*. The significant difference in the number of NT and CRC samples in the highly reactive quadrant can be clearly investigated using a bar chart. In both human and mouse datasets, CRC macrophage cells are highly reactive compared to NT macrophage cells.

With an increase in tumorigenic activity, tumor associated macrophages are inclined to convert from the tolerant polarization state to the reactive, anti-inflammatory, cancer-promoting polarization state [32]. The reactive polarization state allows TAMs and other tumor cells to promote tumor cell proliferation with the excretion of cytokines; the tumor-promoting phenotype induces the growth and metastasis of colon cancer cells. Since TAMs are involved in tumor proliferation, invasion, migration, and angiogenesis along with suppressing anti-tumor immunity and regulating metabolism, there is much interest surrounding TAM-targeted precision diagnostics and therapeutics for CRC.

*Reactivity Gene Signature Refinement and Validation*

Refinement of the SMaRT model's C13 gene signature, consisting of 48 human genes (and 69 equivalent mouse genes) that capture the spectrum of immuno-reactivity, is performed using a list of statistically significant differentially expressed genes from GSE132465. Refinement produces a list of 15 human genes (and equivalent of 40 mouse genes) expressed in tumor associated macrophages in the specialized context of colorectal cancer. The expression of these genes are able to capture the prognosis of normal colon tissue samples and colorectal cancer tissue samples in TAMs: *IFIT2*, *MX1* (*Mx1, Mx2* in mouse), *OAS1* (*Oas1a*, *Oas1e*, *Oas1d*, *Oas1f*, *Oas1c*, *Oas1b*, *Oas1h*, *Oas1g* in mouse), *IFIT3* (*Ifit3b*, *Ifit3* in mouse), *CXCL9*, *XAF1*, *SP110* (*AC125149.5*, *AC133103.5*, *AC168977.1*, *AC133103.4*, *Sp110*, *AC132444.3*, *AC125149.4* in mouse), *STAT1*, *OAS2*, *TAP1*, *APOL1* (*Apol11b*, *Apol7a*, *Apol11a*, *Apol10b*, *Apol7e*, *Apol7c*, *Apol9a*, *Apol9b*, *Apol7b*, *Apol10a*, *Apol8* in mouse), *OAS3*, *TRIM21*, *PML*, and *ISG15*.

The performance of the refined signature is compared to the performance of C13 and the noisy signature (Figure 4.2). By extracting the highly reactive and tolerant macrophages at the single cell level, I am able to manually annotate the sample as Normal Tolerant (NT), Tumor Tolerant (TT), and Tumor Reactive (TR); as shown at the single cell level, I expect the tumor samples (annotated as Tumor Reactive and Tumor Tolerant) to be more reactive compared to the normal samples (Normal Tolerant). This behavior is verified in C13; however, the noisy signature disassembles this expected behavior in CRC. This dynamic is most prominent in Figure 4.2.C, which compares the predictive performance of separating NT, TT, and TR at the pseudo-bulk level of the highly reactive and tolerant macrophage cells. The noisy signature dismantles the expected ordering of TR, TT, NT by incorrectly claiming that the NT samples are more reactive compared to the TT samples as a byproduct of the composite score associated with these genes. The noisy signature provides further evidence to support the necessity for refinement of the macrophage signature. An ROC-AUC score is used to differentiate the performance of the three signatures. Comparing the performance of C13, the noisy subset, and the refined subset at the macrophage specific pseudo-bulk and the pseudo-bulk level with all cells (prior to macrophage extraction) shows that the refined subset of C13 performs equivalent to or better than C13 while only containing a fraction of the total number of genes from the original reactive signature.

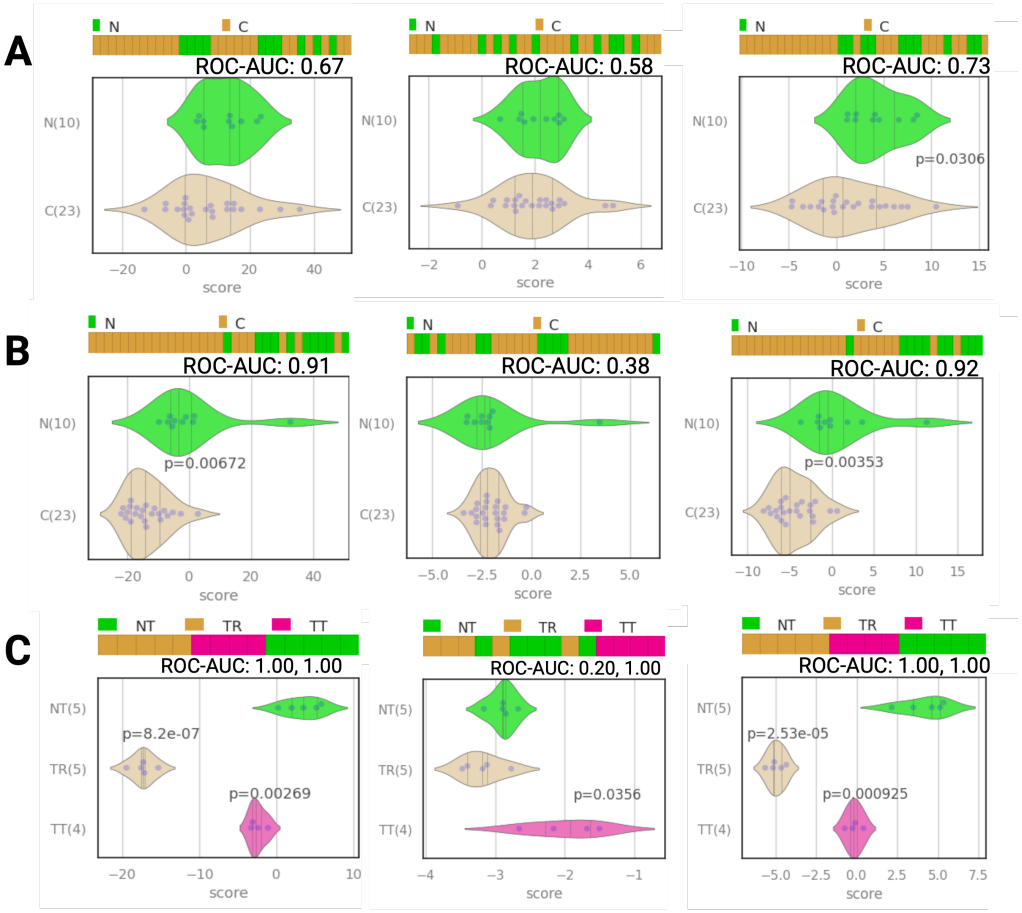



Figure 4.2: Comparison of predictive performance between C13 (48 genes), a noisy signature derived from C13 (6 genes), and a refined TAMs specific C13 signature (15 genes) for GSE132465 (training dataset). When comparing the signatures at the pseudo-bulk level with all cells, the pseudo-bulk level of only macrophage cells, and the pseudo-bulk level of the annotated reactive and tolerant macrophage cells, the noisy signature performs worse than C13 and the refined signature; it adds significant noise to the signature to the extent where the expected relationship of colorectal cancer samples (C) having a lower score compared to the normal samples (N) is scrambled. *A.* Comparison of the predictive performance of distinguishing Normal Colon samples (N) to Colorectal Cancer samples (C) between the three signatures (C13, noisy signature, and refined signature, respectively) at the artificially generated normalized and scaled pseudo-bulk level in which all cells are considered (pre-macrophage purification). *B.* Comparison of the predictive performance of distinguishing Normal Colon samples (N) to Colorectal Cancer samples (C) between the three signatures at the artificially generated normalized and pseudo-bulk level consisting solely of purified macrophage cells. *C.* Comparison of the predictive performance of distinguishing Tolerant Normal Colon samples (NT), Tolerant Colon Cancer samples (TT), and Reactive Colon Cancer samples (TR) between the three signatures at the artificially generated normalized and pseudo-bulk level consisting of a subset of comparable annotated highly reactive and highly tolerant macrophage cells. The incorrect dynamics of the noisy signature is most prevalent within this classification, claiming that NT samples are more reactive compared to the TT samples unlike C13 and the refined signature.

The predictive performance of the refined gene signature to classify normal colon tissue samples and colorectal cancer tissue samples is validated across multiple macrophage-specific datasets used for single-cell analysis and converted to the pseudo-bulk level (macrophage-purified) and at the pseudo-bulk level with all cells present (see Figure 4.3). Preliminary validation is performed across the artificially generated pseudo-bulk datasets with a significantly scalable difference in the number of colorectal cancer samples and normal colon samples in the highly reactive and highly tolerant quadrants at the single cell level. The ROC-AUC metric is used to capture the predictive performance of the logistic regression model; for the refined signature, the ROC-AUC values range between 0.50 - 1.0 when comparing the annotations of normal and tumor. A comparative predictive performance of the current refined signature against the original reactivity signature of the SMaRT model that consists of approximately 4 times more number of genes and the noisy gene signature is provided. The refined signature performs better, or equally as well as the SMaRT's reactivity signature (ROC-AUC: 0.25 - 1.00), whilst only considering a fraction (around 30%) of the genes from the original reactivity signature. As expected, the noisy signature performs worse than C13 and the refined signature at both the all cell pseudo-bulk level and the macrophage specific pseudo-bulk level (ROC-AUC: 0.17 - 1.00). Further validation of the dataset is performed on large microarray and RNA-seq colorectal cancer cohorts with the annotations of Normal Colon Tissue (Normal) and Colorectal Cancer Tissue (Tumor): GPL570, TCGA 2017, GSE20916, GSE44076, and GSE62932 (See Figure 4.4). The refined signature is also capable of lewdly capturing the development of colorectal cancer (from normal colon epithelium to a primary adenoma state to the invasive carcinoma state) in the datasets GPL570, GSE117606, and GSE20916 (See Figure 4.5). Overall, the current results

across datasets are highly indicative of accurate colorectal cancer prognosis and that the robust minimal signature may hold promising implications towards clinical translation.


Figure 4.3: Performance Comparison for Noisy Signature and Refinement of Universal Macrophage Signature for Specialization in Macrophages Originating from Colorectal Cancer Samples. Artificially generated pseudo-bulk representations of validation macrophage specific single-cell datasets are generated by summing the raw counts and performing normalization/ scaling. This was done for the single cell datasets prior to macrophage extraction (all cells) and after macrophage extraction (macrophage purified). ROC-AUC values are used to compare the predictive potential of C13 (SMaRT Reactivity Signature), the Noisy Signature (subset of 6 genes from C13), and the Purified Signature (subset of 15 genes from C13). A general trend to note is that the noisy signature performs worst than or equivalent to C13 and that the refined signature performs better than or equivalent to C13 at both pseudo-bulk levels in capturing the dynamics that we observe at the single-cell RNA-seq level using the SMaRT C13-C14-C3 signature. *A.* Comparison of the predictive power of the three signatures at the all cell pseudo-bulk level. *B.* Comparison of the predictive power of the three signatures at the macrophage purified pseudo-bulk level.

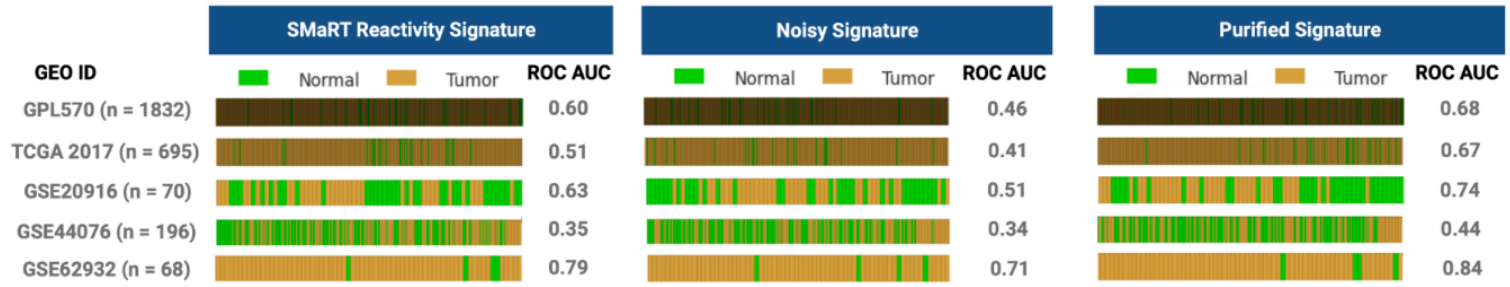



Figure 4.4: Performance Comparison for Validation of Noisy Signature and Refined Signature in large RNA-seq and microarray cohorts. ROC-AUC scores are used to compare the predictive performance of C13, the noisy subset of C13 genes (6 genes), and the purified subset of C13 genes (15 genes). In general, (1) the noisy signature performs worse than the C13 signature and the purified signature and (2) the purified signature is capable of distinguishing the normal and tumor samples and showcasing the expected skewness of tumor samples towards a more reactive composite score (skewness to the left).

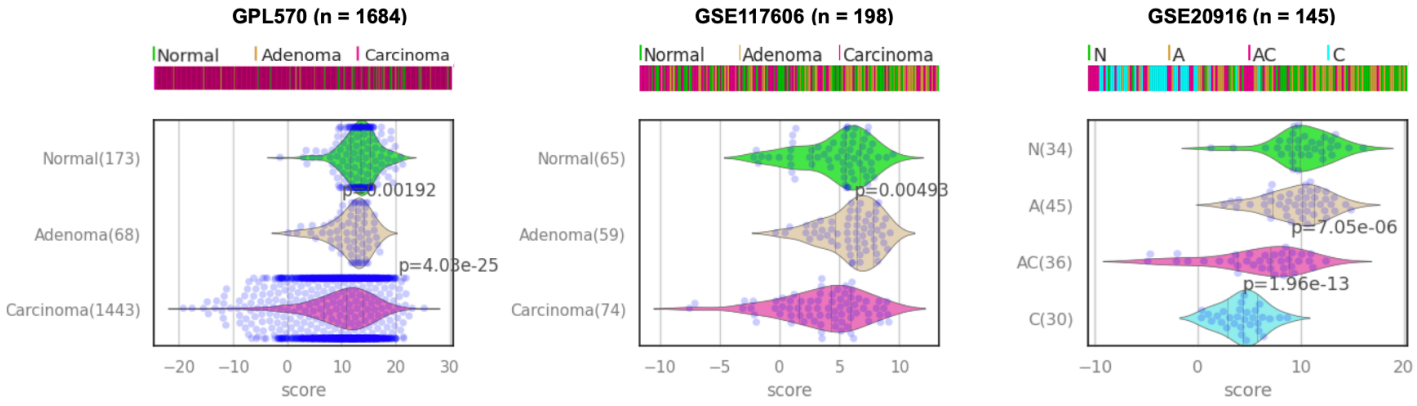



Figure 4.5: Tumor Progression Captured by the Refined Signature in Large RNA-seq and Microarray Datasets. These violin plots showcase that the refined signature is lewdly able to predict the progression of CRC using the signature's immuno-reactive composite scores. As expected, normal (N) colon samples are the least reactive and the invasive carcinoma (C) samples are the most reactive (skewed towards the left). Samples originating from the intermediate stages of adenoma (A) and adenocarcinoma (AC) have a composite score that mostly lies in between that of the normal and carcinoma samples.

Of the 15 genes in the purified signature, a majority of them are involved in the interaction between CRC cells and TAMs. *IFIT2*, *IFIT3*, *ISG15*, *MX1*, *OAS1*, *OAS2*, *OAS3*, *STAT1*, and *TAP1* are crucial interferon-stimulated genes (ISGs) involved in the modulation of immune responses [53–57]. ISGs are activated as a first line antiviral response against pathogens; they are activated by the CSF-1 receptor and EGFR pathways after infections, inflammation, or tissue damage. The activation of reactive (M1) macrophages in the tumor microenvironment of colorectal cancer tissue as response to inflammation triggers intracellular signaling cascades that up regulate the ISGs that are involved in immune activation. *CXCL9* is known for its chemotaxis activity and its ability to enhance the infiltration of of macrophages; it has a vital role in tumor invasion and differentiation, lymph node metastasis, distant metastasis, and vascular invasion for colorectal cancer [58]. *SP110* is involved in encoding for a leukocyte-specific body component that is responsible for generating an inflammatory response and miRNA expression in macrophages [59]. *XAF1* has a role in regulating apoptosis and tumor suppression [60]. *TRIM21*, often highly expressed in immune infiltrates such as macrophages, is involved in the amino acid metabolism responsible for suppressing colorectal cancer tumorigenesis [61]. There is limited research surrounding the specific roles of *APOL1* and *PML* in the context of macrophages or colorectal cancer; however, their broader immuno-modulatory functions may suggest an impact on macrophage behavior. *APOL1* is known to induce inflammation in the kidney and *PML* is known to negatively regulate cell growth through apoptosis and cell-cycle arrest [62, 63]. See Figure 4.6 and Table 4.1 for gene annotations of the refined signature derived from the Reactome and Metascape platforms [64, 65].

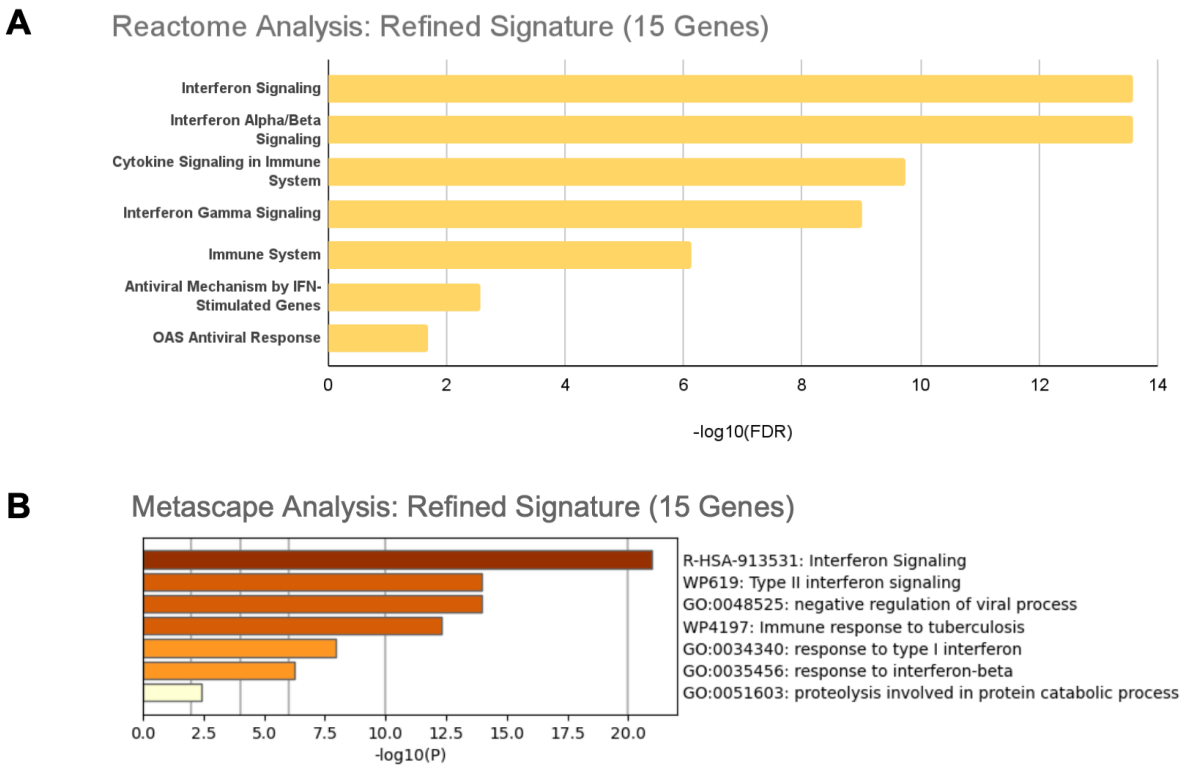



Figure 4.6: Reactome analysis and Metascape are used to perform gene annotations for the refined human signature [64, 65]. As expected, the macrophage genes in the refined signature are involved in the regulation of immune responses and signaling pathways that contribute to immune responses against pathogens (i.e: interferon signaling). A. Report of top functions of genes in the refined signature using Reactome Pathway analysis, based on the false discovery rate (FDR). B. Report of top functions of genes in the refined signature using the Metascape platform, based on the p-value (P).

Table 4.1: Biological gene annotations of refined immuno-reactive signature (Human) derived from the Metascape platform [65]. The table includes Gene Ontology (GO) enrichment analysis results (a database that stores the molecular function, cellular component, and biological process of known genes), Human Protein Atlas functions and sub-cellular location (database where human proteins found in cells, tissues, and organs are mapped using -omics technology), and involvement in known Hallmark genes (well-defined gene sets that represent specific biological states or processes).

| Gene Symbol | Entrez Gene ID | Species | Description | Biological Process (GO) | Protein Function (Protein Atlas) | Subcellular Location (Protein Atlas) | Hallmark Gene Sets |
|---|---|---|---|---|---|---|---|
| IFIT2 | 3433 | H. sapiens | interferon induced protein with tetratricopeptide repeats 2 | GO:0140374 antiviral innate immune response;GO:0008637 apoptotic mitochondrial changes;GO:0032091 negative regulation of protein binding | Predicted intracellular proteins | Vesicles (Approved) | (M5911)HALLMARK INTERFERON ALPHA RESPONSE; (M5890)HALLMARK TNFA SIGNALING VIA NFKB; (M5913)HALLMARK INTERFERON GAMMA RESPONSE |
| MX1 | 4599 | H. sapiens | MX dynamin like GTPase 1 | GO:0070106 interleukin-27-mediated signaling pathway;GO:0140374 antiviral innate immune response;GO:0034340 response to type I interferon | Predicted intracellular proteins | Cytosol;Nuclear membrane (Supported) | (M5911)HALLMARK INTERFERON ALPHA RESPONSE; (M5913)HALLMARK INTERFERON GAMMA RESPONSE; (M5956)HALLMARK KRAS SIGNALING DN |
| OAS1 | 4938 | H. sapiens | 2'-5'-oligoadenylate synthetase 1 | GO:0071659 negative regulation of IP-10 production;GO:0071658 regulation of IP-10 production;GO:2000342 negative regulation of chemokine (C-X-C motif) ligand 2 production | Enzymes; ENZYME proteins:Transferases; Predicted intracellular proteins | Cytosol;Nucleoplasm (Supported) | (M5911)HALLMARK INTERFERON ALPHA RESPONSE |
| IFIT3 | 3437 | H. sapiens | interferon induced protein with tetratricopeptide repeats 3 | GO:0140374 antiviral innate immune response;GO:0051607 defense response to virus;GO:0140546 defense response to symbiont | Predicted intracellular proteins | Cytosol;Mitochondria (Supported) | (M5911)HALLMARK INTERFERON ALPHA RESPONSE; (M5913)HALLMARK INTERFERON GAMMA RESPONSE |

Continued

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CXCL9 | 4283 | H. sapiens | C-X-C motif chemokine ligand 9 | GO:1901741 positive regulation of myoblast fusion;GO:1901739 regulation of myoblast fusion;GO:0060143 positive regulation of syncytium formation by plasma membrane fusion | Human disease related genes:Immune system diseases:Allergies and autoimmune diseases; Predicted secreted proteins; Cancer-related genes:Candidate cancer biomarkers | | (M5897)HALLMARK IL6 JAK STAT3 SIGNALING; (M5913)HALLMARK INTERFERON GAMMA RESPONSE; (M5932)HALLMARK INFLAMMATORY RESPONSE |
| XAF1 | 54739 | H. sapiens | XIAP associated factor 1 | GO:0035456 response to interferon-beta;GO:0032480 negative regulation of type I interferon production;GO:0032479 regulation of type I interferon production | Predicted intracellular proteins | Mitochondria;Nucleoplasm (Supported) | (M5913)HALLMARK INTERFERON GAMMA RESPONSE |
| SP110 | 3431 | H. sapiens | SP110 nuclear body protein | GO:0006357 regulation of transcription by RNA polymerase II;GO:0006355 regulation of DNA-templated transcription;GO:1903506 regulation of nucleic acid-templated transcription | Transcription factors:alpha-Helices exposed by beta-structures; Human disease related genes:Digestive system diseases:Liver diseases; Predicted intracellular proteins; Disease related genes | Nucleoplasm (Supported) | (M5911)HALLMARK INTERFERON ALPHA RESPONSE; (M5913)HALLMARK INTERFERON GAMMA RESPONSE |
| STAT1 | 6772 | H. sapiens | signal transducer and activator of transcription 1 | GO:0046725 negative regulation by virus of viral protein levels in host cell;GO:0072308 negative regulation of metanephric nephron tubule epithelial cell differentiation;GO:0003340 negative regulation of mesenchymal to epithelial transition involved in metanephros morphogenesis | Human disease related genes:Immune system diseases:Primary immunodeficiency; Disease related genes; Human disease related genes:Immune system diseases:Other immune system diseases; Predicted intracellular proteins; Transcription factors:Immunoglobulin fold | Nucleoplasm (Enhanced); Additional: Cytosol | (M5897)HALLMARK IL6 JAK STAT3 SIGNALING; (M5913)HALLMARK INTERFERON GAMMA RESPONSE; (M5950)HALLMARK ALLOGRAFT REJECTION |
| OAS2 | 4939 | H. sapiens | 2'-5'-oligoadenylate synthetase 2 | GO:1903487 regulation of lactation;GO:0070106 interleukin-27-mediated signaling pathway;GO:0060700 regulation of ribonuclease activity | Enzymes; ENZYME proteins:Transferases; Predicted intracellular proteins | Centrosome (Approved) | (M5913)HALLMARK INTERFERON GAMMA RESPONSE |

Continued

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TAP1 | 6890 | H. sapiens | transporter 1, ATP binding cassette subfamily B member | GO:0046967 cytosol to endoplasmic reticulum transport;GO:0019885 antigen processing and presentation of endogenous peptide antigen via MHC class I;GO:0002483 antigen processing and presentation of endogenous peptide antigen | Human disease related genes:Immune system diseases:Primary immunodeficiency; Disease related genes; Transporters:Primary Active Transporters; Potential drug targets; Predicted intracellular proteins; Cancer-related genes:Mutated cancer genes | Centriolar satellite;Endoplasmic reticulum (Approved) | (M5911)HALLMARK INTERFERON ALPHA RESPONSE; (M5941)HALLMARK UV RESPONSE UP; (M5902)HALLMARK APOPTOSIS |
| APOL1 | 8542 | H. sapiens | apolipoprotein L1 | GO:0051838 cytolysis by host of symbiont cells;GO:0051873 killing by host of symbiont cells;GO:1902476 chloride transmembrane transport | Disease related genes; Potential drug targets; Candidate cardiovascular disease genes; Predicted secreted proteins; Transporters:Transporter channels and pores; Human disease related genes:Urinary system diseases:Kidney diseases | | |
| OAS3 | 4940 | H. sapiens | 2'-5'-oligoadenylate synthetase 3 | GO:0035394 regulation of chemokine (C-X-C motif) ligand 9 production;GO:0035395 negative regulation of chemokine (C-X-C motif) ligand 9 production;GO:0039530 MDA-5 signaling pathway | Enzymes; ENZYME proteins:Transferases; Predicted intracellular proteins | Cytosol;Nucleoplasm (Supported); Additional: Plasma membrane | (M5913)HALLMARK INTERFERON GAMMA RESPONSE |
| TRIM21 | 6737 | H. sapiens | tripartite motif containing 21 | GO:0090086 negative regulation of protein deubiquitination;GO:0085020 protein K6-linked ubiquitination;GO:0090085 regulation of protein deubiquitination | Enzymes; ENZYME proteins:Transferases; Predicted intracellular proteins | Nucleoplasm (Approved) | (M5911)HALLMARK INTERFERON ALPHA RESPONSE; (M5913)HALLMARK INTERFERON GAMMA RESPONSE |
| PML | 5371 | H. sapiens | PML nuclear body scaffold | GO:0090402 oncogene-induced cell senescence;GO:0007182 common-partner SMAD protein phosphorylation;GO:0030578 PML body organization | Cancer-related genes:Candidate cancer biomarkers; Human disease related genes:Cancers:Cancers of haematopoietic and lymphoid tissues; Predicted intracellular proteins; Disease related genes | Nuclear bodies (Enhanced) | (M5919)HALLMARK HEDGEHOG SIGNALING; (M5901)HALLMARK G2M CHECKPOINT; (M5913)HALLMARK INTERFERON GAMMA RESPONSE |

Continued

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ISG15 | 9636 | H. sapiens | ISG15 ubiquitin like modifier | GO:0032461 positive regulation of protein oligomerization;GO:0032459 regulation of protein oligomerization;GO:0032020 ISG15-protein conjugation | Predicted secreted proteins; Human disease related genes:Immune system diseases:Primary immunodeficiency; Predicted intracellular proteins; Disease related genes | | (M5911)HALLMARK INTERFERON ALPHA RESPONSE; (M5913)HALLMARK INTERFERON GAMMA RESPONSE |


## Acknowledgements

Chapter 4, in part, is currently being prepared for submission for publication of the material. Dadlani, Ekta; Dash, Tirtharaj; Sahoo, Debashis. The thesis author is the primary researcher and author of this material.

# CHAPTER 5: DISCUSSION AND CONCLUSION

While the refined signature has the capacity to predict the prognosis of colorectal cancer (i.e: differentiate colorectal cancer macrophage cells from normal colon macrophage cells) at the single-cell macrophage level, at the non-macrophage specific pseudo-bulk level, and at the large cohort bulk level, additional analysis can be used to capture the universality of the refined signature. Additional translational predictive analysis performed on MSS/MSI annotated colon cancer datasets, CIMP+/CIMP- annotated colon cancer datasets, and RSCC/LSCC annotated colon cancer datasets confirms that the refined signature also has the capacity to capture the dynamic range of immuno-reactivity within the cancer. The datasets used for translational analysis is provided in Table 5.1.

Table 5.1: General information of datasets used for translational analysis of the refined signature. The two types of translational analysis performed is the predictions for MSS & MSI colorectal cancer sub-types which has the potential to claim immunotherapy treatment outcome in CRC patients, the predictions for CIMP+ & CIMP- colorectal cancer sub-types, and the predictions for tumors localized in the right-side of the colon (RSCC) & left-side of the colon (LSCC).

| GEO ID | Species | n samples | n samples (with required annotations for analysis) | Analysis |
|---|---|---|---|---|
| TCGA 2017 COAD mRNA | Homo Sapien | 521 | 265 (for MSS/MSI)<br>371 (for CIMP+/CIMP-) | MSS/MSI<br>CIMP+/CIMP- |
| GPL570 | Homo Sapien | 1353 | 221 | MSS/MSI |
| GSE42284 | Homo Sapien | 188 | 90 | MSS/MSI |
| Pooled Dataset: GSE13294, GSE13067, GSE35896, GSE26682, GSE24514 | Homo Sapien | 671<br>GSE13294: 155,<br>GSE13067: 74,<br>GSE35896: 62,<br>GSE26682: 331,<br>GSE24514: 49 | 542 | MSS/MSI |
| E-TABM-328 | Homo Sapien | 54 | 51 (for CIMP+/CIMP-)<br>50 (for RSCC/LSCC) | CIMP+/CIMP-<br>RSCC/LSCC |
| GSE39582 | Homo Sapien | 585 | 516 | CIMP+/CIMP- |
| GSE39084 | Homo Sapien | 70 | 69 (for CIMP+/CIMP-)<br>61 (for RSCC/LSCC) | CIMP+/CIMP-<br>RSCC/LSCC |
| GSE31595 | Homo Sapien | 37 | 37 | RSCC/LSCC |
| GSE72970 | Homo Sapien | 124 | 81 | RSCC/LSCC |

*Translational Prediction for Microsatellite Instability (MSS/MSI)*

Colorectal cancer is a complex disease because it can be caused by different genetic and epigenetic alterations; despite this heterogeneity, researchers use tumor classification allows for the identification of unique characteristics of the pathogenesis of the cancer. Tumor classification is based on clinical, pathological, and molecular features, one being microsatellite instability

status [66]. The DNA microsatellite instability captures the evolution of CRC on the basis of the mismatch repair system deficiency (MMR) and epigenetic patterns; a loss in the MMR function, or the fundamental DNA repair mechanism during DNA replication and recombination, leads to a high frequency of frameshift mutations in the microsatellite DNA that are correlated with the emergence of CRC [67]. I hypothesize that the refined TAMs immuno-reactive signature would be able to capture the dynamics of microsatellite instability (MSI) tumors in colorectal cancer because they are associated with an increased infiltration of immune cells in their microenvironments, including TAMs. The MSI status is an approved clinical biomarker associated with the prediction of immunotherapy outcome. The stable microsatellite state (MSS) does not trigger the body's immune response towards the tumor; they typically do not respond to immunotherapy treatments. The unstable microsatellite state (MSI) corresponds to an unstable tumor due to the erroneous DNA. High amounts of microsatellite instability are found in about 15% of all CRC tumors [67]. Comprehensive predictive analysis of the gene signature across several microsatellite instability tumors, or sub-types of colorectal cancer that are studied based on the tumor's DNA, is provided (see Figure 5.1). The refined gene signature has the capacity to predict the differentiation of the microsatellite stable (MSS) and microsatellite instable (MSI) states, thus translating to whether or not the CRC patient would respond to immunotherapy treatment. Predictions of the differentiation of MSI and MSS samples using the CRC TAMs reactive signature is performed for the datasets TCGA 2017 COAD mRNA (Colorectal Adenoma mRNA); GPL570; a pooled dataset consisting of GSE13294, GSE13067, GSE35896, GSE26682, and GSE24514 (see Figure 5.1). In general, a statistically significant and consistent trend shows that colorectal cancer samples annotated as MSI are more immuno-reactive

compared to the samples annotated as MSS using the gene signature. The higher expression of M1 macrophage genes in the MSI samples suggests that immunotherapy works better in microsatellite annotated samples that are associated with a release of pro-inflammatory cytokines, anti-tumor immunity, and the presence of reactive oxygen species. MSI subtype colon cancer patients have inactivated TGF-β and Wnt-β-catenin signaling pathways, which are known to reduce the sensitivity to immune therapies (i.e. PD-1 checkpoint blockade therapy) [68].

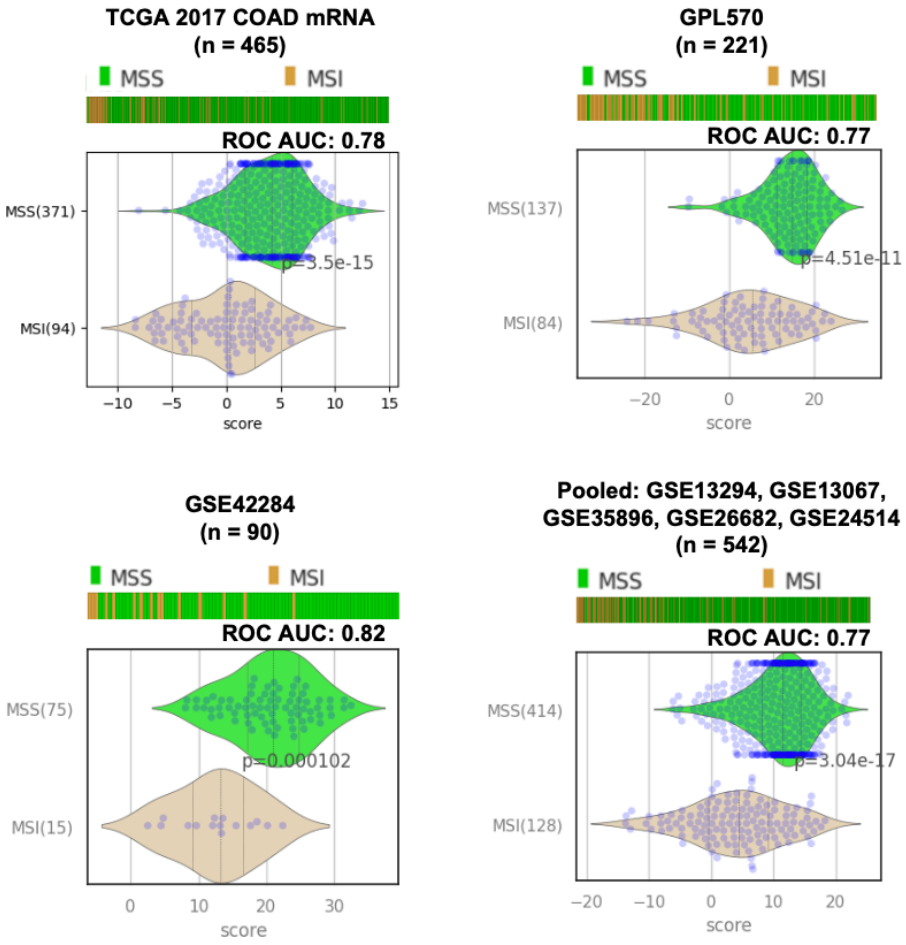



Figure 5.1: Translational Analysis: Predictive Potential For MSS/MSI Annotations. I test for the potential of the refined signature to predict the classifications of MSS and MSI annotations across the datasets TCGA 2017 (COAD mRNA), GPL570, GSE42284, and a pooled dataset (GSE13294, GSE13067, GSE35896, GSE26682, and GSE24514). Using the ROC-AUC metric and the normal (z) test's p-value, a consistent and statistically significant trend that samples annotated as MSI are more reactive than MSS samples using the refined immuno-reactive signature specified for CRC is shown. Literature review relating MSI and macrophage polarization suggests that MSI patients are more sensitive to anti-PD-1 antibodies [68].

*Translational Prediction for CpG Island Methylator Phenotype (CIMP)*

In addition to mutations in the DNA repair mechanism, colorectal cancer can develop through global genome hypermethylation which silences tumor suppressor genes, leads to oncogene activation, and causes chromosomal instability. CpG islands contain promoters used to regulate the activity of human genes; CpG dinucleotides that are methylated in normal cells are often found to be unmethylated or hypermethylated in cancer cells. Colorectal cancer can be classified according to the extreme degrees of hypermethylation. The CpG island methylation phenotype (CIMP) is an epigenetic alteration characterized by the hypermethylation of promoter CpG island sites, resulting in the inactivation or dysregulation of key tumor suppressor genes [69]. CIMP status is determined by the hypermethylation of gene markers, such as *CDKN2A* and *SOCS1*. Cancer samples with a high degree of methylation (CIMP+) are characterized by epigenetic instability and poor prognosis; CIMP+ classified tumors are associated with *BRAF* (high) and *TP53* (low) mutations, proximal tumor location, females, poor differentiation, and high MSI [66, 70]. The epigenetic silencing of MMR genes, which can be caused by hypermethylation of the MLH1 promoter related to CIMP, leads to sporadic MSI tumors [71]. Cancer samples with a no degree of methylation (CIMP-) are associated with wildtype *BRAF* [66]. Higher densities of both M1-like and M2-like macrophages are associated with a CIMP+ phenotype [72]. In a study conducted by Edin et. al, where *NOS2+* is used as a marker for M1 macrophage phenotype and *CD163+* is used as a marker for M2 macrophage phenotype, M2 macrophages showed a significant effect on the prognosis in CIMP- and CIMP+ (specifically CIMP high) cases [73]. The predicted classification of CIMP- and CIMP+ samples using the refined signature is tested across the CRC datasets: E-TABM-328, GSE39582, GSE39084, and

the TCGA 2017 COAD mRNA cohort. The refined TAMs signature proposes that CIMP+ samples are more immuno-reactive compared to the CIMP- samples in a statistically significant manner; this supports the relationship between CIMP+ status and MSI status, which is also more immuno-reactive compared to MSS samples (See Figure 5.2).

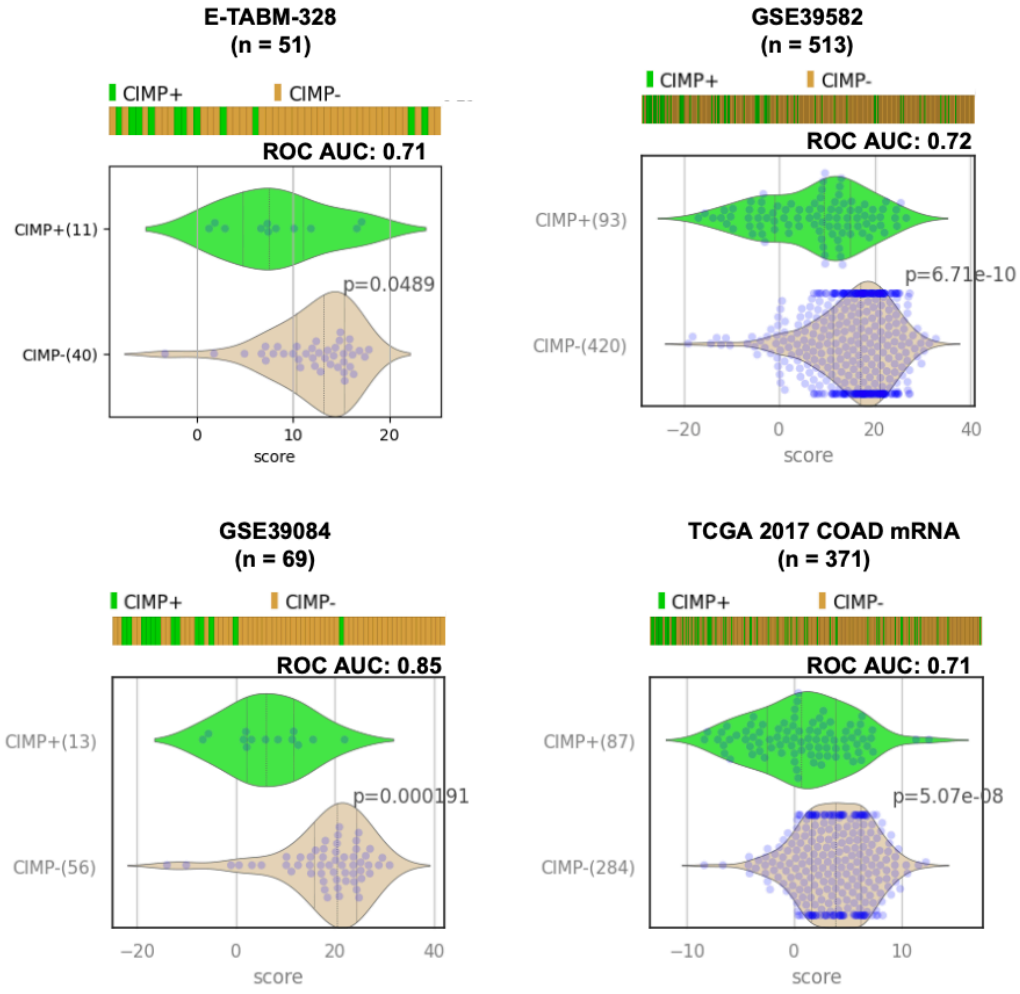



Figure 5.2: Translational Analysis: Predictive Potential For CIMP+/CIMP- Annotations. I test for the potential of the refined signature to predict the classifications of CIMP+ and CIMP- annotations across the datasets E-TABM-328, GSE39582, GSE39084, and TCGA 2017 (COAD mRNA). Using the ROC-AUC metric and the normal (z) test's p-value, a consistent and statistically significant trend that samples annotated as CIMP+ are more reactive than MSS samples using the refined immuno-reactive signature specified for CRC is shown.

*Translational Prediction for Localization of Colorectal Cancer (RSCC/LSCC)*

With gaining evidence that heterogeneity of colorectal cancer spans a diverse range of molecular factors, the location of the tumor within the colon has been noted to play a role in the disease's progression. Right-side colon cancers (RSCC), or proximal tumors, stem from the cecum, ascending colon, and transverse colon (hepatic flexure); left-side colon cancers (LSCC), or distal tumors, occur in the descending colon, sigmoid colon, or splenic flexure. Proximal tumors and distal tumors are suggested to be clinically, pathologically, and transcriptionally different [74]. Recent studies have shown that RSCC patients are correlated with a poor prognosis compared to LSCC patients, due to their likelihood of exhibiting advanced tumor growth, an increase in tumor size, increased hypermethylation, and poorly differentiated tumors [75, 76]. Mechanistically, over expression of *PRAC* (PRAC1), which is a heavily transcribed gene in a healthy prostate, distal colon, and rectum, and upregulation of adjacent genes *HOXB13* and *PRAC2* in RSCC patients suggest regulatory mechanisms that lead to proliferation and tumor growth; RSCC is associated with a genotoxic tumor environment and a more aggressive phenotype [77]. The refined TAMs signature provides further evidence to support this finding in the datasets GSE31595, GSE39084, E-TABM-328, and GSE72970, predicting that CRC originating in the right colon exhibits the M1-like, or immuno-reactive and pro-inflammatory, phenotype compared to CRC originating in the left colon (See Figure 5.3). This finding suggests that there may be some involvement of macrophage polarization (particularly the M1-like phenotype) that contributes to the poor prognosis of right-side originating colorectal cancer.

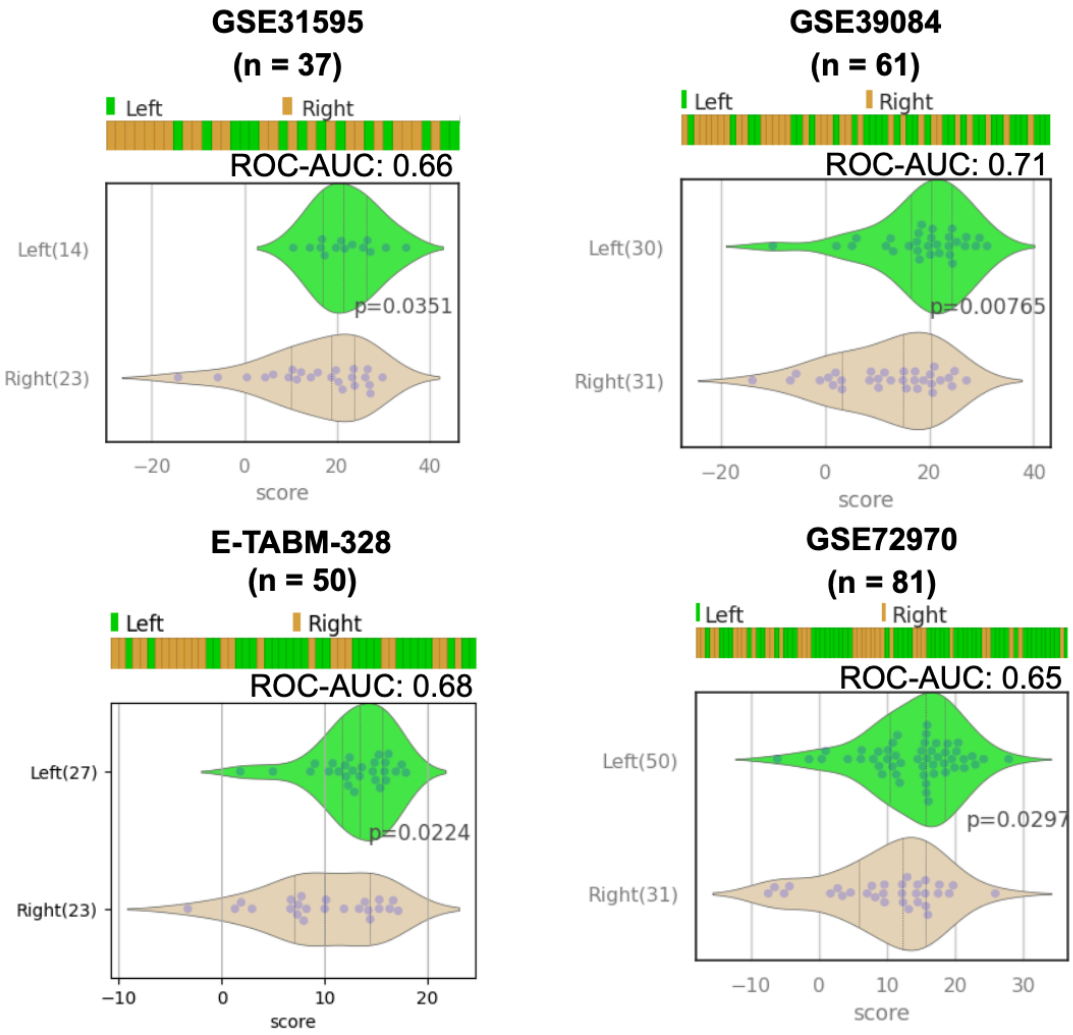



Figure 5.3: Translational Analysis: Predictive Potential For Tumor Localization. Literature review suggests that RSCC patients have a more aggressive phenotype compare to LSCC patients [77]. The refined TAMs lewdly captures an increase immuno-reactivity in RSCC patients, which is consistent with the notion that colorectal cancer tumor macrophages have a poor prognosis and are more immuno-reactive compared to healthy colon macrophages.

*Limitations*

While there is availability of multiple single cell RNAseq human datasets annotated with colorectal cancer tissue samples and normal colon tissue samples for which there is a scalable and comparable number of macrophage cells that fall into these two classifications, a lack of publicly available single cell RNAseq mouse datasets annotated with colorectal cancer tissue samples and normal colon tissue samples prevent me from making the same conclusive remarks with equal confidence. This is because there are not many publicly available datasets with consistent clean boundaries between healthy and cancerous mouse samples, and because within the available datasets, there are a substantially lower number of macrophage specific healthy samples as compared to the cancerous ones.

Limitations of differential gene expression (DEG) analysis includes that the genes available for analysis are limited to the genes expressed under the training dataset. Despite this limitation, I am able to generate a list of 15 genes that are expressed in similar patterns across multiple colorectal cancer datasets. While DEG analysis provides genes that are up-or-down regulated under a specific condition, it does not provide any inferences on the biological processes or pathways for which the highly regulated genes are involved in. Therefore, I utilize secondary tools, the Reactome Pathway and Metascape platforms, and literature review to better characterize the functions of the genes in the refined signature. As a result, further investigations are warranted to elucidate the biological significance of the genes that make up the refined signature and assess their potential as therapeutic targets in personalized medicine for colorectal cancer patients.

**Acknowledgements**

Chapter 5, in full, is currently being prepared for submission for publication of the material. Dadlani, Ekta; Dash, Tirtharaj; Sahoo, Debashis. The thesis author is the primary researcher and author of this material.

# REFERENCES

[1] Mármol, I., Sánchez-de Diego, C., Pradilla Dieste, A., Cerrada, E. & Rodriguez Yoldi, M. J. Colorectal carcinoma: a general overview and future per- spectives in colorectal cancer. International journal of molecular sciences 18, 197 (2017).

[2] Ciardiello, F., Ciardiello, D., Martini, G., Napolitano, S., Tabernero, J. & Cervantes, A. Clinical management of metastatic colorectal cancer in the era of precision medicine. CA: a cancer journal for clinicians 72, 372–401 (2022).

[3] Xi, Y. & Xu, P. Global colorectal cancer burden in 2020 and projections to 2040. Translational oncology 14, 101174 (2021).

[4] Binnewies, M., Roberts, E.W., Kersten, K., Chan, V., Fearon, D.F., Merad, M., Coussens, L.M., Gabrilovich, D.I., Ostrand-Rosenberg, S., Hedrick, C.C. & Vonderheide, R.H. Understanding the tumor immune microenvironment (time) for effective therapy. Nature medicine 24, 541–550 (2018).

[5] Shinji, S., Yamada, T., Matsuda, A., Sonoda, H., Ohta, R., Iwai, T., Takeda, K., Yonaga, K., Masuda, Y. & Yoshida, H. Recent advances in the treatment of colorectal cancer: A review. Journal of Nippon Medical School 89, 246–254 (2022).

[6] Maniewska, J. & Jezëwska, D. Non-steroidal anti-inflammatory drugs in colorectal cancer chemoprevention. Cancers 13, 594 (2021).

[7] Zhong, X., Chen, B. & Yang, Z. The role of tumor-associated macrophages in colorectal carcinoma progression. Karger Cellular Physiology and Biochemistry 45, 356–365 (2018).

[8] Qian, B.-Z. & Pollard, J. W. Macrophage diversity enhances tumor progression and metastasis. Cell 141, 39–51 (2010).

[9] Gordon, S. Macrophages and the immune response. Fundamental immunology (2003).

[10] Mosser, D. M. & Edwards, J. P. Exploring the full spectrum of macrophage activation. Nature reviews immunology 8, 958–969 (2008).

[11] Murray, P.J., Allen, J.E., Biswas, S.K., Fisher, E.A., Gilroy, D.W., Goerdt, S., Gordon, S., Hamilton, J.A., Ivashkiv, L.B., Lawrence, T. & Locati, M. Macrophage activation and polarization: nomenclature and experimental guidelines. Immunity 41, 14–20 (2014).

[12] Martinez, F. O., Sica, A., Mantovani, A. & Locati, M. Macrophage activation and polarization. Frontiers in Bioscience-Landmark 13, 453–461 (2008).

[13] Ghosh, P., Sinha, S., Katkar, G.D., Vo, D., Taheri, S., Dang, D., Das, S. & Sahoo, D. Machine learning identifies signatures of macrophage reactivity and tolerance that predict disease outcomes. bioRxiv, (2022).

[14] Society, A. C. Colorectal cancer (2023).

[15] Noone, A.M., Howlader, N., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D.R. & Chen, H.S. Surveillance, epidemiology, and end results (seer) program cancer statistics review, 1975-2015, national cancer institute. bethesda, md (2018).

[16] Ma, Y., Yang, Y., Wang, F., Zhang, P., Shi, C., Zou, Y. & Qin, H. Obesity and risk of colorectal cancer: a systematic review of prospective studies. PloS one 8, e53916 (2013).

[17] Keum, N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nature reviews Gastroenterology & hepatology 16, 713–732 (2019).

[18] Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. Cell 87, 159–170 (1996).

[19] Moghimi-Dehkordi, B. & Safaee, A. An overview of colorectal cancer survival rates and prognosis in asia. World journal of gastrointestinal oncology 4, 71 (2012).

[20] Fleming, M., Ravula, S., Tatishchev, S. F. & Wang, H. L. Colorectal carcinoma: Pathologic aspects. Journal of gastrointestinal oncology 3, 153 (2012).

[21] Grothey, A., Sobrero, A., Shields, A., Yoshino, T., Paul, J., Taieb, J., Souglakos, J., Shi, Q., Kerr, R., Labianca, R., Meyerhardt, J. & Vernerey, D. Duration of adjuvant chemotherapy for stage iii colon cancer. New England Journal of Medicine 378, 1177–1188 (2018).

[22] André, T., Boni, C., Mounedji-Boudiaf, L., Navarro, M., Tabernero, J., Hickish, T., Topham, C., Zaninelli, M., Clingan, P., Bridgewater, J., Tabah-Fisch, I. & Gramont, A. Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer. New England Journal of Medicine 350, 2343–2351 (2004).

[23] Tournigand, C., André, T., Achille, E., Lledo, G., Flesh, M., Mery-Mignard, D., Quinaux, E., Couteau, C., Buyse, M., Ganem, G., Landi, B., Colin, P., Louvet, C. & Gramont, A. Folfiri followed by folfox6 or the reverse sequence in advanced colorectal cancer: a randomized gercor study. Journal of Clinical Oncology 22, 229–237 (2004).

[24] Rapiti, E., Fioretta, G., Verkooijen, H.M., Zanetti, R., Schmidlin, F., Shubert, H., Merglen, A., Miralbell, R. & Bouchardy, C. Increased risk of colon cancer after external radiation therapy for prostate cancer. International journal of cancer 123, 1141–1145 (2008).

[25] Cummings, B. J. Adjuvant radiation therapy for colorectal cancer. Cancer 70, 1372–1383 (1992).

[26] Krzyszczyk, P., Schloss, R., Palmer, A. & Berthiaume, F. The role of macrophages in acute and chronic wound healing and interventions to promote pro-wound healing phenotype. Frontiers in physiology 9, 419 (2018).

[27] Epelman, S., Lavine, K. J. & Randolph, G. J. Origin and functions of tissue macrophages. Immunity 41, 21–35 (2014).

[28] Mantovani, A., Sozzani, S., Locati, M., Allavena, P. & Sica, A. Macrophage polarization: tumor-associated macrophages as a paradigm for polarized m2 mononuclear phagocytes. Trends in immunology 23, 549–555 (2002).

[29] Quail, D. F. & Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. Nature medicine 19, 1423–1437 (2013).

[30] Tlsty, T. D. & Coussens, L. M. Tumor stroma and regulation of cancer development. Annu. Rev. Pathol. Mech. Dis. 1, 119–150 (2006).

[31] Ruffell, B. & Coussens, L. M. Macrophages and therapeutic resistance in cancer. Cancer cell 27, 462–472 (2015).

[32] Wang, H., Tian, T. & Zhang, J. Tumor-associated macrophages (tams) in col- orectal cancer (crc): from mechanism to therapy and prognosis. International journal of molecular sciences 22, 8470 (2021).

[33] Larionova, I., Tuguzbaeva, G., Ponomaryova, A., Stakheyeva, M., Cherdyntseva, N., Pavlov, V., Choinzonov, E. & Kzhyshkowska, J. Tumor-associated macrophages in human breast, colorectal, lung, ovarian and prostate cancers. Frontiers in oncology 10, 566511 (2020).

[34] Erreni, M., Mantovani, A. & Allavena, P. Tumor-associated macrophages (tam) and inflammation in colorectal cancer. Cancer microenvironment 4, 141–154 (2011).

[35] Krasinskas, A. M. Egfr signaling in colorectal carcinoma. Pathology research international 2011 (2011).

[36] Li, X., Wu, Y. & Tian, T. Tgf-β signaling in metastatic colorectal cancer (mcrc): From underlying mechanism to potential applications in clinical development. International Journal of Molecular Sciences 23, 14436 (2022).

[37] Ghojogh, B. & Crowley, M. The theory behind over fitting, cross validation, regularization, bagging, and boosting: tutorial. arXiv preprint arXiv:1905.12787 (2019).

[38] Li, C., Zhang, S., Qin, Y. & Estupinan, E. A systematic review of deep transfer learning for machinery fault diagnosis. Neurocomputing 407, 121–135 (2020).

[39] Bedre, R. Gene expression units explained: Rpm, rpkm, fpkm, tpm, deseq, tmm, scnorm, getmm, and combat-seq (2017).

[40] Li, P., Piao, Y., Shon, H. S. & Ryu, K. H. Comparing the normalization meth- ods for the differential analysis of illumina high-throughput rna-seq data. BMC bioinformatics 16, 1–9 (2015).

[41] Sahoo, D., Dill, D. L., Tibshirani, R. & Plevritis, S. K. Extracting binary signals from microarray time-course data. Nucleic acids research 35, 3705–3712 (2007).

[42] Sahoo, D., Dill, D. L., Gentles, A. J., Tibshirani, R. & Plevritis, S. K. Boolean implication networks derived from large scale, whole genome microarray datasets. Genome biology 9, 1–17 (2008).

[43] Sahoo, D., Seita, J., Bhattacharya, D., Inlay, M.A., Weissman, I.L., Plevritis, S.K. & Dill, D.L. Midreg: a method of mining developmentally regulated genes using boolean implications. Proceedings of the National Academy of Sciences 107, 5732–5737 (2010).

[44] Dang, D., Taheri, S., Das, S., Ghosh, P., Prince, L.S. & Sahoo, D. Computational approach to identifying universal macrophage biomarkers. Frontiers in physiology 11, 275 (2020).

[45] Sahoo, D., Swanson, L., Sayed, I.M., Katkar, G.D., Ibeawuchi, S.R., Mittal, Y., Pranadinata, R.F., Tindle, C., Fuller, M., Stec, D.L. & Chang, J.T. Artificial intelligence guided discovery of a barrier-protective therapy in inflammatory bowel disease. Nature communications 12, 4246 (2021).

[46] Marzban, C. The roc curve and the area under it as performance measures. Weather and Forecasting 19, 1106–1114 (2004).

[47] Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W. & Edgar, R. Ncbi geo: mining millions of expression profiles—database and tools. Nucleic acids research 33, D562–D566 (2005).

[48] Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M. & Yefanov, A. Ncbi geo: archive for functional genomics data sets—update. Nucleic acids research 41, D991–D995 (2012).

[49] Edgar, R., Domrachev, M. & Lash, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic acids research 30, 207–210 (2002).

[50] Li, B. & Dewey, C. N. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC bioinformatics 12, 1–16 (2011).

[51] Pachter, L. Models for transcript quantification from rna-seq. arXiv preprint arXiv:1104.3889 (2011).

[52] Nahm, F. S. Nonparametric statistical tests for the continuous data: the basic concept and the practical use. Korean journal of anesthesiology 69, 8–14 (2016).

[53] Min, J., Liu, W. & Li, J. Emerging role of interferon-induced noncoding rna in innate antiviral immunity. Viruses 14, 2607 (2022).

[54] Chang, J.J., Woods, M., Lindsay, R.J., Doyle, E.H., Griesbeck, M., Chan, E.S., Robbins, G.K., Bosch, R.J. & Altfeld, M. Higher expression of several interferon-stimulated genes in hiv-1-infected females after adjusting for the level of viral replication. The Journal of infectious diseases 208, 830–838 (2013).

[55] Kane, M., Zang, T.M., Rihn, S.J., Zhang, F., Kueck, T., Alim, M., Schoggins, J., Rice, C.M., Wilson, S.J. & Bieniasz, P.D. Identification of interferon-stimulated genes with antiretroviral activity. Cell host & microbe 20, 392–405 (2016).

[56] Schoggins, J. W. Interferon-stimulated genes: roles in viral pathogenesis. Current opinion in virology 6, 40–46 (2014).

[57] Zhang, D. & Zhang, D.E., 2011. Interferon-stimulated gene 15 and the protein ISGylation system. Journal of interferon & cytokine research, 31(1), pp.119-130.

[58] Wu, Z., Huang, X., Han, X., Li, Z., Zhu, Q., Yan, J., Yu, S., Jin, Z., Wang, Z., Zheng, Q. & Wang, Y. The chemokine cxcl9 expression is associated with better prognosis for colorectal carcinoma patients. Biomedicine & Pharmacotherapy 78, 8–13 (2016).

[59] Gerovska, D., Larrinaga, G., Solano-Iturri, J.D., Márquez, J., García Gallastegi, P., Khatib, A.M., Poschmann, G., Stühler, K., Armesto, M., Lawrie, C.H. & Badiola, I. An integrative omics approach reveals involvement of brca1 in hepatic metastatic progression of colorectal cancer. Cancers 12, 2380 (2020).

[60] Moon, J. R., Oh, S. J., Lee, C. K., Chi, S. G. & Kim, H. J. Tgf-β1 protects colon tumor cells from apoptosis through xaf1 suppression. International journal of oncology 54, 2117–2126 (2019).

[61] Chen, X., Cao, M., Wang, P., Chu, S., Li, M., Hou, P., Zheng, J., Li, Z. & Bai, J. The emerging roles of trim21 in coordinating cancer metabolism, immunity and cancer treatment. Frontiers in Immunology 13 (2022).

[62] Fang, J., Yao, X., Hou, M., Duan, M., Xing, L., Huang, J., Wang, Y., Zhu, B., Chen, Q. & Wang, H. Apol1 induces kidney inflammation through rig-i/nf-κb activation. Biochemical and Biophysical Research Communications 527, 466–473 (2020).

[63] Yamada, N., Tsujimura, N., Kumazaki, M., Shinohara, H., Taniguchi, K., Nakagawa, Y., Naoe, T. & Akao, Y. Colorectal cancer cell-derived microvesicles containing microrna-1246 promote angiogenesis by activating smad 1/5/8 signaling elicited by pml down-regulation in endothelial cells. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 1839, 1256–1272 (2014).

[64] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B. and Milacic, M., 2018. The reactome pathway knowledgebase. Nucleic acids research, 46(D1), pp.D649-D655.

[65] Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C. & Chanda, S.K. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature communications 10, 1523 (2019).

[66] Ogino, S. & Goel, A. Molecular classification and correlates in colorectal cancer. The Journal of Molecular Diagnostics 10, 13–27 (2008).

[67] Zeinalian, M., Hashemzadeh-Chaleshtori, M., Salehi, R. & Emami, M. H. Clin- ical aspects of microsatellite instability testing in colorectal cancer. Advanced biomedical research 7 (2018).

[68] Wang, H., Wang, X., Xu, L., Zhang, J. & Cao, H. Analysis of the transcriptomic features of microsatellite instability subtype colon cancer. BMC cancer 19, 1–16 (2019).

[69] Mojarad, E. N., Kuppen, P. J., Aghdaei, H. A. & Zali, M. R. The cpg island methylator phenotype (cimp) in colorectal cancer. Gastroenterology and hepatology from bed to bench 6, 120 (2013).

[70] Rhee, Y.-Y., Kim, K.-J. & Kang, G. H. Cpg island methylator phenotype-high colorectal cancers and their prognostic implications and relationships with the serrated neoplasia pathway. Gut and liver 11, 38 (2017).

[71] Picard, E., Verschoor, C. P., Ma, G. W. & Pawelec, G. Relationships between immune landscapes, genetic sub-types and responses to immunotherapy in col- orectal cancer. Frontiers in immunology 11, 369 (2020).

[72] Väyrynen, J.P., Haruki, K., Lau, M.C., Väyrynen, S.A., Zhong, R., Dias Costa, A., Borowsky, J., Zhao, M., Fujiyoshi, K., Arima, K. & Twombly, T.S. The prognostic role of macrophage polarization in the colorectal cancer microenvironment macrophage polarization in colorectal cancer. Cancer immunology research 9, 8–19 (2021).

[73] Edin, S., Wikberg, M.L., Dahlin, A.M., Rutegård, J., Öberg, Å., Oldenborg, P.A. & Palmqvist, R. The distribution of macrophages with a M1 or M2 phenotype in relation to prognosis and the molecular characteristics of colorectal cancer (2012).

[74] Mik, M., Berut, M., Dziki, L., Trzcinski, R. & Dziki, A. Right-and left-sided colon cancer–clinical and pathological differences of the disease entity in one organ. Archives of Medical Science 13, 157–162 (2017).

[75] Narayanan, S., Gabriel, E., Attwood, K., Boland, P. & Nurkin, S. Association of clinicopathologic and molecular markers on stage-specific survival of right versus left colon cancer. Clinical Colorectal Cancer 17, e671–e678 (2018).

[76] Koestler, D.C., Li, J., Baron, J.A., Tsongalis, G.J., Butterly, L.F., Goodrich, M., Lesseur, C., Karagas, M.R., Marsit, C.J., Moore, J.H. & Andrew, A.S. Distinct patterns of dna methylation in conventional adenomas involving the right and left colon. Modern Pathology 27, 145–155 (2014).

[77] Mukund, K., Syulyukina, N., Ramamoorthy, S. & Subramaniam, S. Right and left-sided colon cancers-specificity of molecular mechanisms in tumorigenesis and progression. BMC cancer 20, 1–15 (2020).