An AI-assisted Investigation of Tumor-Associated Macrophages and their Polarization in Colorectal Cancer

Ekta Dadlani, Tirtharaj Dash and Debashis Sahoo

Department of Computer Science and Pediatrics
University of California, San Diego
CA 92093, USA
E-mail: {edadlani, tdash, dsahoo}@ucsd.edu

Tumor-associated Macrophages (or TAMs) are amongst the most common cells that play a significant role in the initiation and progression of colorectal cancer (CRC). Recently, Ghosh et al.[1] proposed distinguishing signatures for identifying macrophage polarization states, namely, immuno-reactive and immuno-tolerant, using the concept of Boolean implications and Boolean networks. Their signature, called the Signature of Macrophage Reactivity and Tolerance (SMaRT), comprises of 338 human genes (equivalently, 298 mouse genes). However, SMaRT was constructed using datasets that were not specialized towards any particular disease. In this paper, (a) we perform a comprehensive analysis of the SMaRT signature on single-cell human and mouse colorectal cancer RNA-seq datasets; (b) we then adopt a technique akin to transfer learning to construct a "refined" SMaRT signature for investigating TAMs and their polarization in the CRC tumor microenvironment. Towards validation of our refined gene signature, we use (a) 5 pseudo-bulk RNA-seq datasets derived from single-cell human datasets; and (b) 5 large-cohort microarray datasets from humans. Furthermore, we investigate the translational potential of our refined gene signature in problems related to MSS/MSI (4 datasets) and CIMP+/CIMP- status (4 datasets). Overall, our refined gene signature and its extensive validation provide a path for its adoption in clinical practice in diagnosing colorectal cancer and associated attributes.

Keywords: Tumor Associated Macrophages; Colorectal Cancer; Macrophage Polarization; Boolean Implications; Invariant Relationships; Gene Signature Refinement

1. 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.[2,3] There were an estimated 1.9 million cases of newly diagnosed colorectal cancer and 0.9 million deaths globally in the year 2020.[4] 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.[5] Despite advanced clinical treatments, such as surgical treatment and immunotherapy, CRC has remained to be the second leading cause of cancer-related deaths worldwide.[2] As a result, scientists all around the globe are attempting different methodologies to understand the disease. The initiation and progression of CRC involve genetic alterations in both cancer cells and cells in the surrounding tumor microenvironment.

Preprint of an article submitted for consideration in Pacific Symposium on Biocomputing © 2024 World Scientific Publishing Co., Singapore, http://psb.stanford.edu/

The tumor microenvironment comprises of immune and stromal cells such as macrophages, dendritic cells, and T lymphocytes, all of which can contribute to chronic inflammation.[5] 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 TAMs in the tumor microenvironment support the progression of the normal colonic epithelium to adenomatous polyps, or the gland-like growths on the large intestine's membrane, to invasive colon carcinoma.[6]

Macrophages, or MΦ, 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). M1 macrophages exhibit pro-inflammatory properties in response to the release of inflammatory cytokines and the presence of oxygen species in 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.[7,8] It is conventionally understood that the stimuli in a microenvironment can activate M0 macrophages to adopt a tolerant-like 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. Ghosh et al.[1] propose a gene signature, called the Signature of Macrophage Reactivity and Tolerance (SMaRT), to capture the complex dynamics and clinical implications of macrophage phenotype in various contexts. SMaRT was constructed using a pooled human macrophage transcriptome that is represented across different tissues, organs, species, and immune cells. This makes the gene signature a "general" signature that captures the spectrum of macrophage polarization states. In this paper, we attempt to construct a disease-specific signature by specializing the general SMaRT signature towards CRC. Through computational experiments, we demonstrate that the new CRC-specific signature holds immense potential to be adopted in healthcare research and practice.

Overall, the major contributions of this paper are as follows: (1) We investigate the effectiveness of the original SMaRT signature in capturing the polarization dynamics of macrophage cells, specifically in single-cell RNA-seq colorectal cancer datasets; (2) We propose an idea similar to transfer learning to refine the original SMaRT signature which could increase the specificity and sensitivity of discriminating colorectal cancer samples from healthy colon samples; and (3) We extensively validate our refined signature to large-scale colorectal cancer datasets and demonstrate its translational potential in high-impact clinical studies such as MSS/MSI and CIMP+/CIMP-.

The rest of the paper is organized as follows. In section 2, we provide brief descriptions of the materials and method. We provide and discuss the computational results from single-cell analysis and signature refinement obtained using our experiments in section 3. We conclude our paper and highlight some limitations in section 4. The data, codes, and software packages used in our research are linked and shared at https://github.com/tirtharajdash/TAMs-CRC.

2. Materials and Method

2.1. *Data collection and annotation*

Six publicly available single-cell RNA sequencing (scRNA-seq) databases are downloaded from the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) database.[9,10] scRNA-seq captures the heterogeneity of RNA transcripts across individual cells; it allows for the understanding of a disease with higher resolution compared to bulk gene expression datasets. Gene expression counts are normalized by $\log_2$ scaling of CPM (counts per million) values.[11,12] The datasets are annotated with the classifications of originating from tumor tissue samples or healthy colon tissue samples, as noted in the metadata from the corresponding experiments. For simple nomenclature, colorectal cancer tumor samples are referred to as "colorectal cancer" (or CRC) samples and normal colorectal tissue, adjacent normal colonic tissue, and matched normal mucosa samples are referred as "normal tissue" (or NT) samples. The datasets used in our experiments are diverse with regard to species and data types. A summary of all datasets can be found in Table 1 which includes the usage of each dataset in the context of our experiments.

Table 1. A summary of publicly available single-cell and microarray datasets used in our study. Usage annotations are as follows: Sc: single-cell analysis; Tr: training towards refinement; Va: validation of refinement; Ta: translational analysis

GEO ID	Species	Type	n-samples	n-cells	Usage
GSE132465	Homo Sapien	Single-cell	33	63,689	Sc, Tr
GSE161277	Homo Sapien	Single-cell	13	54,782	Sc, Va
GSE200997	Homo Sapien	Single-cell	23	49,859	Sc, Va
GSE132257	Homo Sapien	Single-cell	10	18,409	Sc, Va
GSE139555	Homo Sapien	Single-cell	8*	12,495	Sc, Va
GSE222300	Homo Sapien	Single-cell	6*	14,852	Sc, Va
GSE198758	Mus Musculus	Single-cell	4	20,849	Sc
GSE224679	Mus Musculus	Single-cell	4	27,656	Sc
GPL570	Homo Sapien	Microarray	1832†, 221†	–	Va, Ta
TCGA 2017	Homo Sapien	Microarray	695†, 465†, 371†	–	Va, Ta
GSE20916	Homo Sapien	Microarray	70†	–	Ta
GSE44076	Homo Sapien	Microarray	19†	–	Ta
GSE62932	Homo Sapien	Microarray	68	–	Va
GSE42284	Homo Sapien	Microarray	90‡	–	Ta
E-TABM-328	Homo Sapien	Microarray	51‡	–	Ta
GSE39084	Homo Sapien	Microarray	69‡	–	Ta
Pooled**	Homo Sapien	Microarray	542‡	–	Ta
GSE39582	Homo Sapien	Microarray	513‡	–	Ta

*Artificially generated technical replicates
†Samples specific to colorectal cancer
‡Colorectal cancer specific samples annotated with status for specific translational analysis
**Source datasets: GSE13294, GSE13067, GSE35896, GSE26682, GSE24514

2.2. *Single-cell analysis using SMaRT*

2.2.1. *Preliminary on the SMaRT Model*

SMaRT[1] is an AI-based technique that utilizes the concept of Clustered Boolean Implication Network (CBIN[13]) to segregate the cellular states in macrophages within the immunologic spectrum (M1-and-M2-like phenotype). Boolean paths that intersect transitive implication relationships[13,14] between large clusters in the CBIN are used to capture the continuum of polarization from the reactive pole to the tolerant pole of the network. Specific enriched paths are selected for testing the classification of samples into the immuno-reactive and immuno-tolerant states in bulk RNA-seq and microarray macrophage datasets annotated with polarization states. The effectiveness of various Boolean paths in classifying the samples is assessed for sample classification. Multivariate analysis of the top Boolean paths shows that the path that connects clusters $C13 \rightarrow C14 \rightarrow C3$ performs the best at discriminating the M1-like and M2-like polarization states. Independently, C13 accurately predicts the reactivity (M1-like) state and the path $C14 \rightarrow C3$ demonstrates close to perfect prediction of the tolerant (M2-like) state. The $C13 \rightarrow C14 \rightarrow C3$, or SMaRT, signature consists of 48 human genes (equivalent to 71 mouse genes) in C13 and 290 human genes in the path $C14 \rightarrow C3$ (equivalent to 227 mouse genes). It successfully identifies the M1-like/M2-like polarization states under a diverse range of tissue-resident macrophages, in both human and mice, and in other immune cells. Figure 1 shows a pipeline for the construction of the SMaRT signature ($C13 \rightarrow C14 \rightarrow C3$) and its use in single-cell analysis, which is described in the following sections.

2.2.2. *Macrophage polarization in scRNA-seq datasets*

Due to the lack of available macrophage-specific CRC sequencing studies, we apply the SMaRT model on computationally filtered macrophage cells stemming from a diverse collection of scRNA-seq colorectal cancer experiments. The filtering of macrophage cells is performed using the universal macrophage biomarkers: FCER1G and TYROBP.[15] We use previously validated thresholds on these two genes; that is, $FCER1G > 2.0$ and $TYROBP > 2.0$. However, to get a reasonable number of cells for analysis in the case of GSE132465 and GSE224679, we apply a condition of greater than 0 on the macrophage biomarkers at the raw gene count level. Filtering macrophages in this manner significantly reduces the number of cells for each dataset resulting in approximately 8.6% of all cells in each dataset. For each dataset, all macrophage cells are annotated as tumor tissue (CRC) or normal colon/rectum tissue (NT). The 338-gene SMaRT signature is then applied to each macrophage-filtered scRNA-seq datasets to compare and analyze the dynamics of the immuno-reactive and immuno-tolerant samples (see Figure 2). That is, for each (macrophage) cell, the relationship between the machine learning-based composite scores[1] for the immuno-reactive cluster ($C13$) and the immuno-tolerant clusters ($C14$ and $C3$) is observed, as shown in the example scatter plots in Figure 2. The StepMiner algorithm[16] is used to binarize these two kinds of composite scores into *high* and *low* values. This approach partitions the space into 4 quadrants (as can be seen in the figure). For each dataset, the cells that are confined to the [C13-low, C14_3-low] quadrant are classified as *highly reactive* macrophages; and the cells that are confined to the [C13-high, C14_3-high] are

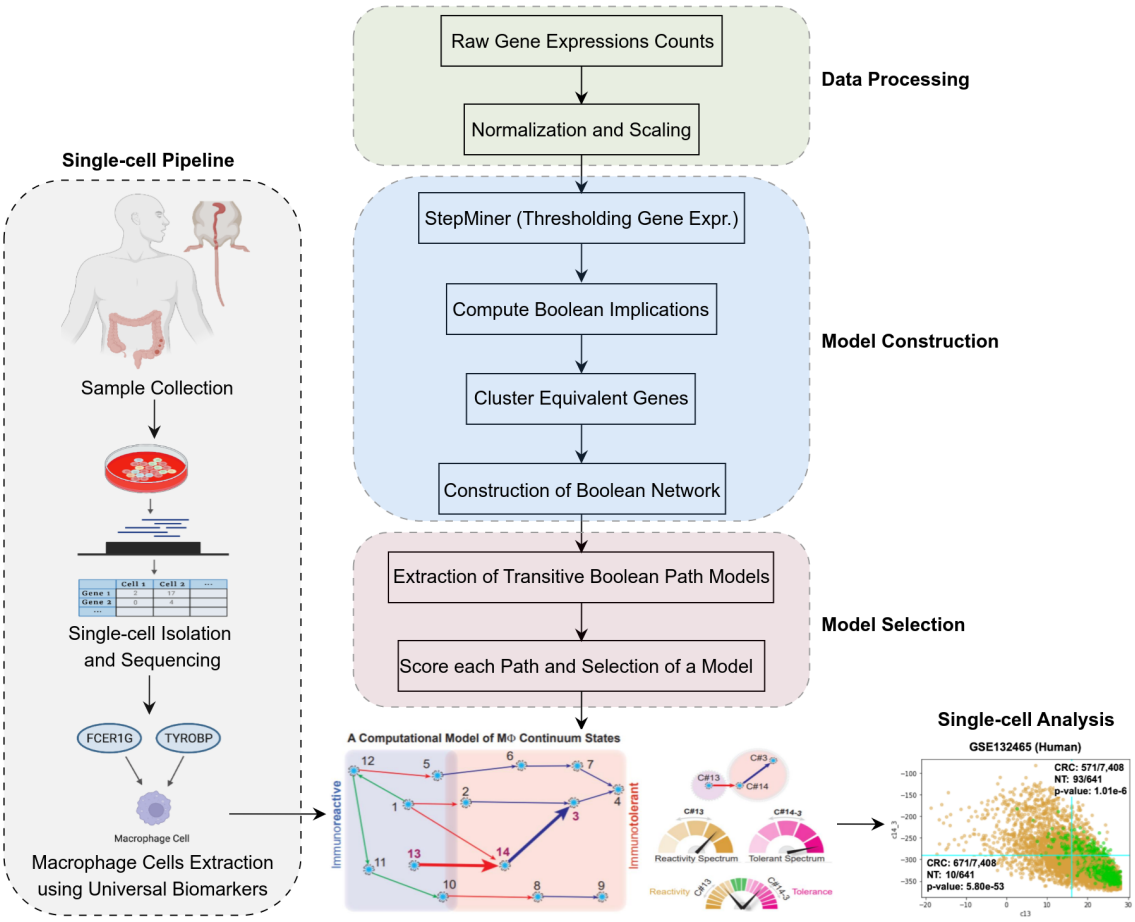


Fig. 1. Utilizing the SMaRT model to understand macrophage polarization in the context of colorectal cancer at the single-cell RNA-seq level.

classified as *highly tolerant* macrophages. To determine whether the number of tumor-specific macrophage cells and the number of normal colon-specific macrophage cells in a quadrant are significantly comparable, a test for proportions based on a normal (z) test is performed. The p-value is used to indicate the statistical significance of the two proportions using the z-test. That is, $p \leq 0.05$ indicates that we can reject the null hypothesis, which states that there is no difference in the proportion of NT samples to CRC samples.

2.3. *Construction of a CRC-specific signature using refinement*

In this section, we describe a methodology used to refine the "general" SMaRT signature[1] to be more specific towards the diagnosis of colorectal cancer. Such a model will represent the signature of the macrophages found in the tumor micro-environment in CRC samples. Mathematically, the TAM-specific signature is a signature obtained from a subset of genes (of the 338 genes in the original SMaRT model) that not only captures the macrophage polarization dynamics but also is predictive towards classifying NT and CRC samples. We use the term "refinement" to refer to the steps used in constructing such a signature, which are as follows: We first identify a scRNA-seq dataset where the composite score versus sample

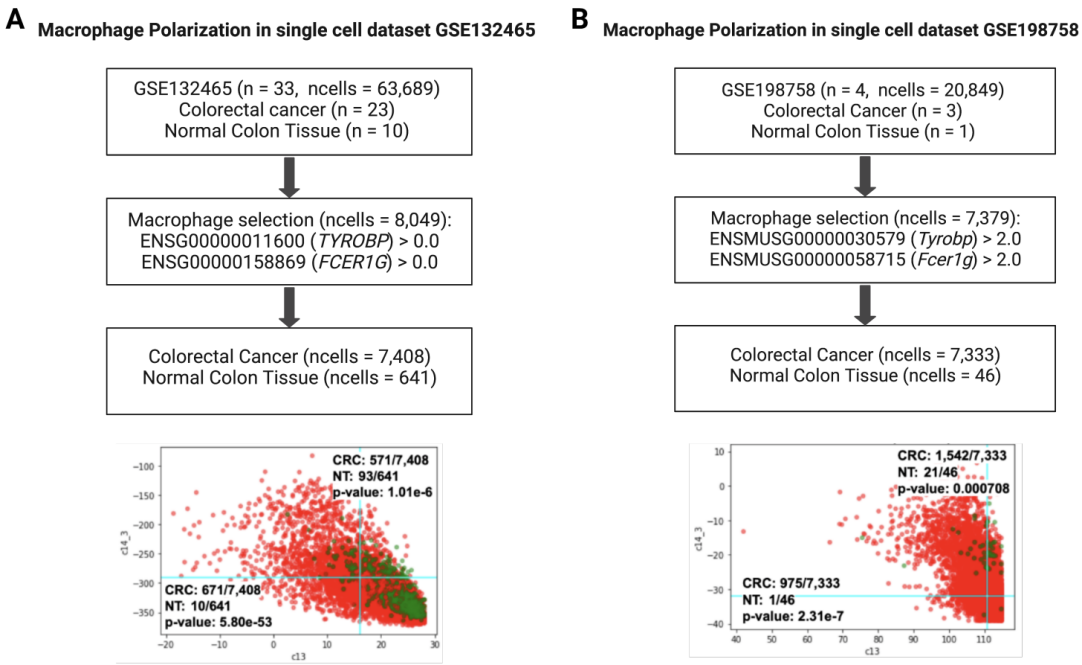


Fig. 2. Workflows showing relationship between macrophage polarization (reactivity and tolerance) and single-cell colorectal cancer samples (normal, NT and tumor, CRC): (A) Human and (B) Mice.

disease annotation relationship shows statistically significant outcomes: high-high and low-low quadrants separate the two categories of samples (NT and CRC) very well. In order to derive a TAM-specific signature, we manipulate the training dataset into different pseudo-bulk representations based on the provided gene expression data and on the conclusions derived from the scRNA-seq analysis pipeline. Here, we would like to note a potential reasoning, owing primarily to domain-knowledge, behind the categorization of cells, we follow subsequently. Due to the capacity and diversity of differentiated cells in the colon,[17] we face the issue of potential contamination of noisy genes in our macrophage refined signature stemming from the colonic epithelium. To ensure that our signature primarily captures macrophage polarization and is not misrepresented by the unique expression pattern of a gene from a differentiated cell, we want to remove the bias of genes that may be expressed differently in epithelial and macrophage cells. Therefore, we consider genes that have similar and consistent expression patterns in both cell types. To computationally extract 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),[18] is applied.

We use differential gene expression (DEG) analysis[19] using t-test to rank the genes within the SMaRT signature that capture the disease annotations in a pattern similar to the results of the scRNA-seq analysis. Specifically, our DEG analysis is performed on 4 types of pseudo-bulk representations (created using $\log_2$-normalized CPM values) of our training dataset, consisting of:

Type 1 All cells in the dataset separating the annotations of NT and CRC samples;

Type 2 A comparable subset of filtered macrophages cells in the dataset separating the annotations of highly-tolerant NT and highly-tolerant CRC samples;

Type 3 A comparable subset of filtered macrophages cells in the dataset separating the annotations of highly-reactive CRC and highly-tolerant CRC samples;

Type 4 All epithelial cells in the dataset separating the annotations of NT and CRC samples.

Overlaying the top rankings from the following differential gene expression results are used to compose the signatures considered for analysis:

Refining $C13$ Let n be the number of genes in $C13$. We take the top k genes of n expressed in Type 4. Then, of the top k, we take the top m genes expressed in Type 2.

Refining $C14_3$ Let n be the total number of genes in $C14$ and $C3$ together. We take the top k (of n) genes expressed in Type 4. Then, of these k genes, we take the top k_1 genes expressed in Type 2. Furthermore, we take the top k_2 (of k_1) genes expressed in Type 3. Of the top k_2 genes, we take the top m genes expressed in Type 1.

Through our experiments, we chose GSE132465 (Homo Sapien, n = 33, n-cells = 63,689, n-macrophage cells = 8,049) to be the training dataset. Using the refinement methodology, as described above, we chose 15 immuno-reactive genes (out of the 48 genes in $C13$) and 9 immuno-tolerant genes (out of the 290 genes in $C14_3$). For the refinement of $C13$, we set $n = 48$, $k = 25$, and $m = 15$. For the refinement of $C14_3$, we set $n = 290$, $k = 116$, $k_1 = 63$, $k_2 = 43$, and $m = 9$. In summary, our refined signature consists of only 24 genes (in contrast to the 338 genes in the original SMaRT signature):

- Immuno-reactive genes (Refined $C13$): $\{OAS2, TAP1, XAF1, IFIT3, IFIT2, OAS1, ISG15, TRIM21, MX1, APOL1, OAS3, CXCL9, SP110, PML, STAT1\}$;
- Immuno-tolerant genes (Refined $C14_3$): $\{RPL3, SLC46A3, CD302, RBM4, METTL7A, CCDC88A, RPL17, CLEC10A, DNASE1L3\}$.

We use the term "Refined-SMaRT" to refer to the SMaRT model constructed using these 24 genes and use the composite score[1] to map the expression values of these 24 genes to a single real-value by using a weight of -1 for the genes in $C13$ and weights of 1 and 2 for the genes in $C14$ and $C3$, respectively. This procedure is similar to building a logistic regression model using the refined set of genes to classify the two classes of samples. We note an interesting observation here: Our process of signature refinement resembles with the way transfer learning[20,21] works, that is, to transfer knowledge from a larger (more general) model to a small (more specific) machine learning model.

2.3.1. *Validation of the Refined-SMaRT model*

The goal of the validation stage is to investigate whether the 24 gene signature obtained in the previous section has any diagnostic benefits. For this, we will use two statistical metrics: (1) AUC-ROC: Area under the receiver-operating characteristic curve[22] to quantify the predictive potential of the composite signature for classifying the NT and CRC samples at the (pseudo)-bulk RNA-seq and microarray levels; and (2) p-value: The composite scores of these genes is

statistically distinctive between NT and CRC groups of samples. We note that for the scRNA-seq datasets, we create two kinds of pseudo-bulks: (a) using all of the cells, and (b) using only the filtered macrophage cells.

3. Results

Using our scRNA-seq and refinement pipelines, we investigate and provide computational evidence for the following: (R1) At the single-cell level, macrophages in colorectal cancer show a more reactive polarization than macrophages in healthy colon cells; (R2) The 24-gene Refined-SMaRT signature is highly predictive in discriminating between the two disease states: healthy and tumorous; and (R3) The refined signature has substantial translational potential, primarily validated in determining MSI-status and CIMP-status.

3.1. *Differential polarization dynamics of macrophages in colorectal cancer*

Figure 3 shows the polarization dynamics in various scRNA-seq datasets, including the dataset used for the construction of our Refined-SMaRT signature (GSE132465). In all of our experiments, we use a statistical significance test ($p \leq 0.05$) for the number of macrophage cells in the immuno-reactive quadrant ([C13-low, C14_3-low]) and the immuno-tolerant quadrant ([C13-high, C14_3-high]). In all cases, we observe a high concentration of CRC macrophage cells in the immuno-reactive quadrant, compared to the concentration of NT macrophage cells. We also find the pattern consistent and statistically significant in mice datasets. 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.[23] 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 CRC cells.

3.2. *Predictive performance of Refined-SMaRT for CRC*

Our primary focus in this paper has been on the refinement of the original SMaRT signature. We adopt two different kinds of comparisons while demonstrating the effectiveness of our Refined-SMaRT signature for CRC: (a) a straightforward comparison with the original SMaRT signature, and (b) a comparison with a "noisy" signature. The comparison with the noisy signature puts much importance on the requirement of refinement in the context of computational research. We first make a note on the noisy signature:

Noisy signature:

We deliberately create a signature by adopting a strategy somewhat opposite to how our refined signature is constructed. In the noisy signature case, the $C13$ genes are noise-refined with filtered macrophages cells in the dataset separating the annotations of highly-tolerant NT and highly-tolerant CRC samples and filtered macrophages cells in the dataset separating the annotations of highly-reactive CRC and highly-tolerant CRC samples. Whereas for the $C14_3$ refined noisy signature, we use all epithelial cells in the dataset separating the annotations of

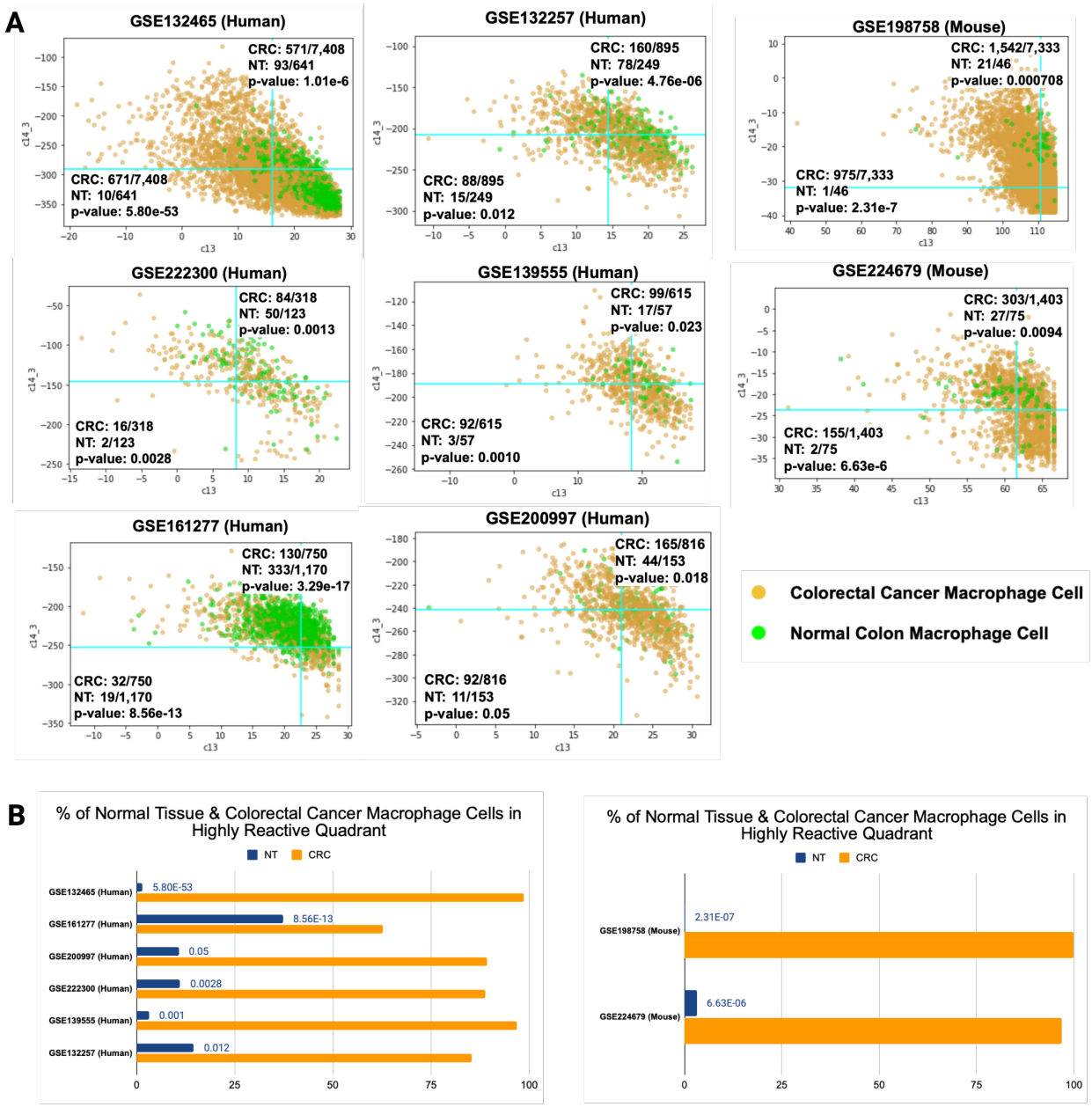


Fig. 3. Macrophage polarization dynamics in scRNA-seq datasets, both humans and mice: (A) Scatter plot between the composite score of $C13$ versus the composite score of $C14_3$ with different quadrants identified clearly with the StepMiner thresholds; (B) A simple visualization of the proportions of macrophage cells in the high-high quadrant for each dataset.

NT and CRC samples and all cells in the dataset separating the annotations of NT and CRC samples.

We now turn to the computational results, as depicted in Figure 4 and Figure 5. First, we show the performance of the refined signature, in Figure 4 on 3 different kinds of pseudo-bulk representations of the training dataset, prepared from all cells, all macrophage cells and all epithelial cells. The results clearly highlights that the very minimal 24-gene signature (92%

reduction in the number of genes in original SMaRT) has a better predictive capability, with AUC-ROC of more than 90%, as the original signature. Furthermore, the results also suggest that a noisy version of the signature (or any randomly drawn subset of genes) will not have a similar performance.

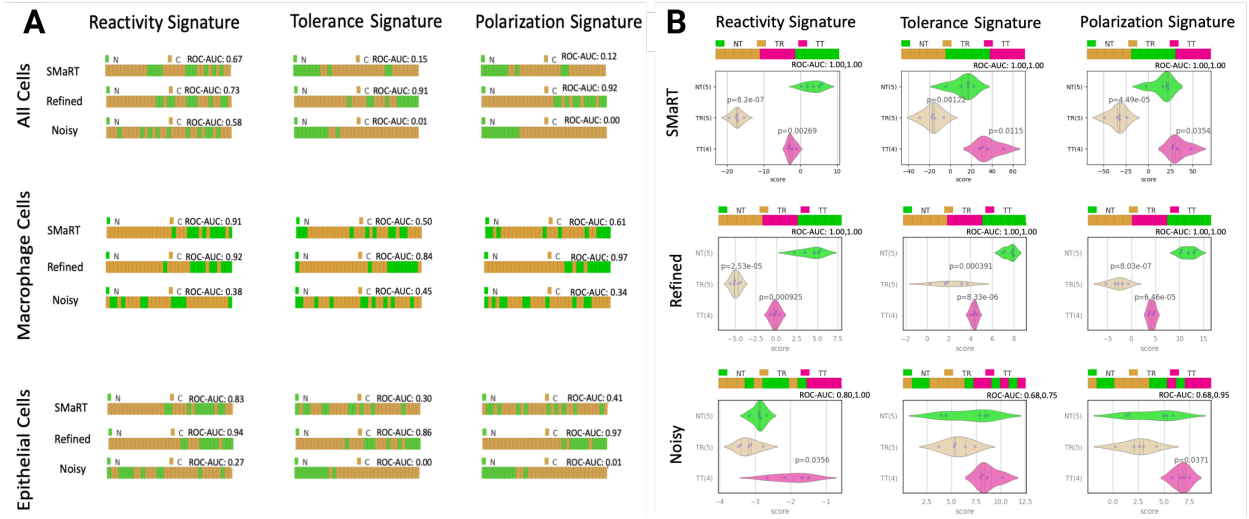


Fig. 4. Comparisons of predictive performance between the SMaRT signature, Refined-SMaRT signature, and a noisy signature for the training dataset, GSE132465. Visualizations in (A) report the predictive analysis of the SMaRT, Refined-SMaRT, and noisy signatures for the pseudo-bulk representations in all cells, computationally filtered macrophage cells, and computationally filtered epithelial cells of GSE132465. The violin plots in (B) showcase the predictions of Tolerant NT (NT), Reactive CRC (TR), and Tolerant Tumor (TT) annotations with the SMaRT, Refined-SMaRT, and noisy signatures. The AUC-ROC metric is reported to quantify the predictive potential of the signatures.

Validation of the predictive performance of the Refined-SMaRT signature is also reported on the pseudo-bulk representations of the five human datasets used for the preliminary scRNA-seq analysis with the SMaRT signature (See Figure 5A). We compare the predictive performance of the signatures in all cells and filtered macrophages. While we report an AUC-ROC broader range of 0.67–1.00 for the predictive potential of our refined signature, our Refined-SMaRT signature performs better than SMaRT or reports a perfect separation of disease states, similar to the SMaRT signature while only considering a fraction (7%) of the entire original gene signature. As one would expect, the noisy signature performs worse than our Refined-SMaRT and the original SMaRT signatures. Secondary large-scale validation of the predictive performance is showcased in five large-cohort microarray colorectal cancer datasets, in which we do not have the capacity to computationally extract macrophage cells (See Figure 5B). Analysis in the bulk datasets allows us to validate the signature without the bias of macrophage-specific data. 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.

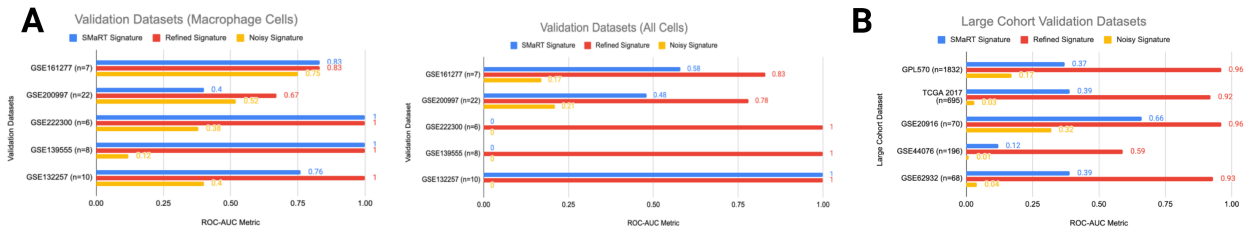


Fig. 5. Comparison of predictive performance between the SMaRT signature, Refined-SMaRT signature, and a noisy signature for testing datasets used for validation: (A) pseudo-bulk representations of five human datasets used for scRNA-seq analysis, reported in filtered macrophage cells and all cells; (B) large-cohort microarray colorectal cancer datasets with the annotations of healthy colon tissue and tumor tissue.

3.2.1. *Reactome pathway analysis*

Reactome pathway analysis[24] is used to identify the enriched signaling and metabolic pathways in our Refined-SMaRT signature. It reveals the potential cellular processes that are impacted by changes in the macrophage polarization dynamics that are likely to occur in the CRC tumor microenvironment as a result of the genes in our signature. We would like to note that many of the genes in our signature are correlated with interferon signaling, which is involved in the modulation of immune responses against pathogens. The activation of reactive macrophages in the tumor microenvironment of CRC tissue, as a response to inflammation, triggers intracellular signaling cascades that up-regulate the interferon-stimulated genes that are involved in immune activation.[25–27]

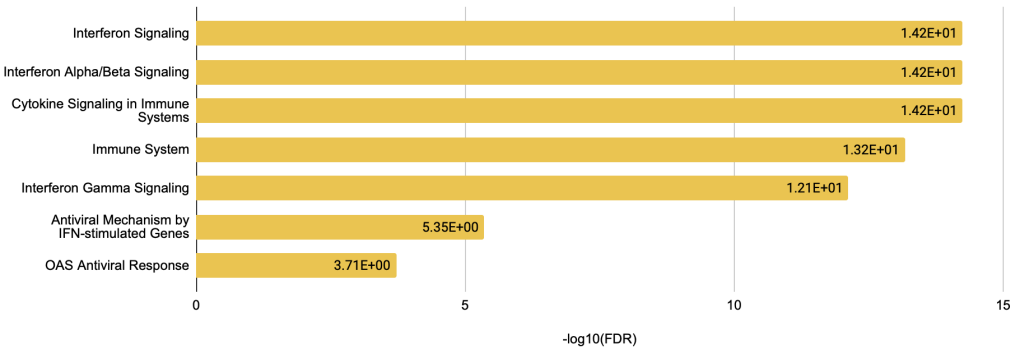


Fig. 6. Reactome pathway analysis report to identify the biological processes that the genes in our Refined-SMaRT signature are involved in.

3.3. *Translation potential of the refined signature*

The MSI status is an approved clinical biomarker associated with the prediction of immunotherapy outcomes. The stable state (MSS) does not trigger the body's immune response toward a tumor and typically does not respond to immunotherapy treatments. The unstable state (MSI) corresponds to an unstable tumor and is associated with increased infiltration of immune cells in their microenvironments (i.e., TAMs). Our refined immuno-reactive signature

shows that MSI colorectal cancer samples are more immuno-reactive compared to the MSS samples (see Figure 7A). The higher expression of M1 macrophage genes in the MSI samples suggests that immunotherapy works better 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).[28] 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.[29] M1 macrophages showed a significant effect on the prognosis in CIMP− and CIMP+ (specifically CIMP high) cases.[30] Our refined immuno-reactive signature proposes that CIMP+ samples are more reactive compared to the CIMP− samples; this supports the relationship between CIMP+ status and MSI status (See Figure 7B).

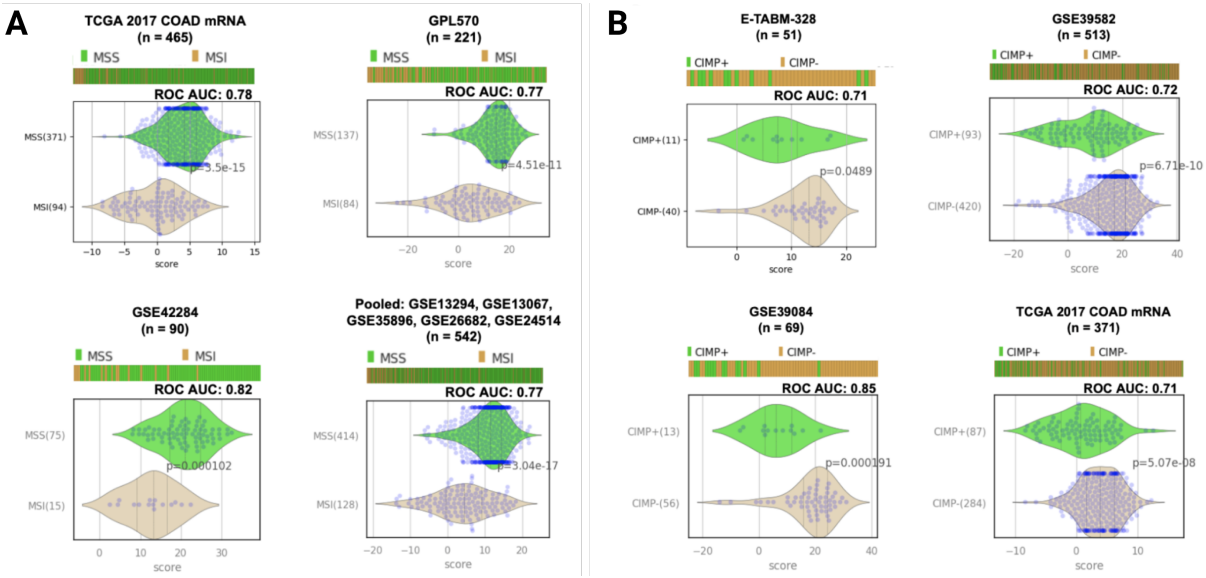


Fig. 7. Predictive potential of our refined signature towards clinical translation studies: (A) MSS/MSI classification, and (B) CIMP+/CIMP- status prediction.

4. Conclusion

We propose a technique for the construction of a minimal-and-refined signature to capture the polarization dynamics of tumor-associated macrophages and associate it with the diagnosis of colorectal cancer, using SMaRT[1]—a recently proposed model to capture macrophage polarization states. The refined signature exhibits substantial predictive capability while dealing with a diverse set of CRC datasets while also showing tremendous translation potential in clinical studies. However, further extensive examination of our 24-gene signature is warranted, at least in a wet-lab setup, to show potential impact in healthcare. Furthermore, on the computational side, the signature will undergo further analysis with newer CRC-specific datasets, especially from spatial transcriptomics. The authors are working on this aspect.

Acknowledgements

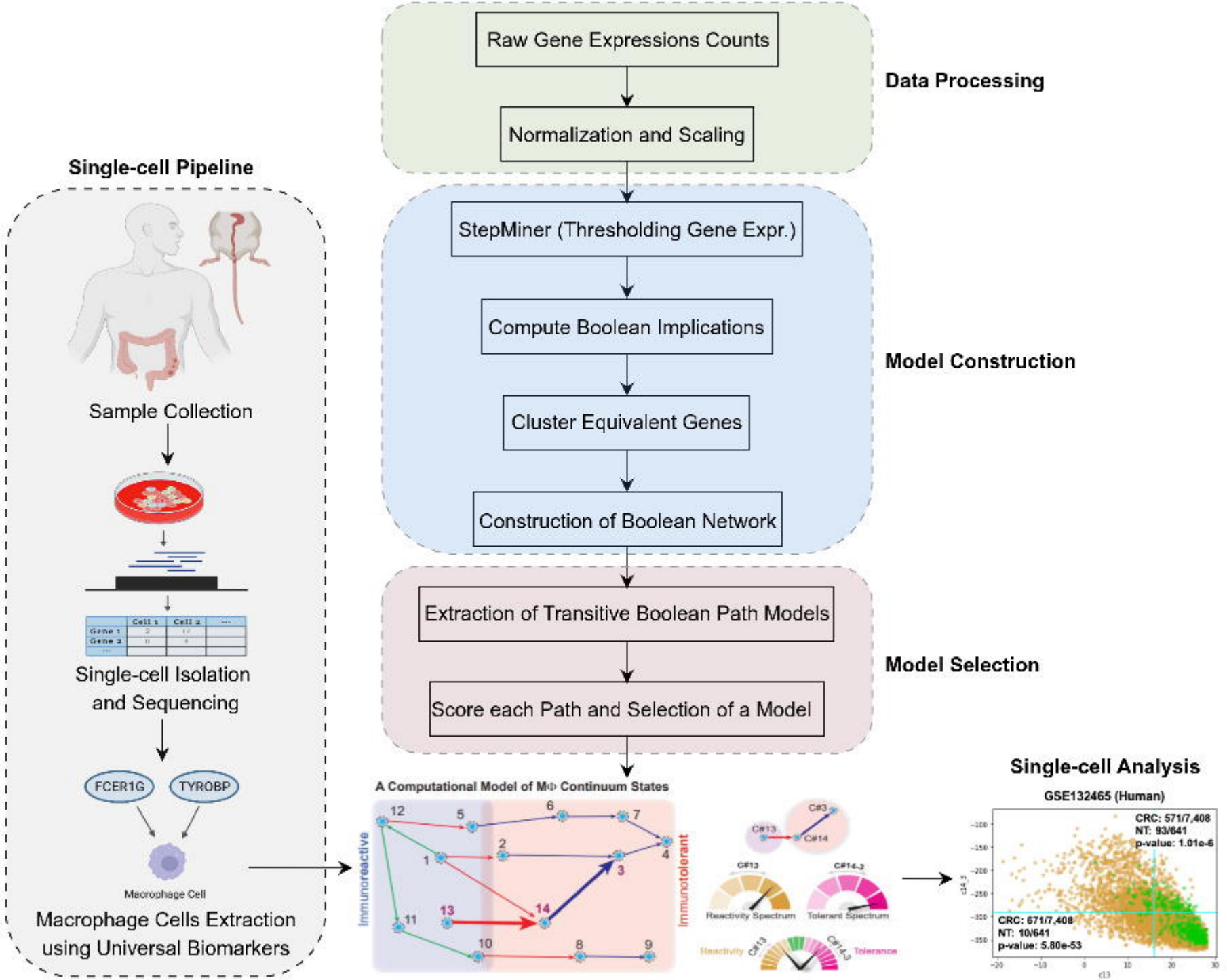
The authors sincerely thank the members of the Boolean Lab for their constructive feedback during this research. This work is supported by the NIH grant R01-AI55696. Other sources of support include R01-GM138385, UG3TR003355, Padres Pedal the Cause/RADY #PTC2017 and #PTC2021.

References

1. Pradipta Ghosh, Saptarshi Sinha, Gajanan D. Katkar, Daniella Vo, Sahar Taheri, Dharanidhar Dang, Soumita Das, and Debashis Sahoo. Machine learning identifies signatures of macrophage reactivity and tolerance that predict disease outcomes. *eBioMedicine*, 94:104719, 2023.
2. Inés Mármol, Cristina Sánchez-de Diego, Alberto Pradilla Dieste, Elena Cerrada, and María Jesús Rodriguez Yoldi. Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. *International journal of molecular sciences*, 18(1):197, 2017.
3. Fortunato Ciardiello, Davide Ciardiello, Giulia Martini, Stefania Napolitano, Josep Tabernero, and Andres Cervantes. Clinical management of metastatic colorectal cancer in the era of precision medicine. *CA: a cancer journal for clinicians*, 72(4):372–401, 2022.
4. Yue Xi and Pengfei Xu. Global colorectal cancer burden in 2020 and projections to 2040. *Translational oncology*, 14(10):101174, 2021.
5. Mikhail Binnewies, Edward W Roberts, Kelly Kersten, Vincent Chan, Douglas F Fearon, Miriam Merad, Lisa M Coussens, Dmitry I Gabrilovich, Suzanne Ostrand-Rosenberg, Catherine C Hedrick, et al. Understanding the tumor immune microenvironment (time) for effective therapy. *Nature medicine*, 24(5):541–550, 2018.
6. Bin-Zhi Qian and Jeffrey W Pollard. Macrophage diversity enhances tumor progression and metastasis. *Cell*, 141(1):39–51, 2010.
7. Peter J Murray, Judith E Allen, Subhra K Biswas, Edward A Fisher, Derek W Gilroy, Sergij Goerdt, Siamon Gordon, John A Hamilton, Lionel B Ivashkiv, Toby Lawrence, et al. Macrophage activation and polarization: nomenclature and experimental guidelines. *Immunity*, 41(1):14–20, 2014.
8. Fernando Oneissi Martinez, Antonio Sica, Alberto Mantovani, and Massimo Locati. Macrophage activation and polarization. *Frontiers in Bioscience-Landmark*, 13(2):453–461, 2008.
9. Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995, 2012.
10. Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.
11. Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12:1–16, 2011.
12. Lior Pachter. Models for transcript quantification from rna-seq. *arXiv preprint arXiv:1104.3889*, 2011.
13. Debashis Sahoo, David L Dill, Andrew J Gentles, Robert Tibshirani, and Sylvia K Plevritis. Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome biology*, 9(10):1–17, 2008.
14. Debashis Sahoo, Jun Seita, Deepta Bhattacharya, Matthew A Inlay, Irving L Weissman, Sylvia K Plevritis, and David L Dill. Midreg: a method of mining developmentally regulated genes using boolean implications. *Proceedings of the National Academy of Sciences*, 107(13):5732–5737, 2010.
15. Dharanidhar Dang, Sahar Taheri, Soumita Das, Pradipta Ghosh, Lawrence S Prince, and De-

bashis Sahoo. Computational approach to identifying universal macrophage biomarkers. *Frontiers in physiology*, 11:275, 2020.

16. Debashis Sahoo, David L Dill, Rob Tibshirani, and Sylvia K Plevritis. Extracting binary signals from microarray time-course data. *Nucleic acids research*, 35(11):3705–3712, 2007.
17. Kaushal Parikh, Agne Antanaviciute, David Fawkner-Corbett, Marta Jagielowicz, Anna Aulicino, Christoffer Lagerholm, Simon Davis, James Kinchen, Hannah H Chen, Nasullah Khalid Alham, et al. Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature*, 567(7746):49–55, 2019.
18. IJ Goossens-Beumer, ECM Zeestraten, A Benard, T Christen, MS Reimers, R Keijzer, CFM Sier, GJ Liefers, H Morreau, H Putter, et al. Clinical prognostic value of combined analysis of aldh1, survivin, and epcam expression in colorectal cancer. *British journal of cancer*, 110(12):2935–2944, 2014.
19. Peng Liang and Arthur B Pardee. Analysing differential gene expression in cancer. *Nature Reviews Cancer*, 3(11):869–876, 2003.
20. Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36. JMLR Workshop and Conference Proceedings, 2012.
21. Soundarya Krishnan, Rishab Khincha, Lovekesh Vig, Tirtharaj Dash, and Ashwin Srinivasan. A case study of transfer of lesion-knowledge. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 138–145, Cham, 2020. Springer International Publishing.
22. David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45:171–186, 2001.
23. Hui Wang, Tian Tian, and Jinhua Zhang. Tumor-associated macrophages (tams) in colorectal cancer (crc): from mechanism to therapy and prognosis. *International journal of molecular sciences*, 22(16):8470, 2021.
24. Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2018.
25. Jie Min, Wenjun Liu, and Jing Li. Emerging role of interferon-induced noncoding rna in innate antiviral immunity. *Viruses*, 14(12):2607, 2022.
26. J Judy Chang, Matt Woods, Robert J Lindsay, Erin H Doyle, Morgane Griesbeck, Ellen S Chan, Gregory K Robbins, Ronald J Bosch, and Marcus Altfeld. 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(5):830–838, 2013.
27. Melissa Kane, Trinity M Zang, Suzannah J Rihn, Fengwen Zhang, Tonya Kueck, Mudathir Alim, John Schoggins, Charles M Rice, Sam J Wilson, and Paul D Bieniasz. Identification of interferon-stimulated genes with antiretroviral activity. *Cell host & microbe*, 20(3):392–405, 2016.
28. Haiwei Wang, Xinrui Wang, Liangpu Xu, Ji Zhang, and Hua Cao. Analysis of the transcriptomic features of microsatellite instability subtype colon cancer. *BMC cancer*, 19(1):1–16, 2019.
29. Ehsan Nazemalhosseini Mojarad, Peter JK Kuppen, Hamid Asadzadeh Aghdaei, and Mohammad Reza Zali. The cpg island methylator phenotype (cimp) in colorectal cancer. *Gastroenterology and hepatology from bed to bench*, 6(3):120, 2013.
30. Sofia Edin, Maria L. Wikberg, Anna M. Dahlin, Jörgen Rutegård, Åke Öberg, Per-Arne Oldenborg, and Richard Palmqvist. The distribution of macrophages with a m1 or m2 phenotype in relation to prognosis and the molecular characteristics of colorectal cancer. *PLOS ONE*, 7(10):1–12, 10 2012.


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
Reactivity
Tolerance
Single-cell Analysis
GSE132465 (Human)
CRC: 571/7,408
NT: 93/641
p-value: 1.01e-6
CRC: 671/7,408
NT: 10/641
p-value: 5.80e-53

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

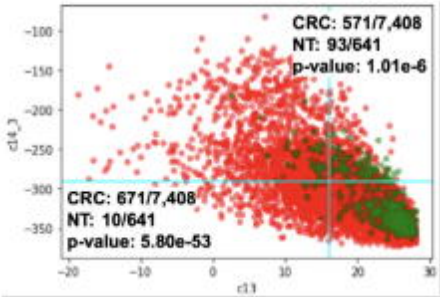
GSE132465 (n = 33, ncells = 63,689)
Colorectal cancer (n = 23)
Normal Colon Tissue (n = 10)

↓

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**

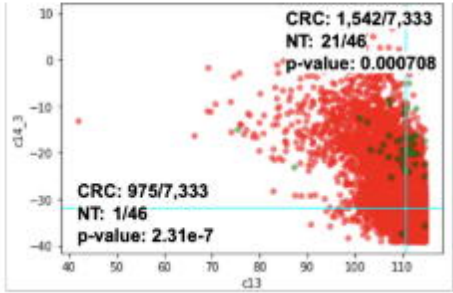
GSE198758 (n = 4, ncells = 20,849)
Colorectal Cancer (n = 3)
Normal Colon Tissue (n = 1)

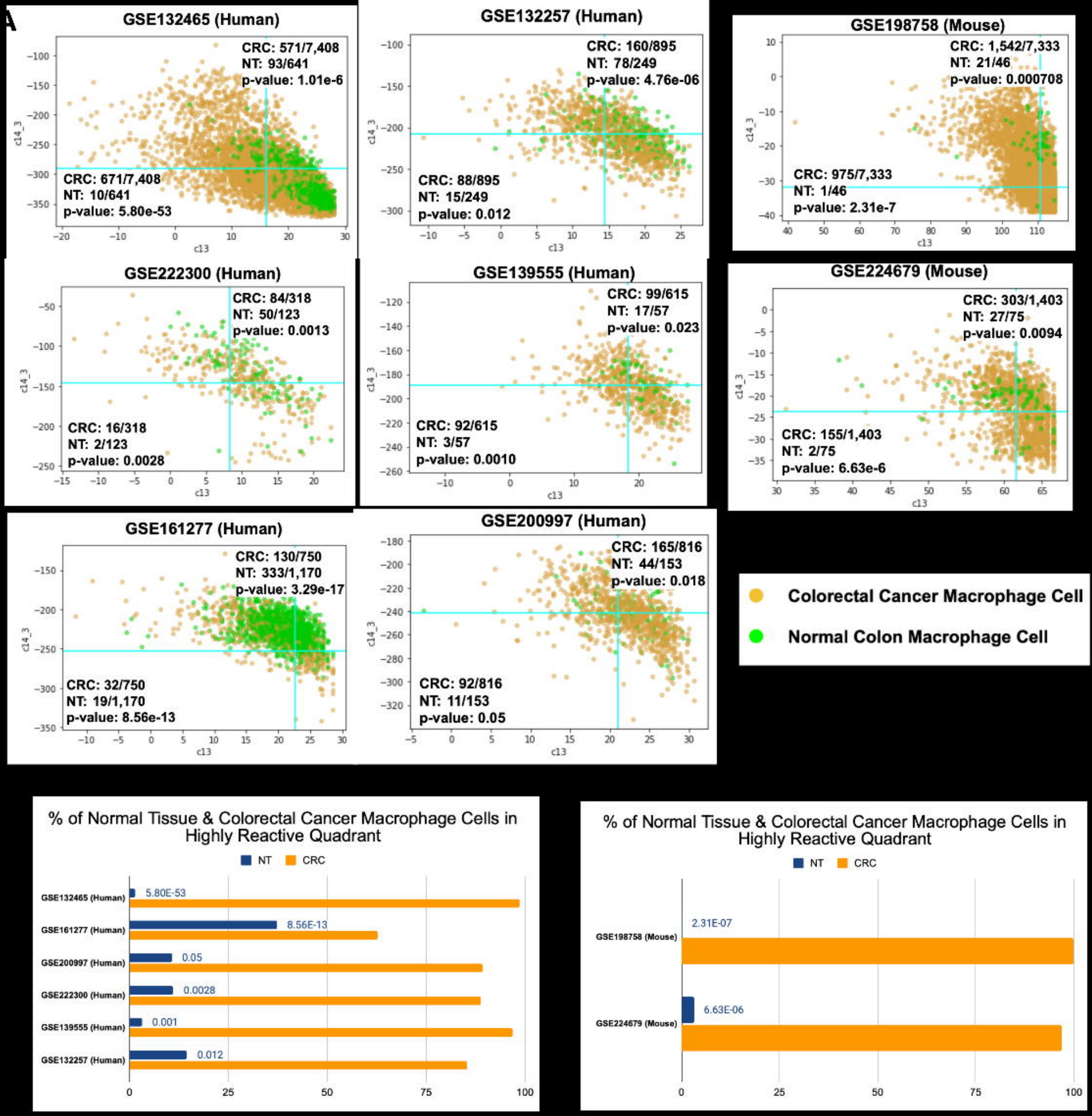
↓

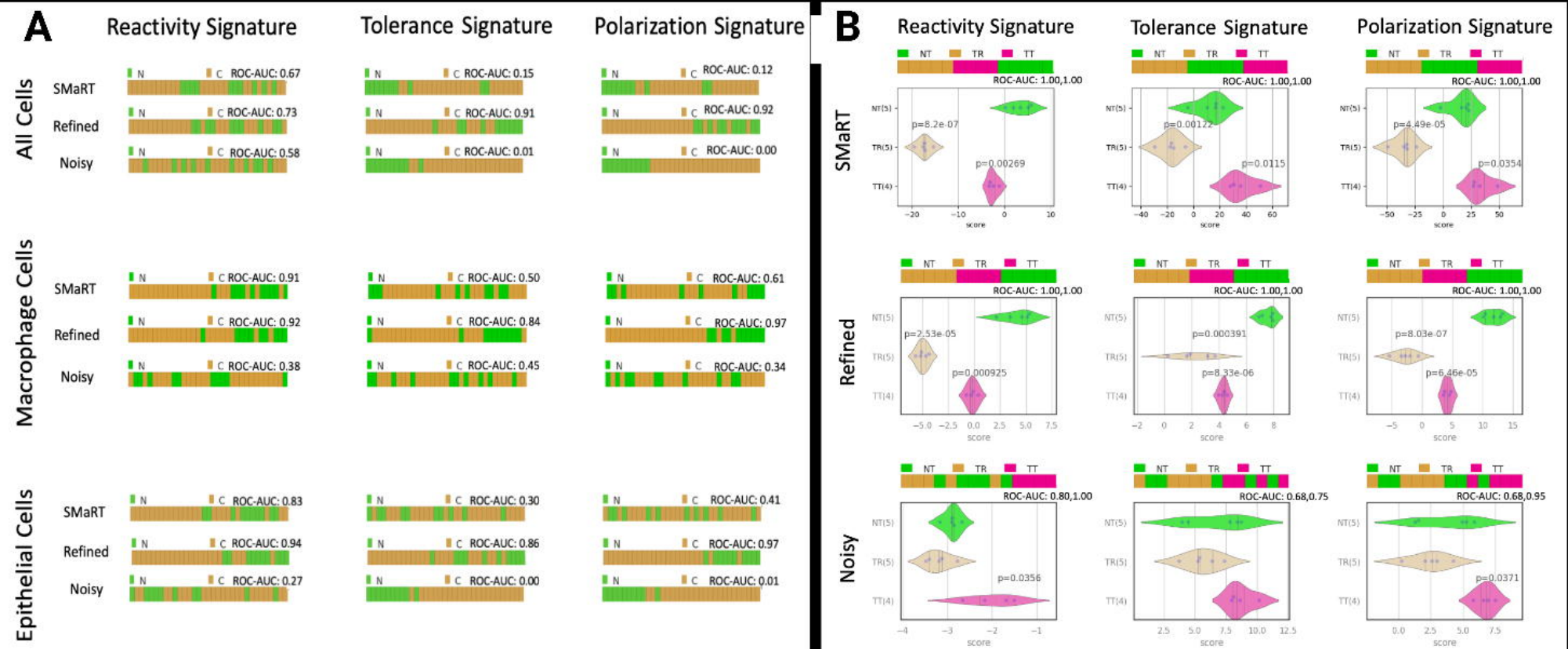
Macrophage selection (ncells = 7,379):
ENSMUSG00000030579 (*Tyrobp*) > 2.0
ENSMUSG00000058715 (*Fcer1g*) > 2.0

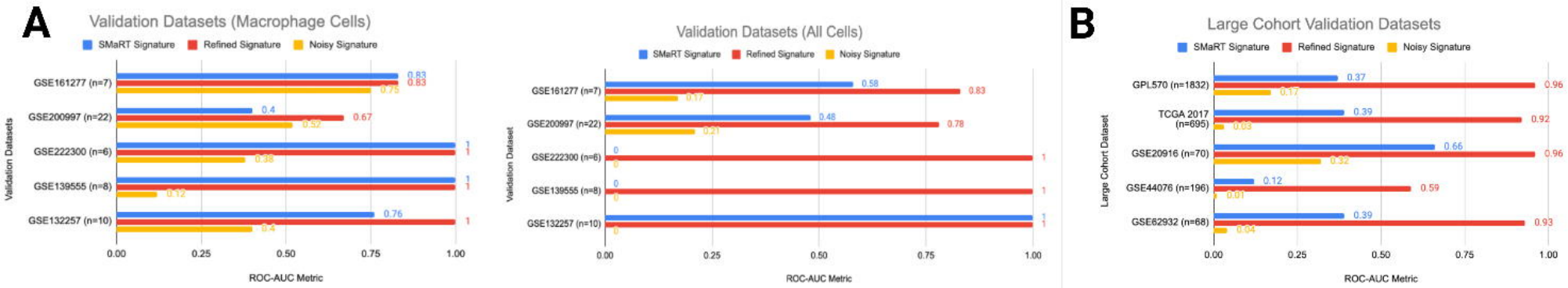
↓

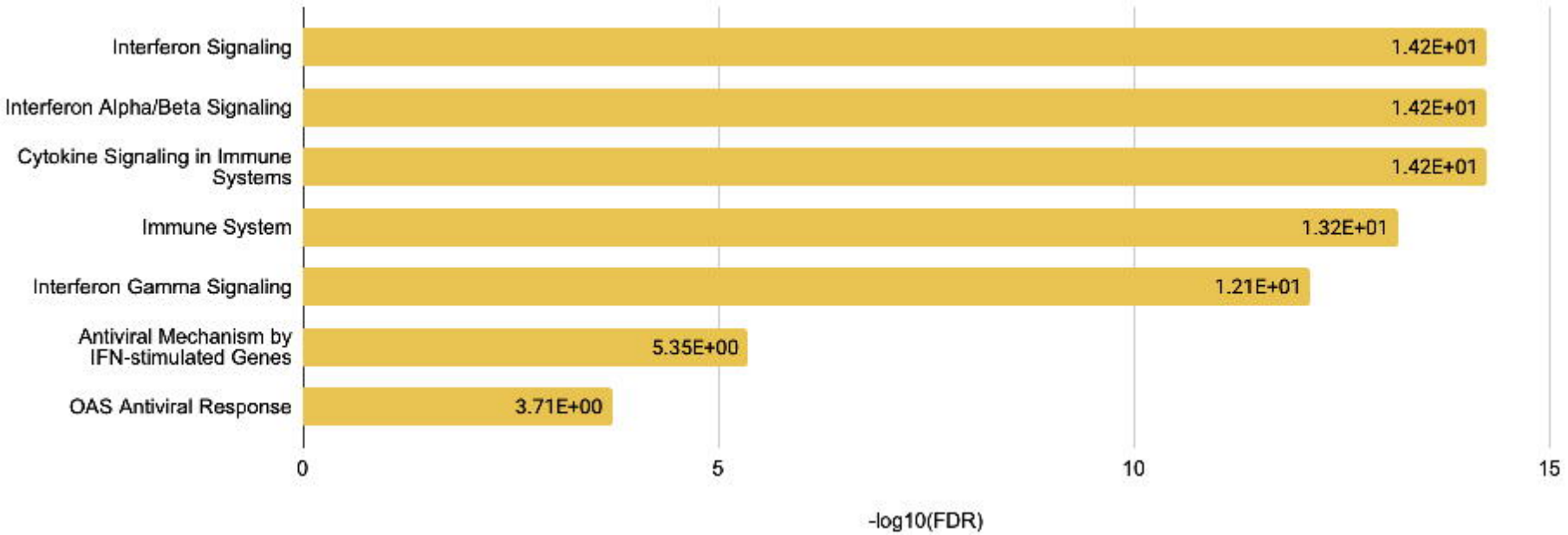
Colorectal Cancer (ncells = 7,333)
Normal Colon Tissue (ncells = 46)

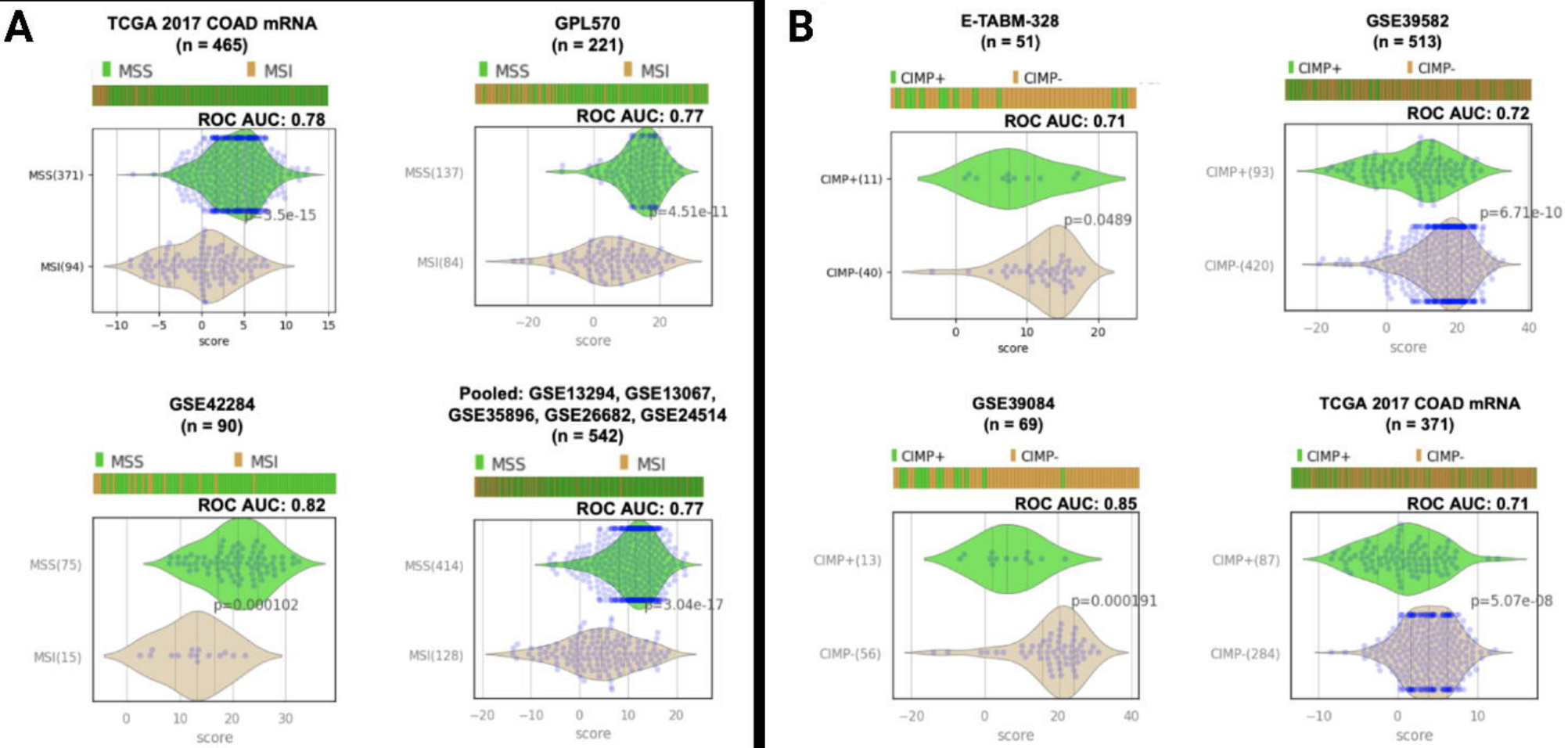



A
GSE132465 (Human)
CRC: 571/7,408
NT: 93/641
p-value: 1.01e-6
CRC: 671/7,408
NT: 10/641
p-value: 5.80e-53
GSE132257 (Human)
CRC: 160/895
NT: 78/249
p-value: 4.76e-06
CRC: 88/895
NT: 15/249
p-value: 0.012
GSE198758 (Mouse)
CRC: 1,542/7,333
NT: 21/46
p-value: 0.000708
CRC: 975/7,333
NT: 1/46
p-value: 2.31e-7
GSE222300 (Human)
CRC: 84/318
NT: 50/123
p-value: 0.0013
CRC: 16/318
NT: 2/123
p-value: 0.0028
GSE139555 (Human)
CRC: 99/615
NT: 17/57
p-value: 0.023
CRC: 92/615
NT: 3/57
p-value: 0.0010
GSE224679 (Mouse)
CRC: 303/1,403
NT: 27/75
p-value: 0.0094
CRC: 155/1,403
NT: 2/75
p-value: 6.63e-6
GSE161277 (Human)
CRC: 130/750
NT: 333/1,170
p-value: 3.29e-17
CRC: 32/750
NT: 19/1,170
p-value: 8.56e-13
GSE200997 (Human)
CRC: 165/816
NT: 44/153
p-value: 0.018
CRC: 92/816
NT: 11/153
p-value: 0.05
c13
c14_3
Colorectal Cancer Macrophage Cell
Normal Colon Macrophage Cell
% of Normal Tissue & Colorectal Cancer Macrophage Cells in Highly Reactive Quadrant
NT
CRC
GSE132465 (Human)
5.80E-53
GSE161277 (Human)
8.56E-13
GSE200997 (Human)
0.05
GSE222300 (Human)
0.0028
GSE139555 (Human)
0.001
GSE132257 (Human)
0.012
0
25
50
75
100
% of Normal Tissue & Colorectal Cancer Macrophage Cells in Highly Reactive Quadrant
NT
CRC
GSE198758 (Mouse)
2.31E-07
GSE224679 (Mouse)
6.63E-06
0
25
50
75
100


A
Reactivity Signature
Tolerance Signature
Polarization Signature
All Cells
Macrophage Cells
Epithelial Cells
SMaRT
Refined
Noisy
N
C
ROC-AUC: 0.67
ROC-AUC: 0.15
ROC-AUC: 0.12
ROC-AUC: 0.73
ROC-AUC: 0.91
ROC-AUC: 0.92
ROC-AUC: 0.58
ROC-AUC: 0.01
ROC-AUC: 0.00
ROC-AUC: 0.91
ROC-AUC: 0.50
ROC-AUC: 0.61
ROC-AUC: 0.92
ROC-AUC: 0.84
ROC-AUC: 0.97
ROC-AUC: 0.38
ROC-AUC: 0.45
ROC-AUC: 0.34
ROC-AUC: 0.83
ROC-AUC: 0.30
ROC-AUC: 0.41
ROC-AUC: 0.94
ROC-AUC: 0.86
ROC-AUC: 0.97
ROC-AUC: 0.27
ROC-AUC: 0.00
ROC-AUC: 0.01
B
Reactivity Signature
Tolerance Signature
Polarization Signature
SMaRT
Refined
Noisy
NT
TR
TT
ROC-AUC: 1.00,1.00
ROC-AUC: 0.80,1.00
ROC-AUC: 0.68,0.75
ROC-AUC: 0.68,0.95
NT(5)
TR(5)
TT(4)
p=8.2e-07
p=0.00269
p=0.00122
p=0.0115
p=4.49e-05
p=0.0354
p=2.53e-05
p=0.000925
p=0.000391
p=8.33e-06
p=8.03e-07
p=6.46e-05
p=0.0356
p=0.0371
score


A
Validation Datasets (Macrophage Cells)
SMaRT Signature
Refined Signature
Noisy Signature
Validation Datasets
GSE161277 (n=7)
GSE200997 (n=22)
GSE222300 (n=6)
GSE139555 (n=8)
GSE132257 (n=10)
0.83
0.83
0.75
0.4
0.67
0.52
1
1
0.38
1
1
0.12
0.76
1
0.4
0.00
0.25
0.50
0.75
1.00
ROC-AUC Metric
Validation Datasets (All Cells)
SMaRT Signature
Refined Signature
Noisy Signature
Validation Dataset
GSE161277 (n=7)
GSE200997 (n=22)
GSE222300 (n=6)
GSE139555 (n=8)
GSE132257 (n=10)
0.58
0.83
0.17
0.48
0.78
0.21
0
1
0
0
1
0
1
1
0
0.00
0.25
0.50
0.75
1.00
ROC-AUC Metric
B
Large Cohort Validation Datasets
SMaRT Signature
Refined Signature
Noisy Signature
Large Cohort Dataset
GPL570 (n=1832)
TCGA 2017 (n=695)
GSE20916 (n=70)
GSE44076 (n=196)
GSE62932 (n=68)
0.37
0.96
0.17
0.39
0.92
0.03
0.66
0.96
0.32
0.12
0.59
0.01
0.39
0.93
0.04
0.00
0.25
0.50
0.75
1.00
ROC-AUC Metric


Interferon Signaling
1.42E+01
Interferon Alpha/Beta Signaling
1.42E+01
Cytokine Signaling in Immune Systems
1.42E+01
Immune System
1.32E+01
Interferon Gamma Signaling
1.21E+01
Antiviral Mechanism by IFN-stimulated Genes
5.35E+00
OAS Antiviral Response
3.71E+00
0
5
10
15
-log10(FDR)


A
TCGA 2017 COAD mRNA
(n = 465)
MSS
MSI
ROC AUC: 0.78
MSS(371)
MSI(94)
p=3.5e-15
score
GPL570
(n = 221)
MSS
MSI
ROC AUC: 0.77
MSS(137)
MSI(84)
p=4.51e-11
score
GSE42284
(n = 90)
MSS
MSI
ROC AUC: 0.82
MSS(75)
MSI(15)
p=0.000102
score
Pooled: GSE13294, GSE13067,
GSE35896, GSE26682, GSE24514
(n = 542)
MSS
MSI
ROC AUC: 0.77
MSS(414)
MSI(128)
p=3.04e-17
score
B
E-TABM-328
(n = 51)
CIMP+
CIMP-
ROC AUC: 0.71
CIMP+(11)
CIMP-(40)
p=0.0489
score
GSE39582
(n = 513)
CIMP+
CIMP-
ROC AUC: 0.72
CIMP+(93)
CIMP-(420)
p=6.71e-10
score
GSE39084
(n = 69)
CIMP+
CIMP-
ROC AUC: 0.85
CIMP+(13)
CIMP-(56)
p=0.000191
score
TCGA 2017 COAD mRNA
(n = 371)
CIMP+
CIMP-
ROC AUC: 0.71
CIMP+(87)
CIMP-(284)
p=5.07e-08
score