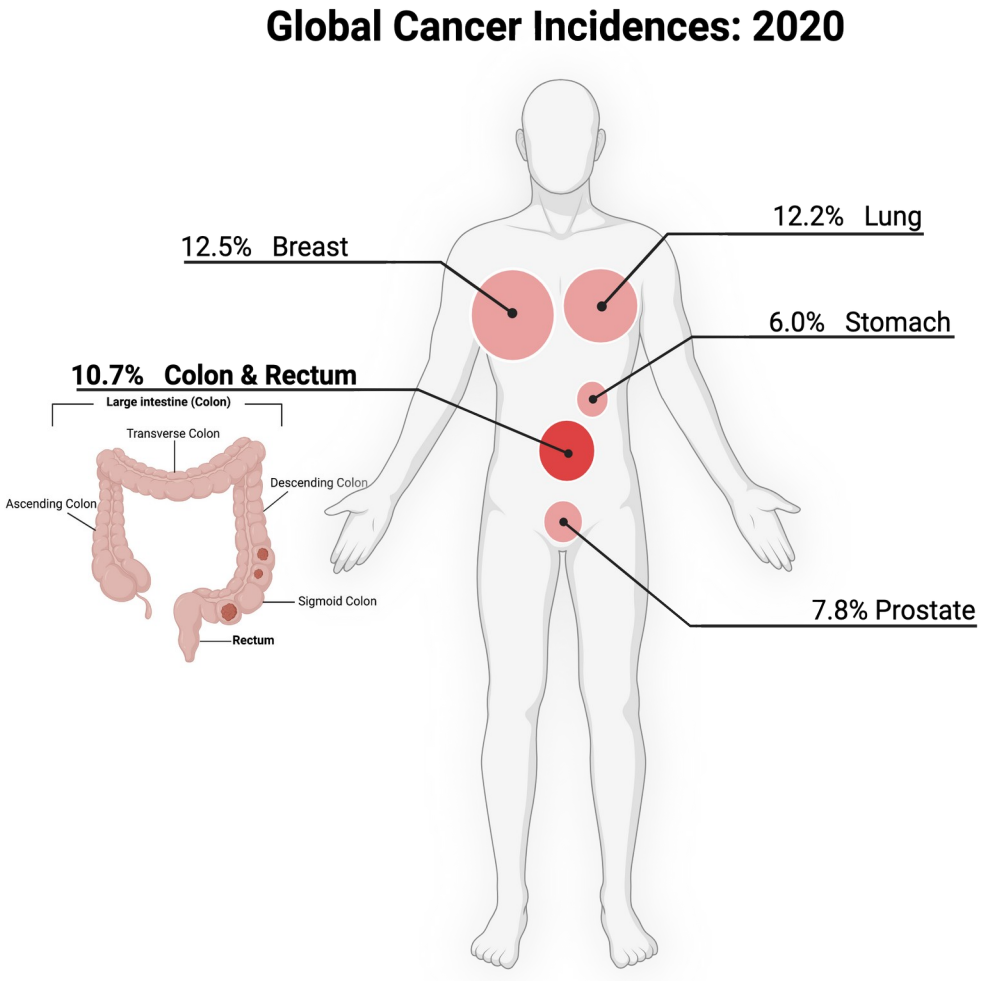
Investigation of Tumor-Associated Macrophages and their Polarization in Colorectal Cancer

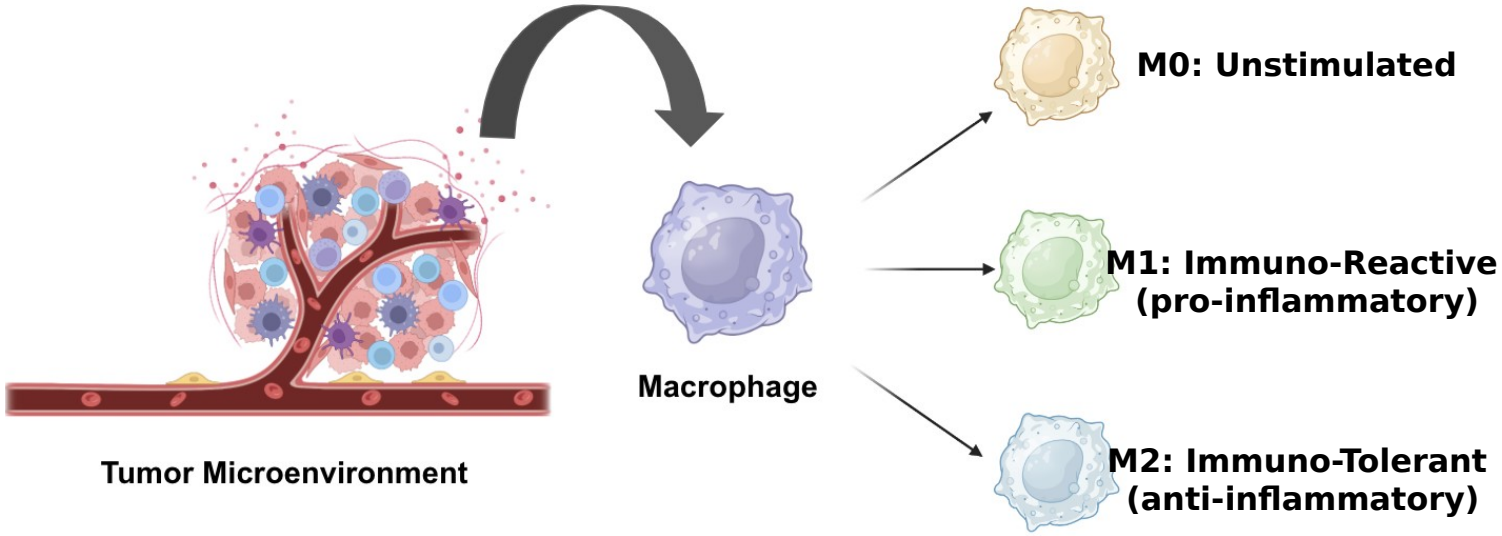
Ekta Dadlani _ Boolean Lab (University of California, San Diego)

Corresponding Authors: Dr. Tirtharaj Dash, Dr. Debashis Sahoo

Primary Advisor: Dr. Debashis Sahoo

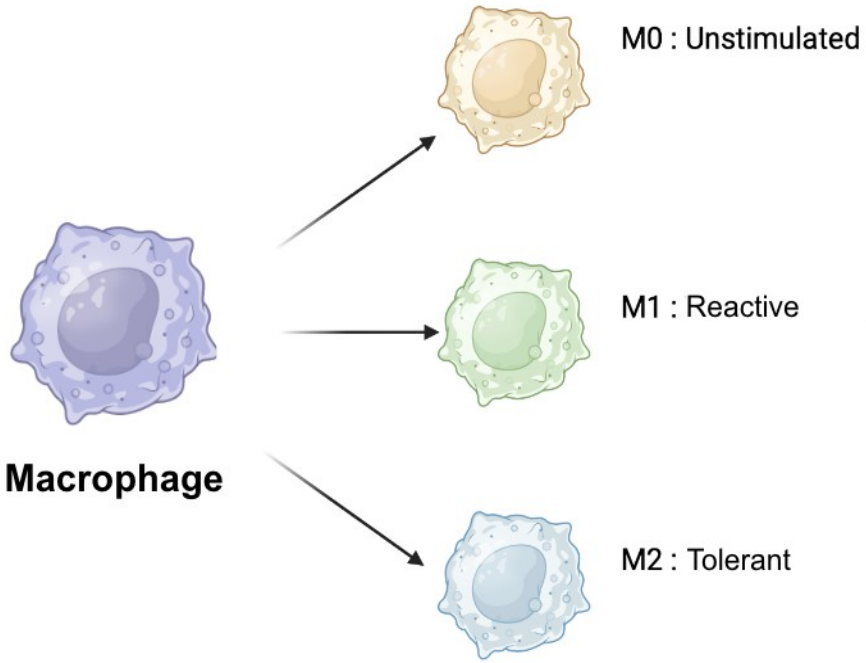
Colorectal Cancer & Tumor Associated Macrophages (TAMs)

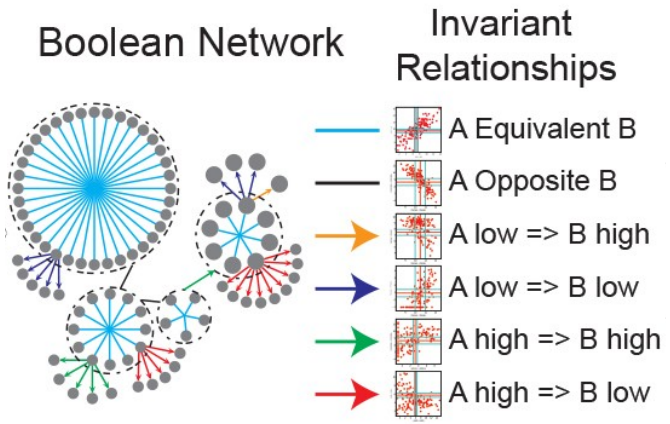


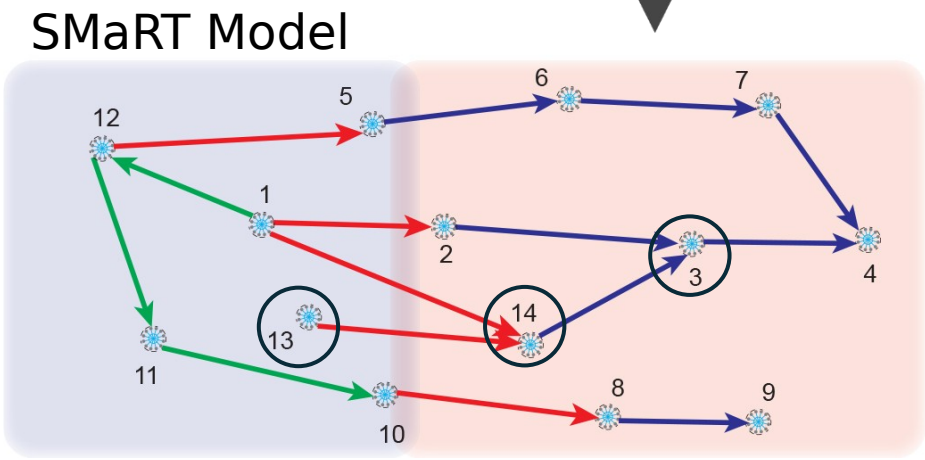


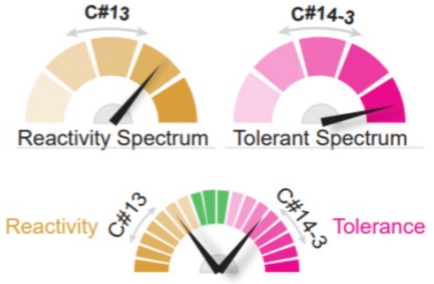
"Worldwide Cancer Data: World Cancer Research Fund International." *WCRF International*, 2022, www.wcrf.org/cancer-trends/worldwide-cancer-data/.

Universal Macrophage Polarization Gene Signature: SMaRT Model



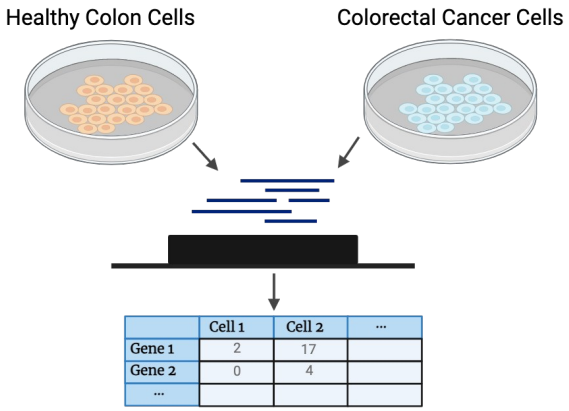




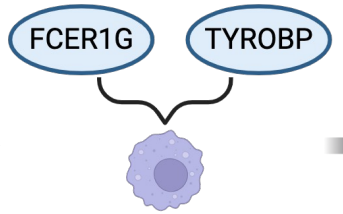


Ghosh P et. al Machine Learning Identifies Signatures of Macrophage Reactivity and Tolerance that Predict Disease Outcomes. bioRxiv; 2022. DOI: 10.1101/2022.06.27.497783.

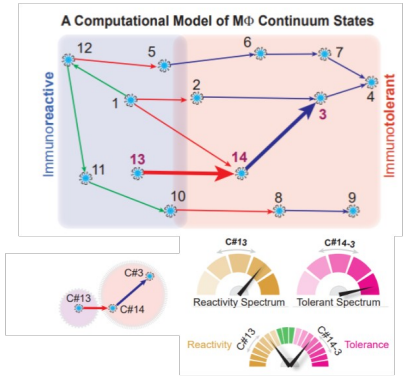
AI-Assisted Investigation of TAM Polarization in Colorectal Canc



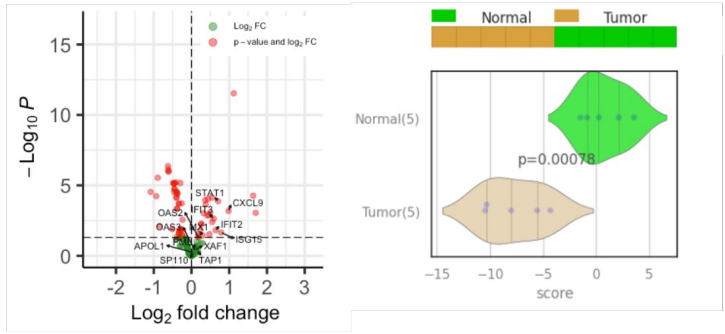
Single Cell Isolation & Sequencing



Macrophage Cell Extraction



SMaRT Computational Model



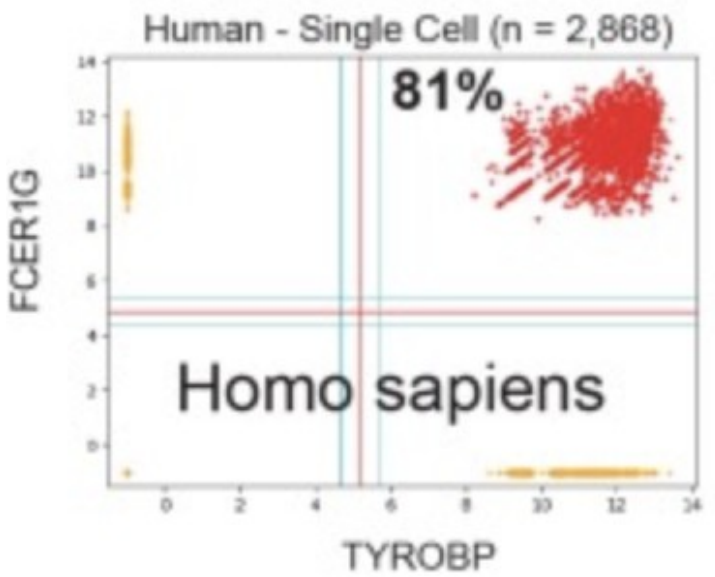
Refined TAMs Signature

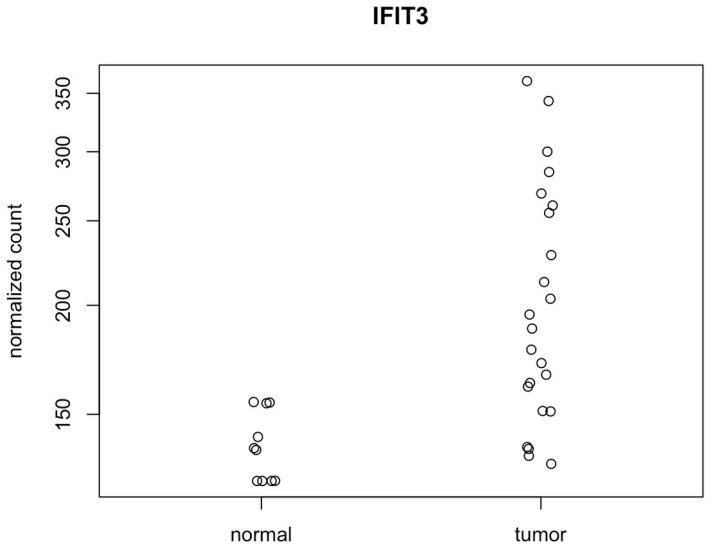
Goal:

Refine the universal macrophage gene signature (SMaRT signature) to capture the polarization dynamics of tumor-associated macrophages in colorectal cancer patients.

Primary Analytical Tools:
Data Collection, Macrophage Extraction, Refinement

	Cell 1	Cell 2	...
Gene 1	2	17	
Gene 2	0	4	
...			





Annotated Data Collection:

NCBI GEO Database

Barrett T et. al NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013; PMCID: PMC3531084.

Macrophage Cell Extraction:

TYROBP & FCER1G Expression

Dang D et. al Computational Approach to Identifying Universal Macrophage Biomarkers. Front Physiol. 2020; PMCID: PMC7156600.

Gene Signature Refinement:

T-test Dependent Differential Gene Expression Analysis

Liang P et. al Analysing differential gene expression in cancer. Nat Rev Cancer. 2003; PMID: 14668817.

Composite Score: Quantifying Gene Signatures

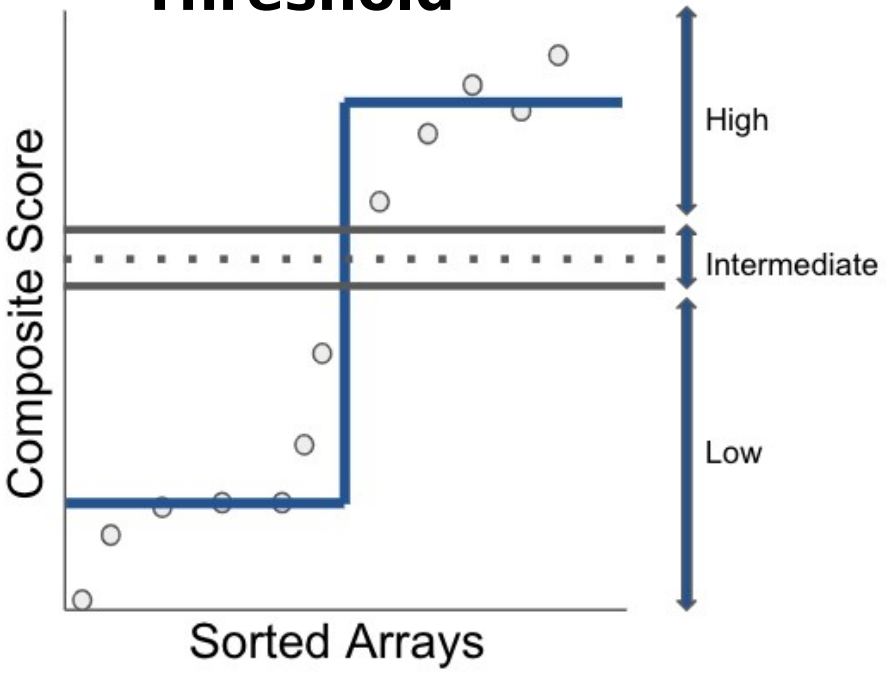
Immuno-Reactive Composite Score

$$-1 * \sum C13_{SMNorm}$$

Immuno-Tolerant Composite Score

$$1 * \sum C14_{SMNorm} + 2 * \sum C3_{SMNorm}$$

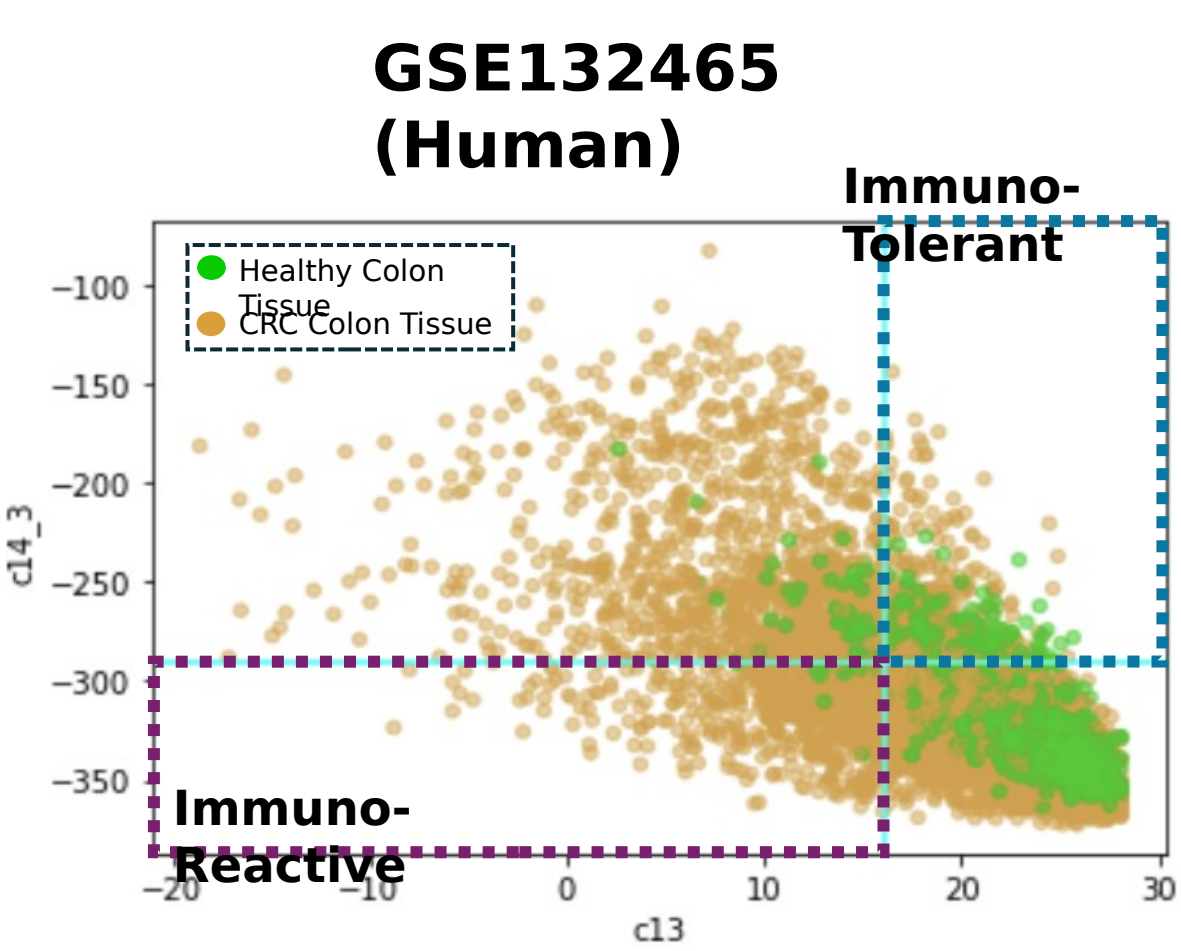
StepMiner Threshold

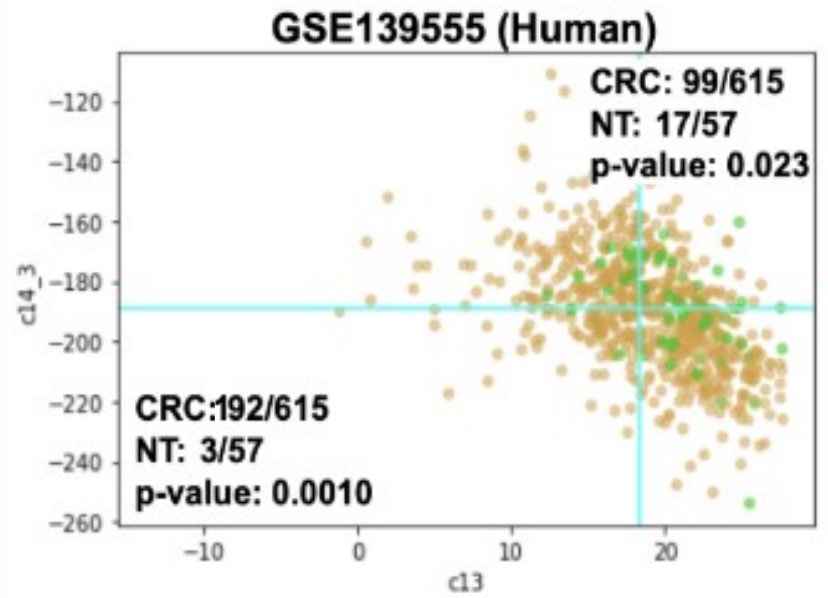


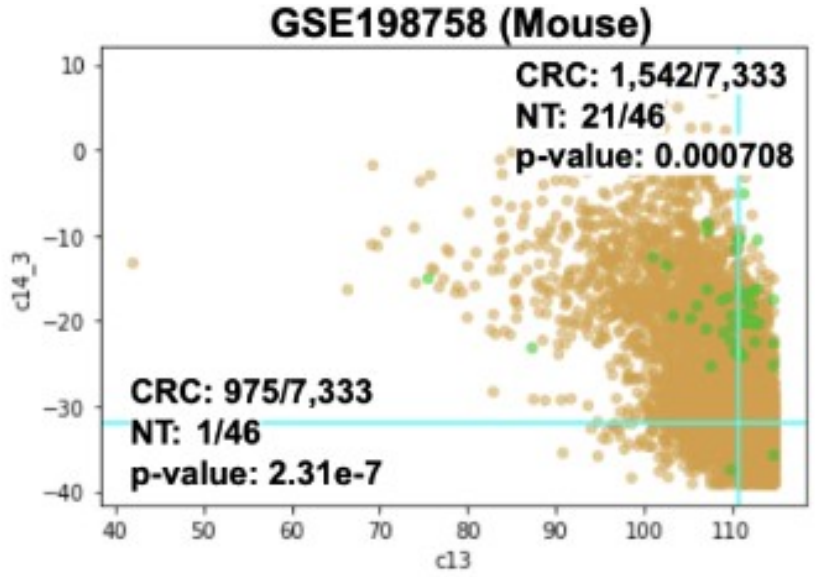
Ghosh P et. al Machine Learning Identifies Signatures of Macrophage Reactivity and Tolerance that Predict Disease Outcomes. bioRxiv; 2022. DOI: 10.1101/2022.06.27.497783.

Sahoo, Debashis. *Boolean analysis of high-throughput biological datasets*. Stanford University, 2008.

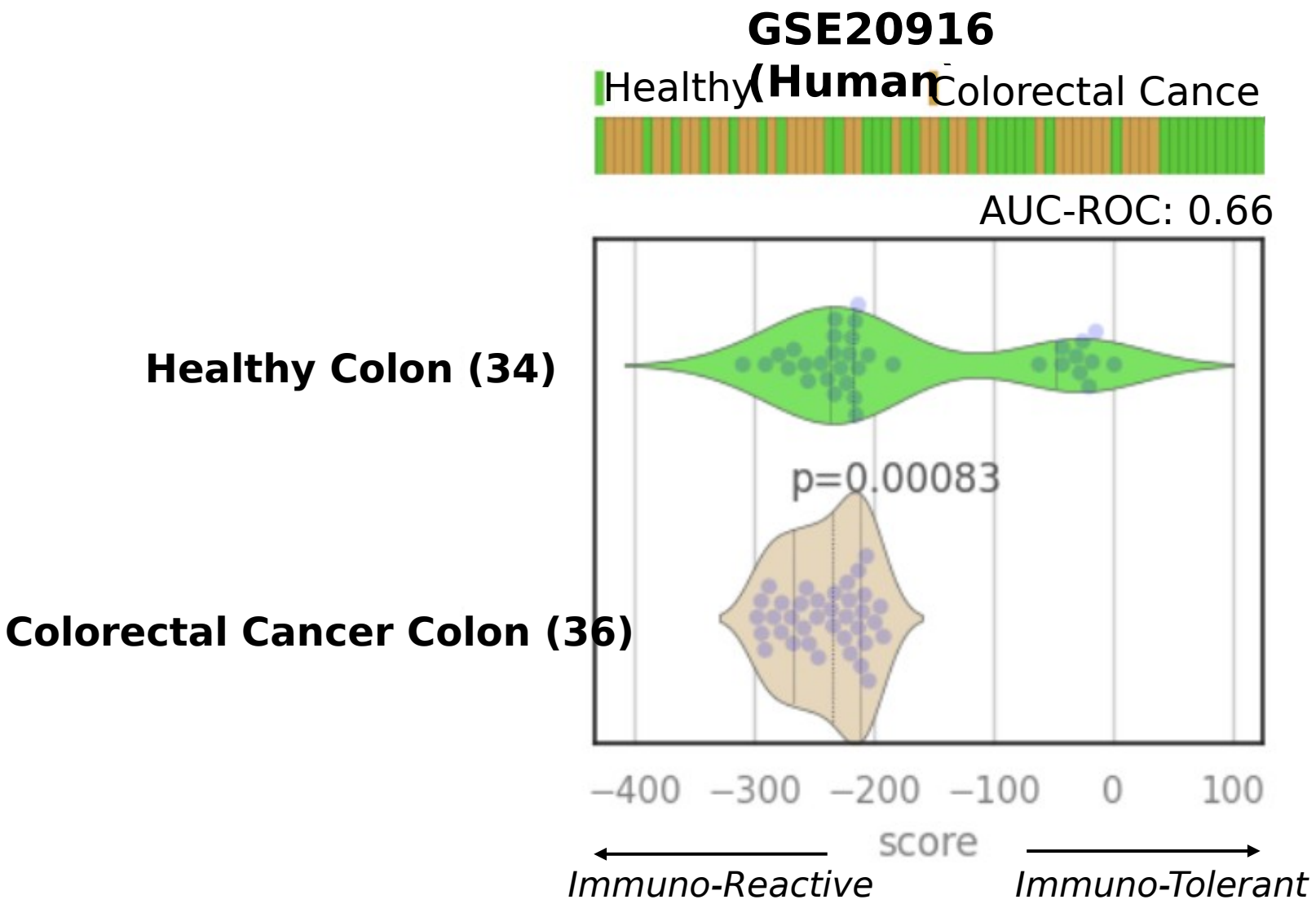
SMaRT Model Application: Immuno-Reactivity & Tumorous Macrophages



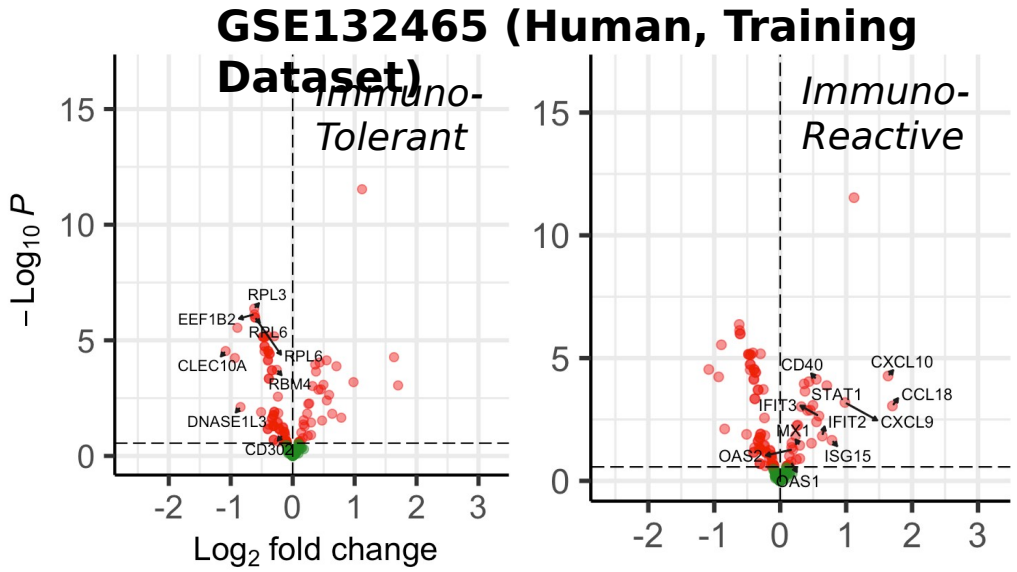


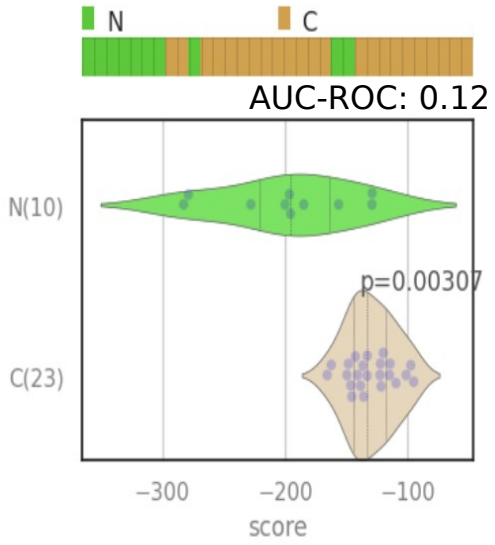


SMaRT Signature on Non-Macrophage Dataset: Lack of Healthy/CRC Separation

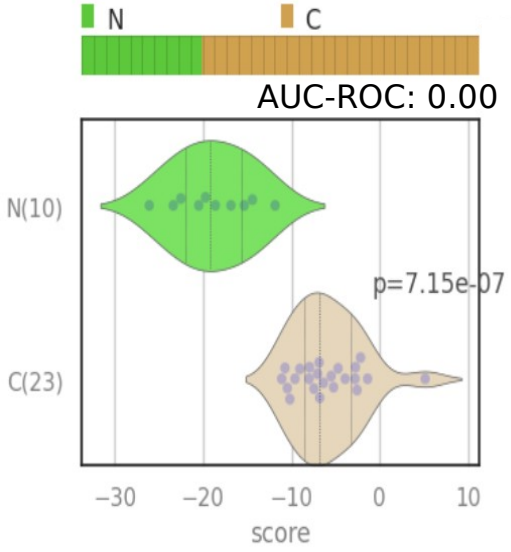


Refinement of SMaRT Signature: Healthy/CRC Separation

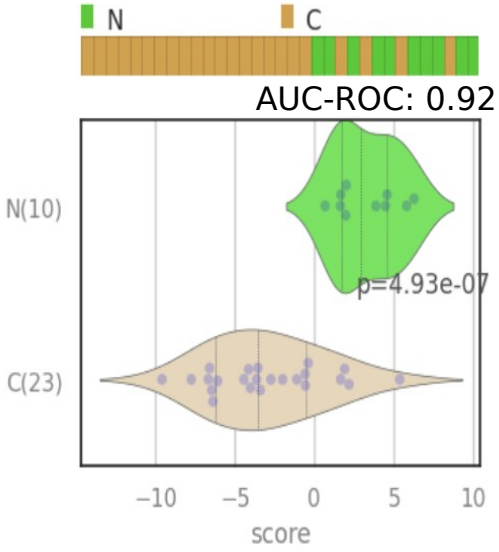




SMaRT Signature



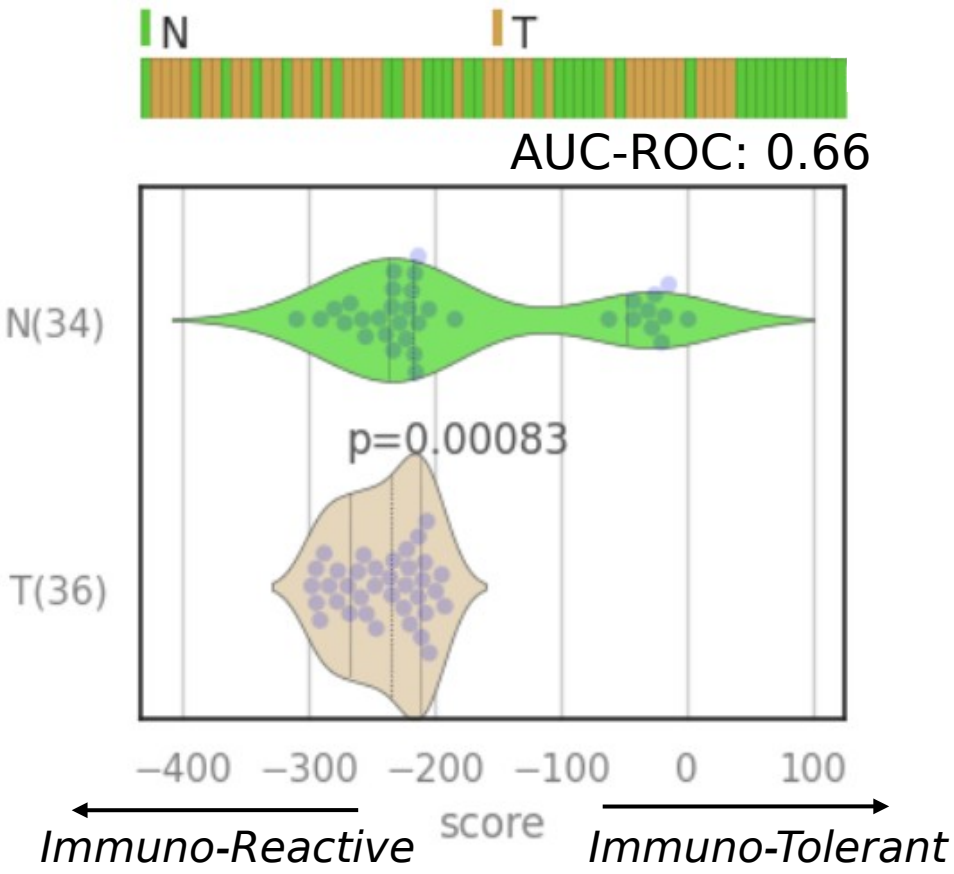
Noisy Signature



Refined Signature

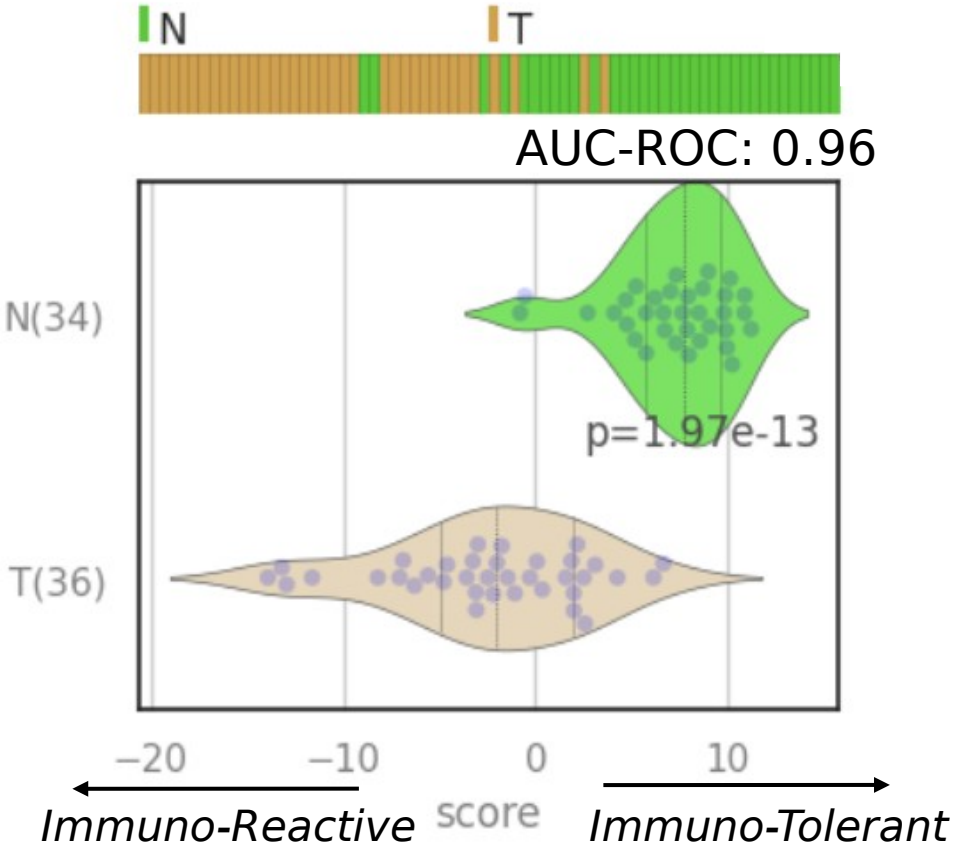
Validation of Refined TAMs Signature on Non-Macrophage Dataset: Healthy/CRC Separation

SMaRT Signature (338 Genes)



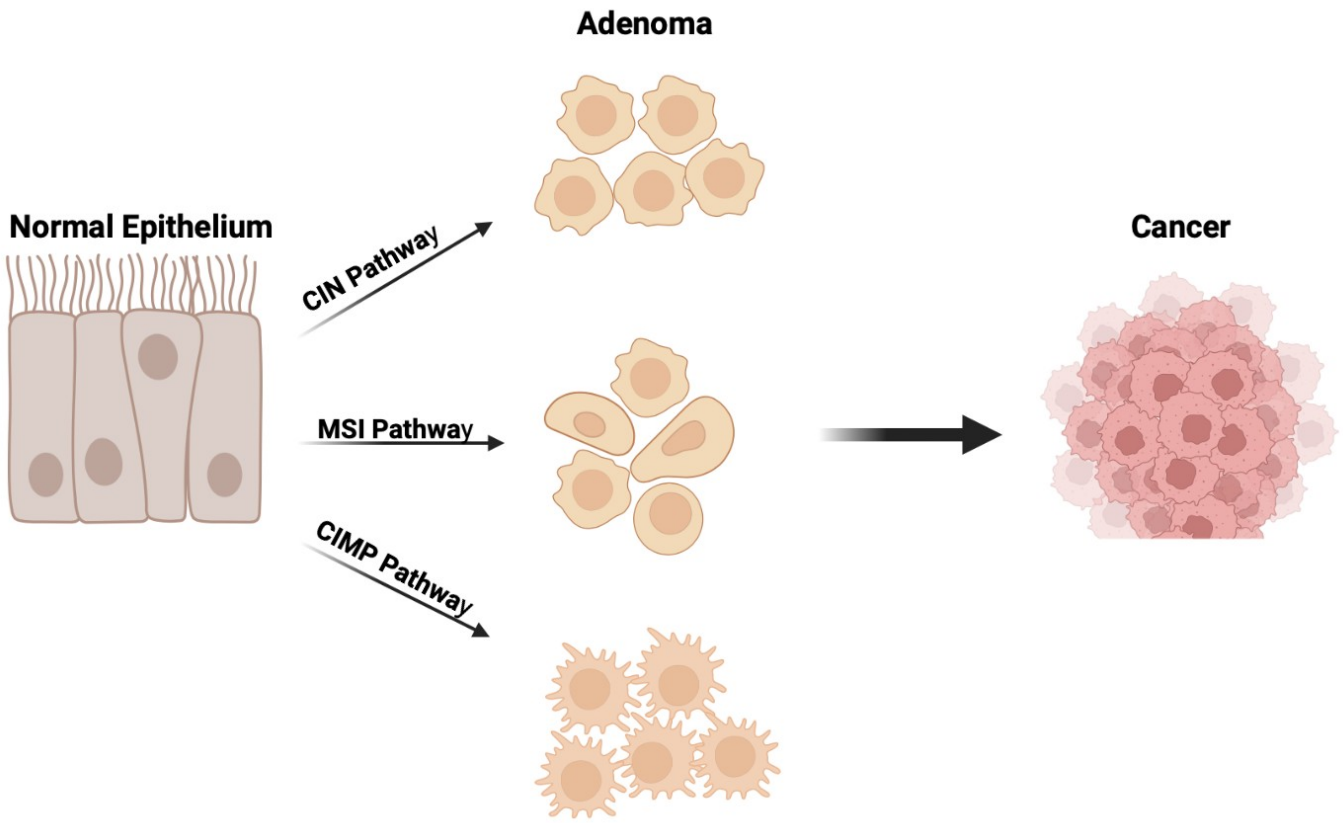
GSE20916 (Human)

Refined Signature (24 Genes)

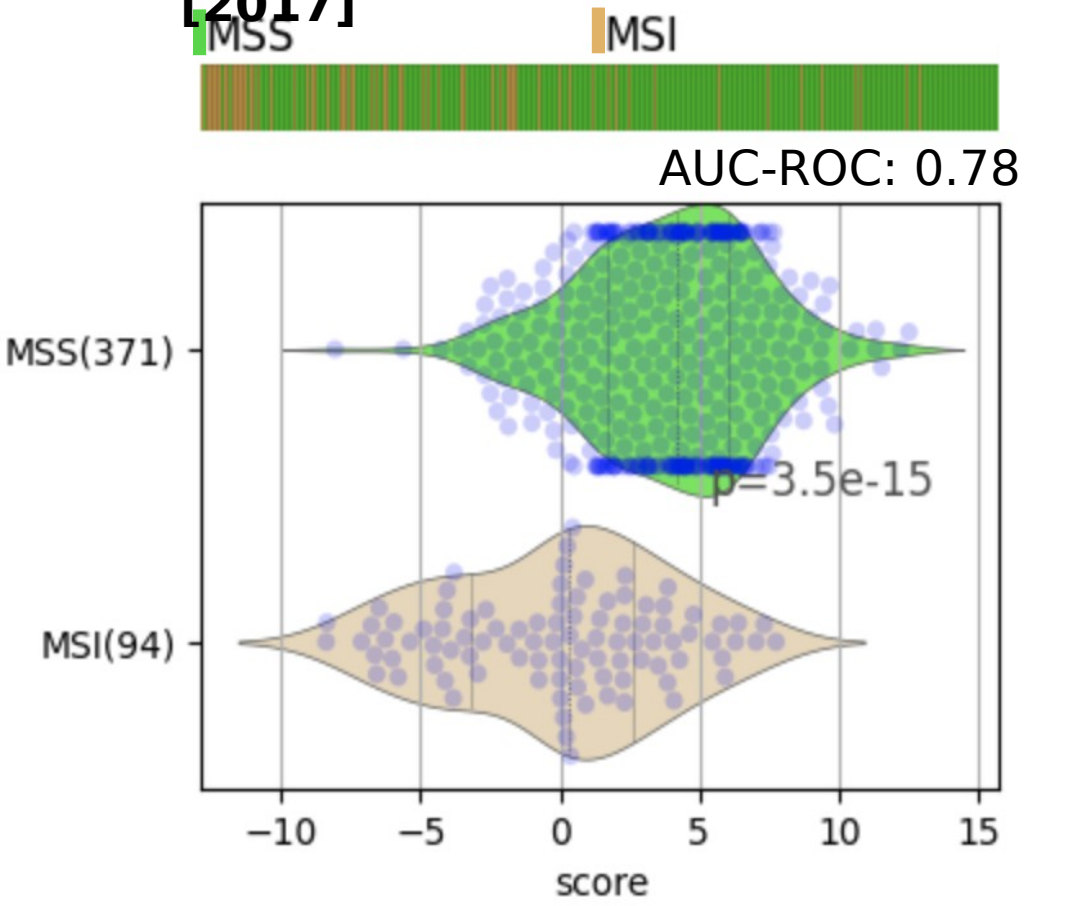


GSE20916 (Human)

Translational Potential of Refined TAMs Signature: MSI Pathway



The Cancer Genome Atlas (TCGA) [2017]



Thank You

Co-Authors & Mentors:

Dr. Debashis Sahoo
Dr. Tirtharaj Dash
Dr. Saptarshi Sinha
Dr. Pradipta Ghosh
Dr. Dharanidhar Dang
Dr. Shankar Subramaniam
Dr. Andrew McCulloch

Funding:

National Institute of Health (NIH)

Boolean Lab:

Daniella Vo
Sahar Taheri
H M Zabir Haque
Sara Safa McCoy
Amitash Nanda
Arya Prabhudesai
Atishna Samantaray
Weixiang Zhao
Manikya Varshney
Jung Liew
Ryan Wang
Rohan Subramaniam

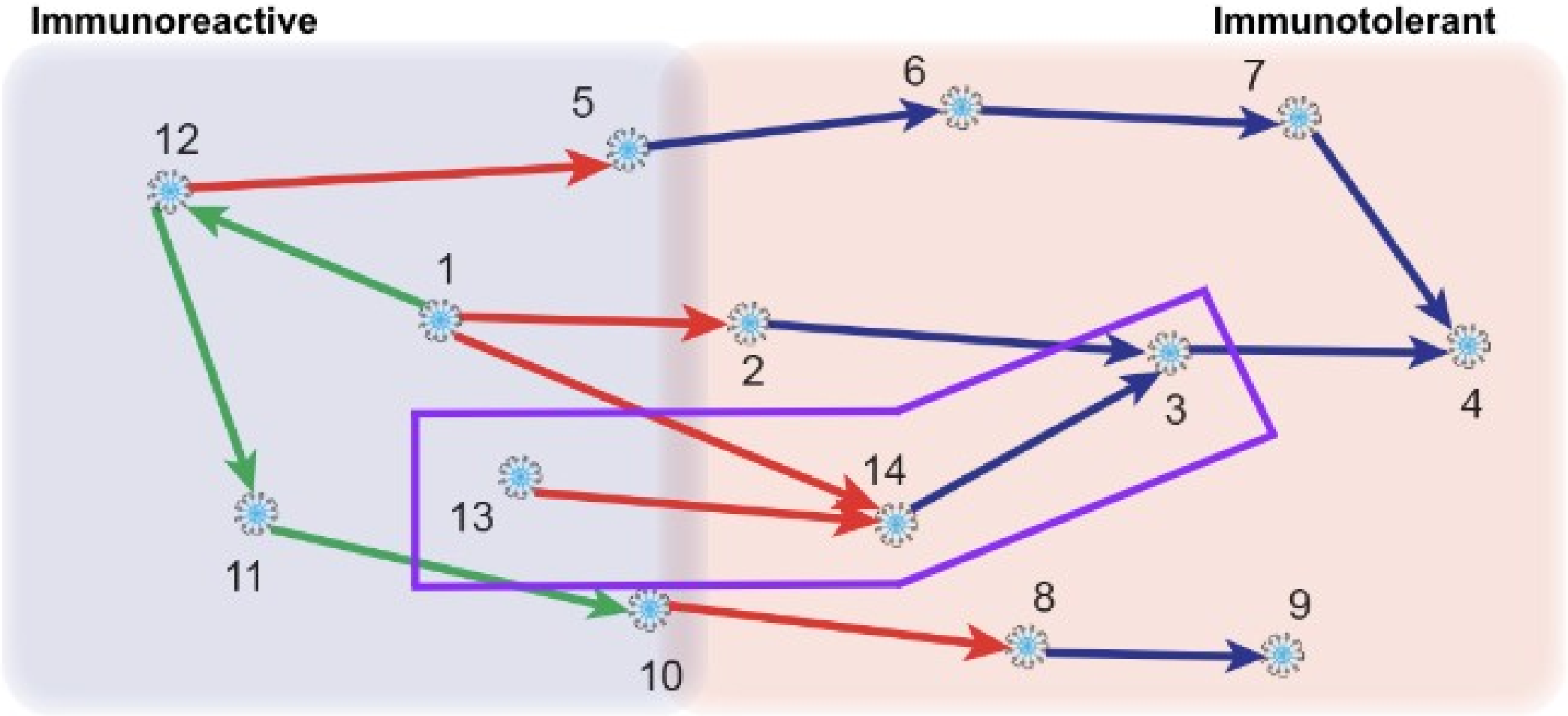
BioSys, 2024
ASPLOS, 2024

University of California, San Diego

Supplementary

SMaRT Model Development

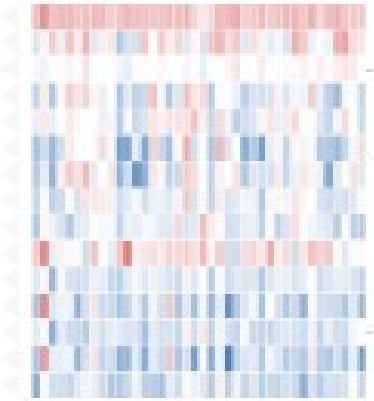
Underlying Boolean Implication Network: SMaRT Model



Built using a Clustered Boolean Implication Network using a pooled dataset (GSE134312, n = 197)

Boolean Logic

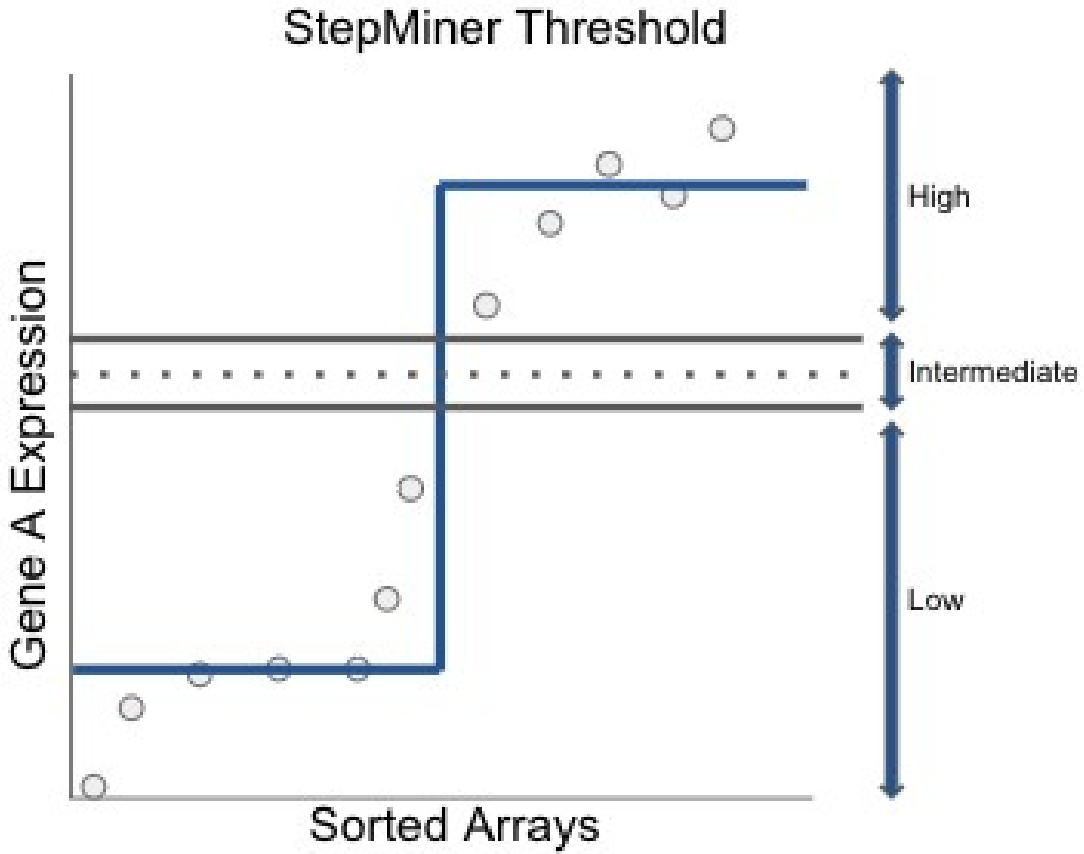
RNASeq or Microarray



Gene Expression Count Matrix

	Sample 1	Sample 2	...
Gene 1	2	17	
Gene 2	0	4	
...			

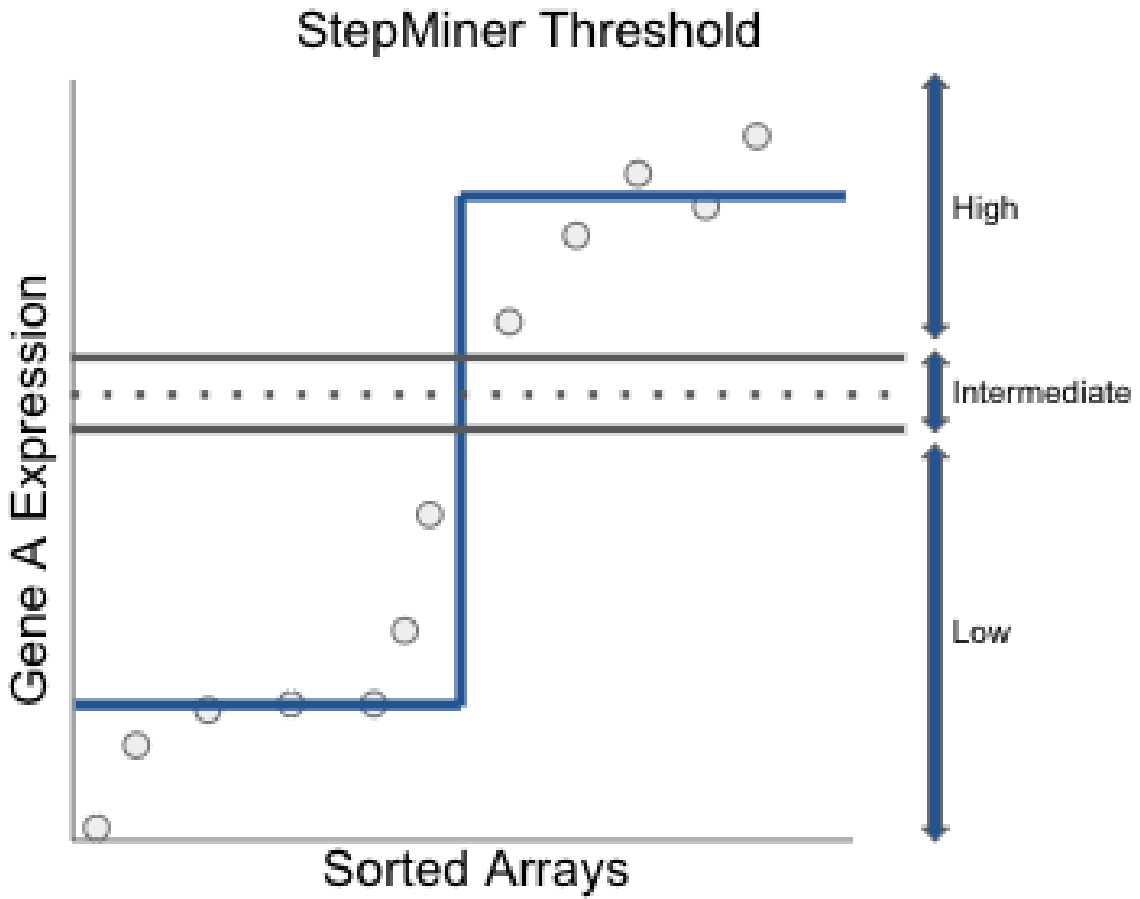
Normalization/Scaling



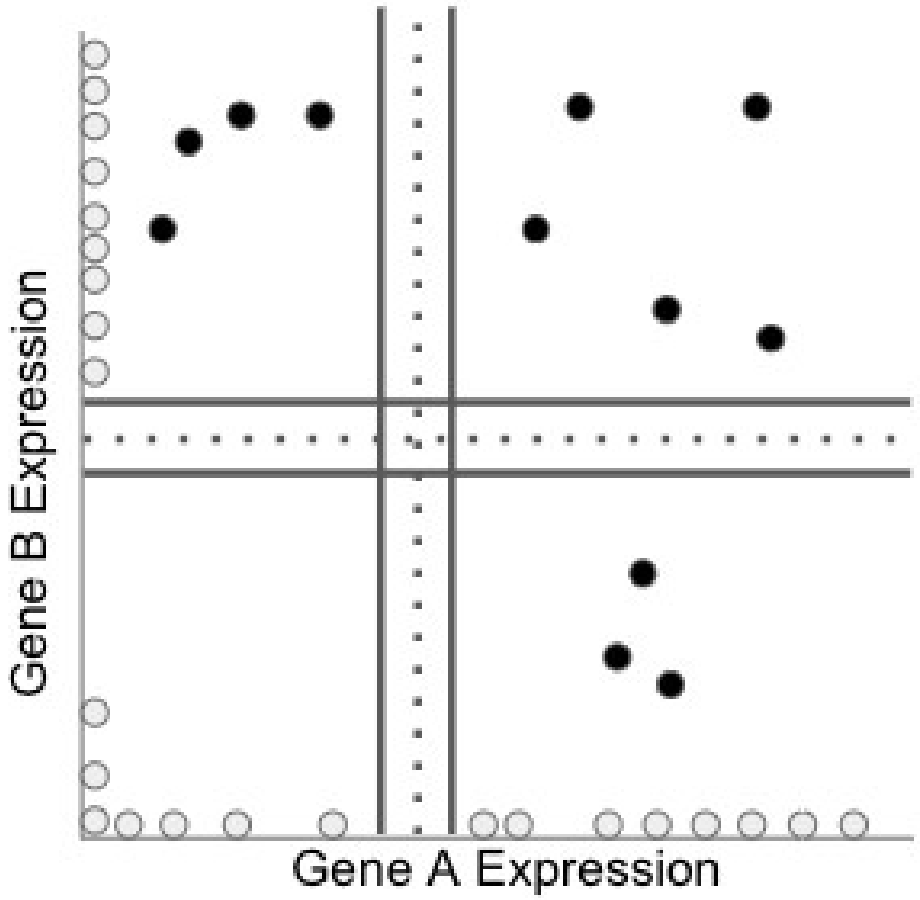
StepMiner

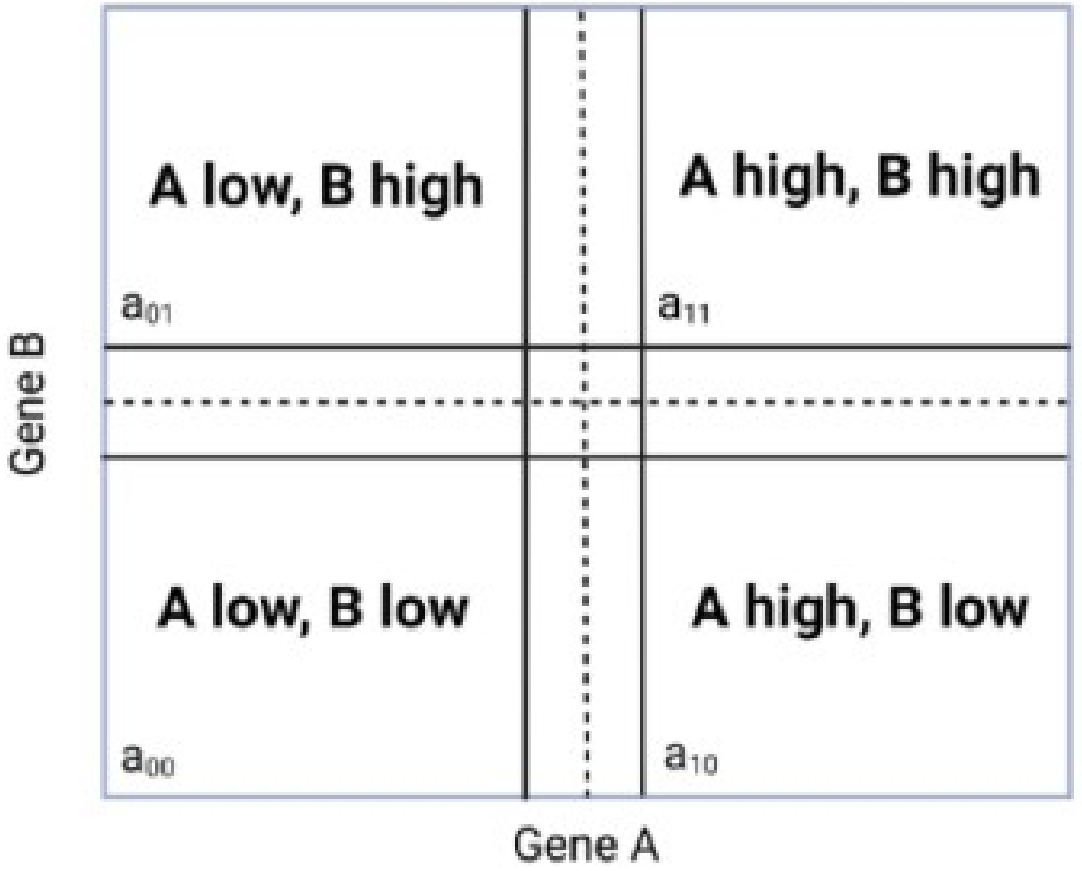
1. All possible step positions are evaluated
2. For each step position: determine the average of gene expression values on each side of the step
3. A regression scheme is used to choose the step position that minimizes the square error with the fitted data
4. A regression statistic (F-statistic, which determines whether the regression model significantly explains the variation in the dependent variables of if the variation can be attributed to random chance) is computed to determine whether the step is significant or not:

   F-stat: MSR/MSE
   = Mean sum of square regression/Mean sum of squares error

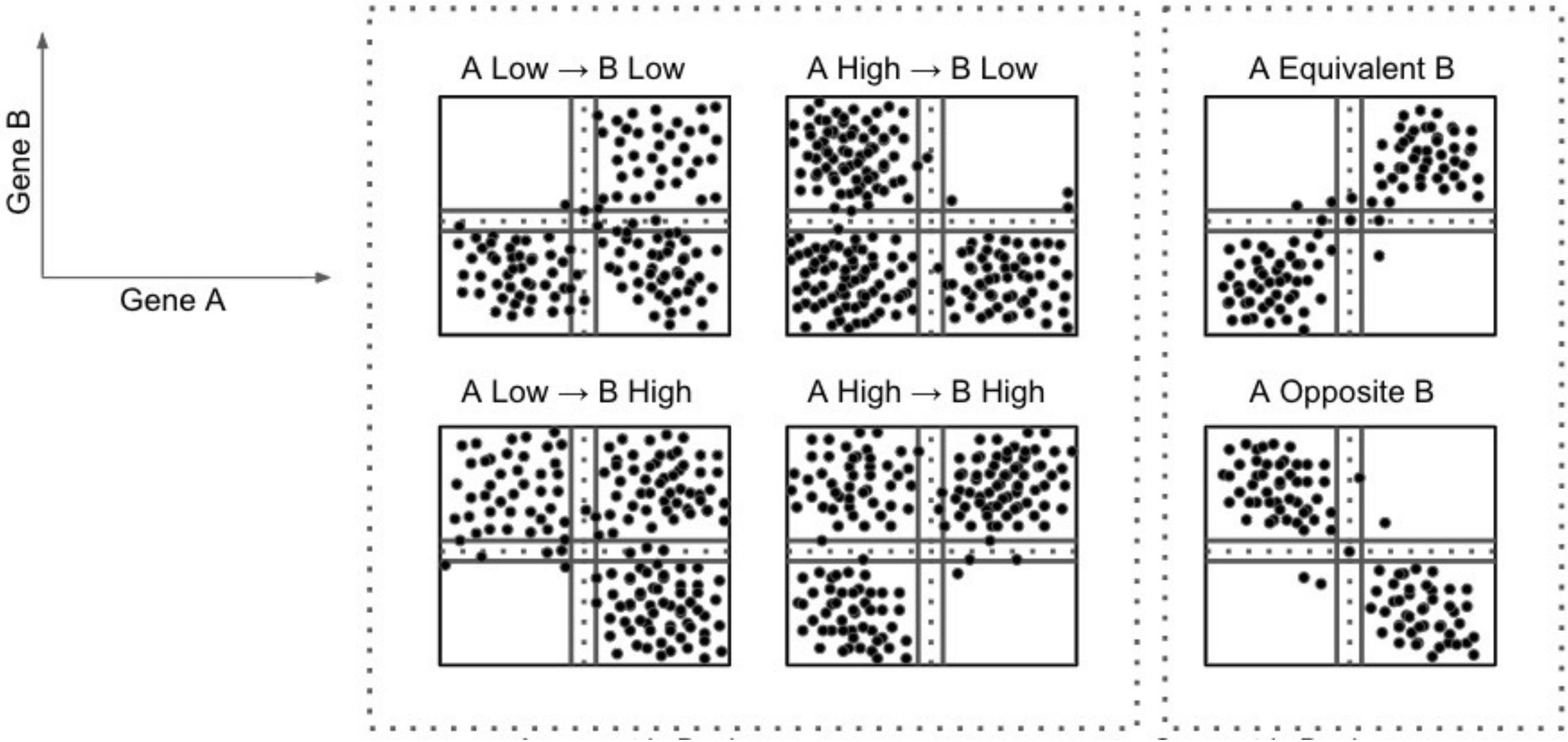


Boolean Implications



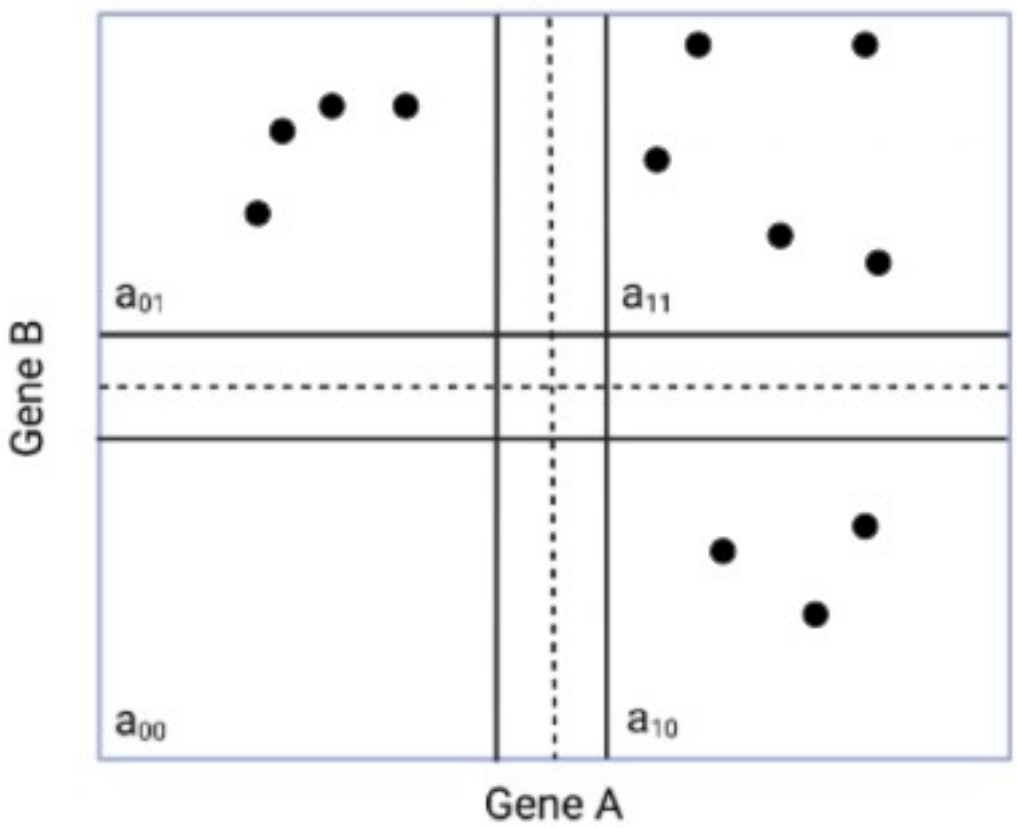


Boolean Implications Relationships



Asymmetric Boolean Relationships Symmetric Boolean Relationships

Boolean Implications Relationships



Boolean Implication: If-then rules

Significance determined on sparsity of quadrant and the metric's error rate:

Example: A low and B low (a_{00}) = sparse:

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

$$nA_{0*}=(a_{00}+a_{01}) \text{ and } p(A_{0*})=nA_{0*} / \text{total}$$

$$nB_{0*}=(a_{00}+a_{10}) \text{ and } p(B_{0*})=nB_{0*} / \text{total}$$

$$n'=p(A_{0*})*p(B_{0*})*\text{total}$$

$$n = \text{number of samples in } a_{00}$$

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

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

If $S_{ij} > 3$ and $p_{ij} < 0.1$, then a_{00} is sparse; implying that **when Gene A is low, Gene B must almost always be high.**

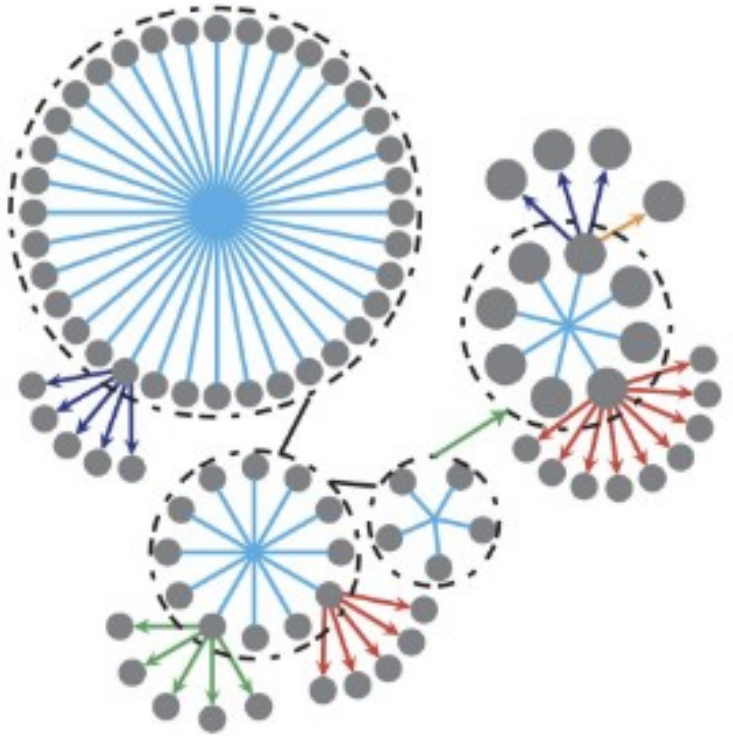
Clustered Boolean Implications Network

Boolean Implication Network

1. StepMiner Algorithm to Convert All Genes Into Binary Values of High Expression/Low Expression
2. Use Boolean Implication Relationships (BIRs) to classify the relationship between each pair of genes
3. Undirected edges = Symmetric Boolean Relationships & Directed edges = Asymmetric Boolean Relationships

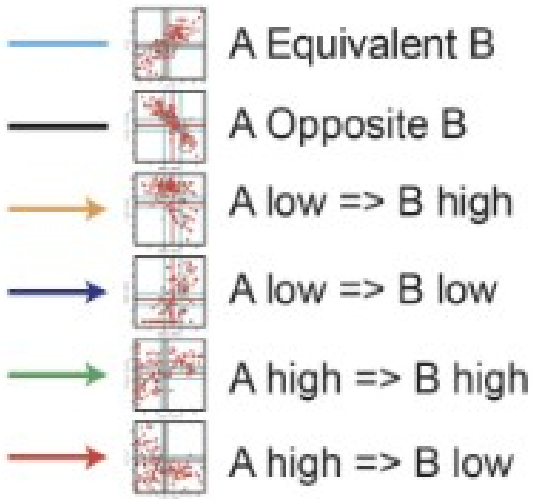
Clustered Boolean Implication Network

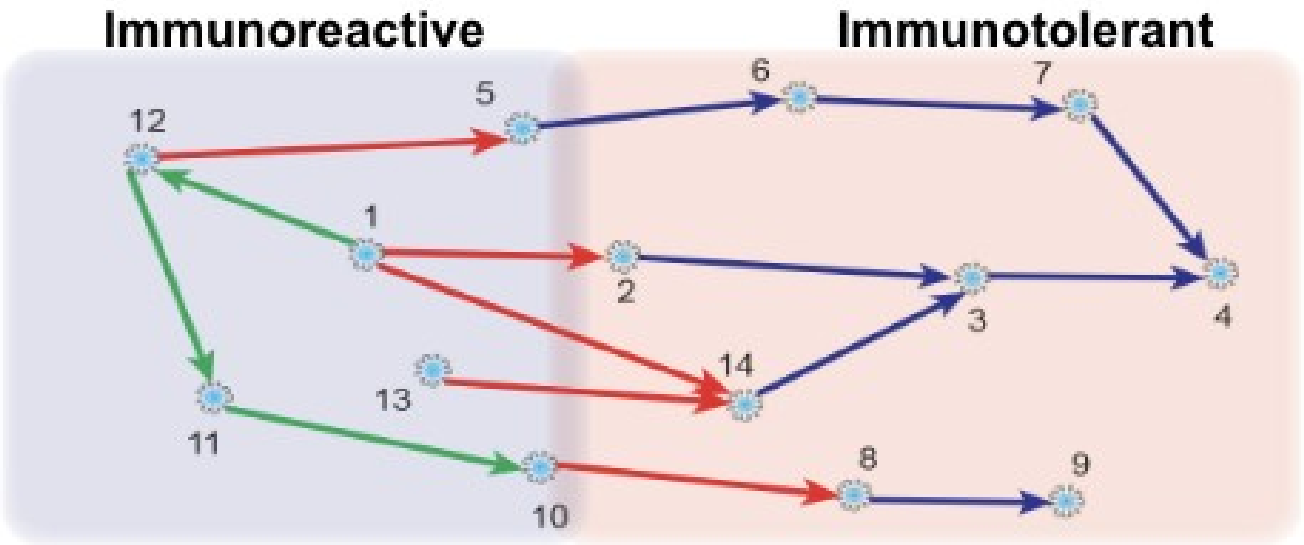
4. Use Equivalent BIRs to cluster genes
5. Use a threshold of >0.5 for the Jaccard similarity coefficient for every gene in every cluster to remove weak links (Opposite)
6. Remove clusters with single gene and repetitive edges (keep transitive edges)
7. Use Asymmetric Boolean Relationships as edges between clusters



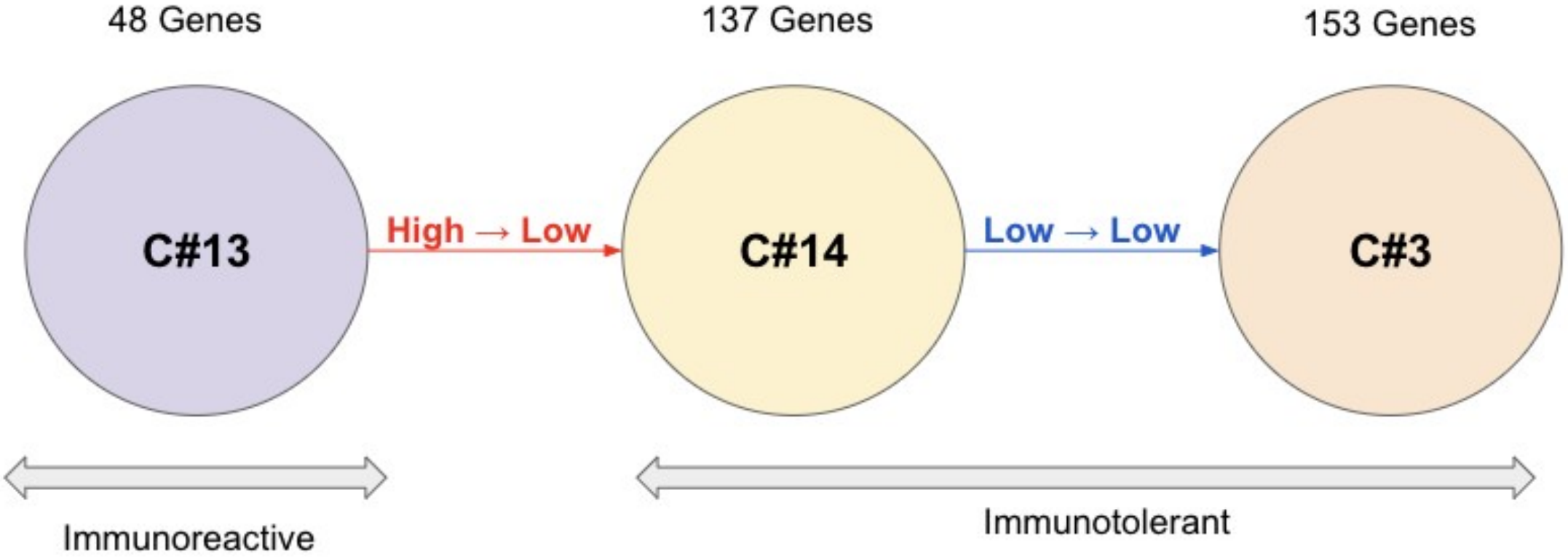
Transitive Boolean Paths

1. Annotate the cluster's macrophage polarization state using the metadata of the dataset
2. Boolean Path minimum: Two nodes (clusters) with a directed edge between them
3. Traversal of path: by Depth First Traversal (DFS)





Composite Score



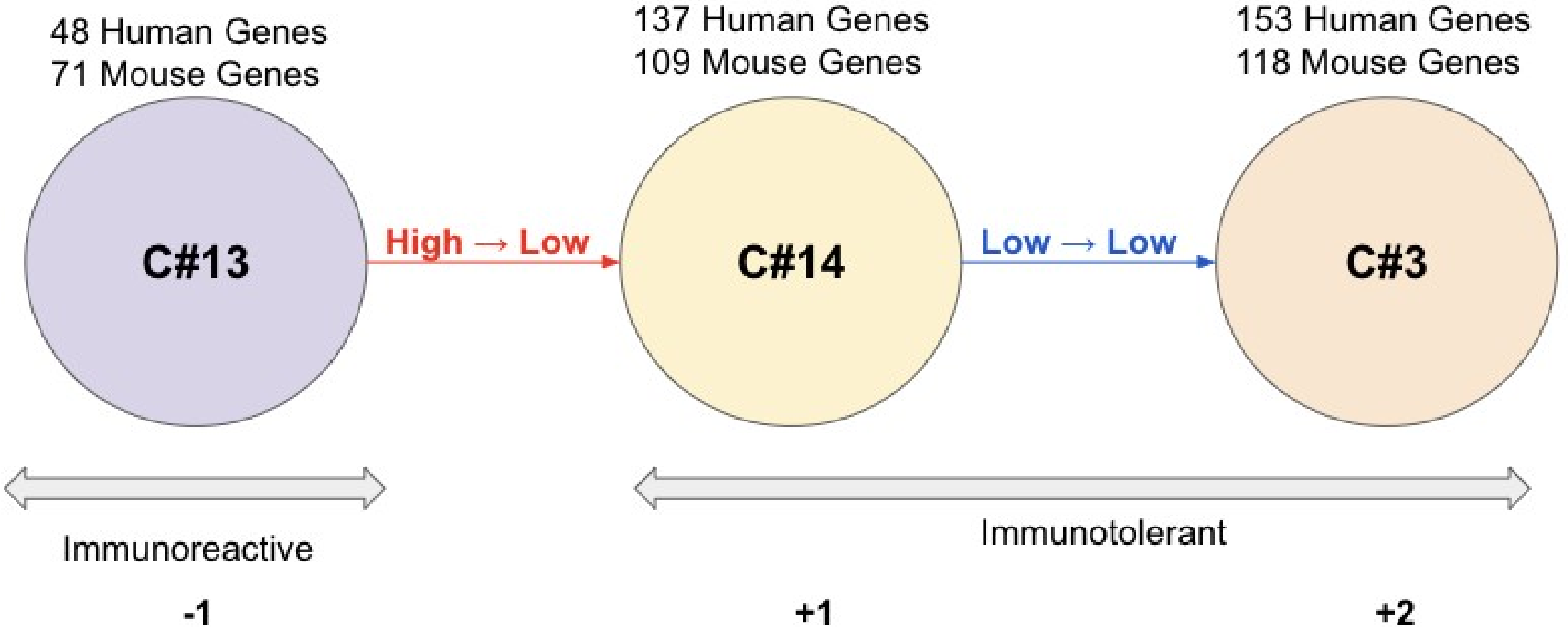
expr = gene expression value for a given sample

θ_{SM} = StepMiner threshold separating low/high expression

σ = standard deviation of expression value from the threshold

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

Composite Score



Composite Score: $-1 * \sum C13_{SMNorm} + 1 * \sum C14_{SMNorm} + 2 * \sum C3_{SMNorm}$

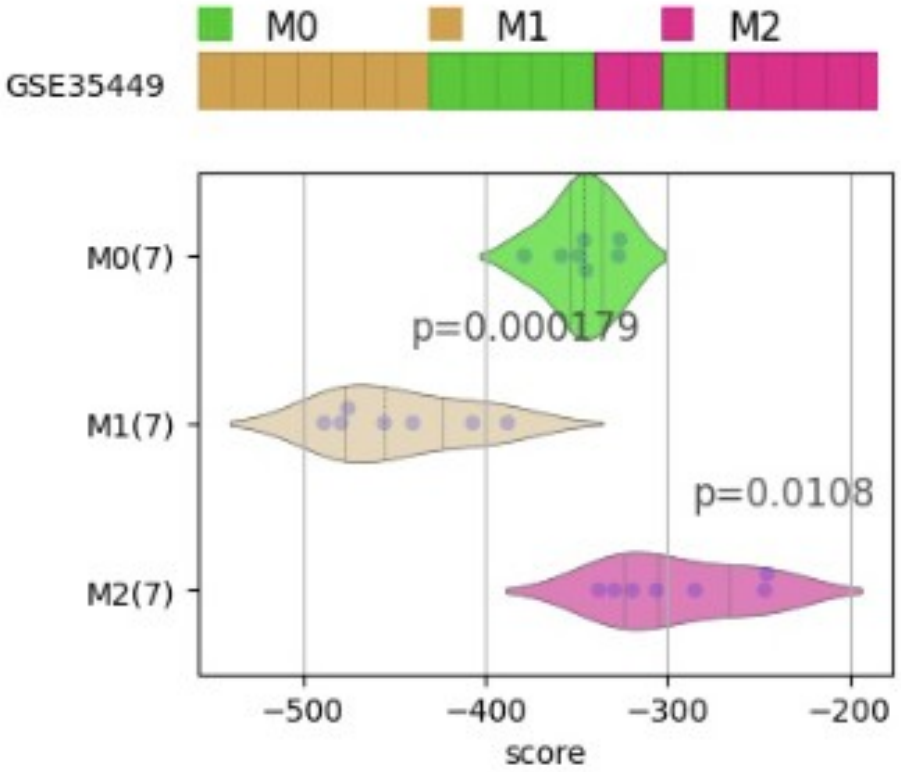
Predictive Model

We know:

1. Composite score for each sample
2. The macrophage polarization state of each sample (annotated)

We can train a simple logistic regression model to predict the macrophage polarization state.

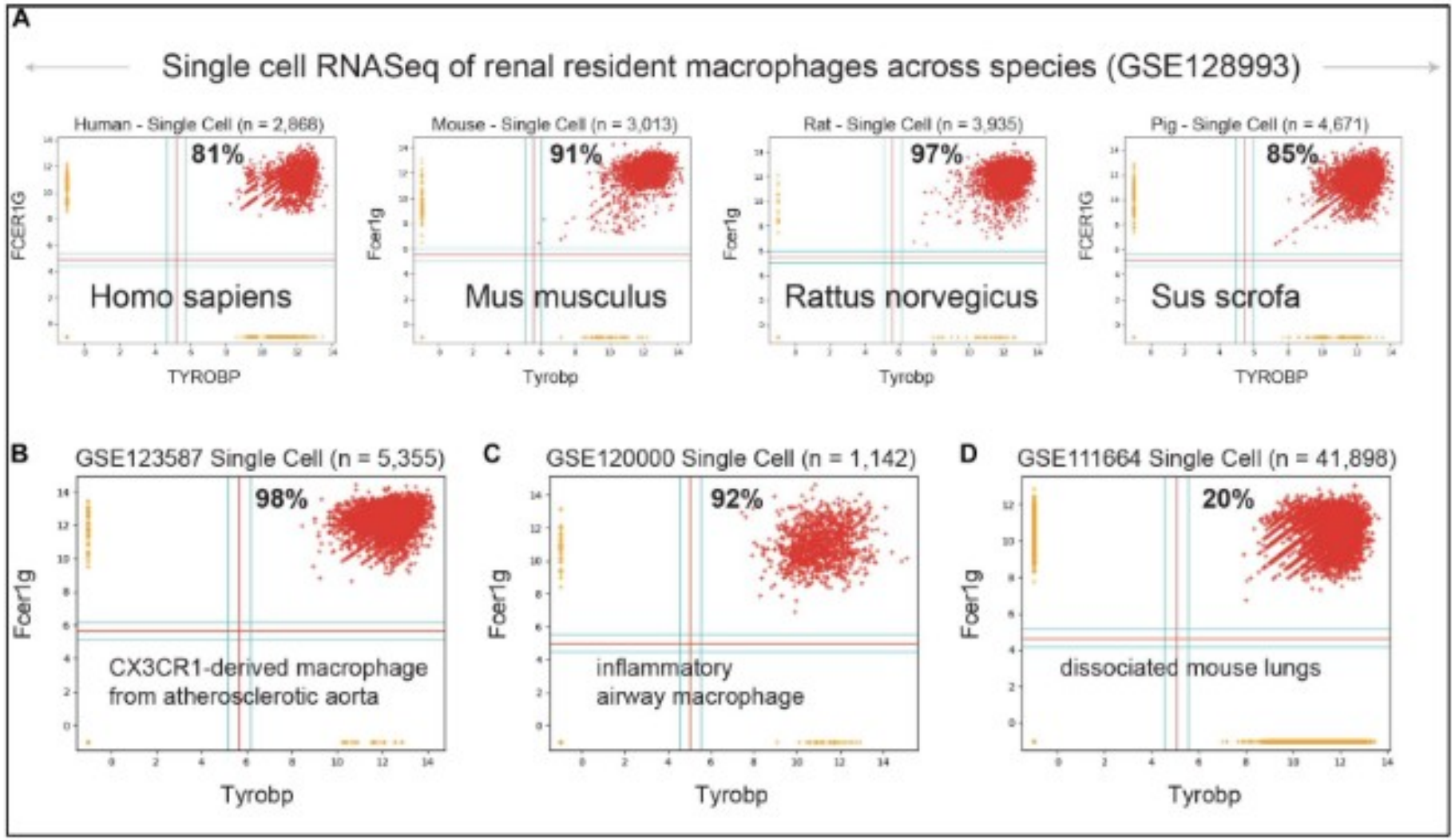
We use the ROC-AUC metric to measure the performance of classification.



ROC-AUC:
Separating M0 and M1: 1.0
Separating M0 and M2: 0.92
Separating M1 and M2: 1.0

Macrophage Extraction

Macrophage Extraction



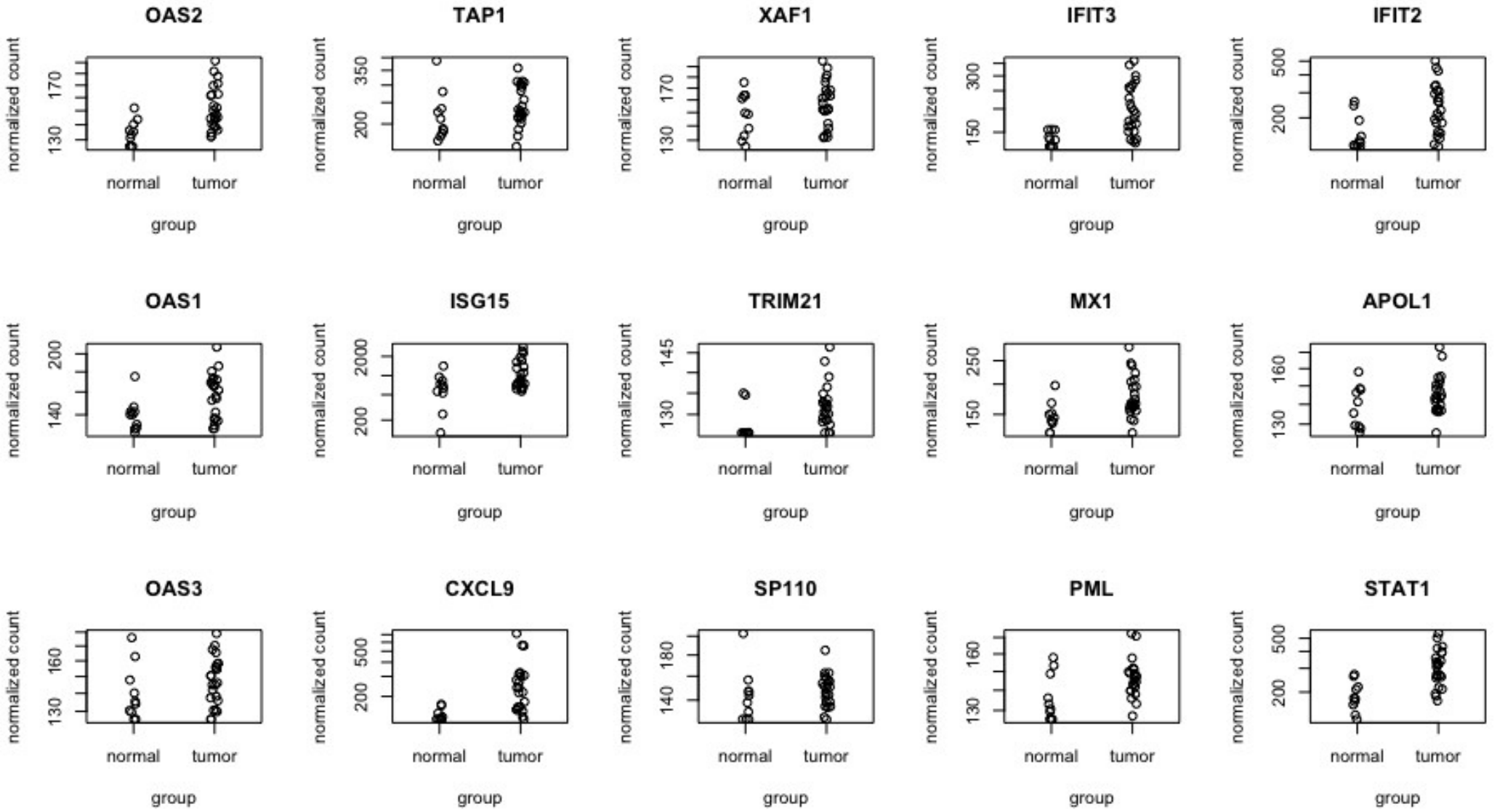
Dang D, Taheri S, Das S, Ghosh P, Prince LS, Sahoo D. Computational Approach to Identifying Universal Macrophage Biomarkers. Front Physiol. 2020 Apr 8;11:275. doi: 10.3389/fphys.2020.00275. PMID: 32322218; PMCID: PMC7156600.

Differential Gene Expression

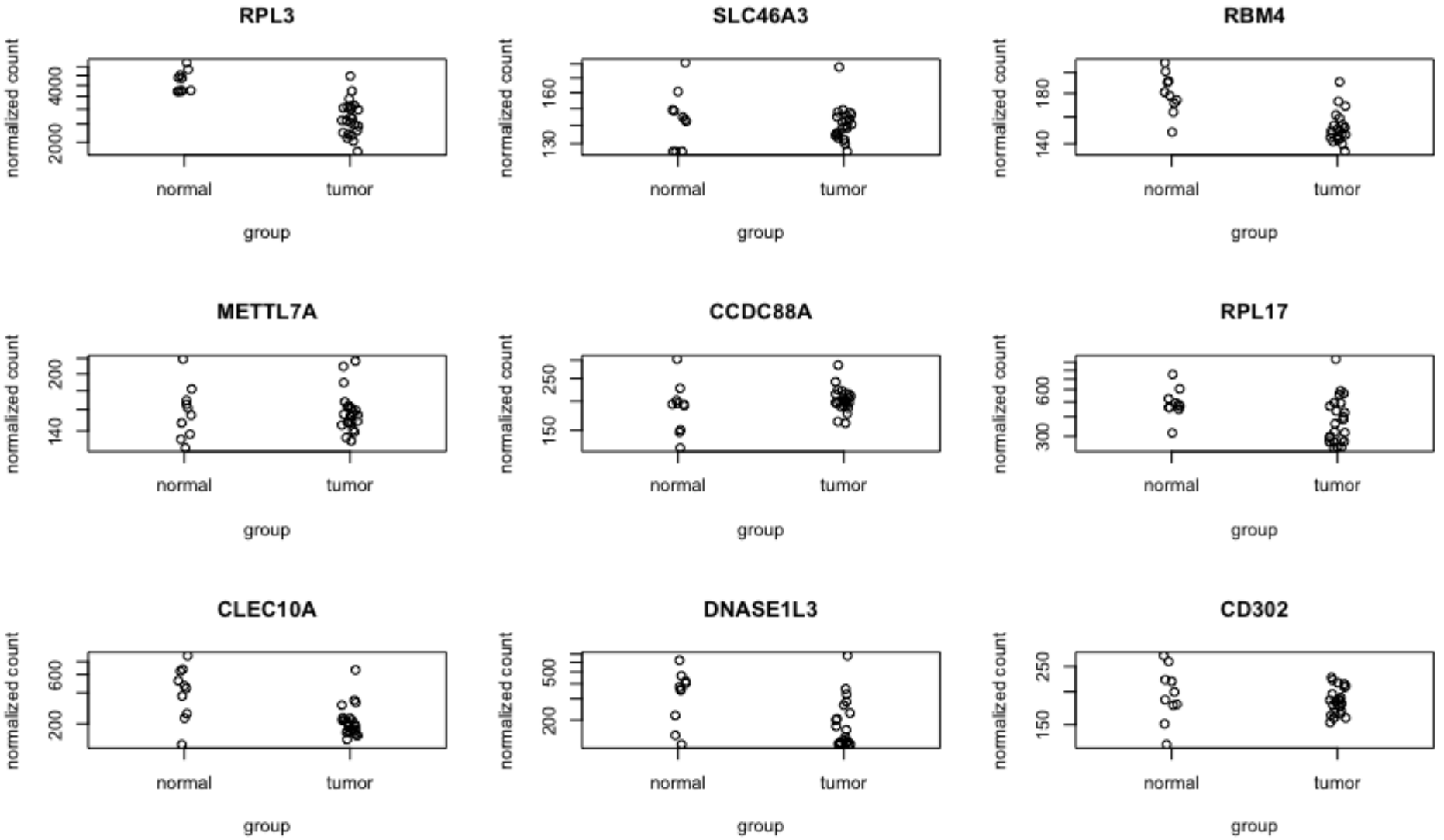
Differential Gene Expression Analysis

- Goal: Identify genes that have different expression patterns across a set of conditions due to a biological phenomenon (not technical variation)
- Technical Variation arises in the form of:
  - Library size: Number of reads that map to a sample (Influenced by sequencing depth)
  - Library composition: Can cause inflation of gene expression in a condition (i.e: organ specific genes where gene is expressed in one organ but not in the other)
  - Both of these factors are accounted for by normalization/log2 scaling of gene expression values
- Output:
  - Difference in the mean expression of a gene between the two binary groups (i.e: M2 - M1): fold-change
  - T-test with the null hypothesis that the two genes have the same average values. Treats the mean values as two independent values : difference in mean/variance

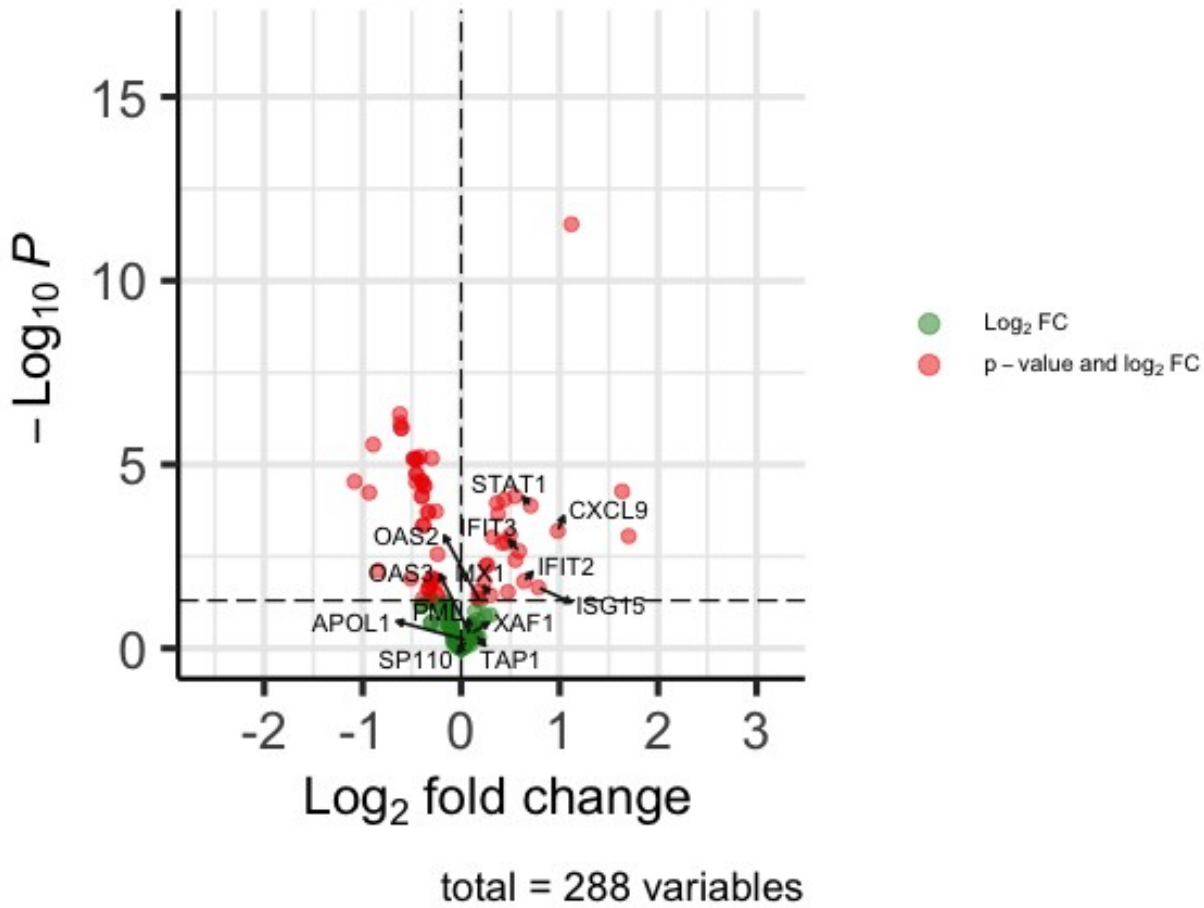
Immuno-Reactive Genes



Immuno-Tolerant Genes

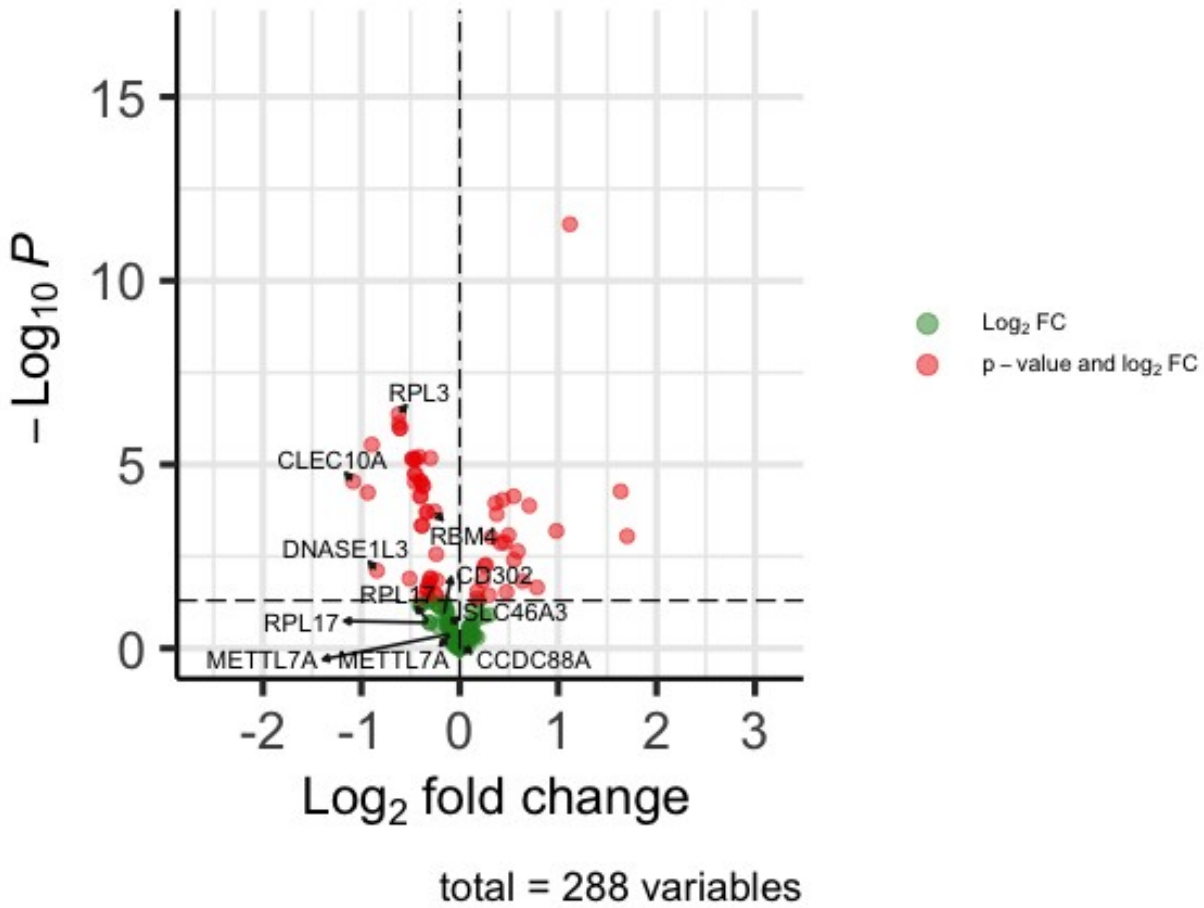


TAM Immuno-Reactive Genes



total = 288 variables

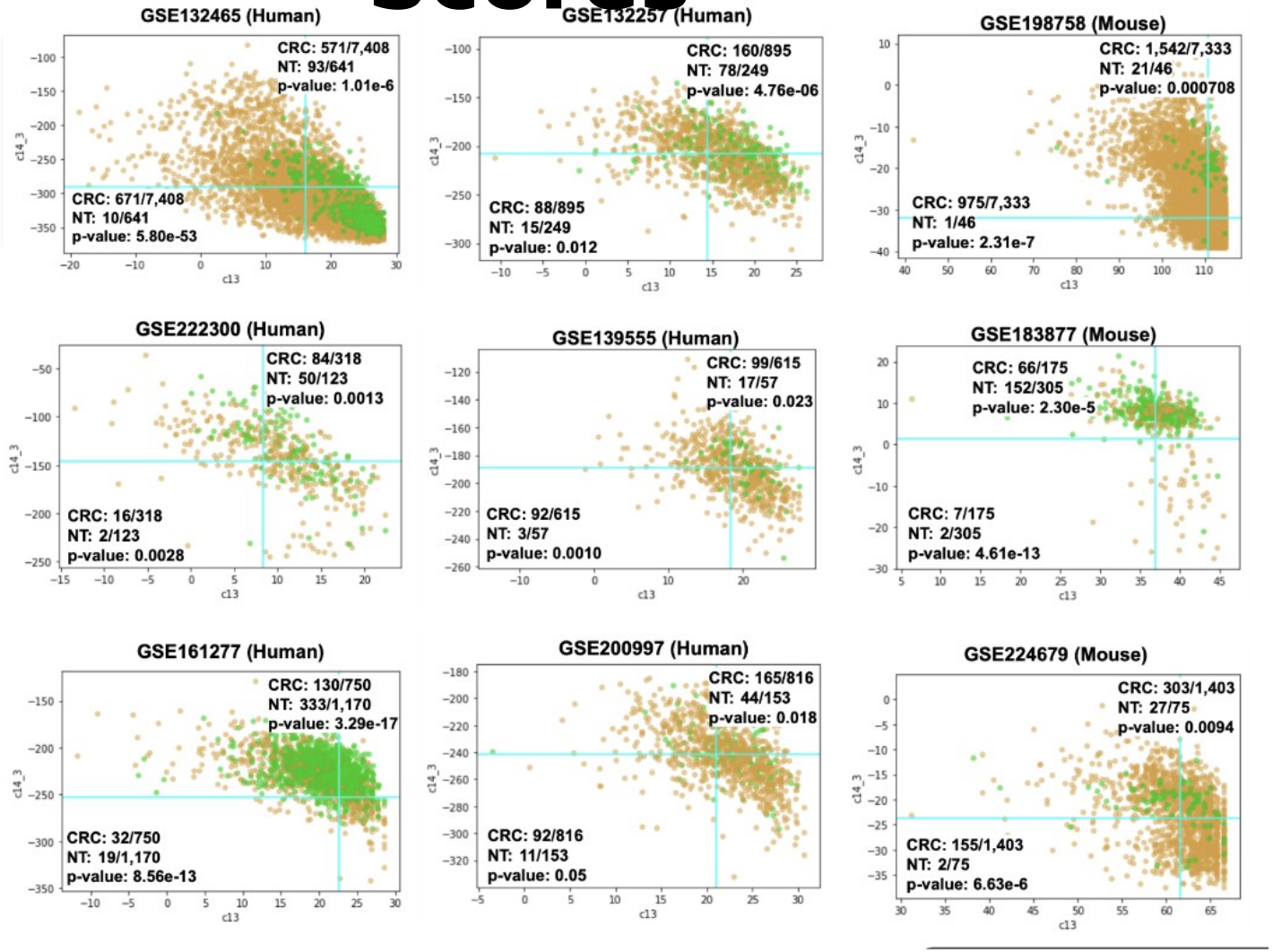
TAM Immuno-Tolerant Genes

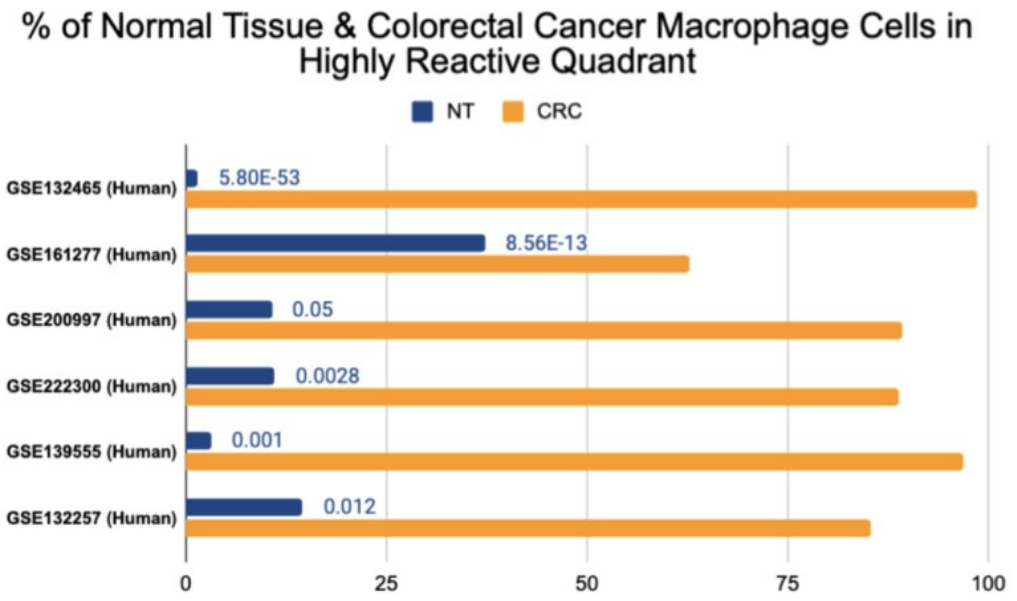


total = 288 variables

SMaRT Application

Using C13-C14-C3 Composite Scores

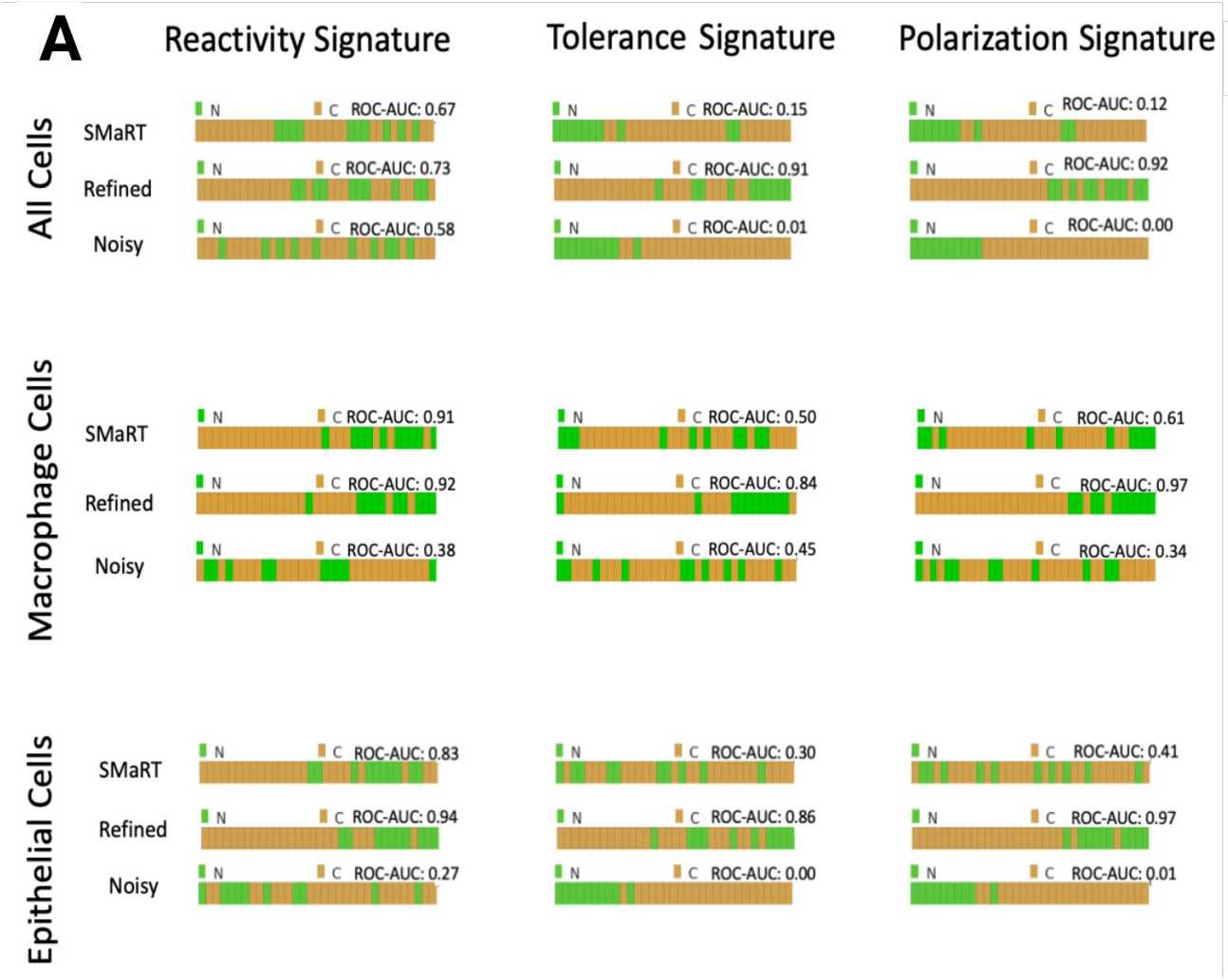


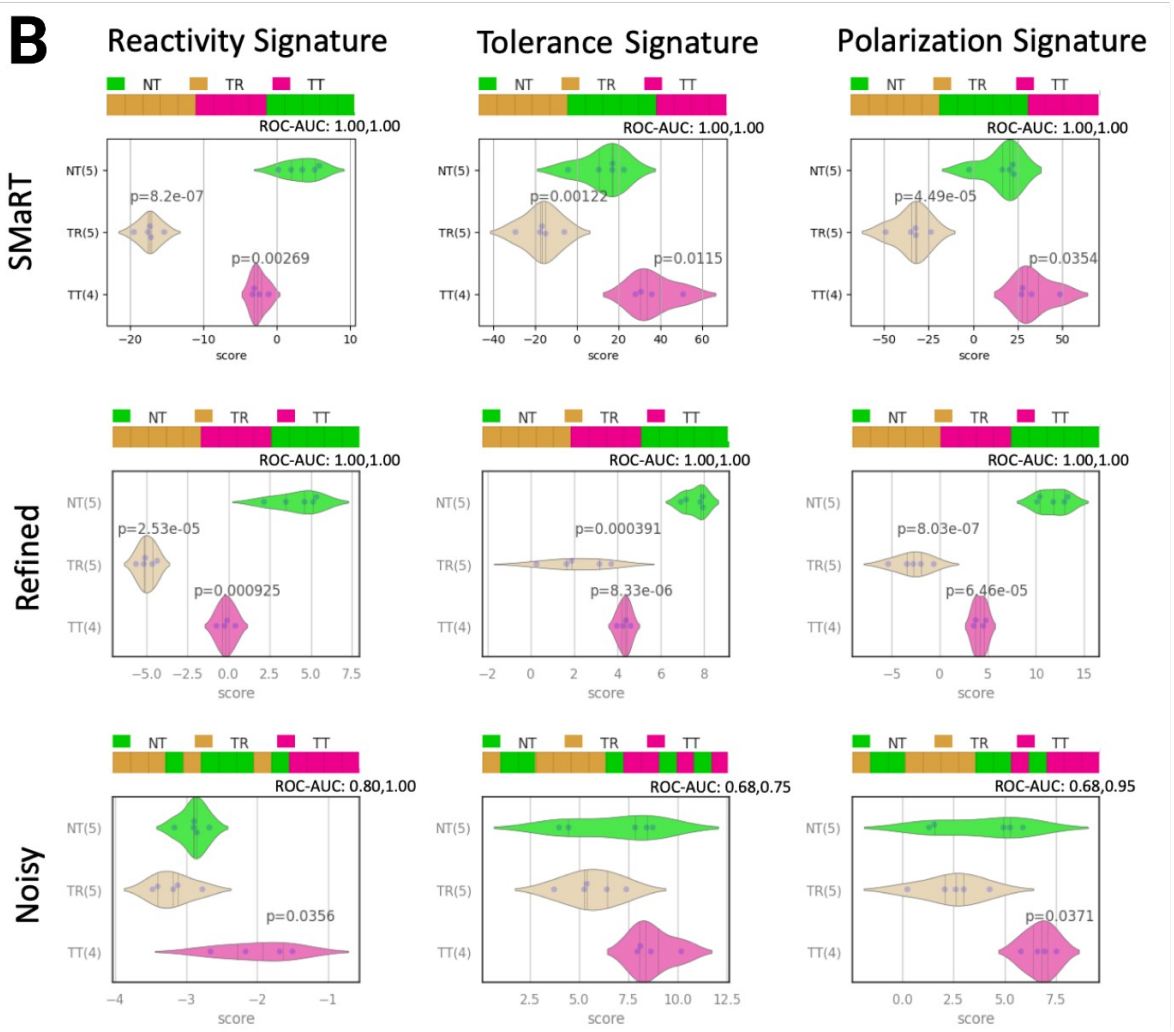


- Colorectal Cancer Macrophage Cell
- Normal Colon Macrophage Cell

Signature Performance

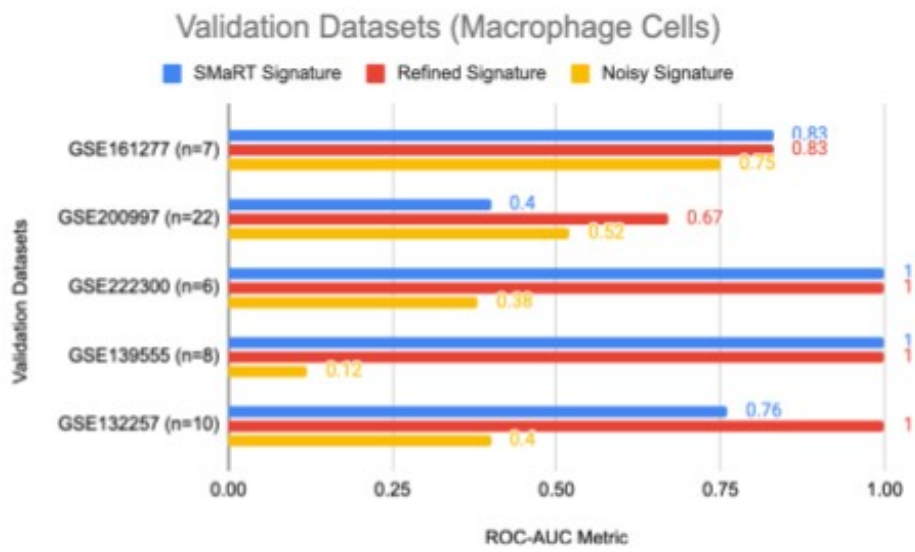
Refinement Training



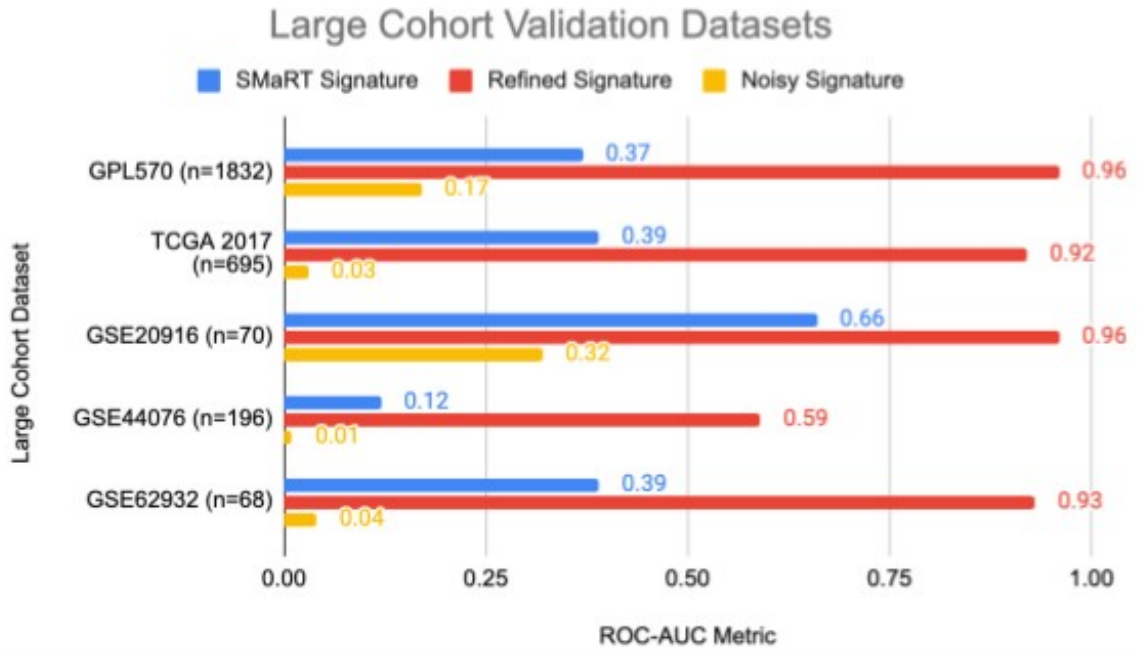


Signature Validation

A



B



MSS/MSI & CIMP+/CIMP- Status