Discrete Ollivier-Ricci Flow Finds Distinct Subpopulations of Patients Treated with PD-1 Inhibition

James Bannon*[1], Charles R. Cantor[2], and Bud Mishra[1]

[1]NYU
[2]Boston University

Abstract

In recent years immune checkpoint inhibitors (ICIs), also called immune checkpoint blockers, have revolutionized the standard of care for patients with cancers of many types. Researchers across many disciplines have endeavored to find biomarkers of response to ICI therapy but so far little consensus has been reached. In this paper we attempt to cluster patients in an unsupervised manner using discrete Ollivier-Ricci Flow (ORF). Our method surfaces populations with distinct survival curves which in turn allows us to find many potential biomarkers, including gene expression modules. We believe the algorithm may be of independent interest for clustering other datasets in a diverse set of research areas.

1 Introduction

In recent years approaches to treating cancer that leverage the immune system — called *immunotherapies* — have revolutionized the standard of care in many types of malignancies [1, 2, 3]. One class of immunotherapy, called immune checkpoint inhibitors (ICIs), has been successful in a diverse range of solid tumors [4]. Checkpoint inhibitors can be tremendously effective, leading to greater tumor shrinkage and longer survival times than typically occur with other therapies [5, 6, 4, 7]. However, not every patient responds to treatment with ICIs. In fact, responders make up the minority of patients. Estimates of ICI response rates (the percentage of patients who meet some response criterion such as RECIST [8]) are in the range of 20-45% [7]. Additionally, like other targeted therapies, it is possible for tumors to become resistant to treatment after an initial period of response [9].

*jjb509@nyu.edu

In order to improve the precision with which ICIs are prescribed, many researchers have focused on identifying biomarkers predictive of patient response to these therapies. Many candidate biomarkers have been identified across many "omics" modalities including bulk RNA-seq data [10, 11], single cell RNA-seq (scRNA-seq) [12], the tumor micro-environment [13, 14, 15], and somatic mutations [16]. Candidate biomarker identification is still an active area of research where so far little consensus has been reached. In fact, a recent large-scale meta-analysis revealed that the expression levels of the genes CXCL9 and CXCL13 and the non-synonymous tumor mutational burden (TMB) are the only consistently effective biomarkers [17].

Most biomarker discovery papers take the response phenotype as a given and attempt to find biological measurements that separate the population of responders from the population of non-responders. In this paper we take a different approach. We essentially ignore the labeling of patient subpopulations as responders and non-responders and aim instead to infer a novel patient stratification in an unsupervised manner using discrete Ollivier-Ricci Flow (ORF) [18, 19, 20]. We evaluate this stratification by looking at the the pairwise log-rank test between the progression free survival (PFS) curves [21] of the resulting partition. In cases where the differences between the PFS curves were statistically significant (log-rank $p \leq 0.1$) we performed extensive analysis of the resulting stratified patient subpopulations and identified potentially prognostic biomarker genes and gene modules.

The chief contributions of this paper are the clustering method deployed and the resulting biomarkers. In the community detection literature the use of ORF for partitioning known graphs (as opposed to our case where the graph is constructed) has theoretical guarantees and has demonstrated state-of-the-art performance on common benchmarks [18, 22]. Noting that for certain classes of graphs the discrete curvature converges to the continuous curvature [23], we endeavored to apply discrete Ricci curvature flow with surgery to split specific data points. Ours is the first use of the ORF algorithm to cluster patients based on RNA-seq measurements. In fact, our approach is, to the best of our knowledge, the first use of ORF to cluster data points in any setting. We also provide confirmatory results on ICI therapy response biomarkers that have only recently been reported in the literature.

2 Results

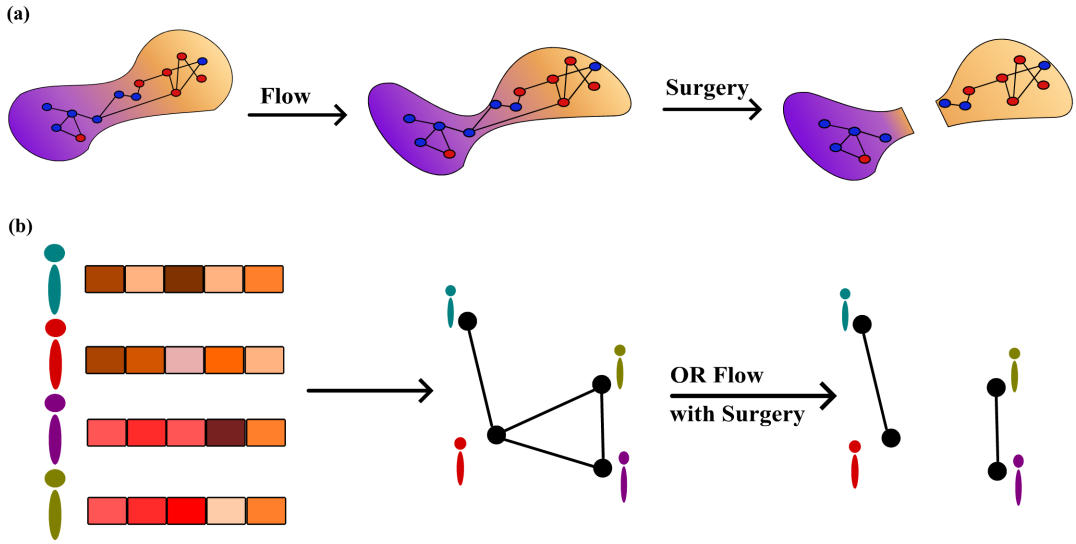

Figure 1: A schematic depicting the underlying point of view of our method **(a)** and the way in which it is implemented on a patient cohort **(b)**. Our assumption **(a)** is that patients' transcriptomic measurements (blue and red dots) live on some unknown abstract surface (purple and orange blob). If responders (blue dots) and non-responders (red dots) are mostly in distinct regions of the surface then cutting the surface via Ricci flow should segregate the responder and non-responder populations. **(b)** In our method we approximate this unknown surface by first reducing the dimensionality of the transcriptomic measurements with locally linear embeddings (LLE) and then constructing a k-nearest neighbors graph where each vertex corresponds to a patient. We then partition the vertices/patients using discrete ORF with surgery.

2.1 Patient Clustering with Discrete Ricci Flow

Our method, which is sketched in Figure 1, is grounded in the so-called "manifold-hypothesis" [24]. Specifically, we assume that high dimensional vectors that correspond to patient-derived gene expression measurements live on some unknown lower dimension surface (see Figure 1,top). We further suppose that, in general, patients who will respond similarly to ICI therapy will be closer together on this hypothetical surface. If the surface were known, we would still need a way to partition it. In the field of Riemannian geometry, the notion of 'Ricci Flow with surgery' is one such method that has been used to great success, most famously being a core tool in proving the

Poincaré conjecture [25, 26, 27]. We base our method on the view that approximating the surface, and the Ricci flow process, should yield a good partition of the data points. Concretely if we approximate the surface in the form of an undirected weighted graph and the Ricci flow process in an iterative way, then we should be able to meaningfully stratify patients in an *unsupervised* manner (without explicit reference to clinical outcome data). We now sketch the method a little more formally. The full definitions are given in Section 4.

We assume access to a *patient cohort*, which we take to mean a collection of patients with cancers from the same tissue of origin who receive the same ICI treatment. A cohort of n patients contains n vectors $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$, each with D dimensions ($\boldsymbol{x}^{(j)} \in \mathbb{R}^D$) where $\boldsymbol{x}^{(j)}$ is the gene expression measurements for patient j. We also assume that each patient has clinical data available in the form of a binary measure of response (responder or non-responder) as well as time and censorship status for progression free survival (PFS).

As our method assumes an underlying surface we perform a manifold learning step to reduce the dimensionality of the data and approximate the assumed surface. Specifically we used locally linear embeddings (LLE) [28] to take the original collection of vectors $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ to a new collection $\tilde{\boldsymbol{x}}^{(1)}, \ldots, \tilde{\boldsymbol{x}}^{(n)}$ where $\tilde{\boldsymbol{x}}^{(j)}$ is the d-dimensional vector (d much smaller than D) corresponding to patient j after LLE.

Using the collection of vectors $\tilde{\boldsymbol{x}}^{(j)}$ we construct a k-nearest neighbor graph where each vertex j corresponds to a specific patient. We assign each edge $\{j, \ell\}$ between patients/vertices j and ℓ in the graph a positive weight $w(j, \ell)$ and the discrete Ollivier-Ricci Flow (ORF) algorithm is then applied to partition the weighted graph.

The ORF algorithm proceeds roughly as follows. For a positive integer T we perform the following update T times:

$$w^{(t+1)}(j, \ell) = \left(1 - \epsilon \kappa^{(t)(j,\ell)}\right) d^{(t)}(j, \ell) \tag{1}$$

where $w^{(t)}(j, \ell)$ is the weight assigned to edge $\{j, \ell\}$ at step t, $\epsilon > 0$ is a step-size parameter, $\kappa^{(t)}(j, \ell)$ is the Olliver-Ricci (OR) curvature of the edge at step t, and $d^{(t)}(j, \ell)$ is the distance between vertices j and ℓ at step t. These terms are fully defined in Section 4. It is known that κ is

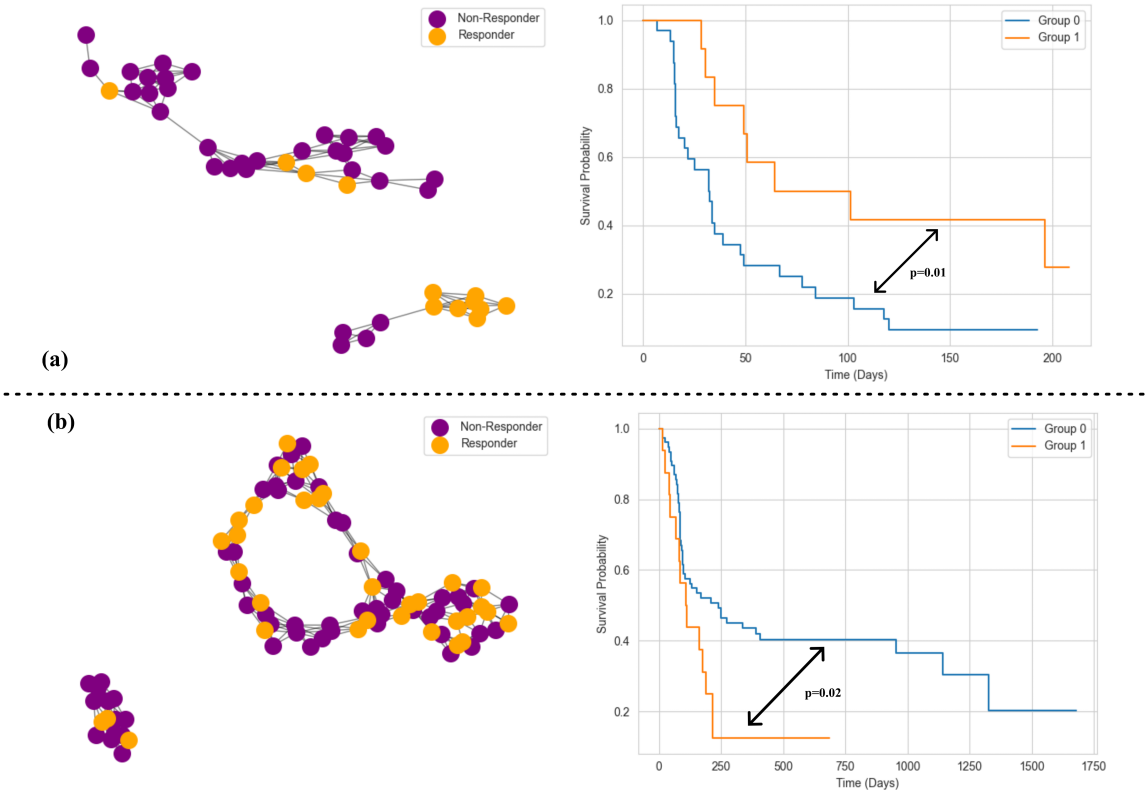

Figure 2: The results of partitioning two pembrolzumab treated cohorts of patients. The top row **(a)** gives the results in terms of the binary response labeled graph and the Kaplan-Meier fit survival curves for the STAD cohort. The bottom row **(b)** gives the same results for the SKCM cohort.

negative on edges between tightly connected communities and positive within those communities [19, 29]. With this in mind the update in Equation (1) increases the weights of edges between communities and decreases the weights of edges within communities. After T steps we remove the edges with the largest weights and assign vertices/patients to groups corresponding to the remaining connected components.

2.2 Application to Two Patient Cohorts

We focused our analysis on two cohorts of patients treated with pembrolizumab (pembro). One cohort consisted of 45 patients with stomach adenocarcinoma (STAD) and the other was a cohort of

129 skin cutaneous melanoma (SKCM) patients. Additional experiments on patients with SKCM treated with nivolumab (nivo), another PD-1 targeted ICI, are presented in the Supplementary Material.

The results for partitioning the two pembro-treated cohorts are presented in Figure 2. Our method finds two subpopulations among both the STAD and SKCM cohorts which have meaningfully different survival profiles (log-rank p-values of 0.01 and 0.02, respectively). While the connected components shown in Figure 2 are mixed and generally reflect poor clustering from the perspective of binary response, our method did succeed in finding subpopulations with different survival profiles. The clustering displayed in Figure 2 suggests that the clustering of patients on the underlying surface is more reflective of the aggressiveness of the disease under treatment rather than response (i.e. tumor shrinkage in response to therapy).

Next we investigated which genes could account for the distinct survival profiles. In order to find those genes we performed differential expression analysis across the discovered groups using DESeq2[30]. In order to summarize the genes that were found to be differentially expressed we performed a network pathway enrichment analysis, the results of which are in Supplementary Table 1. Interestingly we find that the DE genes between the STAD subpopulations yield a collection of pathways involved in antigen presentation. Antigen presentation is a key part of the adaptive immune system that allows cytotoxic T-cells to recognize tumor cells, among other crucial functions [31]. In addition the robustness of antigen processing has been proposed as a marker for ICI effectiveness [32, 33]. Indeed recent work has attempted to create an antigen presentation score for predicting ICI response in gastic cancer [34]. As antigen presentation is the means by which the immune system "discovers" tumor neoantigens, enriched antigen presentation pathways lend credence and nuance to the importance of neoantigen load as an ICI response biomarker [35, 33]. In the SKCM cohort we found a gene set enriched for immune response signaling, among others that are related to protein modification.

To assess if any of the discovered DE genes could be used as predictive response biomarkers we compared the distribution of expression among these DE genes between responders and non-responders. We used the Mann-Whitney U-test to compare the expression distributions among

responders and non-responders with p-values adjusted for multiple hypothesis testing with the Bejamini-Hochberg procedure [36, 37] keeping those genes with an adjusted p value below 0.05. The expression distributions of these genes is displayed in Supplementary Figure 1.

Among the candidate biomarkers we found in this manner, many were known cancer-related genes and some were only recently discovered as being implicated in ICI response. For example *PLAAT4* is a known tumor supressor [38] but recent work implies it has a role in shaping the tumor-immune microenvironment in pancreatic cancer [39]. *MUC20* is another gene that can contribute to sculpting the micronenvironment. In fact in Supplementary Figure 1 the expression levels of *MUC20* are much higher among non-responders, which mirrors existing literature that shows that high *MUC20* expression levels are negatively correlated with positive clinical outcomes [40, 41, 42, 43].

In the interest of finding scalar-valued biomarkers of response, we partitioned the collection of DE genes into modules using hierarchical clustering. Among the DE genes discovered for STAD patients treated with pembrolizumab we found two modules (Figure 3). We performed leave-one-out cross validation using the module score (averaged expression across the module) as a single value predictor for response in a regularized logistic regression model. Both module 1 and module 2 displayed predictive value achieving accuracies of 68% and 66%, and area under the receiver operator curve (ROC/AUC) values of 0.73 and 0.7, respectively.

Similar experiments on the SKCM cohort yielded 3 modules, none of which demonstrated predictive value. In Supplementary Figure 2 and Supplementary Table 1, we present the results for the same methods applied to a cohort of SKCM patients treated with nivolumab. We found four modules with accuracies of 61%,63%, 59%, and 53% and ROC/AUC values of 0.72, 0.69, 0.68, and 0.5.

3 Discussion

We have presented a novel method for data clustering based on discrete Ollivier-Ricci flow. We found that approaching data clustering in the context of predicting patient response to lead to

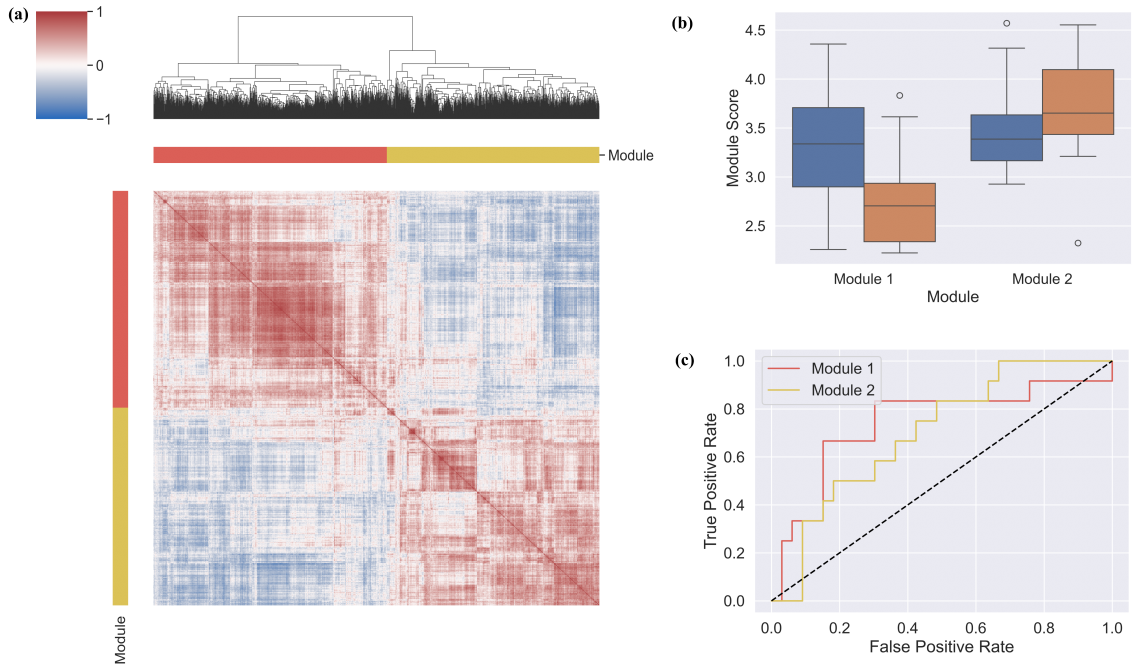

Figure 3: An analysis of the differentially expressed genes between the two discovered subpopulation of STAD patients treated with pembrolizumab. **(a)** A heatmap of the correlations between the gene expression values of the identified differentially expressed genes. The color bars show the modules found via hierarchical clustering performed using the displayed dendrogram above the heatmap. **(b)** Boxplot of the module scores (average expression in $\log_2$(transcripts per million+1) for responders and non-responders for each of the detected modules. In **(c)** we display the ROC curves for ℓ_2-regularized logistic regression fit via leave-one-out cross validation (LOOCV). Module 1 had an area under the curve (AUC) of 0.74 while module 2 had an AUC of 0.7.

meaningful unsupervised discoveries of patient subpopulations. There are a number of potential directions for future work.

First we note that this method did not work for patients with kidney renal clear cell (KIRC) and bladder urothelial carcinoma (BLCA). An immediate next step would be to investigate if the best manifold learning method is drug or tissue dependent, or an artifact of the dataset. It would be valuable to check if other manifold learning techniques, such as t-SNE, UMAP, or Laplacian eigenmaps would be more effective. We also have ongoing work investigating whether the recently developed techniques for building *learnable* geometries [44] can lead to improved performance in ICI response prediction.

Additionally we hope to develop performance guarantees for our method. It is known that in a limited version of the stochastic block model, a generative model of graph formation, the ORF flow approach can perfectly recover communities [18]. One avenue of research that is worthy of exploring would be extending the work of [23], perhaps with reference to the *geometric* block model [45], to determine how the curvature behaves in a randomly generated nearest neighbor graphs. Relatedly, it is, to our knowledge, unknown if the ORF procedure truly reflects the structure of its continuous analogue. That is: do convergence results like those in [23] hold for discrete Olliver-Ricci flow.

We also could consider applying this method in a semi-supervised context. Namely, if only a subset of our data points have labels, the algorithm here could be used to cluster all the data points and assign labels to the unlabeled points by looking at the connected component in which it resides. On the bioinformatic side confirmatory studies of our candidate biomarkers, as well as the discovered modules, performed *in vitro* in mouse models or *in silico* using larger datasets would be useful for giving our work additional translational impact.

4 Methods

Data Collection and Preprocessing We downloaded raw data from the CRI-iAtlas [46] which is a publicly accessible repository of patient-derived bulk RNA-seq data and tumor-immune micro-environment measurements paired with clinical outcome data for humans treated with ICIs. We

only used data from those samples taken **before** ICI treatment was administered and for which progression free survival (PFS) data (time until progression and censorship status) and clinical measures of response data were available.

Clinical response measurements in the iAtlas fall into four distinct categories: complete response, partial response, progressive disease and stable disease. We binned these into two groups: responders — made up of patients with complete or partial response — and non-responders — those with progressive or stable disease.

We curated a collection of genes thought to be related to cancer and ICI response. For cancer-related genes we collected the genes listed in the KEGG database [47] under the heading "KEGG: Pathways in Cancer", available at `https://www.genome.jp/pathway/hsa05200`, and the genes listed in the landmark paper on cancer genome landscapes [48]. For the ICI-related genes we used the genes involved in the computation of the IMPRES score [10]. We kept only those genes that were in our curated set and also had measurements in the CRI-iAtlas. Gene expression values were quantified in terms of transcripts per million (TPM), which were re-scaled via the logarithmic transformation $\log_2(\text{TPM}+1)$. These values were used to construct were the original data vectors $\boldsymbol{x}^{(j)}$ in each cohort.

Graph Theory We consider only undirected weighted graphs. Specifically a graph G, written $G = (V, E)$ consists of a finite set V of elements called *vertices* and $E \subset V \times V$ of edges. A graph is undirected if the edges "go both ways", i.e: $(u, v) \in E$ implies $(v, u) \in E$. We write edges as $\{u, v\}$. If there is an edge between two vertices we write this as $u \sim v$.

The weight of an edge is a positive value which we write as $w(u, v)$. The *degree* of a vertex v, written d_v, is the sum of the weights that are incident to it, that is: $d_v = \sum_{u:u\sim v} w(v, u)$. A *path* $P(u, v)$ between two vertices v, u is a sequence of vertices $q_1, q_2, \ldots, q_r$ where no vertex appears more than once and $q_1 = v, q_r = u$ and $q_i \sim q_{i+1}$. The *length* of a path $P(u, v)$, $\mathfrak{L}(P(u, v))$ is taken to be the sum of the edge weights. Specifically:

$$\mathfrak{L}(P(u, v)) = \sum_{i=1}^{r-1} w(q_i, q_{i+1})$$

Lastly, for the distance between two vertices u, v we consider the all-pairs shortest path (APSP) distance. If $\mathfrak{P}(u,v)$ is the set of all paths between u and v, then we define the distance $d(u,v)$ between u, v as

$$d(u,v) = \min_{P(u,v)\in\mathfrak{P}(u,v)} \mathfrak{L}(P(u,v))$$

Constructing the K-nearest Neighbors Graph We have a collection of n patients represented as feature vectors $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ where each $\boldsymbol{x}^{(j)}$ has D dimensions representing expression levels for D distinct genes. We denote the expression level of gene g in patient j with $\boldsymbol{x}_g^{(j)}$. The LLE algorithm proceeds in two steps. The first step involves learning a set of weights Ω_{ij} between the vectors such that each $\boldsymbol{x}^{(j)}$ can be approximately reconstructed as the sum of its neighbors. Specifically the weights are learned to minimize the function:

$$\mathcal{L}(\Omega) = \sum_{j=1}^{n} \left|\left| \boldsymbol{x}^{(j)} - \sum_{r=1}^{n} \Omega_{jr} \boldsymbol{x}^{(r)} \right|\right|^2$$

where $||\cdot||^2$ represents the standard Euclidean distance between vectors, i.e:

$$\left|\left| \boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)} \right|\right|^2 = \sum_{g=1}^{D} \left(x_g^{(j)} - x_g^{(k)} \right)^2$$

The set of weights Ω is found by solving an optimization problem subject to some constraints. Specifically one must specify the number L of neighbors to be considered in reconstructing each $\boldsymbol{x}^{(j)}$ and then Ω is forced to satisfy: $\Omega_{ji} = 0$ if $\boldsymbol{x}^{(i)}$ is not one of $\boldsymbol{x}^{(j)}$'s L nearest neighbors and that $\sum_{j=1} \Omega_{ij} = 1$. We fixed $L = 5$ in our experiments.

The second step of LLE involves choosing an embedding dimension $d < D$ and constructing reduced vectors $\tilde{\boldsymbol{x}}^{(j)} \in \mathbb{R}^d$ using the learned weights Ω. The embedding is found such that the *reconstruction error*

$$e(d;\Omega) = \sum_{i=i}^{n} \left|\left| \tilde{\boldsymbol{x}}^{(i)} - \sum_{j=1}^{n} \Omega_{ij} \tilde{\boldsymbol{x}}^{(j)} \right|\right|^2$$

is minimized. Details on this procedure can be found in [49]. In our experiments we tried every d

in the range $2, \ldots, 15$ and used the one for which the reconstruction error was the smallest, which lead to $d = 2$ for all cohorts.

To construct the patient-patient graph we apply a symmetrized K-nearest neighbors. Let $N_K(j)$ be collection of indices $i \in \{1, \ldots, n\}$ for which $\left|\left|\tilde{\boldsymbol{x}}^{(j)} - \tilde{\boldsymbol{x}}^{(i)}\right|\right|^2$ is the smallest. We then add a vertex for each patient and an edge $\{j, \ell\}$ between patients j and ℓ according to the rule

$$\{j, \ell\} \in E \text{ if and only if } \ell \in N_K(j) \text{ or } j \in N_k(\ell)$$

The initial weights of the edges were all set to $w(j, \ell) = 1$ as setting the weights to distances made no difference in our experiments.

Discrete Ollivier-Ricci Flow with Surgery The notion of graph curvature that we use was developed in a landmark paper by Ollivier [50] in the more general setting of Markov chains on metric spaces. In this work we adopt the form of this curvature, known as Ollivier's Ricci Curvature (ORF), developed by Lin and Yau [51]. We now describe how to construct this curvature and perform graph partitioning with ORF flow on a general undirected weighted graph. The steps are performed identically on the specific patient-patient interaction graph.

We begin with a value $\alpha \in [0, 1]$ and assign a one-step random walk probability to each vertex v. For any given α and any vertex v the probability distribution μ_v^α is given by

$$\mu_v^\alpha(u) = \begin{cases} \alpha & u = v \\ (1-\alpha)\frac{w(v,u)}{d_v} & u \sim v \\ 0 & \text{otherwise} \end{cases}$$

The curvature $\kappa(u, v; \alpha)$, for any given α and edge u, v, is compute via:

$$\kappa(u, v; \alpha) = 1 - \frac{\mathcal{W}\left(\mu_v^\alpha, \mu_u^\alpha\right)}{d(u, v)}$$

where $\mathcal{W}$ is the well-known Wasserstein distance [52, 53]. Since in all our experiments $\alpha = 0$ we

now drop α from the notation and write the curvature as $\kappa(u, v)$. Let $d_{\max}$ be the largest **number** of neighbors that any vertex in the graph has. The computational complexity of computing $\kappa(u, v))$ is $O(d_{\max}^3)$ as $d(u, v)$ can be pre-computed in $O(|V|^3)$ time and $O(d_{\max}^3)$ is the computational complexity of computing the Wasserstein distance via a linear programming.

The base of our method, Algorithm 1, is the discrete Ollivier-Ricci Flow (ORF) procedure [19, 20, 18] which is based on the update given in Equation (1). At each iteration we must recompute the pairwise distances and curvature values, which for T iterations has complexity $O(T(|V|^3 + d_{\max}^3)) = O(T|V|^3)$. For all drugs and tissues we set $\epsilon = 0.3$, $K = 5$, $T = 50$.

The 'surgery' portion of the algorithm involves cutting the weights that remain after the T iterations. We use the heuristic procedure of [22]. Let be $w_{\max}$ be the largest edge weight in the graph and $w_{\min}$ a smallest weight value which is left as a parameter and $\delta > 0$ a step size. We iterate over a grid of values w between $w_{\max}$ and $w_{\min}$ equally spaced by δ in decreasing order, removing edges with weight greater than w. At each step we compute the modularity $Q(w)$ [54] of assigning each node to a community that corresponds to the connected component in which it lies.

Differential Expression and Pathway Enrichment For differential expression we used the DESeq method from the `R` package of the same name [30]. We kept all genes which were differentially expressed with an adjusted Wald p-value below 0.01 as output by the DESeq2 software. For network-based enrichment we used the WebGestalt tool [55] to perform network-based pathway analysis. The "method of interest," "organism of interest," and "functional database" parameters were set to "Network Topology-based Analysis", "homo sapiens", and "network"/"PPI BIOGRID" respectively.

Implementation Code to perform these experiments was written in python. We used the library `scikit-learn` [56] to perform LLE and fit the logistic regression models. Code supporting this can be found at `https://github.com/jbannon/ORFlowICIs`.

Algorithm 1 Data Clustering with Ollivier Ricci Flow

Require: Collection of n vectors $\left\{\boldsymbol{x}^{(j)}\right\}_{j=1}^{n}$, $\boldsymbol{x}^{(j)} \in \mathbb{R}^D$, $\epsilon > 0$, $\alpha \in [0,1]$, $T > 0$, $K > 0$,$\delta > 0$

Construct the K nearest neighbor graph $G = (V, E, w)$ for the dataset.
Set $w^0(i,j) = w(i,j)$ for all edges $\{i,j\}$.
for t $= 1, \ldots, T$ **do**
  Construct μ_v^α for all $v \in V$.
  Compute the path distance $d^{(t)}(u,v)$ for all u, v using the weights $w^{(t)}$
  Compute $\kappa^{(t)}(v,u;\alpha)$ for all edges $\{v,u\} \in E$
  for $\{u,v\} \in E$ **do**

$$w^{(t+1)}(u,v) \leftarrow \left(1 - \epsilon\kappa^{(t)}(u,v)\right) d^{(t)}(u,v)$$

  end for
end for
initialize $Q_{\text{best}} = -\infty$, w_{best}
for $w \in \{w_{\max}, w_{\max} - \delta, w_{\max} - 2\delta, \ldots, w_{\min}\}$ **do**
  Remove edges with weight greater than w
  Compute the connected components of the graph that results
  Compute $Q(w)$
  if $Q(w) > Q_{\text{best}}$ **then**
    $Q_{\text{best}} = Q(w)$
    $w_{\text{best}} = w$
  end if
end for

5 Data Availability Statement

The data underpinning our experiments is publicly available through the CRI-iAtlas. Specifically one can navigate to the portal at `https://cri-iatlas.org` and download the data from there. The supporting code is at `https://github.com/jbannon/ORFlowICIs`.

6 Acknowledgements

JB and BM are grateful to NYU's departmental support during the early stages of this manuscript. JB is also grateful to many fruitful discussions with Dr. Andrew Hall.

7 Author Contributions

JB and CC conceived of the experiments. JB implemented the code and performed the analysis and figure generation and wrote the majority of the manuscript. JB, CC, and BM all edited the manuscript. All authors have read and approved the manuscript.

8 Competing Interests

The authors declare no competing interests.

References

[1] Robert C Sterner and Rosalie M Sterner. Car-t cell therapy: current limitations and potential strategies. *Blood cancer journal*, 11(4):69, 2021.

[2] Haiying Ding, Wenxiu Xin, Yinghui Tong, Jiao Sun, Gaoqi Xu, Ziqi Ye, and Yuefeng Rao. Cost effectiveness of immune checkpoint inhibitors for treatment of non-small cell lung cancer: a systematic review. *PLoS One*, 15(9):e0238536, 2020.

[3] Patrick Hwu. Treating cancer by targeting the immune system, 2010.

[4] Syed Hussaini, Rania Chehade, Ronald Gabriel Boldt, Jacques Raphael, Phillip Blanchette, Saman Maleki Vareki, and Ricardo Fernandes. Association between immune-related side effects and efficacy and benefit of immune checkpoint inhibitors–a systematic review and meta-analysis. *Cancer Treatment Reviews*, 92:102134, 2021.

[5] Jonathan J Havel, Diego Chowell, and Timothy A Chan. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nature Reviews Cancer*, 19(3):133–150, 2019.

[6] Willy Hugo, Jesse M Zaretsky, LU Sun, Chunying Song, Blanca Homet Moreno, Siwen Hu-Lieskovan, Beata Berent-Maoz, Jia Pang, Bartosz Chmielowski, Grace Cherry, et al. Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. *Cell*, 165(1):35–44, 2016.

[7] Rilan Bai, Zheng Lv, Dongsheng Xu, and Jiuwei Cui. Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors. *Biomarker research*, 8:1–17, 2020.

[8] C Carl Jaffe. Measures of response: Recist, who, and new alternatives. *Journal of Clinical Oncology*, 24(20):3245–3251, 2006.

[9] Russell W Jenkins, David A Barbie, and Keith T Flaherty. Mechanisms of resistance to immune checkpoint inhibitors. *British journal of cancer*, 118(1):9–16, 2018.

[10] Noam Auslander, Gao Zhang, Joo Sang Lee, Dennie T Frederick, Benchun Miao, Tabea Moll, Tian Tian, Zhi Wei, Sanna Madan, Ryan J Sullivan, et al. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nature medicine*, 24(10):1545–1549, 2018.

[11] JungHo Kong, Doyeon Ha, Juhun Lee, Inhae Kim, Minhyuk Park, Sin-Hyeog Im, Kunyoo Shin, and Sanguk Kim. Network-based machine learning approach to predict immunotherapy response in cancer patients. *Nature communications*, 13(1):3703, 2022.

[12] Saeed Khodayari, Hamid Khodayari, Elnaz Saeedi, Habibollah Mahmoodzadeh, Alireza Sadrkhah, and Karim Nayernia. Single-cell transcriptomics for unlocking personalized cancer

immunotherapy: toward targeting the origin of tumor development immunogenicity. *Cancers*, 15(14):3615, 2023.

[13] Jérôme Galon and Daniela Bruni. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nature reviews Drug discovery*, 18(3):197–218, 2019.

[14] Jérôme Galon, Helen K Angell, Davide Bedognetti, and Francesco M Marincola. The continuum of cancer immunosurveillance: prognostic, predictive, and mechanistic signatures. *Immunity*, 39(1):11–26, 2013.

[15] Jérôme Galon, Wolf-Herman Fridman, and Franck Pages. The adaptive immunologic microenvironment in colorectal cancer: a novel perspective. *Cancer research*, 67(5):1883–1886, 2007.

[16] Mark Lee, Robert M Samstein, Cristina Valero, Timothy A Chan, and Luc GT Morris. Tumor mutational burden as a predictive biomarker for checkpoint inhibitor immunotherapy. *Human vaccines & immunotherapeutics*, 16(1):112–115, 2020.

[17] Kevin Litchfield, James L Reading, Clare Puttick, Krupa Thakkar, Chris Abbosh, Robert Bentham, Thomas BK Watkins, Rachel Rosenthal, Dhruva Biswas, Andrew Rowan, et al. Meta-analysis of tumor-and t cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell*, 184(3):596–614, 2021.

[18] Chien-Chun Ni, Yu-Yao Lin, Feng Luo, and Jie Gao. Community detection on networks with ricci flow. *Scientific reports*, 9(1):9984, 2019.

[19] Chien-Chun Ni, Yu-Yao Lin, Jie Gao, Xianfeng David Gu, and Emil Saucan. Ricci curvature of the internet topology. In *2015 IEEE conference on computer communications (INFOCOM)*, pages 2758–2766. IEEE, 2015.

[20] Chien-Chun Ni, Yu-Yao Lin, Jie Gao, and Xianfeng Gu. Network alignment by discrete ollivier-ricci flow. In *International symposium on graph drawing and network visualization*, pages 447–462. Springer, 2018.

[21] Nathan Mantel et al. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50(3):163–170, 1966.

[22] Yu Tian, Zachary Lubberts, and Melanie Weber. Mixed-membership community detection via line graph curvature. In *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, pages 219–233. PMLR, 2023.

[23] Pim van der Hoorn, William J Cunningham, Gabor Lippner, Carlo Trugenberger, and Dmitri Krioukov. Ollivier-ricci curvature convergence in random geometric graphs. *Physical Review Research*, 3(1):013211, 2021.

[24] Alexander N Gorban and Ivan Yu Tyukin. Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2118):20170237, 2018.

[25] Bennett Chow, Peng Lu, and Lei Ni. *Hamilton's Ricci flow*, volume 77. American Mathematical Society, Science Press, 2023.

[26] Grisha Perelman. The entropy formula for the ricci flow and its geometric applications. *arXiv preprint math/0211159*, 2002.

[27] Grisha Perelman. Ricci flow with surgery on three-manifolds. *arXiv preprint math/0303109*, 2003.

[28] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[29] Jürgen Jost and Shiping Liu. Ollivier's ricci curvature, local clustering and curvature-dimension inequalities on graphs. *Discrete & Computational Geometry*, 51(2):300–322, 2014.

[30] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

[31] Judith A Owen. Kuby immunology 7th edition, 2013.

[32] Kailin Yang, Ahmed Halima, and Timothy A Chan. Antigen presentation in cancer—mechanisms and clinical implications for immunotherapy. *Nature Reviews Clinical Oncology*, 20(9):604–623, 2023.

[33] Anastasia Mpakali and Efstratios Stratikos. The role of antigen processing and presentation in cancer and the efficacy of immune checkpoint inhibitor immunotherapy. *Cancers*, 13(1):134, 2021.

[34] Ke-wei Wang, Mei-dan Wang, Zi-xi Li, Ben-shun Hu, Jun-jie Wu, Zheng-dong Yuan, Xiao-long Wu, Qin-fang Yuan, and Feng-lai Yuan. An antigen processing and presentation signature for prognostic evaluation and immunotherapy selection in advanced gastric cancer. *Frontiers in Immunology*, 13:992060, 2022.

[35] Peipei Wang, Yueyun Chen, and Chun Wang. Beyond tumor mutation burden: tumor neoantigen burden as a biomarker for immunotherapy and other types of therapy. *Frontiers in oncology*, 11:672677, 2021.

[36] Nadim Nachar et al. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20, 2008.

[37] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[38] Christine Sers, Urban Emmenegger, Knut Husmann, Katharina Bucher, Ann-Catherine Andres, and Reinhold Schäfer. Growth-inhibitory activity and downregulation of the class ii tumor-suppressor gene h-rev107 in tumor cell lines and experimental tumors. *The Journal of cell biology*, 136(4):935–944, 1997.

[39] Tian-Hao Li, Yuan-Yang Wang, Bang-Bo Zhao, Cheng Qin, Ze-Ru Li, and Wei-Bin Wang. Phospholipase a/acyltransferase 4 is a prognostic biomarker and correlated with immune infiltrates in pancreatic cancer. *Heliyon*, 8(9), 2022.

[40] Xiuying Xiao, Lisha Wang, Ping Wei, Yayun Chi, Dali Li, Qifeng Wang, Shujuan Ni, Cong Tan, Weiqi Sheng, Menghong Sun, et al. Role of muc20 overexpression as a predictor of recurrence and poor outcome in colorectal cancer. *Journal of translational medicine*, 11:1–12, 2013.

[41] Fei Zheng, Huimin Yu, and Jingjing Lu. High expression of muc20 drives tumorigenesis and predicts poor survival in endometrial cancer. *Journal of cellular biochemistry*, 120(7):11859–11866, 2019.

[42] Syue-Ting Chen, Ting-Chun Kuo, Ying-Yu Liao, Mei-Chun Lin, Yu-Wen Tien, and Min-Chuan Huang. Silencing of muc20 suppresses the malignant character of pancreatic ductal adenocarcinoma cells through inhibition of the hgf/met pathway. *Oncogene*, 37(46):6041–6053, 2018.

[43] Chi-Hau Chen, Shu-Wei Wang, Chih-Wei Chen, Miao-Ruei Huang, Ji-Shiang Hung, Hsiu-Chin Huang, Ho-Hsiung Lin, Ruey-Jien Chen, Ming-Kwang Shyu, and Min-Chuan Huang. Muc20 overexpression predicts poor prognosis and enhances egf-induced malignant phenotypes via activation of the egfr–stat3 pathway in endometrial cancer. *Gynecologic oncology*, 128(3):560–567, 2013.

[44] Haitz Sáez de Ocáriz Borde and Anastasis Kratsios. Neural snowflakes: Universal latent graph inference via trainable latent geometries. *arXiv preprint arXiv:2310.15003*, 2023.

[45] Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. The geometric block model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[46] James A Eddy, Vésteinn Thorsson, Andrew E Lamb, David L Gibbs, Carolina Heimann, Jia Xin Yu, Verena Chung, Yooree Chae, Kristen Dang, Benjamin G Vincent, et al. Cri iatlas: an interactive portal for immuno-oncology research. *F1000Research*, 9, 2020.

[47] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[48] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[49] Lawrence K Saul and Sam T Roweis. An introduction to locally linear embedding. *unpublished. Available at: http://www. cs. toronto. edu/˜ roweis/lle/publications. html*, 2000.

[50] Yann Ollivier. Ricci curvature of metric spaces. *Comptes Rendus Mathematique*, 345(11):643–646, 2007.

[51] Yong Lin, Linyuan Lu, and Shing-Tung Yau. Ricci curvature of graphs. *Tohoku Mathematical Journal, Second Series*, 63(4):605–627, 2011.

[52] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[53] Jan van den Brand. A deterministic linear program solver in current matrix multiplication time. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 259–278. SIAM, 2020.

[54] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[55] John M Elizarraras, Yuxing Liao, Zhiao Shi, Qian Zhu, Alexander R Pico, and Bing Zhang. Webgestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Research*, page gkae456, 2024.

[56] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

A Supplementary Material

Additional Results for Pembrolizumab

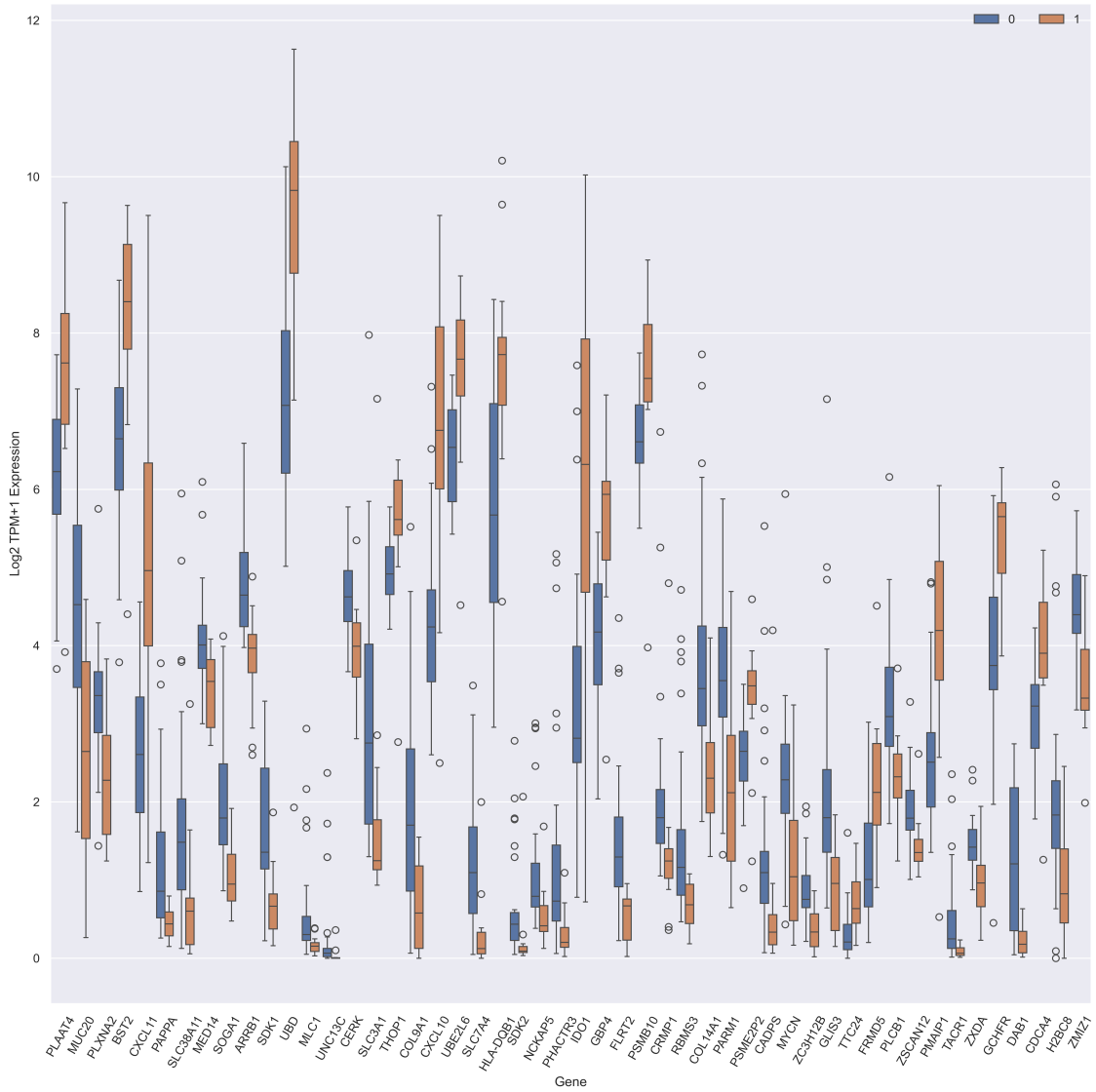

Figure 4: Genes found by performing differential expression analysis between our discovered groups in STAD patients treated with pembrolizumab. Displayed are only those genes which had statistically significant differences between expression distributions in responders and non-responders (multiple-testing adjusted Mann-Whitney p-value ≤ 0.05)

Pembro - STAD		Pembro - SKCM	
GO ID	Description	GO ID	Description
GO:0002478	antigen processing and presentation of exogenous peptide antigen	GO:0043170	macromolecule metabolic process
GO:0048002	antigen processing and presentation of peptide antigen	GO:0036211	protein modification process
GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II	GO:0006807	nitrogen compound metabolic process
GO:0002495	antigen processing and presentation of peptide antigen via MHC class II	GO:0002764	immune response-regulating signaling pathway
GO:0002399	MHC class II protein complex assembly	GO:0002757	immune response-activating signaling pathway
GO:0002503	peptide antigen assembly with MHC class II protein complex	GO:0002253	activation of immune response
GO:0002396	MHC protein complex assembly	GO:0043412	macromolecule modification
GO:0002501	peptide antigen assembly with MHC protein complex	GO:0090304	nucleic acid metabolic process
GO:0019884	antigen processing and presentation of exogenous antigen	GO:0044238	primary metabolic process
GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	GO:0071704	organic substance metabolic process

Table 1: Network-based pathway enrichment analysis for pembrolizumab patients. Enriched pathways for patients with STAD are the left and those for SKCM patients are on the right.

Results for SKCM Patients Treated with Nivolumab

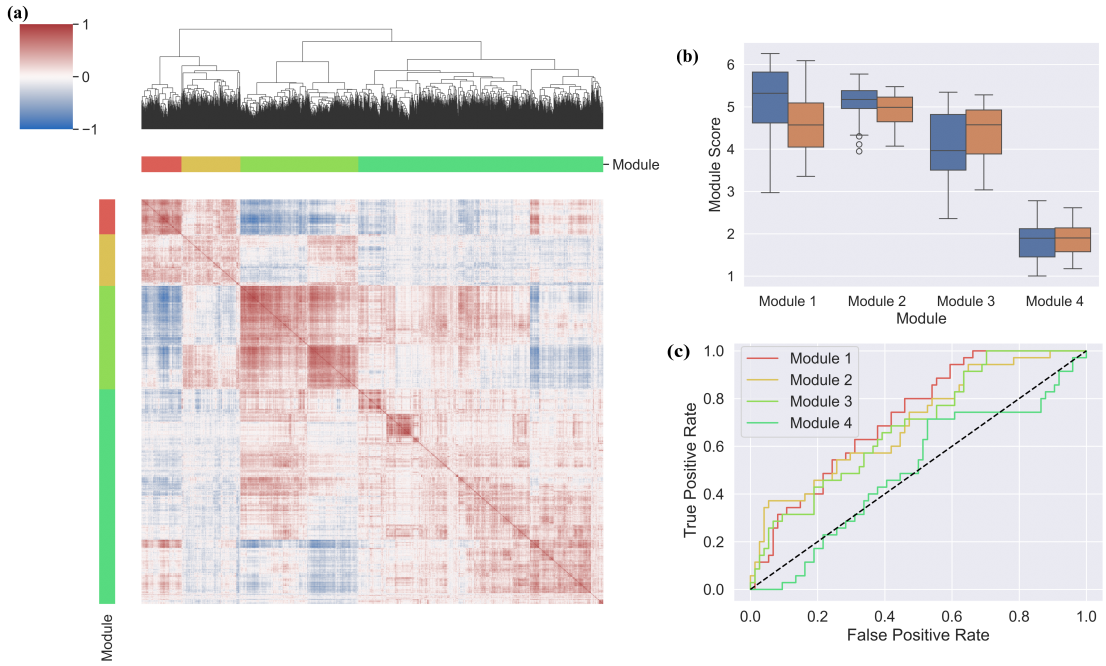

Figure 5: Results on module discovery for SKCM patients treated with nivolumab. **(a)** The the correlation heatmap for the expression of the genes found differentially expressed with the modules highlighted by the colored bars. **b** The module scores (average expression of genes in the module in $\log_2$(transcripts per million)+1) between responders and non-resonders. **(c)** The leave-one-out ROC/AUC for logistic regression using each module score as a single predictor. Module 1 had an ROC/AUC 0.72, and the values for Modules 2, 3, and 4 were 0.0.69, 0.68, 0.5, respectively

GO ID	Description
GO:0048583	regulation of response to stimulus
GO:0034097	response to cytokine
GO:0071345	cellular response to cytokine stimulus
GO:0006915	apoptotic process
GO:0071310n	cellular response to organic substance
GO:0019538	protein metabolic process
GO:0070887	cellular response to chemical stimulus
GO:0008219	cell death
GO:0006412	translation
GO:0006950	response to stress

Table 2: Gene ontology (GO) pathway enrichment for SKCM patients treated with nivolumab.