MIAMI: Mutual Information-based Analysis of Multiplex Imaging data

Souvik Seal and Debashis Ghosh

Department of Biostatistics and Informatics, Colorado School of Public Health,
University of Colorado CU Anschutz Medical Campus, Aurora, Colorado

Abstract

Motivation: Studying the interaction or co-expression of the proteins or markers in the tumor microenvironment (TME) of cancer subjects can be crucial in the assessment of risks, such as death or recurrence. In the conventional approach, the cells need to be declared positive or negative for a marker based on its intensity. For multiple markers, manual thresholds are required for each marker, which can become cumbersome. The performance of the subsequent analysis relies heavily on this step and thus suffers from subjectivity and lack of robustness.

Results: We present a new method where different marker intensities are viewed as dependent random variables, and the mutual information (MI) between them is considered to be a metric of co-expression. Estimation of the joint density, as required in the traditional form of MI, becomes increasingly challenging as the number of markers increases. We consider an alternative formulation of MI which is conceptually similar but has an efficient estimation technique for which we develop a new generalization. With

the proposed method, we analyze a lung cancer dataset and a triple-negative breast cancer dataset finding the co-expression of the markers, HLA-DR and CK and the co-expression of markers, HLA-DR, CD45RO, HLA-Class-1, H3K27me3 and H3K9ac to be associated with survival in the two datasets respectively. We also demonstrate the robustness of our method through different simulation studies.

Availability: The associated *R* package can be found here,
https://github.com/sealx017/MIAMI.

Contact: souvik.seal@cuanschutz.edu

Supplementary information: The supplementary material is attached.

1 Introduction

In recent years, multiplex tissue imaging (Bataille *and others*, 2006) technologies like, imaging mass cytometry (IMC) (Ali *and others*, 2020), multiplex immunohistochemistry (mIHC) (Tan *and others*, 2020) and multiplexed ion beam imaging (MIBI) (Angelo *and others*, 2014) have become increasingly popular for probing single-cell spatial biology. The technologies help in understanding the biological mechanisms underlying cellular and protein interactions in a wide array of scientific contexts. MIBI (Ionpath Inc.) (Angelo *and others*, 2014) platform and the mIHC platforms such as Vectra 3.0 (Akoya Biosciences) (Huang *and others*, 2013) and Vectra Polaris (Akoya Biosciences) (Pollan *and others*, 2020) produce images of similar structure. In particular, each image is two-dimensional, collected at cell- and nucleus-level resolution and proteins in the sample have been labeled with antibodies that attach to cell membranes. We will refer to the antibodies as markers in the paper. Typically, mIHC images have 6-8 markers, whereas MIBI images can have 40 or more markers.

Many of the above markers are surface or phenotypic markers (Zola *and others*, 2007;

Shipkova and Wieland, 2012) which are primarily used for cell type identification. Additionally, there are functional markers such as HLA-DR (Saraiva *and others*, 2018), CD45RO (Lee *and others*, 2008) and H3K27me3 (Hansen *and others*, 2008). that dictate or regulate important cell-functions. Both surface and functional markers are quantified as continuous valued marker intensities. However, the interaction effects of the markers are usually studied by binarizing them. In the traditional method, for every marker, a threshold is chosen to indicate whether a cell in the tumor microenvironment (TME) (Binnewies *and others*, 2018) is positive or negative for that marker. If a cell is positive for two markers, it implies that the markers have co-expressed or co-occurred in that cell. Next, for all the subjects considering every unique pair of markers, the proportion of cells positive for both the markers, the proportion of cells positive for only the first marker and the proportion of cells positive for only the second marker are computed. These proportions can then be used in hiearchical clustering (Murtagh and Legendre, 2014) to group the subjects. We can then test to see if the cluster labels correlate with clinical outcomes, like time to death or disease recurrence (Koguchi *and others*, 2015; Johnson *and others*, 2020; Jackson *and others*, 2020; Patwa *and others*, 2021). It should be pointed out that such expression-thresholding is also popular in cell-cell communication or interaction analysis (Jin *and others*, 2021) in the context of single-cell RNA sequencing (scRNA-seq) data. For example, Armingol *and others* (2021) define a pair of cells to be interacting if the expressions of both ligand and receptor (Wong *and others*, 1997) in those cells exceed certain thresholds for multiple pairs of ligands and receptors.

The process of threshold selection for multiple markers is extremely challenging and suffers from subjectivity. The problem gets exacerbated as the number of markers increases. Also, the subsequent association analysis relies heavily on the chosen thresholds and can

perform poorly, as we show in the simulations. In this paper, we present a threshold-free approach for studying marker co-expression in the subjects. We treat different marker intensities as continuous random variables and use their marginal and joint probability density functions (PDF) to construct a metric of co-expression based on mutual information (MI) (Duncan, 1970). Unlike the correlation coefficient, MI is capable of capturing non-linear dependence between the markers and is easily extendable for more than two markers. However, as the number of markers increases, computing the joint PDF becomes increasingly challenging which makes the computation of MI infeasible as well. Therefore, we use a slightly different formulation of MI known as Euclidean quadratic mutual information (EQMI) (Principe *and others*, 2000) which has a similar interpretation but can be computed much efficiently. The efficient computation algorithm is discussed in Principe (2010) with a simpler assumption that we further generalize. The vector of estimated values of the EQMI of all the subjects is then tested for association with clinical outcomes. With the proposed method, we analyze an mIHC lung cancer dataset finding the co-expression of markers, HLA-DR and CK to be significantly associated with the overall survival. We also analyze a MIBI triple-negative breast cancer dataset where the interaction or co-expression of multiple proteins, such as HLA-DR, CD45RO, HLA-Class-1, H3K27me3 and H3K9ac is found to be influential on both survival and recurrence. We demonstrate the robustness of our method over the existing ones through different simulation studies.

2 Materials and Methods

Suppose there are p markers and N subjects with the j-th subject having n_j cells. Let X_{kij} denote the scaled expression, between 0 and 1, of the k-th marker in the i-th cell of j-th

subject for $k = 1, 2, \ldots, p$, $i = 1, 2, \ldots, n_j$, and $j = 1, 2, \ldots, N$. Let $Y = (Y_1, Y_2, \ldots, Y_N)^T$ denote a subject-level outcome vector. Note that we focus on cell-level data in this paper but the framework is readily usable on pixel-level data as well.

2.1 Traditional thresholding-based approach to study co-expression of the markers

For every marker k, we choose a cut-off t_k to define cell i of subject j as positive for that marker if $X_{kij} > t_k$. The choice of t_k can be guided by prior biological insight or by careful inspection of the marker intensity profile. For example, extreme quantiles like (e.g., the 90th or 95th percentiles) can serve as viable thresholds, as we see in the simulations. However, in most real datasets, it requires user-defined thresholds to obtain an appropriate threshold that leads to meaningful and interpretable conclusion. A cell can be positive for multiple markers. If a cell is positive for a pair of markers, (k_r, k_s), it would imply that these markers have co-expressed in that particular cell. For every subject, compute the proportion of such double-positive cells, denoted by $k_r^+ k_s^+$, the proportion of cells positive for only the first marker, denoted by $k_r^+ k_s^-$, and the proportion of cells positive for only the second marker, denoted by $k_r^- k_s^+$, for $k_r, k_s \in \{1, 2, \ldots, p\}$ and $k_r \neq k_s$. Next, the subjects are classified into two or more groups based on their vectors of proportions using hierarchical clustering (Murtagh and Legendre, 2014). Denote the clustering vector as $\mathbf{Z} = (Z_1, \ldots, Z_N)^T$. When Y is a continuous/categorical outcome, a standard multiple linear regression model with $\mathbf{Z}$ as a predictor can be written as

$$Y = C\boldsymbol{\beta} + \mathbf{Z}\gamma + \epsilon,$$

where $\boldsymbol{\beta}, \gamma$ are fixed effects and ϵ is an $N \times 1$ error vector following multivariate normal distribution (MVN) with mean $\mathbf{0}$ and identity covariance matrix $\sigma^2 \mathbb{I}_N$. After estimating the parameters, the null hypothesis, $H_0 : \gamma = 0$, can be tested using the Wald test (Gourieroux *and others*, 1982).

Next, we consider the case of Y being a a right-censored failure time outcome. Let the outcome of the j-th individual be $Y_j = min(T_j, U_j)$, where T_j is the time to event and U_j is the censoring time. Let $\delta_j \equiv I(T_j \leq U_j)$ be the corresponding censoring indicator. Assuming that T_j and U_j are conditionally independent given the covariates for $j = 1, 2, \ldots, N$, the hazard function for the Cox proportional hazards (PH) model (Andersen and Gill, 1982) with fixed effects can be written as,

$$\lambda_j(t|C_j, Z_j) = \lambda_0(t) \exp(C_j^T \beta + Z_j \gamma), \quad j = 1, 2, \ldots, N \tag{1}$$

where $\lambda_j(t|C_j, Z_j)$ is the hazard of the j-th subject at time t, given the vector of covariates C_j and the cluster label Z_j and $\lambda_0(t)$ is an unspecified baseline hazard at time t. To test the null hypothesis: $H_0 : \gamma = 0$, a likelihood ratio test (LRT) (Therneau, 1997) can be considered.

2.2 Proposed Method: Mutual Information based analysis of marker co-expression

2.2.1 Theory of Mutual Information

Mutual information (MI) is an information theoretic measure of dependence between two or more random variables. In contrast to the linear correlation coefficient, it captures dependences which do not manifest themselves in the covariance (Kraskov *and others*, 2004).

For two random variables, R_1 and R_2 with sample space $\mathcal{D}$, marginal probability density functions (PDFs), f_1, f_2, and joint PDF f_{12}, the MI is defined as the following:

$$\mathrm{MI}(R_1, R_2) = \int\int f_{12}(r_1, r_2) \log \frac{f_{12}(r_1, r_2)}{f_1(r_1) f_2(r_2)} dr_1 dr_2.$$

MI has been used as a tool for feature selection in many different contexts (Yang and Moody, 1999; Liu *and others*, 2009; Hoque *and others*, 2014; Song *and others*, 2021). The measure can be easily generalized for more than two random variables. However, it becomes extremely difficult to estimate the joint density and hence, the MI nonparametrically in high dimensional spaces because of the curse of dimensionality (Langrené and Warin, 2019).

Xu (1998) and Principe *and others* (2000) looked into alternative definitions of MI that would remain conceptually similar but can be computed efficiently. They discussed two measures, Euclidean quadratic mutual information (EQMI) and Cauchy-Schwarz quadratic mutual information (CSQMI), defined as below,

$$\begin{aligned}
\mathrm{EQMI}(R_1, R_2) &= \int\int f_{12}(r_1, r_2)^2 dr_1 dr_2 - 2\int\int f_{12}(r_1, r_2) f_1(r_1) f_2(r_2) dr_1 dr_2 \\
&\quad + \int\int f_1(r_1)^2 f_2(r_2)^2 dr_1 dr_2, \qquad (2) \\
\mathrm{CSQMI}(R_1, R_2) &= \log \frac{(\int\int f_{12}(r_1, r_2)^2 dr_1 dr_2)(\int\int f_1(r_1)^2 f_2(r_2)^2 dr_1 dr_2)}{(\int\int f_{12}(r_1, r_2) f_1(r_1) f_2(r_2) dr_1 dr_2)}
\end{aligned}$$

It is trivial to verify that $\mathrm{EQMI}(R_1, R_2) \geq 0$ with equality occurring if and only if R_1, R_2 are independent i.e., $f_{12}(r_1, r_2) = f_1(r_1) f_2(r_2)$ for any $r_1, r_2 \in \mathbb{R}$. Using the Cauchy-Schwarz

inequality, we have,

$$\left(\int\int f_{12}(r_1,r_2)^2 dr_1 dr_2\right)\left(\int\int f_1(r_1)^2 f_2(r_2)^2 dr_1 dr_2\right)$$
$$\geq \left(\int\int f_{12}(r_1,r_2) f_1(r_1) f_2(r_2) dr_1 dr_2\right)^2$$

implying that CSQMI$(R_1, R_2) \geq 0$ with equality happening if and only if R_1, R_2 are independent. Xu (1998) argued for the strengths of EQMI versus CSQMI as a dependence measure. It should also be noticed that the form of EQMI is very similar to the distance covariance measure proposed by Székely *and others* (2007). In this paper, we focus on EQMI and propose its usage in our context.

2.2.2 Formulation of EQMI

For every subject j, we assume that the expression of marker k is a continuous random variable, denoted by X_{kj}, with sample space $\mathcal{D} = [0, 1]$. X_{kj} is observed in n_j cells as, $X_{k1j}, X_{k2j}, \ldots, X_{kn_jj}$. Suppose there are $p = 2$ markers and for every subject j, denote their joint PDF as $f_{12j}(x_1, x_2)$ and their marginal PDFs as $f_{1j}(x_1)$ and $f_{2j}(x_2)$ respectively. Following Equation (2), the EQMI between the markers 1 and 2 for subject j can be defined as,

$$\text{EQMI}(X_{1j}, X_{2j}) = V_J - 2V_C + V_M, \tag{3}$$

where $V_J = \int\int f_{12j}(x_1,x_2)^2 dx_1 dx_2$, $V_C = \int\int f_{12j}(x_1,x_2) f_{1j}(x_1) f_{2j}(x_2) dx_1 dx_2$, and $V_M = \int\int f_{1j}(x_1)^2 f_{2j}(x_2)^2 dx_1 dx_2$. EQMI$(X_{1j}, X_{2j})$ can be interpreted as a generalized measure of co-expression of the markers, capable of capturing non-linear dependences. A large

value of $\text{EQMI}(X_{1j}, X_{2j})$ will imply that the markers $1, 2$ have significantly co-expressed in the TME of subject j. $\text{EQMI}(X_{1j}, X_{2j})$ is bounded below by 0 for every j but there is no common upper bound for different j's. Therefore, to compare $\text{EQMI}(X_{1j}, X_{2j})$'s across different subjects, we need to appropriately standardize the values so that they lie in the same scale. We define a new measure as, $\text{EQMI}^*(X_{1j}, X_{2j}) = (V_J + V_M)^{-1}(V_J - 2V_C + V_M)$. We observe that $\text{EQMI}^*(X_{1j}, X_{2j})$ lies between $[0, 1]$ because $V_C \geq 0$, and has a similar interpretation as $\text{EQMI}(X_{1j}, X_{2j})$.

It is intuitive to generalize the measure for any $p \geq 2$ markers. For subject j, letting $f_{12\ldots pj}(x_1, x_2, \ldots, x_p)$ to be the joint PDF and $f_{1j}(x_p), f_{2j}(x_p), \ldots, f_{pj}(x_p)$ to be the marginal PDFs, EQMI^* can be defined as,

$$
\begin{aligned}
&\text{EQMI}^*(X_{1j}, X_{2j}, \ldots, X_{pj}) = \frac{V_J - 2V_C + V_M}{V_J + V_M} \\
&V_J = \int\int\ldots\int f_{12\ldots pj}(x_1, x_2, \ldots, x_p)^2 dx_1 dx_2 \ldots dx_p \\
&V_C = \int\int\ldots\int f_{12\ldots pj}(x_1, x_2, \ldots, x_p) f_{1j}(x_1) f_{2j}(x_2) \ldots f_{pj}(x_p) dx_1 dx_2 \ldots dx_p \\
&V_M = \int\int\ldots\int f_{1j}(x_1)^2 f_{2j}(x_2)^2 \ldots f_{pj}(x_p)^2 dx_1 dx_2 \ldots dx_p.
\end{aligned}
$$

Notice that to estimate $\text{EQMI}^*(X_{1j}, X_{2j}, \ldots, X_{pj})$ when the true PDFs are unknown, a naive approach would be to estimate the PDFs first by using a kernel density estimation (KDE) approach (Silverman, 1981). Then, we use the estimated PDFs to compute the terms V_J, V_C and V_M via numerical integration (Davis and Rabinowitz, 2007). However, such an approach would be computationally infeasible for a large p and will defeat the purpose of considering EQMI instead of the standard form of MI.

$\text{EQMI}^*(X_{1j}, X_{2j}, \ldots, X_{pj})$ can be estimated efficiently using the following formulae,

$$
\begin{aligned}
\widehat{\text{EQMI}}^*(X_{1j}, X_{2j}, \ldots, X_{pj}) &= \frac{\hat{V}_J - 2\hat{V}_C + \hat{V}_M}{\hat{V}_J + \hat{V}_M} \\
\hat{V}_J &= \frac{1}{n_j^2} \sum_{i=1}^{n_j} \sum_{s=1}^{n_j} \prod_{k=1}^{p} \hat{V}_k(i,s) \\
\hat{V}_C &= \frac{1}{n_j} \sum_{i=1}^{n_j} \prod_{k=1}^{p} \hat{V}_k(i); \hat{V}_k(i) = \frac{1}{n_j} \sum_{s=1}^{n_j} \hat{V}_k(i,s), \\
\hat{V}_M &= \prod_{k=1}^{p} \hat{V}_k; \hat{V}_k = \frac{1}{n_j^2} \sum_{i=1}^{n_j} \sum_{s=1}^{n_j} \hat{V}_k(i,s),
\end{aligned}
\tag{4}
$$

where $\hat{V}_k(i,s) = G_{\sqrt{2}h_k}(X_{kij} - X_{ksj})$ and G_h stands for a Gaussian kernel with bandwidth parameter h. This clever way of estimating the terms, $\hat{V}_J$, $\hat{V}_C$ and $\hat{V}_M$ is described in Principe (2010) with a simpler assumption that h_k's are equal for all k i.e., $h_k = h$ for $k = 1, 2, \ldots, p$. We provide the general derivation (i.e., $h_k \neq h_{k'}$, for $k \neq k'$) and other associated details in the Supplemetary material. For choosing the optimal values of h_k's, we generally consider the diagonal multivariate plug-in bandwidth selection procedure (Wand *and others*, 1994; Chacón and Duong, 2010). However, when p is large ($p > 6$), to avoid computational deadlock we suggest using Silverman's rule of thumb (Silverman, 1981) for choosing h_k's individually.

2.2.3 Using Mutual Information in association analysis

Denote the vector of estimated values of EQMI^* as $\mathbf{E} \equiv (E_1, \ldots, E_N)^T$, where $E_j = \widehat{\text{EQMI}}^*(X_{1j}, X_{2j}, \ldots, X_{pj})$. Our goal is to test if $\mathbf{E}$ is associated with the clinical outcome Y. When Y is a continuous/categorical outcome, a standard multiple linear regression

model with $\mathbf{E}$ as a predictor can be written as,

$$Y = C\boldsymbol{\beta} + \mathbf{E}\gamma + \epsilon,$$

where $\boldsymbol{\beta}, \gamma$ are fixed effects and ϵ is an $N \times 1$ error vector following multivariate normal distribution (MVN) with mean $\mathbf{0}$ and identity covariance matrix $\sigma^2 \mathbb{I}_N$. After estimating the parameters, the null hypothesis, $H_0 : \gamma = 0$, can be tested using the Wald test. Note that we can also add higher order terms of E_j, such as $E_j^2, E_j^3, \ldots$ as in a polynomial regression model (Ostertagová, 2012) to account for non-linear relationship between Y and E_j.

Next, we consider the case of Y being a survival or recurrence outcome. Using the same definitions and conditional independence assumptions of T_j, U_j and covariates as in Section 2.1, the hazard function for the Cox proportional hazards (PH) model can be written as,

$$\lambda_j(t|C_j, E_j) = \lambda_0(t) \exp(C_j^T \beta + E_j \gamma), \quad j = 1, 2, \ldots, N \tag{5}$$

where $\lambda_j(t|C_j, E_j)$ is the hazard of the j-th subject at time t, given the vector of covariates C_j and the predictor E_j and $\lambda_0(t)$ is an unspecified baseline hazard at time t. To test the null hypothesis: $H_0 : \gamma = 0$, a likelihood ratio test (LRT) can be considered.

In a two-marker scenario (i.e., $p = 2$), one can also treat the absolute value of the Pearson correlation as a measure of marker co-expression and use it as γ in the earlier equations for testing association with the outcome. However, as we later demonstrate, using correlation instead of EQMI can be sub-optimal in many cases. Also, it is difficult to generalize for more than two markers.

3 Real Data Analysis

We apply our method on two real datasets, an mIHC lung cancer dataset (Johnson *and others*, 2021; Seal *and others*, 2021) from Vectra 3.0 platform and a MIBI triple-negative breast cancer dataset (Keren *and others*, 2018). In all the subsequent analyses, the marker intensities were scaled to have expression value between zero and one.

3.1 Application to mIHC Lung Cancer data

In the lung cancer dataset, there are 153 subjects each with 3-5 images (in total, 761 images). The subjects have varying number of cells identified (from 3,755 to 16,949). The cells come from two different tissue regions: tumor and stroma. They are classified into either of the six different cell types: CD14+, CD19+, CD4+, CD8+, CK+ and Other, based on the expression of the phenotypic markers, CD19, CD3, CK, CD8 and CD14. A functional marker, HLA-DR (also known as MHCII), is also measured in each of the cells. Johnson *and others* (2021) classified the subjects into two groups, a) MHCII: High and b) MHCII: Low based on the proportion of CK+ tumor cells that are also positive for HLA-DR. They found out that there is significant difference in five-year overall survival between the groups.

Here, we were interested in a similar but slightly more general question, namely whether the degree of co-expression of the markers HLA-DR and CK in tumor cells is associated with the survival. Considering the tumor cells of every subject j, we first estimated EQMI*, $E_j = \widehat{\text{EQMI}}^*(X_{1j}, X_{2j})$ between the two markers, HLA-DR and CK. Next, we tested the association of E_j with five-year overall survival using the Cox-PH model from Equation (5). The p-value of the LRT came out to be 0.0286 indicating the survival to be associated with the degree of co-expression of the markers, HLA-DR and CK. The result was consistent with

Johnson *and others* (2021)'s finding. An important point to keep in mind is that unlike their approach, our approach did not require identifying tumor cells that were CK+, and identifying how many of these cells are also positive for the marker HLA-DR. Both of these steps would require manual threshold selection and hence, are prone to subjectivity.

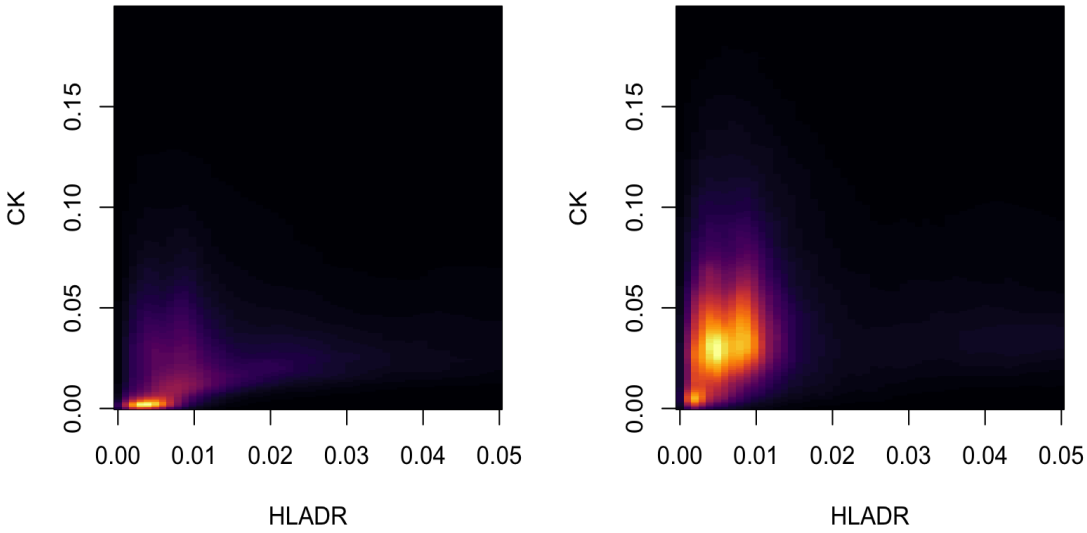

Figure 1: The left panel shows the mean of the estimated joint density of the markers, HLADR and CK across ten subjects who died before the study ended. The right panel shows the mean of the estimated joint density of those markers across ten subjects who were alive till the end of the study. A lighter color indicates high value of the estimated density.

We also compared the co-expression or dependence pattern of the two markers between the subjects who died before the end of the study (i.e., events) and those who survived. We found that the former group of subjects, on an average, had significantly lower values of EQMI* (p-value was 0.003 for a one-sided two-sample t-test based on the difference of the mean EQMI* values between the two groups). Figure 1 shows the mean of the estimated joint density of the two markers across ten subjects who died and across ten subjects who survived.

It was observed that the dependence between the markers was much more prominent in the second density than the first one. Overall, the significant difference of the estimated EQMI* between the deceased and alive subjects and the mean joint density plots further supported our earlier finding that 5-year survival was associated with the degree of co-expression of the markers, HLA-DR and CK.

3.2 Application to TNBC MIBI data

The triple-negative breast cancer (TNBC) MIBI dataset (Keren *and others*, 2018) has 43 subjects, each with one image. There are 201,656 cells in total with each of the images having varying numbers of cells (between 1217 and 8212). There are several functional proteins, such as HLA-DR, CD45RO, H3K27me3, H3K9ac and HLA-Class-1. Using a pixel-level thresholding-based method, the pairwise co-expression of these proteins have been shown to have an impact on the clinical outcomes, survival and recurrence in Patwa *and others* (2021). Here, we wanted to check if the co-expression of these markers in terms of EQMI* would be associated with survival and recurrence. Since there were five markers, we also studied if the co-expression of a smaller set of markers, such as the co-expression of only HLA-DR and CD45RO, would be associated with the clinical outcomes. We again used the Cox-PH model from Equation (5) and listed the p-values in Table 1. We noticed that the joint co-expression of all the five markers was significantly associated with both survival and recurrence. It was interesting to note that even the co-expression of three markers HLA-DR, CD45RO and HLA-Class-1 is significantly associated with both the clinical outcomes. This shows that with multiple markers, studying the co-expression pattern of a smaller set of markers could often be informative.

Table 1: p-values for testing association of EQMI* with survival and recurrence for different marker combinations in the MIBI dataset.

Marker Combination	Survival	Recurrence
HLA-DR, CD45RO	0.059	0.044
HLA-DR, CD45RO, HLA-Class-1	0.032	0.017
HLA-DR, CD45RO, HLA-Class-1, H3K27me3	0.032	0.006
HLA-DR, CD45RO, HLA-Class-1, H3K27me3, H3K9ac	0.010	0.008

4 Simulation

Next, we compared the performance of the EQMI*-based association analysis with the correlation-based association analysis and the traditional thresholding-based method in different simulation setups. We assumed that there were two groups of subjects, one with subjects having high marker co-expression and the other with subjects having low or almost zero marker co-expression. In Section 4.1, we considered two markers sharing linear and non-linear patterns of co-expression relationship. In Section 4.2, we considered three markers sharing varying degree of linear co-expression relationship. In the thresholding-based method, it was unknown which value of the threshold (t_k) would be optimal for marker k. Therefore, for every marker, we concatenated the intensity data of all the subjects and computed the 95% and 99% quantiles. Next, two different thresholding-based methods were considered with these quantiles as the threshold (t_k) and were referred to as Threshold 1 and Threshold 2 respectively.

4.1 Simulation with two markers

4.1.1 Simulation using Gaussian copula

We replicated the characteristics of the lung cancer dataset in this simulation. The mean marginal distribution of the markers HLA-DR and CK across the subjects could be approximated by Beta distribution respectively with parameters, $(1.5, 170)$ and $(1.6, 35)$ (refer to the Supplementary material). We used a Gaussian copula (Masarotto and Varin, 2012) to simulate correlated intensity data for two markers which had the above marginal Beta distributions. The simulation strategy was as follows,

1. Choose a random number I_j between 0 and 1, respectively standing for group (1), where subjects have high co-expression of the markers and group (2), where subjects have mild to none co-expression of the markers. It assigns j-th subject to either of the two groups.

2. The intensity vector of two markers, $(X_{1ij}, X_{2ij})^T$ for every individual j is simulated as follows,

   (a) If $I_j = 0$, simulate a correlation parameter ρ_j from Unif$(0.75, 0.9)$, or else simulate ρ_j from Unif$(0, 0.15)$.

   (b) Consider a correlation matrix, $R_j = \begin{bmatrix} 1 & \rho_j \\ \rho_j & 1 \end{bmatrix}$ and simulate $(u_{1ij}, u_{2ij})^T \sim N_2(0, R_j)$ for $i = 1, \ldots, n_j$.

   (c) Compute $(v_{1ij}, v_{2ij})^T = (\Phi(u_{1ij}), \Phi(u_{2ij}))^T$, where $\Phi()$ denotes the cumulative distribution function (CDF) of the standard normal distribution.

   (d) Perform inverse transformation as,
   $(X_{1ij}, X_{2ij})^T = (F_1^{-1}(v_{1ij}), F_2^{-1}(v_{2ij}))^T$, where F_1 and F_2 are Beta distributions

with parameters $(1.5, 170)$ and $(1.6, 35)$. Refer to the Supplementary material for the plots of joint densities of the markers for two different groups of subjects i.e., $I_j = 1$ vs. $I_j = 0$.

3. The clinical outcome of j-th subject is simulated as, $Y_j = I_j\beta + \epsilon_j$, where $\epsilon_j \sim N(0, \sigma^2)$. This step is repeated 100 times to generate 100 different datasets having different Y vectors but the same intensity vectors, X_1 and X_2. All the methods are applied on these 100 datasets and empirical power is computed.

Steps $1-3$ were repeated 20 times and in Figure 1 the mean empirical power of different methods were displayed for varying values of the number of cells (n_{cells}) and the number of individuals (N). The EQMI*-based association analysis (EQMI) and the correlation-based association analysis (Corr) achieved comparable performance in all the cases. This particular simulation strategy inherently assumed that the dependence between the markers were linear. Thus, the estimated values of the EQMI* and the correlation shared an almost one-to-one relationship making the association analysis using either of them equivalent. Threshold 1 also showed similar performance, whereas Threshold 2 had consistently lower power. It showed the importance of choosing a proper threshold in the traditional thresholding-based method for achieving a reasonable performance. All the methods expectedly had the least power when N was the smallest, whereas the value of n_{cells} only seemed to have a major impact on Threshold 2. It implied if the co-expression pattern was well captured even through a smaller number of cells, most of the methods would perform well.

4.1.2 Simulation with non-linear marker co-expression relationship

The last simulation strategy essentially assumed a linear pattern of co-expression between the markers. Next, we simulated a non-linear pattern of co-expression between the markers

where we would expect the correlation-based association analysis (Corr) to perform worse since it could only capture a linear dependence pattern. The simulation strategy was as follows,

1. Choose a random number I_j between 0 and 1, respectively standing for group (1), where subjects have high co-expression of the markers and group (2), where subjects have mild to none co-expressionof the markers.

2. The intensity vector of two markers, $(X_{1ij}, X_{2ij})^T$ for every individual j is simulated as follows,

   (a) If $I_j = 0$, simulate X_{1ij} from Unif$(0, 0.1)$, and e_{ij} from Unif$(0, 0.0005)$. Construct X_{2ij} as,
   $$X_{2ij} = (X_{1ij} - 0.05)^2 + e_{ij}.$$

   (b) If $I_j = 1$, independently simulate x_{1ij}, x_{2ij} both from Unif$(0, 0.1)$.

   Refer to the Supplementary material for the plots of joint densities of the markers for two different groups of subjects i.e., $I_j = 1$ vs. $I_j = 0$.

3. The clinical outcome of j-th subject is simulated as, $Y_j = I_j\beta + \epsilon_j$, where $\epsilon_j \sim N(0, \sigma^2)$. This step is repeated 100 times to generate 100 different datasets having different Y vectors but the same intensity vectors, X_1 and X_2. All the methods are applied on these 100 datasets and empirical power is computed.

Steps $1 - 3$ were repeated 20 times and in Figure 1 the mean empirical power of different methods were displayed. We noticed that EQMI*-based association analysis performed the best in all the cases. Thresholding 1 performed similarly, particularly for large n_{cells}. Note

that in both the groups of subjects the correlation between the markers were close to zero as the dependence pattern was non-linear, squared to be specific. Expectedly, the correlation-based association analysis (Corr) had almost no power in every case. The simulation strategy demonstrated why using a generalized measure of co-expression such as EQMI* would be more optimal in many cases.

4.1.3 Simulation with circular marker co-expression relationship

In the last simulation setup, Thresholding 1 performed well despite the marker co-expression pattern being non-linear. Next, we looked into a slightly more complicated co-expression relationship for which the thresholding-based approach would suffer. The simulation strategy was as follows,

1. Choose a random number I_j between 0 and 1, respectively standing for group (1), where subjects have high co-expression of the markers and group (2), where subjects have mild to none co-expression of the markers.

2. The intensity vector of two markers, $(X_{1ij}, X_{2ij})^T$ for every individual j is simulated as follows,

   (a) If $I_j = 0$, simulate X_{1ij} from Unif$(0, 0.1)$, e_{ij} from Unif$(0, 0.0005)$ and a random number s_{ij} between -1 and 1. Construct X_{2ij} as,

   $$X_{2ij} = 0.05 + s_{ij}\sqrt{0.05^2 - (X_{1ij} - 0.05)^2} + e_{ij}.$$

   (b) If $I_j = 1$, independently simulate X_{1ij}, X_{2ij} both from Unif$(0, 0.1)$.

Refer to the Supplementary material for the plots of joint densities of the markers for two different groups of subjects i.e., $I_j = 1$ vs. $I_j = 0$.

3. The clinical outcome of j-th subject is simulated as, $Y_j = I_j\beta + \epsilon_j$, where $\epsilon_j \sim N(0, \sigma^2)$. This step is repeated 100 times to generate 100 different datasets having different Y vectors but the same intensity vectors, X_1 and X_2. All the methods are applied on these 100 datasets and empirical power is computed.

Steps $1 - 3$ were repeated 20 times and in Figure 1 the mean empirical power of different methods were displayed. In all the cases, the EQMI*-based association analysis performed the best. The correlation-based association analysis (Corr) expectedly performed the worst. Threshold 1, unlike the last simulation, performed significantly worse. In this simulation setup, the subjects of one group had a circular pattern of marker co-expression and the others had almost zero co-expression. Recall that in a two-marker scenario, the thresholding-based method depends on computing the proportions of the cells positive for both the markers and of the cells positive for only one of the markers (Section 2.1). The difference between these proportions across the two groups of subjects became negligible under this setup making it difficult distinguishing them. This explains the poor performance of the thresholding-based methods.

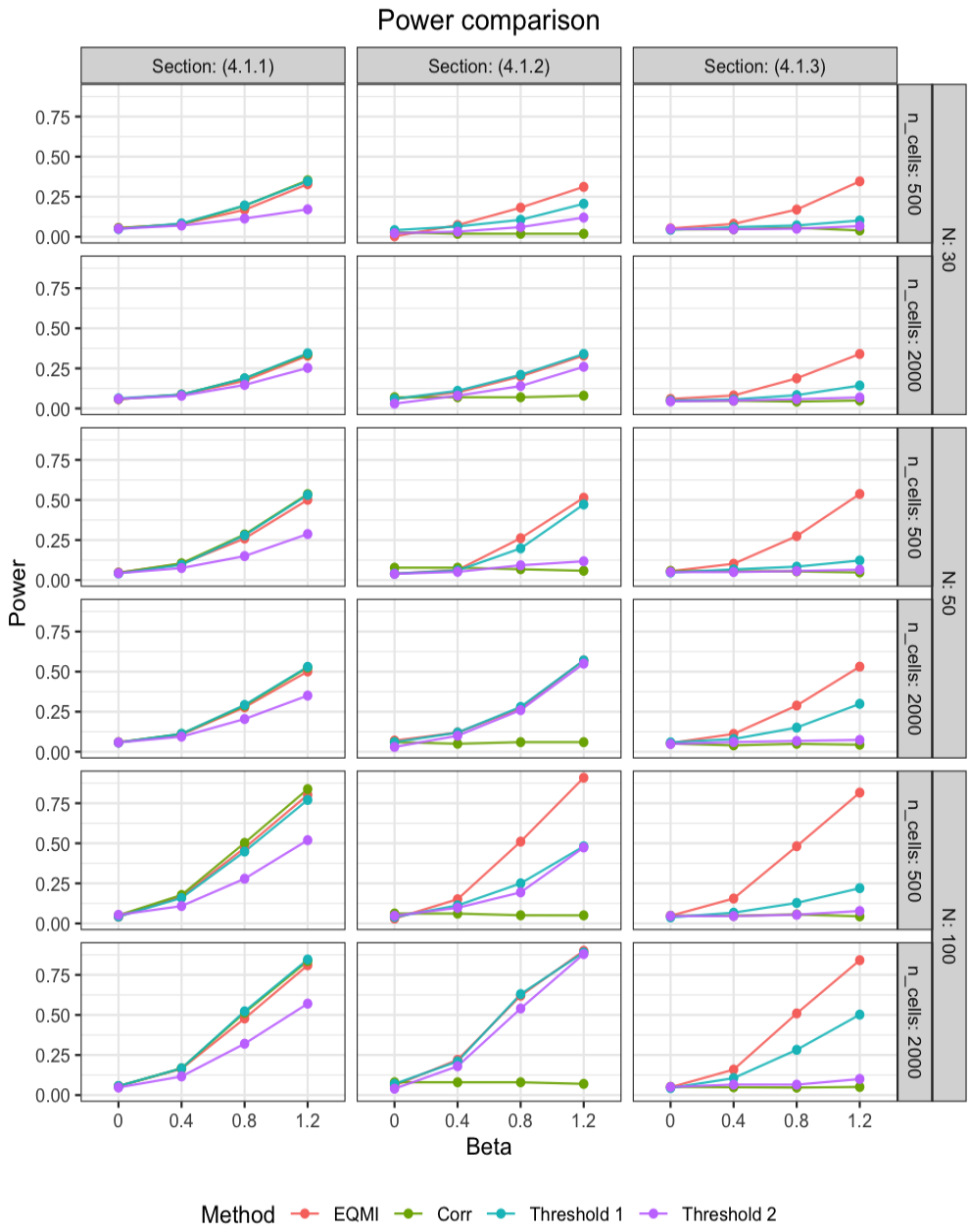

Figure 2: The figure displays the power of different methods under different simulation scenarios from Section 4.1 with two markers for varying numbers of subjects (N) and cells (n_{cells}). On the x-axis, the fixed effect size β was varied from low to high.

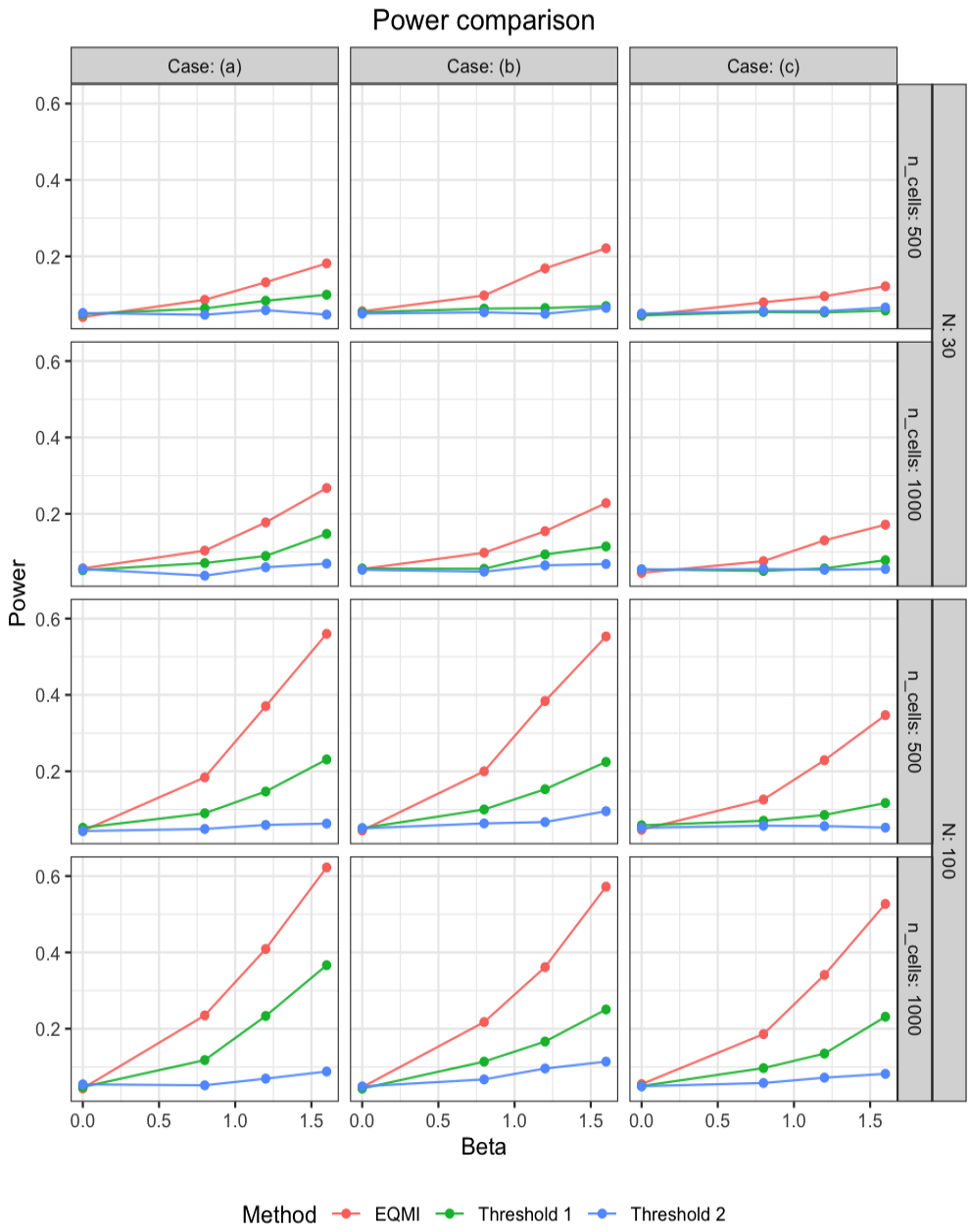

Figure 3: The figure displays the power of different methods under different cases from Section 4.2 with three markers for varying numbers of subjects (N) and cells (n_{cells}). On the x-axis, the fixed effect size β was varied from low to high.

4.2 Simulation with three markers

Next, we considered three markers and simulated varied degree of linear dependence between them using Gaussian copula. We only performed the EQMI*-based association analysis and the two thresholding-based methods Threshold 1 and 2 in this case. The simulation strategy was as follows,

1. Choose a random number I_j between 0 and 1, respectively standing for group (1), where subjects have high co-expression of the markers and group (2), where subjects have mild to none co-expression of the markers.

2. The intensity vector of three markers, $(X_{1ij}, X_{2ij}, X_{3ij})^T$ for every individual j is simulated as follows,

   (i) Consider a correlation matrix, $R_j = \begin{bmatrix} 1 & \rho_{12j} & \rho_{13j} \\ \rho_{12j} & 1 & \rho_{23j} \\ \rho_{13j} & \rho_{23j} & 1 \end{bmatrix}$. For the subjects in group (1), the off-diagonal elements of the correlation matrix would have high values, whereas they would have low values for the subjects in group (2). We consider three different cases with varying differences between the correlation matrices of the two groups.

   Case (a) : If $I_j = 0$, the correlation parameters ρ_{12j}, ρ_{23j} and ρ_{13j} are independently simulated from Unif$(0.4, 0.6)$. Otherwise, they are independently simulated from Unif$(0.2, 0.4)$.

   Case (b) : Regardless of the value of I_j, ρ_{13j} is kept to be 0. If $I_j = 0$, ρ_{12j} and ρ_{23j} are independently simulated from Unif$(0.4, 0.6)$. Otherwise, these two are independently simulated from Unif$(0.2, 0.4)$.

Case (c) : Regardless of the value of I_j, ρ_{13j} and ρ_{23j} are kept to be 0. If $I_j = 0$, ρ_{12j} is simulated from Unif$(0.4, 0.6)$. Otherwise, it is simulated from Unif$(0.2, 0.4)$.

(ii) Simulate $(u_{1ij}, u_{2ij}, u_{3ij})^T \sim N_3(0, R_j)$ and compute,
$(v_{1ij}, v_{2ij}, v_{3ij})^T = (\Phi(u_{1ij}), \Phi(u_{2ij}), \Phi(u_{3ij}))^T$.

(iv) Perform inverse transformation as,
$(X_{1ij}, X_{2ij}, X_{3ij})^T = (F_1^{-1}(v_{1ij}), F_2^{-1}(v_{2ij}), F_3^{-1}(v_{3ij}))^T$, where F_1, F_2 and F_3 are the Beta distributions with parameters $(1.5, 170)$, $(1.6, 35)$ and $(1.6, 35)$.

3. The clinical outcome of j-th subject is simulated as, $Y_j = I_j\beta + \epsilon_j$, where $\epsilon_j \sim N(0, \sigma^2)$. This step is repeated 100 times to generate 100 different datasets having different Y vectors but the same intensity data, X_1, X_2 and X_3. All the methods are applied on these 100 datasets and empirical power is computed.

Steps $1 - 3$ were repeated 20 times and in Figure 2 the mean empirical power of the methods were displayed. The power of the methods were quite low when N was small. The EQMI*-based method outperformed both the thresholding-based methods, Threshold 1 and 2 in every case. Threshold 2 had little to no power in most of the cases. Note that, the cases (a), (b), and (c) differed in how different the marker co-expression pattern of the two groups were. The difference between the marker co-expression pattern of the two groups was the largest in case (a) since all the three correlation parameters, ρ_{12j}, ρ_{23j} and ρ_{13j} were different across the groups. The difference was the smallest in case (c) as two of the three correlation parameters, ρ_{23j} and ρ_{13j} were kept to be 0 in both the groups. Quite expectedly, the power of the methods decreased going from case (a) to case (c), as the difference between the groups of subjects reduced. The decrease was more prominent with Threshold 1 suggesting the method's lack of robustness.

5 Discussion

In multiplex imaging data, studying the interaction or co-expression of multiple functional markers in the cells of the tumor microenvironment (TME) can be crucial for subject-specific assessment of risks. The traditional approach requires subjective threshold selection. This becomes more complex as the number of markers increases. In this paper, we propose a method for studying co-expression or co-expression of multiple markers based on the theory of mutual information (MI). We treat the subject-specific intensity or expression of every marker as a continuous random variable. We determine how much the markers have co-expressed in the TME of a particular subject by computing a measure known as Euclidean quadratic mutual information (EQMI), comparing the estimated marginal and joint probability density functions (PDFs) of the markers. The formula of EQMI has a similar interpretation as the standard formula of MI but allows a much efficient computation. We adopt and generalize an existing algorithm for computing EQMI that does not require explicitly estimating the joint PDF of the markers, a step which becomes increasingly intractable as the number of markers increases. Next, the subject-level EQMI values are tested for association with the clinical outcomes. The proposed method is free from the subjectivity bias of the traditional thresholding-based method and is readily applicable with any number of markers.

We have applied the proposed method to two real datasets, one mIHC lung cancer dataset and one MIBI triple negative breast cancer dataset. In the former, we find that the co-expression of the markers, HLA-DR and CK is associated with the five-year overall survival of the subjects. In the latter, we find that the co-expression of the markers, HLA-DR, CD45RO, H3K27me3, H3K9ac and HLA-Class-1 is associated with both risk of recurrence and risk of death. We evaluate the performance of our method through several simulation scenarios with two or three markers. In the scenarios with two markers, we show that all the methods

perform well and close to each other if the pattern of dependence (co-expression) between the markers is linear. However, with a more complex non-linear dependence pattern, only the proposed method can achieve respectable power. In the scenarios with three markers, we find that the proposed method performs consistently better than the thresholding-based method and shows superior robustness.

As we have shown in the simulation studies, the EQMI can capture both linear and non-linear patterns of co-expression between the markers very well. However, the measure is not well suited for capturing the differences between the patterns. For example, it may happen that one subject has a linear pattern of co-expression, whereas some other subject has a non-linear pattern. The EQMI for both the subjects can be very similar, making it hard to distinguish between them. As a part of our future direction, we would like to improve the method by detecting and incorporating the type of the co-expression pattern. When there are a large number of markers, it may be sensible to focus on a smaller set of markers and study their co-expression pattern for the sake of better interpretability and computational rapidness. In that light, we would like to explore the causal direction between the markers which can then be used to determine a smaller and optimal set of markers.

Our method is available as an *R* package named MIAMI at this link, https://github.com/sealx017/MIAMI. The package is readily applicable to any multiplex imaging dataset which has cell-level intensity data on two or more markers. In future, we would like to further augment the package's capability by incorporating a pixel-level analysis as well.

Acknowledgments

S. Seal acknowledges funding from the Grohne-Stepp Endowment from the University of Colorado Cancer Center, NCI R01 CA129102 and NSF DMS 1914937. D. Ghosh acknowledges funding from NCI R01 CA129102 and NSF DMS 1914937.

References

Ali, H Raza, Jackson, Hartland W, Zanotelli, Vito RT, Danenberg, Esther, Fischer, Jana R, Bardwell, Helen, Provenzano, Elena, Rueda, Oscar M, Chin, Suet-Feung, Aparicio, Samuel *and others*. (2020). Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nature Cancer* **1**(2), 163–175.

Andersen, Per Kragh and Gill, Richard D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 1100–1120.

Angelo, Michael, Bendall, Sean C, Finck, Rachel, Hale, Matthew B, Hitzman, Chuck, Borowsky, Alexander D, Levenson, Richard M, Lowe, John B, Liu, Scot D, Zhao, Shuchun *and others*. (2014). Multiplexed ion beam imaging of human breast tumors. *Nature Medicine* **20**(4), 436–442.

Armingol, Erick, Officer, Adam, Harismendy, Olivier and Lewis, Nathan E. (2021). Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics* **22**(2), 71–88.

Bataille, Frauke, Troppmann, Sabine, Klebl, Frank, Rogler, Gerhard, Stoelcker, Benjamin, Hofstadter, Ferdinand, Bosserhoff, Anja-Katrin

and Rümmele, Petra. (2006). Multiparameter immunofluorescence on paraffin-embedded tissue sections. *Applied Immunohistochemistry & Molecular Morphology* **14**(2), 225–228.

Binnewies, Mikhail, Roberts, Edward W, Kersten, Kelly, Chan, Vincent, Fearon, Douglas F, Merad, Miriam, Coussens, Lisa M, Gabrilovich, Dmitry I, Ostrand-Rosenberg, Suzanne, Hedrick, Catherine C *and others*. (2018). Understanding the tumor immune microenvironment (time) for effective therapy. *Nature Medicine* **24**(5), 541–550.

Chacón, José E and Duong, Tarn. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test* **19**(2), 375–398.

Davis, Philip J and Rabinowitz, Philip. (2007). *Methods of numerical integration*. Courier Corporation.

Duncan, Tyrone E. (1970). On the calculation of mutual information. *SIAM Journal on Applied Mathematics* **19**(1), 215–220.

Gourieroux, Christian, Holly, Alberto and Monfort, Alain. (1982). Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. *Econometrica: journal of the Econometric Society*, 63–80.

Hansen, Klaus H, Bracken, Adrian P *and others*. (2008). A model for transmission of the h3k27me3 epigenetic mark. *Nature Cell Biology* **10**(11), 1291–1300.

Hoque, Nazrul, Bhattacharyya, Dhruba K and Kalita, Jugal K. (2014). Mifs-nd: A mutual information-based feature selection method. *Expert Systems with Applications* **41**(14), 6371–6385.

Huang, Wei, Hennrick, Kenneth and Drew, Sally. (2013). A colorful future of quantitative pathology: validation of vectra technology using chromogenic multiplexed immunohistochemistry and prostate tissue microarrays. *Human Pathology* **44**(1), 29–38.

Jackson, Hartland W, Fischer, Jana R *and others*. (2020). The single-cell pathology landscape of breast cancer. *Nature* **578**(7796), 615–620.

Jin, Suoqin, Guerrero-Juarez, Christian F, Zhang, Lihua, Chang, Ivan, Ramos, Raul, Kuan, Chen-Hsiang, Myung, Peggy, Plikus, Maksim V and Nie, Qing. (2021). Inference and analysis of cell-cell communication using cellchat. *Nature Communications* **12**(1), 1–20.

Johnson, Amber M, Boland *and others*. (2021). Cancer cell-specific mhcii expression as a determinant of the immune infiltrate organization and function in the non-small cell lung cancer tumor microenvironment. *Journal of Thoracic Oncology*.

Johnson, Amber M, Bullock *and others*. (2020). Cancer cell–intrinsic expression of mhc class ii regulates the immune microenvironment and response to anti–pd-1 therapy in lung adenocarcinoma. *The Journal of Immunology* **204**(8), 2295–2307.

Keren, Leeat, Bosse, Marc, Marquez, Diana, Angoshtari, Roshan, Jain, Samir, Varma, Sushama, Yang, Soo-Ryum, Kurian, Allison, Van Valen, David, West, Robert *and others*. (2018). A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell* **174**(6), 1373–1387.

Koguchi, Yoshinobu, Hoen, Helena M, Bambina, Shelly A, Rynning, Michael D, Fuerstenberg, Richard K, Curti, Brendan D, Urba, Walter J,

Milburn, Christina, Bahjat, Frances Rena, Korman, Alan J *and others*. (2015). Serum immunoregulatory proteins as predictors of overall survival of metastatic melanoma patients treated with ipilimumab. *Cancer Research* **75**(23), 5084–5092.

Kraskov, Alexander, Stögbauer, Harald and Grassberger, Peter. (2004). Estimating mutual information. *Physical Review E* **69**(6), 066138.

Langrené, Nicolas and Warin, Xavier. (2019). Fast and stable multivariate kernel density estimation by fast sum updating. *Journal of Computational and Graphical Statistics* **28**(3), 596–608.

Lee, HE, Chae, SW, Lee, YJ, Kim, MA, Lee, HS, Lee, BL and Kim, WH. (2008). Prognostic implications of type and density of tumour-infiltrating lymphocytes in gastric cancer. *British Journal of Cancer* **99**(10), 1704–1711.

Liu, Huawen, Sun, Jigui, Liu, Lei and Zhang, Huijie. (2009). Feature selection with dynamic mutual information. *Pattern Recognition* **42**(7), 1330–1339.

Masarotto, Guido and Varin, Cristiano. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics* **6**, 1517–1549.

Murtagh, Fionn and Legendre, Pierre. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of Classification* **31**(3), 274–295.

Ostertagová, Eva. (2012). Modelling using polynomial regression. *Procedia Engineering* **48**, 500–506.

Patwa, Aalok, Yamashita, Rikiya, Long, Jin, Risom, Tyler, Angelo, Michael, Keren, Leeat and Rubin, Daniel L. (2021). Multiplexed imaging analysis of the tumor-immune microenvironment reveals predictors of outcome in triple-negative breast cancer. *Communications Biology* **4**(1), 1–14.

Pollan, Sara, Hanifi, Arezoo, Nagy, Mate, Stavrou, Nicholas, Parnell, Erinn, Gozo, Maricel, Attanasio, Nickolas, William, Josette and Au, Qingyan. (2020). Profiling exhausted t cells using vectra® polaris™ multiplex immunofluorescence assay in hnscc.

Principe, Jose C. (2010). *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media.

Principe, Jose C, Xu, Dongxin, Zhao, Qun and Fisher, John W. (2000). Learning from examples with information theoretic criteria. *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology* **26**(1), 61–77.

Saraiva, Diana P, Jacinto, António, Borralho, Paula, Braga, Sofia and Cabral, M Guadalupe. (2018). Hla-dr in cytotoxic t lymphocytes predicts breast cancer patients' response to neoadjuvant chemotherapy. *Frontiers in Immunology* **9**, 2605.

Seal, Souvik, Wrobel, Julia, Johnson, Amber M, Nemenoff, Raphael A, Schenk, Erin L, Bitler, Benjamin G, Jordan, Kimberly R and Ghosh, Debashis. (2021). On clustering for cell phenotyping in multiplex immunohistochemistry (mihc) and multiplexed ion beam imaging (mibi) data.

Shipkova, Maria and Wieland, Eberhard. (2012). Surface markers of lymphocyte activation and markers of cell proliferation. *Clinica Chimica Acta* **413**(17-18), 1338–1349.

Silverman, Bernard W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)* **43**(1), 97–99.

Song, Xian-fang, Zhang, Yong, Gong, Dun-wei and Sun, Xiao-yan. (2021). Feature selection using bare-bones particle swarm optimization with mutual information. *Pattern Recognition* **112**, 107804.

Székely, Gábor J, Rizzo, Maria L and Bakirov, Nail K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**(6), 2769–2794.

Tan, Wei Chang Colin, Nerurkar, Sanjna Nilesh, Cai, Hai Yun, Ng, Harry Ho Man, Wu, Duoduo, Wee, Yu Ting Felicia, Lim, Jeffrey Chun Tatt, Yeong, Joe and Lim, Tony Kiat Hon. (2020). Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications* **40**(4), 135–153.

Therneau, Terry M. (1997). Extending the cox model. In: *Proceedings of the First Seattle Symposium in Biostatistics*. Springer. pp. 51–84.

Wand, Matt P, Jones, M Chris *and others*. (1994). Multivariate plug-in bandwidth selection. *Computational Statistics* **9**(2), 97–116.

Wong, Joyce Y, Kuhl, Tonya L, Israelachvili, Jacob N, Mullah, Nasreen and Zalipsky, Samuel. (1997). Direct measurement of a tethered ligand-receptor interaction potential. *Science* **275**(5301), 820–822.

Xu, Dongxin. (1998). *Energy, entropy and information potential for neural computation*. University of Florida.

Yang, H and Moody, John. (1999). Feature selection based on joint mutual information. In: *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, Volume 23. Citeseer.

Zola, Heddy, Swart, Bernadette, Nicholson, Ian and Voss, Elena. (2007). *Leukocyte and stromal cell molecules: the CD markers*. John Wiley & Sons.