Identifying Spatial Biomarkers from Cellular Imaging Data

Casey Bradshaw [*1] **Wesley Tansey** [*2]

Abstract

The tumor microenvironment (TME) plays an important role in cancer development and progression. High-resolution spatial data describing cell phenotypes and locations within tumor tissue can offer insights into the structures that form among cells in the TME. Inferring clinically-important spatial TME structures remains an open problem. We outline a strategy for detecting recurring cellular structures in tumor tissue samples, and apply this method to imaging mass cytometry data from a breast cancer study, revealing several recurring spatial patterns. We demonstrate that two of these discovered patterns are unfavorable markers for overall survival, and cause a reduction in median survival time.

1. Introduction

The composition of the TME impacts tumor evolution, growth, immune surveillance, and response to therapy in cancer (Quail & Joyce, 2013; Chaudhary & Elkord, 2016). For example, the composition and phenotypic states of immune cells in the TME impact response to immunotherapy (Kalbasi & Ribas, 2020). More recent work suggests that not only cellular composition but *spatial organization* of the TME determines immunotherapy response (Herbst et al., 2014; Echarti et al., 2019).

Advances in multiplexed imaging technologies such as imaging mass cytometry (IMC) (Giesen et al., 2014) have created a wealth of fine-grained information on the spatial composition of tumor tissue. It is now possible to measure individual cell types *in situ*–preserving and recording the precise location of each cell within a tissue sample. This information presents an opportunity for spatial modeling approaches to improve our understanding of the spatial architecture of the tumor microenvironment.

However, in this high-dimensional data, it can be challenging to synthesize organizational patterns among groups of cells. Recent work (Danenberg et al., 2022; Kim et al., 2022) explores computational methods for detecting recurring microenvironment structures in IMC data. Danenberg et al. (2022)'s methods incorporate input from expert pathologists at several stages; in practice, this approach may be too labor-intensive for routine use. The UTAG method (Kim et al., 2022) takes an unsupervised approach to discovering spatial structures using clustering on local graph convolutions. While UTAG identifies recurring structures, it requires several subjective hyperparameter choices and hand-labeling of cell types and structure markers.

In this paper, we present `SpaceMarkers`, a multiscale spatial factor modeling approach to discovering recurrent spatial patterns in IMC data. We apply `SpaceMarkers` to 681 breast cancer tissue samples from participants in the METABRIC study (Curtis et al., 2012; Danenberg et al., 2022), in search of spatial biomarkers. `SpaceMarkers` first infers the cell types for each individual cell in the tissue images in a spatially-agnostic way. Using the cell type labels, `SpaceMarkers` then segments each image into regions of high similarity by solving a graph-based optimization problem. From this set of smaller regions, `SpaceMarkers` detects recurring spatial regions that appear across patient samples. `SpaceMarkers` then analyzes these recurrent spatial patterns for their potential as novel biomarkers in a causal survival analysis framework. On the METABRIC data, `SpaceMarkers` discovers two spatial patterns which, after controlling for confounders, are indicative of negative prognosis with respect to overall survival. The presence of the two recurring patterns correspond to differences in median survival times of 47 and 71 months, respectively.

2. Biomarker discovery

`SpaceMarkers` discovers spatial biomarkers in the form of cellular communities. A cellular community is defined as a spatially-contiguous region of tissue in which the proportions of cell types are spatially invariant. This is motivated by the insight that tumors with CD8+ T-cells interspersed

[*]Equal contribution [1]Department of Statistics, Columbia University, New York, USA [2]Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA. Correspondence to: Casey Bradshaw <cb3431@columbia.edu>, Wesley Tansey <tanseyw@mskcc.org>.

The 2022 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

with tumor cells respond better to immunotherapy than tumors where T-cells are excluded around the periphery. In the former case, moderate levels of T-cells and tumor cells are seen throughout the tumor. In the latter case, tumor cells are the dominant cell type in one portion of the tissue, with T and other cell types occupying separate, excluded regions. SpaceMarkers formalizes and generalizes this biological insight.

2.1. Setup and preprocessing

We assume image s contains $n^{(s)}$ cells $y_1^{(s)}, \dots, y_{n^{(s)}}^{(s)}$, for $s = 1, \dots, S$ tumor images. For ease of exposition, we drop the superscript s where it is clear that we are referring to a single image. For each cell, we follow the standard preprocessing pipeline to segment cells and summarize the mean expression level for each protein (Danenberg et al., 2022). After preprocessing, we are left with measurements y_{ij}, $j = 1, \dots, P$ corresponding to P different protein signals.

We convert the nonnegative raw signal into continuous measurements by log-transforming them. We then clip the signals at the 99th percentile to remove outliers and z-score each signal to place each signal and image on a common scale. After standardization, signals typically exhibit two dominant modes, corresponding to the (noisy) signal being "on" or "off" in the cell. See the appendix for an example of preprocessed signals in the METABRIC dataset.

To discover spatial patterns, SpaceMarkers constructs a neighbor graph for each image. Each cell is a node labeled with its prepossessed signal and edges $(e_1, e_2) \in \mathcal{E}$ connect each cell to its 10 nearest neighbors. This graph simultaneously encodes information about cell types and proximity.

2.2. Generative model for cellular communities

For a given image, SpaceMarkers poses the following generative model for the preprocessed graph,

$$\begin{aligned} (y_{ij} \mid h_i = k) &\sim \mathcal{N}(\hat{\mu}_{jm_{kj}}, \hat{\sigma}^2_{jm_{kj}}) \,, \quad m \in \{0,1\} \\ h_i &\sim Cat(\boldsymbol{\theta}_i) \\ ||\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i'}|| &\sim Laplace(\lambda) \,, \quad (i, i') \in \mathcal{E} \,, \end{aligned} \tag{1}$$

where each cell signal y_{ij} comes from either the on ($m = 1$) or off ($m = 0$) component. Each cell is assumed to have a latent cell type $h_i = k$ for $k = 1, \dots, K$. Each signal component's parameters $(\hat{\mu}, \hat{\sigma})$ are estimated in the preprocessing phase and assume fixed.

SpaceMarkers places a graph fusion spatial prior (Equation (1), last line) on the distribution of cell types (Kyung et al., 2010). This fusion prior captures the idea of cellular communities. Cell type distributions between neighboring nodes are shrunk to be exactly equal, creating connected regions of constant probability.

2.3. Inference

The METABRIC dataset contains over 1M cells. To ensure scalability to these large cohorts, SpaceMarkers takes a two-step *maximum a posteriori* (MAP) estimation approach.

Stage 1: Inferring cell types. SpaceMarkers performs a spatially-invariant nonnegative matrix factorization to infer the latent cell types,

$$\underset{H \in \{1,\dots,K\}^N, M \in \{0,1\}^{K \times P}}{\text{minimize}} - \sum_i \sum_j \log \mathcal{N}(\hat{\mu}_{m_{h_i,j}}, \hat{\sigma}^2_{m_{h_i,j}}) \,, \tag{2}$$

where H is the matrix of all N cells across all images and M is the $K \times P$ matrix of cell types; likelihood parameters $(\hat{\mu}, \hat{\sigma})$ are assumed to correspond the image for the corresponding cell.

Each cell type is modeled as having each protein either expressed or not. SpaceMarkers fits Equation (2) via Expectation Maximization (EM) and chooses the number of cell types K through 5-fold cross-validation. On the METABRIC dataset, CV indicated a reasonable choice was between 11 and 20 cell types; we chose $K = 15$.

Figure 1 (top) shows the inferred cell types; red dots correspond to the cell type being on for that signal. Biological interpretation of these cell types is left for future work. A point of interest for future exploration is that cell types 2 and 3 both express panCK and CD4. However, this is implausible, as panCK is indicative of epithelial tumor cells, while CD4 is indicative of immune cells. One possibility is that there are in fact two cells interacting with each other in such locations (e.g. CD4+ T cells attacking panCK+ tumor cells), and the cell segmentation algorithm mistakenly labeled them as a single cell.

Stage 2: Inferring cellular communities. SpaceMarkers discovers cellular communities by fixing the inferred cell types and solving a second optimization problem,

$$\underset{\boldsymbol{\beta}_i \in \mathbb{R}^K,\, i=1,\dots,n}{\text{minimize}} -1/n \sum_i \log \beta_{ih_i} + \frac{\lambda}{\mid \mathcal{E} \mid} \sum_{i,i' \in \mathcal{E}} ||\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}|| \,, \tag{3}$$

where the left side of Equation (3) is the cross-entropy loss and the right side is the group graph-fused lasso penalty (Hallac et al., 2015). The objective function is convex and we solve it via L-BFGS-B. The hyperparameter λ controls the smoothness of the resulting graph. We set $\lambda = 0.5$ as it yielded visibly reasonable communities in a few test images; tuning this automatically, e.g. via BIC, is left for future work. After fitting, we recursively merge

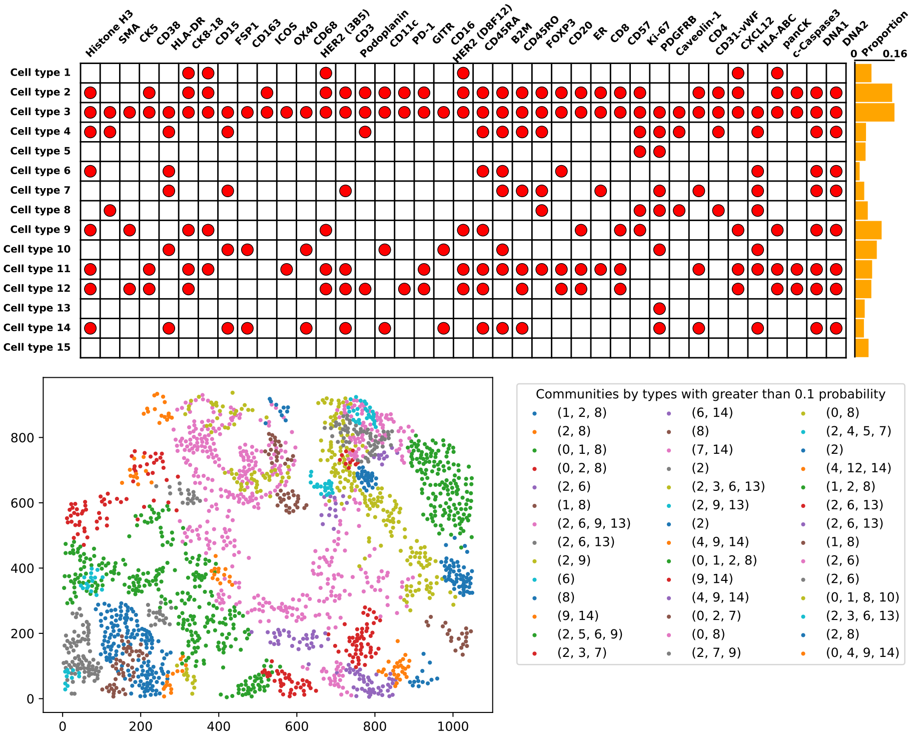

Figure 1. Top: Antibody profiles of the 15 identified cell types; *Bottom*: Cellular community structures identified in one patient's imaging. Colors indicate community, and are annotated with the cell types having probability greater than 0.1 within that community

adjacent nodes with cell type distributions within 10^-2 of each other. This identifies regions in which the probability distribution over cell types is effectively constant. Figure 1 (bottom) demonstrates the result of this process for a sample image, with the cells in an image highlighted according to their segmented communities. Applying SpaceMarkers to the METABRIC dataset process yields approximately 16,000 communities across all 681 images.

2.4. Identifying recurring communities

Each of the 16,000 cellular communities is characterized by a probability distribution over cell types. SpaceMarkers pools together these communities from different images to find canonical community archtypes that recur across images. To do this, SpaceMarkers performs K-means clustering on the probability vectors. The number of clusters K is chosen by maximizing the silhouette score (Rousseeuw, 1987); on the METABRIC dataset, the score is maximized at $K = 15$ clusters.

3. Survival Analysis

SpaceMarkers identified 15 recurring structures of cellular organization in the METABRIC IMC images. Whether these structures are meaningful depends on what they can tell us about a patient's disease. In particular, we assess whether these identified structures have implications for overall survival time.

3.1. Hazard Ratios

Having identified 15 recurring community structures in the imaging data, we record the presence or absence of each community in each individual image. Several patients had multiple samples in the dataset; for these patients, we record whether a particular community is present in any of the patient's samples, or is absent from all of them. To evaluate the implications of the identified structures, we fit a Cox proportional hazards model to the community indicators. The clinical data includes variables that could conceivably influence both the tumor microstructure and survival time, so we include these as covariates in the model. These variables are:

- **Hormone receptor status**: HER2, ER, and PR status
- **Treatments types received**: chemotherapy, radiotherapy, hormone therapy
- **Inferred menopausal status**: in this dataset, equiva-

lent to an indicator variable for age 50).

Figure 2 displays the hazard ratios for each community indicator, as well as their 95% confidence intervals. Community 7 and Community 13 emerge as significant, with estimated hazard ratios of 1.51 and 1.35 respectively. Because these hazard ratios are greater than one, the presence of either of these structures is a negative prognostic factor for overall survival.

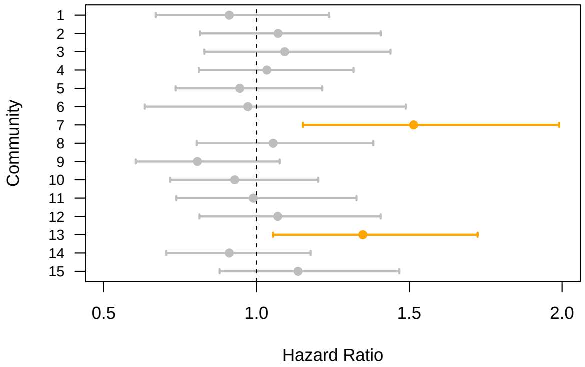

Figure 2. Hazard Ratios for the presence of each of the 15 community structures, with associated 95% confidence intervals

3.2. Causal Effects

Next we focus our attention on communities 7 and 13, and investigate their effect on overall survival. This is a question of causality rather than association. Because the available data are observational, we must make adjustments to accommodate causal conclusions.

We posit that the clinical covariates from Figure 2 affect both tumor structure and survival, and furthermore, that these covariates are sufficient to block any backdoor paths connecting tumor structure and survival (Pearl et al., 2016). Then, controlling for these clinical variables allows us to infer the effect of tumor structure on survival.

Suppose that X_i indicates the presence of community 7 in patient i, and S_i is patient i's set of clinical covariates (hormone receptor status, treatments received, menopausal status). Then, we can equivalently control for $\mathbb{P}(X_i = 1|S_i)$, the propensity score for patient i (Rosenbaum & Rubin, 1983). We estimate these propensity scores from the data using logistic regression.

Having calculated propensity scores, we want to compare the outcomes of patients exhibiting community structure 7 with their counterparts who were equally likely to exhibit that structure, but did not, due to chance. To facilitate this comparison, follow the strategy of Rosenbaum & Rubin (1985) and construct a resampled dataset from the original. For each patient with community 7 present, we identify their match: the patient with the closest propensity score among those with community 7 absent. Repeated propensity score values were jittered by a small, random amount to break ties and encourage diversity in the set of matches. The resampled dataset consists of patients with community 7 present in their imaging, together with their respective matches. In this resampled dataset, differences in outcome between the community-7-present and -absent groups can be attributed to the community structure itself.

Fitting a Kaplan-Meier estimator to the survival curves in this resampled dataset illustrates the effect of community 7. As shown in Figure 4 in the appendix, median survival time is estimated to be 125.6 months for patients with community 7 in their imaging, 47.3 months shorter than those without.

Repeating this analysis for the community 13 structure, we find that the presence of community 13 reduces the estimated median survival time from 205.6 months to 134.5 months. This 71 month reduction should be viewed with caution, as it is quite large within the landscape of breast cancer survival times. Figure 4 in the appendix shows that the survival curve for the group without community 13 has an irregular shape near the 50% threshold, which contributes to the large difference in the estimated medians.

4. Discussion

We have presented `SpaceMarkers`, a method for detecting recurring spatial structures in cellular imaging data. `SpaceMarkers` begins with identifying cell types from measured protein abundances. These cell type labels are then used to segment images into distinct regions via a graph-structured penalized optimization algorithm. Clustering of the regions from all images yields recurring spatial structures. On the METABRIC dataset, `SpaceMarkers` identified 15 recurring structures. Two of these structures are associated with significant hazard relative to baseline. Further analysis indicated that the presence of these two structures in a patient's tumor tissue causes a reduction in median survival time.

This initial result warrants further exploration of the biological interpretation of the identified cell types and spatial communities. Community 13 showed no significant associations with any categorical features in the clinical data set after correction for multiple comparisons, suggesting that the presence of this structure conveys prognostic information not captured by the clinical data. Further investigation is needed to establish whether clinical subtypes, genomics, or cell composition would account for this signal or if the recurrent communities represent novel and informative biomarkers. Overall, the results here demonstrate the potential of `SpaceMarkers` to uncover informative, recurrent spatial patterns that carry prognostic value for cancer patients.

References

Chaudhary, B. and Elkord, E. Regulatory T cells in the tumor microenvironment and cancer progression: Role and therapeutic targeting. *Vaccines*, 4(3):28, 2016.

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A. L., Brenton, J. D., Tavaré, S., Caldas, C., and Aparicio, S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

Danenberg, E., Bardwell, H., Zanotelli, V. R. T., Provenzano, E., Chin, S., Rueda, O., Green, A., Rakha, E., Aparicio, S., Ellis, I. O., Bodenmiller, B., Caldas, C., and Ali, H. R. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nature Genetics*, pp. 1–10, 2022.

Echarti, A., Hecht, M., Büttner-Herold, M., Haderlein, M., Hartmann, A., Fietkau, R., and Distel, L. CD8+ and regulatory T cells differentiate tumor immune phenotypes and predict survival in locally advanced head and neck cancer. *Cancers*, 11(9):1398, 2019.

Giesen, C., Wang, H. A., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., and Brandt, S. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4):417–422, 2014.

Hallac, D., Leskovec, J., and Boyd, S. Network lasso: Clustering and optimization in large graphs. In *International Conference on Knowledge Discovery and Data Mining*, 2015.

Herbst, R. S., Soria, J.-C., Kowanetz, M., Fine, G. D., Hamid, O., Gordon, M. S., Sosman, J. A., McDermott, D. F., Powderly, J. D., and Gettinger, S. N. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature*, 515(7528):563–567, 2014.

Kalbasi, A. and Ribas, A. Tumour-intrinsic resistance to immune checkpoint blockade. *Nature Reviews Immunology*, 20(1):25–39, 2020.

Kim, J., Rustam, S., Mosquera, J. M., Randell, S. H., Shaykhiev, R., Rendeiro, A. F., and Elemento, O. Unsupervised discovery of tissue architecture in multiplexed imaging. *bioRxiv: 2022.03.15.484534*, 2022.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

Pearl, J., Glymour, M., and Jewell, N. (eds.). *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

Quail, D. F. and Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. *Nature Medicine*, 19(11):1423–1437, 2013.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rosenbaum, P. R. and Rubin, D. B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

A. Example of preprocessed IMC signals

We fit a two-component Gaussian mixture model via Expectation Maximization on the raw signal for each image. Figure 3 shows an example image and the GMM fits.

B. Kaplan-Meier survival curves for recurrent communities

Figure 4 shows the survival curves for patients stratified by the two recurrent communities identified as significant in our Cox survival model.

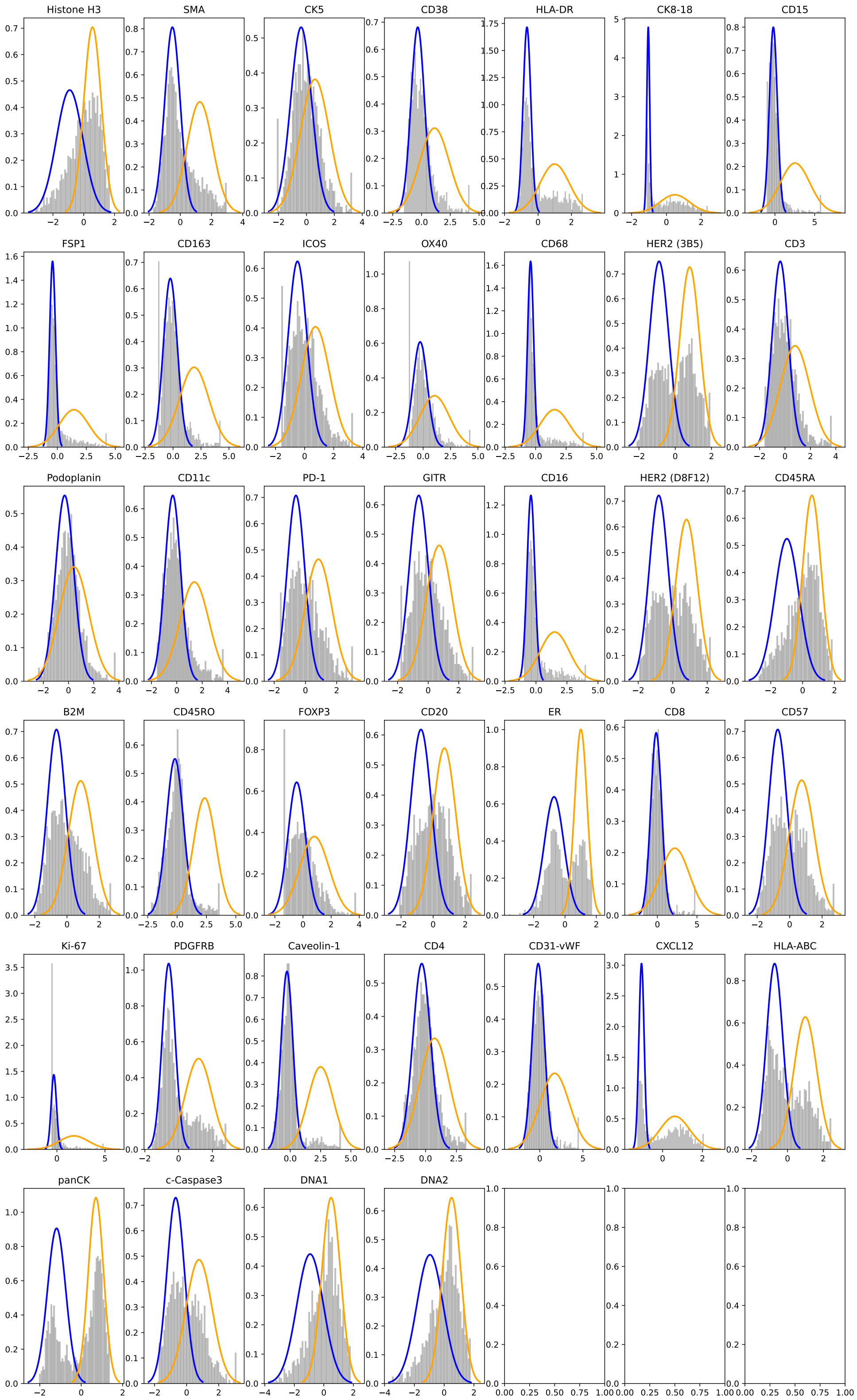

Figure 3. Standardized signals fit well to a two-component mixture model. Each histogram corresponds to the marginal distribution of each protein signal for a single image from the METABRIC dataset. Blue and orange densities correspond to the off and on components fit via EM.

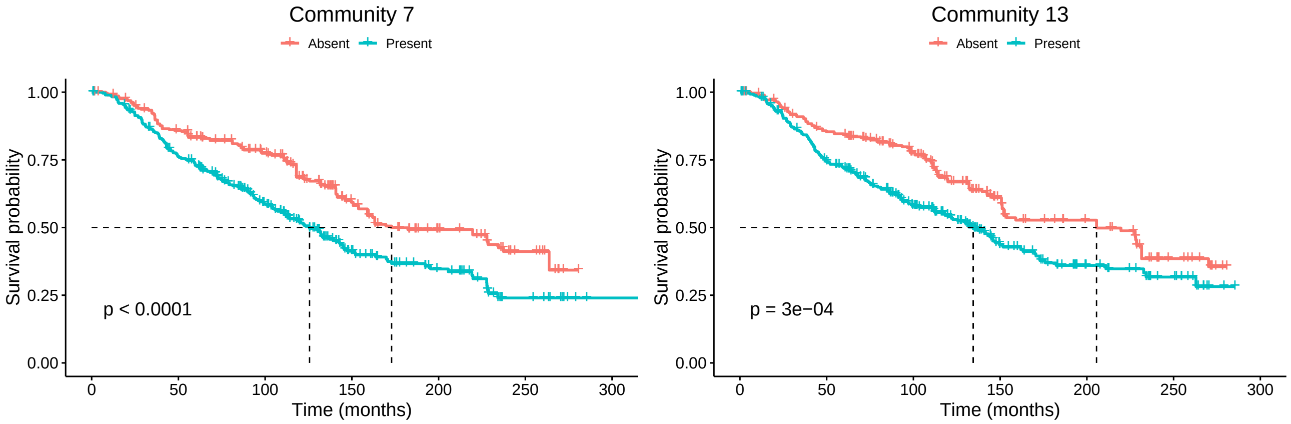

Figure 4. Left: Kaplan-Meier survival curve for community 7 (p-value corresponds to log-rank test and is not adjusted for multiple comparisons); *Right*: Kaplan-Meier survival curve for community 13