A Novel Integrative Multiomics Method Reveals a Hypoxia-Related Subgroup of Breast Cancer with Significantly Decreased Survival

Maryam Pouryahya[a], Jung Hun Oh[a], Pedram Javanmard[b], James C. Mathews[a], Zehor Belkhatir[a], Joseph O. Deasy[a,*], and Allen R. Tannenbaum[a,b,*]

[a]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, NY; [b]Department of Medicine, Division of Endocrinology Diabetes and Bone Disease, Icahn School of Medicine at Mount Sinai, NY; [c]Departments of Computer Science and Applied Mathematics & Statistics, Stony Brook University, NY; [*]Co-senior authors

This manuscript was compiled on February 24, 2019

The remarkable growth of multi-platform genomic profiles has led to the multiomics data integration challenge. The effective integration of such data provides a comprehensive view of the molecular complexity of cancer tumors and can significantly improve clinical outcome predictions. In this study, we present a novel network-based integration method of multiomics data as well as a clustering technique involving the Wasserstein (Earth Mover's) distance from the theory of optimal mass transport. We applied our proposed method of integrative Wasserstein-based clustering (iWCluster) to invasive breast carcinoma from The Cancer Genome Atlas (TCGA) project. The subtypes were characterized by the concordant effect of mRNA expression, DNA copy number alteration, and DNA methylation as well as the interaction network connectivity of the gene products. iWCluster is substantially more effective in distinguishing clusters with different survival rates as compared to isolated one-dimensional conventional omics analysis. Applying iWCluster to breast cancer TCGA data successfully recovered the known PAM50 molecular subtypes. In addition, iWCluster preserves the gene-specific data, which enables us to interpret the results and perform further analysis of significant genes for a specific cluster. The gene ontology enrichment analysis of significant genes in our substantially low survival subgroup leads to the well-known phenomenon of tumor hypoxia and the transcription factor ETS1 whose expression is induced by hypoxia. Increased expression of ETS1 is associated with an increased risk of recurrence and worse prognosis in breast cancer. Consequently, we believe iWCluster has the potential to discover novel subtypes by accentuating the genes that have concordant multiomics measurements in their interaction network, which are challenging to find without the network inference or with single omics analysis.

Multiomics Integration| Cancer Network | Clustering | Optimal Transport | Hypoxia | Breast Cancer

The molecular complexity of neoplastic development occurs at a number of genomic, transcriptomic and epigenomic scales. Integrating multiomics provides a comprehensive view of the etiology of oncogenesis and tumor progression at different molecular levels. Many sophisticated mathematical and statistical methods of multiomics data integration have been proposed, yet the need for an effective technique to improve the clinical outcome prediction remains a challenge in the era of precision medicine. Large scale cancer genome projects, such as The Cancer Genome Atlas (TCGA) project include an unprecedented amount of multi-dimensional data to explore the entire spectrum of genomic abnormalities in human cancer (1). With the exponential growth of such multi-dimensional data, the great need to obtain an integrated view of their interplay is becoming ever more pressing. A standard approach considers clustering each one-dimensional omics data separately, then finding the consensus of these clusterings (2). Such methods result in loss of information and accumulation of noise in the database due to single omics treatment of data sets and not considering the impact of different layers on each other. A more elegant approach that we focused on in this work is integrating these multiomics in the very beginning followed by a single integrated clustering of the samples. Our integration method regards the gene expression of any transcript as a function of the gene copy number alteration and DNA methylation. Moreover, the gene expression can be modulated in any level by the other genes and their protein products in their interaction network. Our integrative measurements of genes in iWCluster, which is based on the interplays of these different layers in their interaction network, represent the functional concordance of these multiomics data.

Genomic instability can be caused by DNA copy number alteration and DNA methylation of CpG islands. In fact, amplification of oncogene and deletion of tumor suppressors give rise to malignant neoplasm (3). Furthermore, epigenetic aberration caused by DNA methylation has a central role in tumor progression (4). Our work in iWCluster is based on characterizing the genes that are upregulated due to hypomethylation and amplification, and the genes that are downregulated as the result of deletion and hypermethylation in the neighborhood of their interaction network. Combining the values of copy number alteration and DNA methylation with the gene expression in the network-based manner significantly improves the statistical accuracy in characterizing the genes that have an essential role in cancer progression. Our method of integration in iWCluster is akin to CNAmet in aggregating the copy number, methylation, and gene expression data (5). However, CNAmet does not consider the interaction network and does not include a clustering algorithm. We also include the continuous value of the multiomics data rather than the binary values $\{0, 1\}$ used in CNAmet to avoid the sensitive dependency on such thresholds.

In order to include the gene interaction network in iWCluster, we consider the weighted network defined by an underlying stochastic process. For each sample, we construct a distribution of integrative omics measures, which is closely related to the invariant (stationary) measures of the associated Markov

Author contributions: M.P., J.-H.O., J.M., Z.B., J.D. and A.T. designed the research; M.P. performed research; M.P. and J.-H.O. analyzed data; M.P. wrote the paper; J.-H.O., J.M., Z.B., J.D. and A.T. edited the paper; J.D. and A.T. directed research.

chain. We have employed these integrative measures to cluster TCGA breast cancer data. Our clustering method in iWCluster is based on optimal mass transport (OMT) methods, utilizing the 1-Wasserstein distance (6), also known as Earth Mover's Distance (EMD), applied to the invariant distributions of integrative measures computed between samples (7). Clustering via the Wasserstein distance in iWCluster is significantly more effective compared to clustering with Euclidean distance between samples. We applied several measurements to evaluate the clustering results of iWCluster, which we will discuss in more detail in the 'Materials and Methods' section. The invariant measure of the Markov chain considered in this work has an explicit closed formula which makes our method more efficient and interpretable than the spectral clustering methods used in other integrative methods such as similarity network fusion (SNF) (8). Also, even though the iterative model of SNF is network-based, the biological interaction network and the mechanistic interplay of multiomics layers has not been considered. The difficulty of the interpretability also occurs with the latent variables in matrix factorization methods such as iCluster (9). One of the advantages of iWCluster is its interpretability. Using the proposed iWCluster methodology, we are able to identify the significant genes in the specific cluster with a significantly low survival rate. We further perform gene ontology (GO) enrichment analysis on these genes utilizing a curated bioinformatics database (MetaCore) to discover the significantly correlated biological processes/pathways that could be related to the high mortality in this cluster.

The complexity of cancer etiology, the advent of large scale diverse genome-wide data, and the significant improvement in mathematical/statistical data analysis tools has resulted in considerable progress in the field of multiomics integration (10). In addition, recent efforts in network-based analysis of 'omics' allow identification of new disease genes, pathways and rational drug targets that were not easily detectable by isolated gene analysis (11). An effective bridge between the biology of cancer and mathematical techniques can bring about a comprehensive, meaningful and predictive model of multiomics integration with a medical target treatment plan for cancer patients. In this work, we strived to fulfill this goal and intertwine the mechanistic function of multiple regularity levels and the gene (protein) interaction network with advanced mathematical techniques for analyzing and clustering the data. The present study illustrates a novel mathematical way to integrate and cluster the multiple genomics data of TCGA breast cancer. Our clustering result is concordant with existing cancer subtypes of PAM50 in breast cancer, is effective in predicting survival rate, and facilitates the identification of the important driver genes in each sub-cluster, which in turn allow researchers to glean new therapeutic approaches from the database.

Materials and Methods

We applied methods from OMT (6, 12, 13) to measure the similarity of the integrative multiomics profiles between samples. To this end, we calculated the Wasserstein distance between the probability distributions of the integrative measures for the samples. The integrative measure is derived from the invariant measure of the stochastic matrix associated with the interaction network of genes. More precisely, we assigned the probability measures Π^1 and Π^2 to every two samples (Figure 1). We then utilized the Wasserstein distance to measure the similarity between every pair of samples. This similarity score measures the difference between the two distributions of the integrative measures assigned to every two samples and is applied in the hierarchical clustering of the samples. In the following section, we first discuss our method of integration in iWCluster, then we review the clustering of samples via Wasserstein distance on the gene interaction network.

Constructing Sample Specific Integrative Measures. Our integration method is based on characterizing those genes with high expression level, upregulated due to hypomethylation and amplification, and the genes that have low expression level, downregulated as the result of deletion and hypermethylation in a network-based manner. Moreover, gene expression levels are governed by the adjacent genes (proteins) in their interaction networks. Therefore, we consider the network effect in our integrative formula so that a specific gene integrative measure is not only dependent on the expression level, methylation and copy number of that gene, but also of its neighboring genes in the network.

We applied our method to integrate the multiomics data of TCGA breast carcinoma. The integrative measure is defined via the mRNA expression (z-score of RNA Seq V2 RSEM), copy number alteration (relative linear values from Affymetrix SNP6) and methylation (HM450) data of TCGA, and the interaction network from the Human Protein Reference Database (HPRD, http://www.hprd.org) (14). After removing some missing data, our breast cancer TCGA data consists of mRNA expression of 18,022 genes for 1,100 cases, copy number values of 15,213 genes for 1,080 cases and methylation of 15,585 genes in 741 cases. The intersection of all three data resulted in 7,737 genes and 726 samples. Moreover, this gene set has 3,426 genes in common with HPRD in the largest connected component. Subsequently, we have all the three multiomics of 3,426 genes with their connected interaction network for 726 samples. We further considered making the method computationally less expensive by limiting the genes with the ones only in common with OncoKB (Precision Oncology Knowledge Base) genes (http://oncokb.org/) which is discussed extensively in the 'Supplementary Information' (SI) of the present paper.

We constructed a weighted graph by considering the given gene interaction network as a Markov chain (7). Considering a gene i and its neighbor genes $j \in N(i)$ from a given sample, the principle of mass action allows us to compute the probability of the interaction of gene i to gene j to be proportional to their expression (15). Therefore, we can define the stochastic matrix of the Markov chain of the network as follows:

$$p_{ij} := \frac{ge_j}{\sum_{k \in N(i)} ge_k}, \qquad [1]$$

where ge stands for the gene expression and p_{ij} has been normalized such that $\Sigma_j p_{ij} = 1$. The Markov chain given by Eq. (1) defines the flow of information among genes that reaches a stationary distribution which is invariant under a right multiplication by p, i.e. $\pi p = \pi$, (7). Solving this formula for the special stochastic matrix p defined by Eq. (1), π has the explicit expression:

$$\pi_i = \frac{1}{Z}(ge_i \times \sum_{j \in N(i)} ge_j). \qquad [2]$$

Z is a normalization factor forcing π to be a probability vector. Of note, this normalization is necessary since we want to consider the invariant measure to be a probability distribution over all genes for each specific sample. Also, the invariant measure defined by Eq. (2), gives a value to each gene which is not only dependent on the gene expression of the gene i, but also on the total gene expressions of the neighboring genes $j \in N(i)$.

We have extended this invariant measure to the integrative measure, which not only considers gene expression, but also copy number alteration and methylation. Our approach roughly follows that of CNAmet (5), however, our method is network-based. We characterize the genes that are upregulated with high expression, amplification and hypomethylation, and also have many connections with similarly upregulated genes in their interaction network. In fact, for TCGA data, the Spearman's correlations between the gene expression and the copy number alteration across the samples are mostly significantly positive (see Figure S1 in the SI).

Similar positive correlations exist between gene expression and 1−methylation (one *minus* methylation). Here, we considered the values of 1−methylation (which are positive) since the high values of methylation have a reverse effect on the upregulation of genes. In CNAmet, the gene expression (mean across all samples) values have been multiplied by binary values assigned to copy number and methylation. In the proposed iWCluster method, we utilized the actual normalized values of copy number and methylation in the integrative measures for each sample.

More precisely, for the gene i, we define the vectors W_{ge}, W_{cn} and W_{me} in a manner similar to the invariant measure formula given in Eq. (2):

$$W_{\xi}^{i} = \xi_i \times \sum_{j \in N(i)} \xi_j, \qquad [3]$$

where ξ is any of the ge, cn, and me. Similar to CNAmet, we have a scaling factor ε, which accounts for the concordance of copy number alteration and methylation. Yet again, we consider the actual normalized value of the data in the network-based fashion (Figure 1). For the gene i, the scaling factor is defined as follows:

$$\varepsilon_i = \frac{1}{d+1} \sum_{j \in N(i) \cup \{i\}} (me_j) \times (cn_j). \qquad [4]$$

where d is the node degree of the gene i in the interaction network. We next add the normalization factor Z in order to complete the formulation of the integrative measure as a probability vector. Consequently, the network integrative measure is defined as follows:

$$\pi_i{}^* = \frac{1}{Z} W_{ge}^{i} (W_{me}^{i} + W_{cn}^{i}) \varepsilon_i. \qquad [5]$$

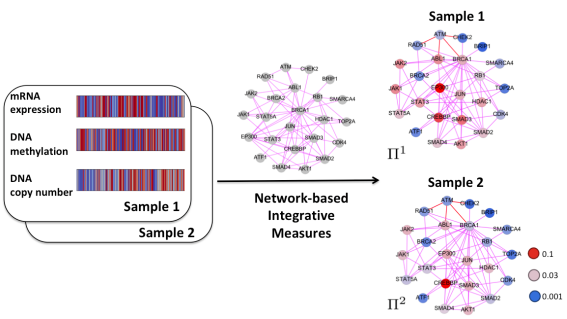


Fig. 1. The integration of multiomics in iWCluster is network-based. We assigned an integrative measure to the nodes of the interaction network using the network connectivity and the values of gene expression, copy number alteration and methylation. These integrative measures are exploited by iWCluster to find the distances between samples and to cluster the sample space of TCGA breast cancer data. Here, we only show a small network for the purpose of illustration.

For each sample, we calculated a vector $\Pi = (\pi_i{}^*)_{i=1,\ldots,n}$ for all the n genes in our data sets. A closed form of the formula can be presented using the adjacency matrix associated with the network as follows:

$$\Pi := \frac{1}{\mathbf{Z}} W(ge) \circ \Big(W(me) + W(cn)\Big) \circ \varepsilon \qquad [6]$$

where

$$W(\xi) := \xi \circ (\mathrm{Adj} \times \xi)$$

and

$$\varepsilon := \frac{1}{\mathbf{d}+\mathbf{1}} (\mathrm{Adj} + \mathbb{I}_n) \times (me \circ cn).$$

In these expressions, $\circ$ denotes the component-wise (Hadamard) product, whereas $\times$ is the standard matrix multiplication. Also, $\mathbf{d}$ is the vector of all node degrees and $\mathbf{1}$ is all-ones vector. Here, every row of the matrix $\mathrm{Adj} + \mathbb{I}_n$ is divided by the component of $\mathbf{d}+\mathbf{1}$. Given the fact that we used the factor $\mathbf{Z}$ to normalize the Π in Eq. (6), we can treat them as probability distributions defined along the genes for a given sample.

We then applied methods from OMT (6, 12) to find the distance between a pair of vectors in the form of π assigned to every two samples. In fact, we measured the similarity between samples by finding the Wasserstein distance between the distributions of the integrative measures which we define in more detail in the following section.

Clustering via Wasserstein distance. The iWCluster uses the theory of OMT to define distances among samples. OMT is a rapidly developing area of research that deals with the geometry of probability densities (6). The subject began with the work of Gaspard Monge in 1781 (16) who formulated the problem of finding minimal transportation cost to move a pile of soil ("deblais"), with mass density ρ^0, to an excavation ("remblais"), with a mass density ρ^1. A relaxed version of the problem was introduced by Leonid Kantorovich in 1942 (17). We will only give a special case of the general problem that we will need in the present work.

Let $\rho^0,\ \rho^1 \in P(\Omega)$ where $\Omega \subseteq \mathbb{R}^m$ and $P(\Omega) = \{\rho(x) : \int_\Omega \rho(x)dx = 1, \rho(x) \geq 0\}$. The *1-Wasserstein distance*, also known as the *Earth Mover's Distance (EMD)*, is defined as follows:

$$W_1(\rho^0, \rho^1) = \inf_{\pi \in \Pi(\rho^0, \rho^1)} \int_{\mathbb{R}^n \times \mathbb{R}^m} \|x - y\| \, d\pi(x, y), \qquad [7]$$

where $\Pi(\rho^0, \rho^1)$ denotes the set of all couplings between ρ^0 and ρ^1, that is the set of all joint probability measures π on $\Omega \times \Omega$ whose marginals are ρ^0 and ρ^1 (see Figure S2 in the SI). Here, the cost function of the transportation is defined as the ground distance $d(x, y) = \|x - y\|$.

An alternative and computationally more efficient formulation of the EMD (18) is defined by optimizing the flux vector $\mathbf{m} : \Omega \to \mathbb{R}^m$ in the following manner:

$$W_1(\rho^0, \rho^1) = \inf_{\mathbf{m}} \left\{ \int_{\mathbb{R}^n} \|\mathbf{m}(x)\| \, dx \mid \rho^0 - \rho^1 - \nabla_x \,.\, \mathbf{m} = 0 \right\}, \qquad [8]$$

where $\|.\|$ is the standard Euclidean distance based norm.

The optimization problem Eq. (7) has an analogous formulation on a weighted graph. Indeed, we consider a connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with n nodes in $\mathcal{V}$ and m edges in $\mathcal{E}$. Given two probability densities $\rho_0, \rho_1 \in \mathbb{R}^n$ on the graph, the EMD problem seeks a joint distribution $\rho \in \mathbb{R}^{n \times n}$ with marginals ρ^0 and ρ^1 minimizing the total cost $\sum c_{ij}\rho_{ij}$:

$$W_1(\rho^0, \rho^1) = \min_{\rho} \left\{ \sum_{i,j=1}^{n} c_{ij}\rho_{ij} \ \mid \ \sum_k \rho_{ik} = \rho_i^0, \ \sum_k \rho_{kj} = \rho_j^1, \ \forall i, j \right\}. \qquad [9]$$

Here c_{ij} is the cost of moving unit mass from node i to node j and is taken to be the minimum of the number of steps to go from i to j, namely, c is the ground metric on the graph. The minimum of this optimization problem defines a metric W_1 (the Earth Mover's Distance) on the space of probability densities on $\mathcal{G}$. Note that this optimization problem consists of n^2 variables.

Now we can give a graph-theoretic formulation of Eq. (8) as follows which gives an alternative way to compute the EMD on the graph $\mathcal{G}$:

$$W_1(\rho^0, \rho^1) = \min_{u} \left\{ \sum_{i=1}^{m} \|u_i\| \ \mid \ \rho^0 - \rho^1 - D\,u = 0 \right\}. \qquad [10]$$

On the graph $\mathcal{G}$, the *fluxes* u_i are defined on the m edges, and $D \in \mathbb{R}^{n \times m}$ denotes the incidence matrix of $\mathcal{G}$ with an assigned orientation. More precisely, the incidence matrix D is a matrix with rows and columns indexed by the vertices and edges of G such that every entry (i, k) is equal to 1 if the vertex i is assigned to be the head of the edge k and is equal to -1 if it is the tail of k. Very importantly, note that the optimization problem in Eq. (10) depends on m variables, while the primal node based version of OMT on a graph in Eq. (9) depends on $n \times n$ variables. Thus formula Eq. (10) is certainly much more efficient especially when the graph is sparse, that is, $m << n^2$.

In our computer implementations, we used the CVX optimization package (19) to solve the optimization problem in Eq. (10). We used the pairwise EMD of samples for clustering TCGA data of breast cancer. Each sample is represented as a vector of its pairwise distances to all other samples (its distance to itself is zero). We

then applied the hierarchical agglomerative clustering to the sample vectors. Hierarchical clustering imposes a hierarchical structure on the samples and their stepwise clusters. To achieve a certain number of clusters, the hierarchy is cutoff at the relevant depth. The optimal number of clusters can be determined with several techniques. In the present work, we validated the clustering via the following approach. First, we measured the homogeneity of clusters with silhouette mean values (20). The optimal number of clusters is based on the silhouette values. The silhouette value for each sample is a measure of how similar that sample is to other samples in its own cluster compared to samples in other clusters. The *silhouette value* for the sample i, $s(i)$ is defined as:

$$s(i) := (b(i) - a(i))/(\max(a(i), b(i)) \quad [11]$$

where $a(i)$ denotes the average distance of the sample i to all samples within its own cluster (squared Euclidean distance between sample vectors), and $b(i)$ denotes the minimum average distance of the sample i to samples of other clusters. Silhouette values range from -1 to +1. A high silhouette value indicates that the data are appropriately clustered. Therefore, we chose the optimal number of clusters by analyzing the average silhouette values of samples to make sure they stay close to 1.

For further evaluation of breast cancer clustering via iWCluster, we studied the difference of survival counts and survival rates among the clusters using the chi-squared test and log-rank test, respectively. We also compared our clustering results with the known PAM50 molecular subtypes of breast cancer (Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like), which were first described in 2000 by Perou *et al.* (21). This classification begins with the analysis of microarray expressions of 50 genes (known as the "PAM50" gene signatures) to cluster breast tumors into one of the subtypes.

In summary, iWCluster consists of three parts: integrating the multiomics, finding the Wasserstein distance among samples, and applying the hierarchical clustering to the samples space of the vectors of Wasserstein distances. In this work, we implemented iWCluster via MATLAB.

Results

We primarily applied iWCluster to TCGA breast cancer data consisting of 3,426 genes and 726 samples. The visualization of the interaction network of these 3,426 genes is included in Figure S3 of SI. The hierarchical clustering of 726 samples of TCGA breast cancer via iWCluster is shown in Figure 2(a). The survival counts in these clusters are given in the contingency Table 1. The p-value of the Pearson's chi-squared test for this table is 0.02, indicating the significant separation of survival counts in our four breast cancer clusters. Moreover, the Kaplan-Meier plot of the clusters' survival rate (truncated at 5 years) is illustrated in Figure 2(b). The log-rank test indicates a significantly different survival time between the four Kaplan-Meier curves (p-value=0.00187). We chose the number of clusters based on the silhouette values we discussed previously in the 'Materials and Methods' section. Here, choosing the 3 or 4 clusters had a very small effect on the silhouette values, however, clusters 1 and 2 (which are combined in the presence of three clusters) have very different survival rates. Therefore, we chose the number of clusters to be four.

We also compared our clustering with the well-known PAM50 molecular subtypes of breast cancer (Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like). As shown in Table 2, our clustering substantially recovers the major PAM50 subtypes (chi-squared test's p-value $\ll 10^{-4}$). We have the PAM50 subtype classification for 649 out of 726 samples in the database. Clusters 3 and 4 significantly distinguish Luminal A from the Basal-like subtype. Cluster 4 also includes most of the Luminal B subtype. Moreover, many of the Her2 enriched tumor types are in cluster 1 whereas the normal-like subtype is clustered together with Luminal A tumors in cluster 3. Despite this consistency, our set of 3,426 genes only includes 19 genes in common with PAM50. Notably, as we expected, the survival of cluster 3 which mainly consists of the Luminal A subtype is higher than cluster 4 which includes many of the Luminal B and Basal-like subtypes (p-value=0.014).

Figure 2(a) also illustrates the heatmap of integrative measures of 150 selected genes for 726 samples. The values of integrative measures are visually distinguishable between clusters which assures the accuracy of the clustering method. We utilized Analysis of Variance (ANOVA) to choose 150 top genes that have significantly different mean values of integrative measures across the four clusters. For better visualization, we further reordered these 150 genes based on the highest to lowest mean values in cluster 1. The highest values of integrative measures in cluster 1 (which has many HER2-enriched tumors) are by far ERBB2, GRB7, and PIK3C2A. These genes also have very significantly different mean values across the 4 clusters (after sorting based on ANOVA's p-values). This result is consistent with the known co-amplification of ERBB2 and GRB7 in HER2-enriched breast cancer (22). The complete list of these 150 genes has been provided in 'Additional data table S1' in the SI. We also provided the heatmap of gene expression, copy number alteration, and methylation in Figure 2(a). We still see the pattern in the values of each single omics. In fact, observing these patterns is expected, since the integrative measures are built of these single measures. However, these single omics patterns are not as clear as the one with the integrative values. It appears that the integrative procedure of iWCluster de-noises and smoothens the data and amplifies the contrast among extreme values of the genes in the single omics.

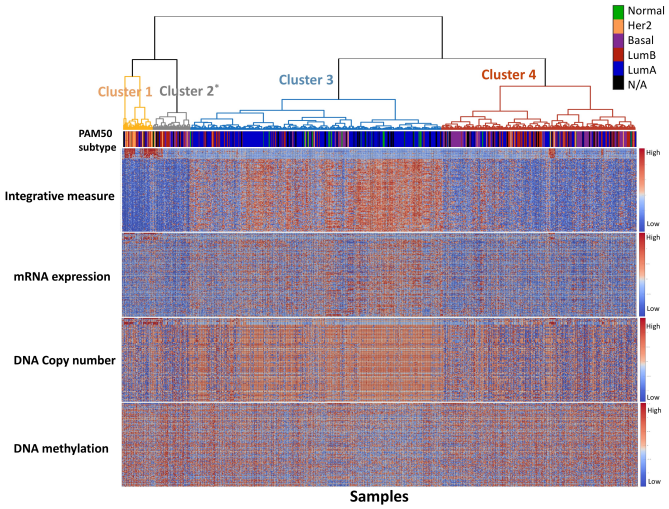
One of the advantages of iWCluster is its interpretability. The integrative measures of iWCluster are defined explicitly for all genes in the study, therefore, we can investigate genes that contribute the most in separating a specific cluster. Here, we are interested in the GO enrichment analysis of the significant genes in the clusters to see which biological processes/ pathways are related to the clusters with distinguished survival rates. As we see in Figure 2(b), cluster 2 has the lowest survival rate compared to all other clusters. This subgroup consists of 53 cases of mostly Luminal B and Basal-like subtypes. We identified the genes that have significantly different mean values in this cluster compared to the other three clusters using the t-test. We discovered that the significant genes genes in cluster 2 notably correspond with tumor hypoxia. Tumor hypoxia, a well-known phenomenon where tumor cells have been deprived of oxygen, is a prominent issue in tumor physiology and cancer treatment (23). Specifically, hypoxia appears to be strongly associated with tumor malignancy, resistance to treatment and the metastatic phenotype of cancer (24, 25). We chose 166 significant genes in cluster 2 based on the (t-test) Bonferroni corrected p-value less than 0.01 (Additional data table S2). The GO enrichment analysis of these genes was performed via MetaCore software (Thomson Reuters). MetaCore is an integrated software based on a manually-curated database of molecular interactions, molecular pathways, gene-disease associations, chemical metabolism, and toxicity information. The top ten biological processes presented with very small p-values ($\ll 10^{-6}$) are shown in Figure 3. These p-values correspond

Table 1. Survival counts in the four clusters of 726 samples with breast cancer using 3,426 genes. The p-value of the chi-squared test for this contingency table is 0.02. Even clusters 3 and 4 which includes most of the samples (631 of the 726 samples) have separated the samples with a significantly different survival count (chi-squared test's p-value = 0.014).

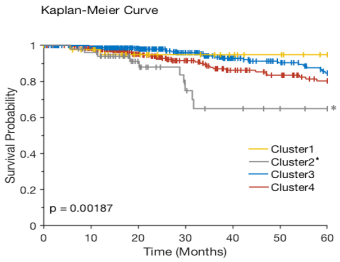
Survival	Cluster 1	Cluster 2*	Cluster 3	Cluster 4	Total
Living	39	42	319	228	628
Deceased	3	11	37	47	98
Total	42	53	356	275	726

Table 2. iWCluster is significantly concordant with the PAM50 subtypes. The chi-squared test's p-value $\ll 10^{-4}$ (even after excluding the normal subtype due to small counts). Cluster 3 includes most of the Luminal A subtype, whereas cluster 4 consists of the Basal-like and Luminal B subtypes. Moreover, many of the Her2 tumor types are in cluster 1.

PAM50	Cluster 1	Cluster 2*	Cluster 3	Cluster 4	Total
Lum A	7	8	240	83	338
Lum B	11	18	32	69	130
Her 2	17	4	5	14	40
Basal	4	16	17	76	113
Normal	0	1	21	6	28
Total	39	47	315	248	649



(a)



(b)

Fig. 2. (a) Hierarchical clustering of 726 samples using 3,426 genes via iWCluster. The first bar underneath the hierarchical clustering corresponds to the PAM50 subtypes. The black color represents the samples whose PAM50 subtypes are not available. Furthermore, the heatmaps represent the integrative measures, gene expression, copy number alteration, and methylation of 150 genes selected based on ANOVA. Here, these top 150 genes are sorted (descending order) for cluster 1. The difference in the integrative measures between clusters is visually detectable in the heatmap. Of note, the samples are ordered based on clustering via integrative measures and if we started with one of these single omics we would not be able to achieve this clustering. We still see the pattern in the values of each single omics especially for gene expression and copy number alteration, however, they are not as clear as the one with the integrative values. (b) Graphical display of survival rate using the Kaplan-Meier curves. Sample survival times (months) are plotted on the x-axis (truncated at 5 years), and the probability of survival calculated according to the Kaplan-Meier method is plotted on the y-axis. Cluster 2* has a significantly lower survival rate compared to the other three clusters.

to the hypergeometric test performed by MetaCore using the number of input genes, ontology related genes, and the total number of genes in the database. Here, the top biological processes (first and second and also fifth) strongly correspond to hypoxia (p-value $\ll 10^{-6}$) which is a central issue in tumor physiology and cancer treatment. The protein/gene interaction network of these 166 genes is also presented in Figure 1. As we see in this figure, the network is very much connected which indicates that many of these 166 genes are related to each other. The three hub nodes, ETS1, AP-1, and STAT3 located within the nucleus are highly connected to many other proteins in the network. We will discuss these proteins and the genes related to them in more detail in the discussion section. Of note, we similarly chose the significant genes in the other three clusters using the t-test's p-value, however, hypoxia did not appear as a significant top biological process in the other GO enrichment analysis.

We finally used iWCluster to cluster TCGA breast cancer data using the genes that are also included in the OncoKB database. This smaller network with 290 genes is computationally much more efficient. The clustering of the 726 samples using this gene set is significantly consistent with our previous clustering using the 3,426 genes (chi-squared test's p-value $\ll 10^{-4}$ for contingency Table S3) indicating the robustness of iWCluster's methodology. We included the details of this clustering along with more clinical outcome analysis in the SI.

In order to investigate the importance of integrating the multiomics in iWCluster, we repeated all the steps of iWCluster with only the gene expression data (similarly with only copy number alteration/ methylation). The invariant measures given in Eq. (2) only depend on the gene expressions. After excluding copy number alteration and methylation, the Kaplan-Meier curves of clusters do not have a significant survival rate difference where the log-rank's p-value is 0.09. Even though the difference in survival counts between two major clusters (4 and 5) is still significant (chi-squared test's p-value= 0.02), it is not as strongly separated as in the integrative method (p-value= 0.014). Considering the copy number alteration or methylation solely results in even worse clustering in terms of survival difference between clusters. Therefore, the integrative procedure of iWCluster is crucial in the clustering of TCGA data.

Discussion

The effective integration of large-scale cancer genome data can significantly improve clinical outcome predictions. Aggregat-

Table 3. Top 10 biological processes obtained from the gene ontology enrichment analysis of the significant genes for the subgroup with the worst survival outcome (via MetaCore). Top biological processes (first, second and also fifth) are strongly correlated with hypoxia.

	Biological Processes
1	Response to decreased oxygen levels
2	Response to hypoxia
3	Regulation of cellular biosynthetic process
4	Cellular macromolecule metabolic process
5	Response to oxygen levels
6	Regulation of macromolecule biosynthetic process
7	Regulation of biosynthetic process
8	Positive regulation of macromolecule biosynthetic process
9	Positive regulation of cellular biosynthetic process
10	Positive regulation of nucleobase-containing compound metabolic process

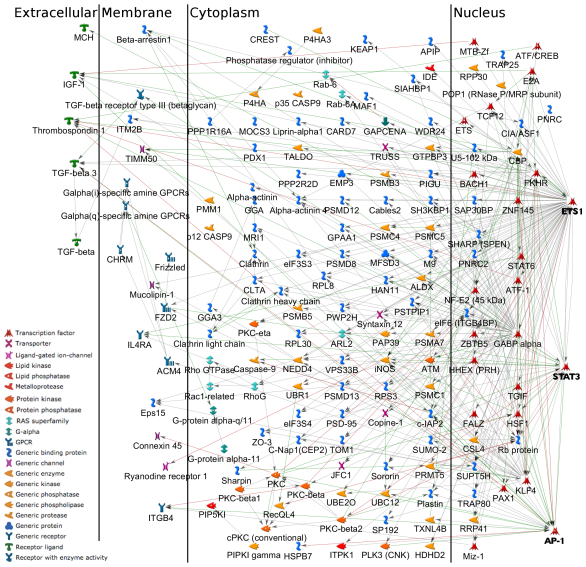


Fig. 3. The protein (gene) interaction network of the significant genes for the subgroup with the worst survival outcome (via MetaCore). The network includes the hub node ETS1 as well as AP-1 and STAT3 located within the nucleus.

ing multiomics data leads to detection of biomarkers that are perturbed in a certain subpopulation of cancer patients. In addition, cancer-related genes do not act individually, but within an interaction network, which must be explicitly incorporated into the integration procedure. Therefore, network-based integration of all the genomic, transcriptomic, epigenomic data increases the information content and accuracy of the results more than any of the single omics studies separately. Here, we propose the use of iWCluster for integrating the multiomics and clustering samples (patients). The network-based integration of iWCluster considers the mRNA expression, DNA methylation, and copy number alteration of the genes in its neighborhood of the interaction network. The concordance of the mutiomic values for a gene and connectivity to many similar genes results in a high integrative measure of the gene. iWCluster defines the similarity measured among samples by the Wasserstein distance between the samples' distributions of the integrative measures along the genes. Applying iWCluster to TCGA breast cancer data successfully clusters data with significantly different survival rates. Moreover, we were able to recover the well-known PAM50 subtypes. Our clustering results also reveal a subgroup of breast cancer patients with substantially poor survival outcome. We performed the GO enrichment analysis on the genes that have significantly different values in this cluster compared to others. The analysis discovered that this gene set is significantly related to the biological process of hypoxia. Also, as we see in Figure 1, the network of these genes is very densely connected, with a hub node ETS1 which is a transcription factor included in the list of 166 significant genes of the subgroup with the worst survival rate.

Tumor hypoxia, a well-known phenomenon where tumor cells have been deprived of oxygen, is a prominent issue in tumor physiology and cancer treatment (24). Notably, tumor hypoxia is strongly associated with tumor propagation, malignant progression, and tumor angiogenesis, with cells in hypoxic regions of a tumor more resistant to the effects of radiation and chemotherapy (26). In hypoxic environments, cellular adaptive processes are provoked which are mediated by the elevated expression of hypoxia-inducible genes thereby enabling tumor cells to adapt and survive under harsh conditions (27). The hypoxia-inducible factor (HIF-1) is the principal transcription factor related to hypoxia, and it has been demonstrated that HIF-1 activity is increased in various tumors relative to that found within normal tissues (28, 29). Furthermore, in the setting of hypoxia, HIF-1 allows for adaptations of cancer cells that are essential for establishment of a primary tumor and its progression (28). Along with HIF-1, members of the v-ets erythroblastosis virus E26 oncogene homolog (ETS) family transcription factors, most prominently the proto-oncogene ETS1, participate in the upregulation of hypoxia-inducible genes (27). Oikawa *et al.* (23) showed for the first time that ETS1 is induced in the setting of hypoxia via the transcriptional activity of HIF-1. Of note, increased expression of ETS1 is seen in a variety of solid cancers including lung, colorectal, sarcoma, and squamous cell carcinomas, and higher levels have correlated with a higher incidence of lymph node metastasis and overall worse prognosis (30). This gene is also involved in tumor progression in breast cancer, where in the setting of hypoxia, increased expression by mammary epithelial cells contributes to aggressive tumor phenotypes by activating the transcription of genes involved in angiogenesis, extracellular matrix remodeling, cell adhesion, and invasion (23, 31). Moreover, increased expression of ETS1 is associated with increased risk of recurrence and worse prognosis in human breast cancers (32).

In addition, Activator Protein 1 (AP-1), another hub node in our network of significant genes (Figure 1) has also been identified as a hypoxia-inducible transcription factor (33). C-jun, a proto-oncogene, encodes a major component of AP-1 transcription factors, which are key regulators of immediate-early signals directing cellular proliferation, differentiation, survival, and environmental stress responses (34). AP-1 appears to be involved in the modulation of the apoptotic pathway and also plays a protective role in cellular response to DNA damage (35, 36). Piret *et al.* (37) demonstrated an AP-1 mediated protective role of hypoxia against cell death induced by the chemotherapeutic agent etoposide. Similarly, in the setting of

hypoxia, an anti-apoptotic role of AP-1 was seen in paclitaxel exposed breast cancer cells (38).

Furthermore, STAT3 is another hub node in our network and is included in the set of 166 significant genes related to the subgroup with the worst survival. Hypoxia can induce the activation of transcription 3 (STAT3) protein, with the hypoxia-induced biochemical alterations likely contributing to drug resistance under hypoxic conditions (39). Notably, hypoxia-induced STAT3 accelerates the accumulation of HIF-1 protein and has been shown to prolong its half-life in solid tumor cells (40). Moreover, in a triple-negative breast cancer cell line, STAT3 has been shown to play a key role in hypoxia-induced chemoresistance to the chemotherapeutic agent cisplatin (41).

Employing iWCluster, we combined the mRNA expression, DNA methylation and DNA copy number of alteration in a network-based fashion. Our method considers the interactive relationship among different omics layers and identifies the genes that have concordant values of these multiomics data in the neighborhood of its interaction network. We used the theory of OMT to find the clusters of samples with concordant genes' integrative measures. One important advantage of our approach is preserving the integrative measure for all the genes, thereby enabling us to identify significant genes in each cluster. In the era when cancer multiomics/ biological data is increasing exponentially every year, the need for such methods of integration is essential. Adding other layers of information such as mutation, single-nucleotide polymorphism (SNP), and microRNA data can depict the complexity of the real picture even further (42).

One of the limitations of this work is considering an undirected form of the interaction network. In addition to the direction, we previously studied the type of control (e. g. activator and repressor in transcription networks) in biological networks (43). For future work, we are interested to include the effect of both the network direction and regulation control type in our integrative measures. We would also like to consider more general versions of OMT (44) to directly work with the vector-valued data. Finally, we are also interested to integrate omics data with other data types such as imaging and drug response data.

We believe that the integration of multiomics/ biological data paves the way for precision medicine in treating sophisticated diseases such as cancer. To this end, we proposed a novel integrating method that considers the interactive relationship among different omics layers, and accurately clusters breast cancer samples with significantly different survival rates.

ACKNOWLEDGMENTS. This study was supported by AFOSR grant (FA9550-17-1-0435), a grant from National Institutes of Health (R01-AG048769), MSK Cancer Center Support Grant/Core Grant (P30 CA008748), and a grant from Breast Cancer Research Foundation (BCRF-17-193).

1. Network CGAR, et al. (2013) The cancer genome atlas pan-cancer analysis project. *Nature genetics* 45(10):1113–1120.
2. Lock EF, Dunson DB (2013) Bayesian consensus clustering. *Bioinformatics* 29(20):2610–2616.
3. Tang YC, Amon A (2013) Gene copy-number alterations: a cost-benefit analysis. *Cell* 152(3):394–405.
4. Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics* 3:415 EP.
5. Louhimo R, Hautaniemi S (2011) Cnamet: an r package for integrating copy number, methylation and expression data. *Bioinformatics* 27(6):887–888.
6. Villani C (2009) *Optimal Transport: Old and New*, Grundlehren der mathematischen Wissenschaften. (Springer-Verlag Berlin Heidelberg) Vol. 338.
7. Chen Y, et al. (2017) Pediatric sarcoma data forms a unique cluster measured via the earth mover's distance. *Scientific Reports* 7(1):7035.
8. Wang B, et al. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* 11:333 EP.
9. Shen R, Olshen AB, Ladanyi M (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25(22):2906–2912.
10. Kristensen VN, et al. (2014) Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14(5):299–313.
11. Pouryahya M, Oh JH, Mathews JC, Deasy JO, Tannenbaum AR (2018) Characterizing cancer drug response and biological correlates: A geometric network approach. *Scientific Reports* 8(1):6402.
12. Rachev ST, Rüschendorf L (1998) *Mass Transportation Problems: Volume I: Theory*. (Springer Science & Business Media).
13. Evans L (1999) Partial differential equations and monge–kantorovich mass transfer. *Current Developments in Mathematics* pp. 65–126.
14. Keshava Prasad TS, et al. (2009) Human protein reference database–2009 update. *Nucleic acids research* 37(Database issue):D767–D772.
15. Teschendorff AE, Banerji CRS, Severini S, Kuehn R, Sollich P (2015) Increased signaling entropy in cancer requires the scale-free property of protein interaction networks. *Scientific Reports* 5:9646 EP.
16. Monge G (1781) *Mémoire sur la théorie des déblais et des remblais*. (De l'Imprimerie Royale).
17. Kantorovich LV (1942) On the transfer of masses in *Dokl. Akad. Nauk. SSSR*. Vol. 37, pp. 227–229.
18. Evans LC, Gangbo W (1999) *Differential Equations Methods for the Monge-Kantorovich Mass Transfer Problem*. (American Mathematical Soc.) Vol. 653.
19. Boyd S, Vandenberghe L (2004) *Convex Optimization*. (Cambridge University Press).
20. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20:53–65.
21. Perou CM, et al. (2000) Molecular portraits of human breast tumours. *Nature* 406:747 EP –.
22. Pradip D, Bouzyk M, Dey N, Leyland-Jones B (2013) Dissecting grb7-mediated signals for proliferation and migration in her2 overexpressing breast tumor cells: Gtp-ase rules. *American Journal of Cancer Research* 3(2):173–195.
23. Oikawa M, et al. (2001) Hypoxia induces transcription factor ets-1 via the activity of hypoxia-inducible factor-1. *Biochemical and Biophysical Research Communications* 289(1):39–43.
24. Höckel M, Vaupel P (2001) Tumor hypoxia: Definitions and current clinical, biologic, and molecular aspects. *JNCI: Journal of the National Cancer Institute* 93(4):266–276.
25. Brown JM (2007) *Tumor Hypoxia in Cancer Therapy*. (Academic Press) Vol. 435, pp. 295–321.
26. Vaupel P, Mayer A (2005) Hypoxia and anemia: effects on tumor biology and treatment resistance. *Transfus. Clin. Biol.* 12(1):5–10.
27. Salnikow K, et al. (2008) Regulation of hypoxia-inducible genes by ETS1 transcription factor. *Carcinogenesis* 29(8):1493–1499.
28. Bos R, van Diest PJ, van der Wall E (2001) RESPONSE: Re: Levels of Hypoxia-Inducible factor-1 during breast carcinogenesis. *JNCI Journal of the National Cancer Institute* 93(15):1177–1177.
29. Dachs GU, et al. (1997) Targeting gene expression to hypoxic tumor cells. *Nat. Med.* 3(5):515–520.
30. Dittmer J (2003) The biology of the ets1 proto-oncogene. *Mol. Cancer* 2:29.
31. Behrens P, Rothe M, Wellmann A, Krischler J, Wernert N (2001) The ets-1 transcription factor is up-regulated together with MMP 1 and MMP 9 in the stroma of pre-invasive breast cancer. *J. Pathol.* 194(1):43–50.
32. Span PN, et al. (2002) Expression of the transcription factor ets-1 is an independent prognostic marker for relapse-free survival in breast cancer. *Oncogene* 21(55):8506–8509.
33. Cummins EP, Taylor CT (2005) Hypoxia-responsive transcription factors. *Pflugers Arch.* 450(6):363–371.
34. Laderoute KR, et al. (2002) The response of c-Jun/AP-1 to chronic hypoxia is Hypoxia-Inducible factor 1α dependent. *Mol. Cell. Biol.* 22(8):2515–2523.
35. Park S, Hahm ER, Lee DK, Yang CH (2004) Inhibition of AP-1 transcription activator induces myc-dependent apoptosis in HL60 cells. *J. Cell. Biochem.* 91(5):973–986.
36. Potapova O, Basu S, Mercola D, Holbrook NJ (2001) Protective role for c-jun in the cellular response to DNA damage. *J. Biol. Chem.* 276(30):28546–28553.
37. Piret JP, Cosse JP, Ninane N, Raes M, Michiels C (2006) Hypoxia protects HepG2 cells against etoposide-induced apoptosis via a HIF-1-independent pathway. *Exp. Cell Res.* 312(15):2908–2920.
38. Flamant L, Notte A, Ninane N, Raes M, Michiels C (2010) Anti-apoptotic role of HIF-1 and AP-1 in paclitaxel exposed breast cancer cells under hypoxia. *Mol. Cancer* 9:191.
39. Soleymani Abyaneh H, et al. (2017) STAT3 but not HIF-1α is important in mediating Hypoxia-Induced chemoresistance in MDA-MB-231, a triple negative breast cancer cell line. *Cancers* 9(10).
40. Jung JE, et al. (2008) STAT3 inhibits the degradation of HIF-1alpha by pVHL-mediated ubiquitination. *Exp. Mol. Med.* 40(5):479–485.
41. Soleymani Abyaneh H, et al. (2018) Hypoxia induces the acquisition of cancer stem-like phenotype via upregulation and activation of signal transducer and activator of transcription-3 (STAT3) in MDA-MB-231, a triple negative breast cancer cell line. *Cancer Microenviron.* 11(2-3):141–152.
42. Huang S, Chaudhary K, Garmire LX (2017) More is better: Recent progress in Multi-Omics data integration methods. *Front. Genet.* 8:84.
43. Pouryahya M, Mathews J, Tannenbaum A (2017) Comparing three notions of discrete ricci curvature on biological networks.
44. Chen Y, Georgiou T, Tannenbaum A (2108) Vector-valued optimal mass transport. *SIAM Journal Applied Math* 78(3):1682–1696.