PLOS ONE

RESEARCH ARTICLE

vWCluster: Vector-valued optimal transport for network based clustering using multi-omics data in breast cancer

Jiening Zhu[1☯], **Jung Hun Oh**[2☯], **Joseph O. Deasy**[2], **Allen R. Tannenbaum**[1,3*]

1 Department of Applied Mathematics & Statistics, Stony Brook University, New York, NY, United States of America, **2** Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, United States of America, **3** Departments of Computer Science, Stony Brook University, New York, NY, United States of America

☯ These authors contributed equally to this work.
* allen.tannenbaum@stonybrook.edu

Citation: Zhu J, Oh JH, Deasy JO, Tannenbaum AR (2022) vWCluster: Vector-valued optimal transport for network based clustering using multi-omics data in breast cancer. PLoS ONE 17(3): e0265150. https://doi.org/10.1371/journal.pone.0265150

Editor: Serdar Bozdag, University of North Texas, UNITED STATES

Received: November 2, 2021

Accepted: February 23, 2022

Published: March 14, 2022

Data Availability Statement: All data is publicly available. The multi-omics data used in this study can be downloaded from the cBioPortal database (https://www.cbioportal.org/). The CIBERSORT data are available at https://gdc.cancer.gov/about-data/publications/panimmune for the TCGA study and https://github.com/cclab-brca/Patterns-Immune-Infiltration for the METABRIC study.

Funding: A.R.T. is funded by AFOSR grants FA9550-17-1-0435, FA9550-20-1-0029, NIH grant R01AT011419. J.O.D. is funded by NIH grant R21-CA234752. J.O.D. and A.R.T. are funded by Breast

Abstract

In this paper, we present a network-based clustering method, called ***vector Wasserstein clustering*** (vWCluster), based on the vector-valued Wasserstein distance derived from optimal mass transport (OMT) theory. This approach allows for the natural integration of multi-layer representations of data in a given network from which one derives clusters via a hierarchical clustering approach. In this study, we applied the methodology to multi-omics data from the two largest breast cancer studies. The resultant clusters showed significantly different survival rates in Kaplan-Meier analysis in both datasets. CIBERSORT scores were compared among the identified clusters. Out of the 22 CIBERSORT immune cell types, 9 were commonly significantly different in both datasets, suggesting the difference of tumor immune microenvironment in the clusters. vWCluster can aggregate multi-omics data represented as a vectorial form in a network with multiple layers, taking into account the concordant effect of heterogeneous data, and further identify subgroups of tumors in terms of mortality.

Introduction

Current large-scale cancer genome projects, such as The Cancer Genome Atlas (TCGA), provide a comprehensive molecular portrait of human cancers, including gene expression, copy number variation (CNV), and DNA methylation profiles. These offer unprecedented opportunities for exploring cancer biology that is characterized through various molecular functions and their complex interactions. Several computational methods for multi-omics data integration and further clustering have been proposed to identify tumor subgroups associated with distinct clinical outcomes, leveraging complementary information of multi-omics data [1]. iCluster uses a joint latent variable method across multi-omics types to model integrative clustering [2]. Recently, Alkhateeb *et al.* proposed a deep learning method to predict the 5-year interval survival of breast cancer based on multi-omics data integration [3]. Network based

Cancer Research Foundation Grant BCRF-17-193. J.O. and J.O.D. are funded by NIH/NCI Cancer Center Support grant P30 CA008748. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

clustering methods using integrated multi-omics data have been proposed. Similarity network function (SNF) is a technique for combining multiple networks of each omics type into a single network, which is followed by a spectral clustering method to identify subtypes of tumors [4]. On the other hand, aWCluster utilizes a prior known network of gene products to integrate multi-omics data [5]. First, the data integration method yields an invariant measure for each node (gene). After repeating this process for each omics type, the invariant measures are integrated at each node. The Wasserstein distance, derived from optimal mass transport (OMT) theory [6, 7], is then computed between all pairs of samples on the network using the integrative invariant measure. The distance matrix is then input into a hierarchical clustering algorithm, resulting in clusters.

Representing data, e.g. as latent variables or weighted graphs, is essential to efficiently integrate multi-omics data while minimizing information loss. In this study, we propose a new method, called ***vector Wasserstein clustering*** (vWCluster), in which we employ a vector-valued version of the Wasserstein distance [8]. First, multi-omics data are represented as a multi-layer biological network, forming a layer for each single omics type. The Wasserstein distance is then computed on the vector-valued data in the network between all pairs of samples. The resulting distance matrix is then input into a hierarchical clustering method to identify subtypes of tumors. This method that represents multi-omics data vectorially on a network appears to be more straightforward to handle heterogeneous data compared to previously proposed methods while minimizing information loss.

The Wasserstein distance from OMT has increasingly received attention in data analysis due to its attractive property of (weak) continuity [6, 7]. In the present work, we will only use the $\mathcal{W}_1$ version, also known as the *Earth Mover's distance* (EMD). Other metrics commonly used on distributions, such as Kullback Leibler, Jensen-Shannon, or total variation, do not have the property [9, 10], which makes the metrics much more susceptible to the noise that is typically observed in medical data. Moreover, the Wasserstein distance is a metric for distributions defined on a metric space, which is essential for us to include the information from the weighted graphs used in this paper. Due to its attractive properties, OMT is becoming more and more widely used in signal processing, machine learning, computer vision, meteorology, statistical physics, quantum mechanics, and network theory [9, 11–15]. To even strengthen its power, several works deal with various extensions of the theory; see [11, 13, 16–18] and the references therein. In the present work, we empoly vector-valued extension of the Wasserstein distance [8, 19].

To the best of our knowledge, we are the first to use a vector-valued OMT methodology for the multi-omics data integration. We propose a general pipeline to analyze heterogeneous data in a multi-layer structure, employing a known protein-protein interaction network, and then cluster samples based on the resulting Wasserstein distance matrix. In the present work, our method is applied to multi-omics data from the two largest breast cancer studies: the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and TCGA studies [20, 21]. In the following section, we describe the proposed method and data in detail.

Background and methods

We developed a vector-valued OMT approach that integrates multi-omics data represented in a multi-layer network, on which we applied the $\mathcal{W}_1$ Wasserstein distance (EMD). Accordingly, we will only outline the OMT theory in this special case. In the following, we first describe the basic concept of Wasserstein distance and then introduce the proposed method.

Scalar-valued optimal transport

The $\mathcal{W}_1$ Wasserstein distance (EMD) was first formulated by the French civil engineer and mathematician Gaspard Monge in 1781 [6, 7, 22, 23]. Originally, this subject was inspired by the problem of finding the optimal plan, relative to a given cost, for moving a pile of soil from a given location to another in a mass preserving manner. The original Monge's formulation of OMT (in which the cost function is defined by the distance) may be given a modern expression as follows [6, 7]:

$$\mathcal{W}_M(\rho_0, \rho_1) = \inf_T \left\{ \int_S ||x - T(x)||\rho_0(x)dx \mid T_\# \rho_0 = \rho_1 \right\}, \tag{1}$$

where S denotes a subdomain of $\mathbb{R}^n$, T is the transport map, and ρ_0, ρ_1 are two marginals. Here $T_\#$ denotes the push-forward of T. Therefore, the $\mathcal{W}_1$ Wasserstein distance is the optimal cost with respect to the norm among all possible T.

As pioneered by Leonid Kantorovich [24], the Monge formulation of OMT may be relaxed by replacing transport maps T by couplings π:

$$\mathcal{W}_K(\rho_0, \rho_1) = \inf_{\pi \in \Pi(\rho_0, \rho_1)} \int_S ||x - y||\pi(dx, dy), \tag{2}$$

where $\Pi(\rho_0, \rho_1)$ denotes the set of all the couplings between ρ_0 and ρ_1 (joint distributions whose two marginal distributions are ρ_0 and ρ_1). Despite the relaxation, one may show that Kantorovich and Monge formulations are equivalent in a number of cases under certain continuity constraints; see [6, 7] and the references therein.

One of the benefits of Eq (2) is that it amounts to a linear programming problem. Via duality theory, an equivalent form may be expressed as follows (see [22] for the proof):

$$\mathcal{W}_{\tilde{K}}(\rho_0, \rho_1) = \inf_u \int_S ||u(x)||dx \tag{3a}$$

$$\mathrm{div}_x u(x) = \rho_0(x) - \rho_1(x), \tag{3b}$$

where $u = (u_1, u_2, \ldots, u_n) : S \to \mathbb{R}^n$ is the flux, and div_x denotes the divergence operator.

It is straightforward to extend Eq (3) to the discrete case by simply replacing the integral by an appropriate summation and replacing div_x by the discrete divergence operator:

$$\mathcal{W}_{\mathcal{G}}(\rho_0, \rho_1) = \min_u \sum_{i=1}^{|E_{\mathcal{G}}|} |u_i| \tag{4a}$$

$$\rho_0 - \rho_1 - Du = 0. \tag{4b}$$

On the graph $\mathcal{G} = (V_{\mathcal{G}}, E_{\mathcal{G}})$, the fluxes u_i now are defined on the edges $E_{\mathcal{G}}$, and $D \in \mathbb{R}^{|V_{\mathcal{G}}| \times |E_{\mathcal{G}}|}$ denotes the incidence matrix of $\mathcal{G}$ with directionality, namely we need to specify the directions of the fluxes. Thus, in the matrix D, each column has two nonzero entries, where one is 1 whose row number is the starting point of an edge while the other nonzero entry is -1 whose row number is the ending point of that edge.

Vector-valued optimal transport

A vector-valued density $\vec{\rho} = [\rho^1(x), \rho^2(x), \cdots, \rho^m(x)]^T$ on a given space S (continuous or discrete) may represent a physical entity that can mutate or transition between alternative manifestations, e.g., power reflected off a surface at different frequencies or polarizations. More

formally, in the continuous setting, an m-layer *vector-valued density* $\vec{\rho}$ on $S \subseteq \mathbb{R}^n$ is a map from S to $\mathbb{R}_+^m$ whose total mass is defined as $\sum_{i=1}^m \int_S \rho^i(x)dx$. As a distribution, we require its total mass to be 1. Note that the integral over S is just in general. If the space S is a discrete space, then the integral is replaced by summation.

Vector-valued optimal transport studies such distributions, which is of great theoretical and practical interest since it does not simply consider each layer separately, but explicitly models the relationships among layers [19]. A relationship is expressed as an additional graph structure that connects each layer. Specifically, each component of ρ is represented by a node of a graph $\mathcal{F} = (V_\mathcal{F}, E_\mathcal{F})$ and an edge between two nodes allows for direct transport between the corresponding layers. So $|V_\mathcal{F}| = m$, which represents the cardinality of all the channels (layers), and $E_\mathcal{F}$ is the set of all the direct connections between the layers.

Thus, the vector-valued optimal transport problem may be written as follows:

$$\mathcal{W}_V(\vec{\rho}_0, \vec{\rho}_1) = \inf_{\vec{u},\vec{w}} \int_S ||\vec{u}(x)|| + \gamma||\vec{w}(x)||dx \tag{5a}$$

$$\mathrm{div}_x\vec{u}(x) + \mathrm{div}_\mathcal{F}\vec{w}(x) = \vec{\rho}_0(x) - \vec{\rho}_1(x), \tag{5b}$$

where $\vec{u}, \vec{w}$ are both vector-valued, div_x is the spatial divergence which is taken componentwise for each layer, and $\mathrm{div}_\mathcal{F}$ is the discrete divergence on the graph $\mathcal{F}$ which takes the flows between channels into account. Here $\gamma \geq 0$ is a parameter to control flow between channels.

As in the scalar-valued case, we can extend the definition for distributions to a discrete graph $\mathcal{G}$. The vector-valued formulation on a graph is then the following:

$$\mathcal{W}_{\tilde{V}}(\vec{\rho}_0, \vec{\rho}_1) = \min_{u,w} \sum_{i=1}^{|E_\mathcal{G}|}\sum_{j=1}^{|V_\mathcal{F}|}|u_i^j| + \gamma\sum_{i=1}^{|V_\mathcal{G}|}\sum_{j=1}^{|E_\mathcal{F}|}|w_i^j| \tag{6a}$$

$$\vec{\rho}_0 - \vec{\rho}_1 - D_1u - D_2w = 0, \tag{6b}$$

where u is the flux within each layer, w is the flux across layers, and D_1 and D_2 are two matrices of the discrete divergence operators for two graphs.

On the one hand, this is a generalized form of Eq (4) derived by replacing each original node in the graph $\mathcal{G}$ by another graph. On the other hand, this formulation may be understood as a distribution on a super-graph $\mathcal{G} \times \mathcal{F}$. This super-graph is an irregular grid version of the Kronecker product. A slight difference from directly computing OMT distance on such a super-graph is that vector OMT on a graph here gives two different weights for the two different sets of edges. It is weighted vector-valued OMT. We later will see that two different kinds of fluxes via two graphs have different meanings.

Multi-omics data from two breast cancer studies

Multi-omics data for METABRIC and TCGA breast cancer studies were downloaded from the cBioPortal database [25, 26]. The METABRIC dataset contains microarray gene expression of 24,368 genes from 1,904 samples and copy number variation (CNV) of 22,544 genes from 2,173 samples. The intersection of the two omics data resulted in 16,195 genes from 1,904 samples. The TCGA breast cancer dataset consists of RAN-Seq gene expression of 18,022 genes from 1,100 samples, CNV of 15,213 genes from 1,080 samples, and methylation of 15,585 genes from 741 samples. The intersection of the three omics data resulted in 7,737 genes from 726 samples.

vWCluster requires all the nodal values in the network to be positive, because of the Markov chain process (see Section Markov chain and stationary distribution). The only data preprocessing was to exponentiate CNV values to ensure their positive.

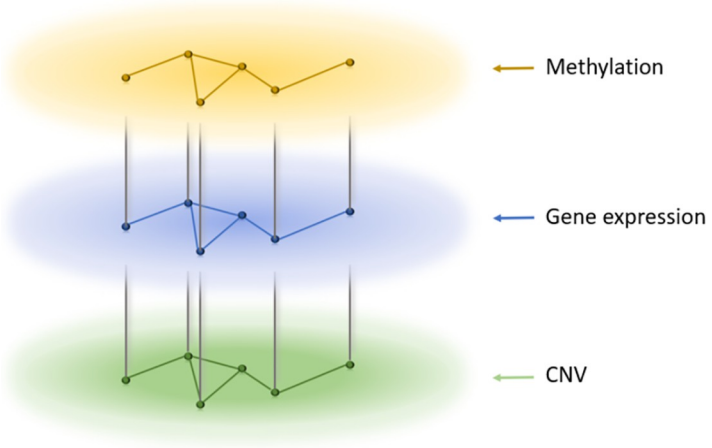

Fig 1. Graph structure for the TCGA breast cancer data.

https://doi.org/10.1371/journal.pone.0265150.g001

Graph structures for analysis

We represented multi-omics data as vector-valued distributions on the gene (product) interaction network. The interaction network was derived from the Human Protein Reference Database (HPRD) [27]. The largest connected network component was found in the interaction of the HPRD and the gene list of METABRIC or TCGA breast cancer data, separately, resulting in 3,147 and 3,426 genes, respectively. As multi-omics data in the TCGA breast cancer cohort, gene expression, CNV, and methylation data were used, whereas in the METABRIC cohort, only gene expression and CNV data were available, thereby forming 3-vector and 2-vector distributions, respectively.

More specifically, the network for METABRIC consisted of two layers (gene expression and CNV), each of which had the same topology (the largest connected network component) derived from HPRD. The connection between the two layers was formed by connecting the two nodes for the same gene in each layer, yielding the graph $\mathcal{F}$ structure.

For the network with the TCGA data, the layer for gene expression was connected with both layers for CNV and methylation since CNV and methylation may affect the level of gene expression. There was no connection between the CNV and methylation layers. See Fig 1.

Markov chain and stationary distribution

One problem of applying the vector-valued optimal transport method to multi-omics data is that the scale of individual omics data varies. For example, CNV data consists of integer values, while gene expression and methylation data have continuous values. To tackle this issue, we use the invariant (stationary) distribution derived from a Markov process of the gene network.

A *Markov process* is a stochastic process such that the probability of a given event depends only on the state of the previous event. To put it simpler in our graph setting, one starts with a certain distribution. At each time step, the probability at each node redistributes to all its neighbors with predefined weights. In the gene network setting [28], we set the probability of moving from a node i to its neighbor j to be:

$$p_{ij} = \frac{g_j}{\sum_{k \in N(i)} g_k}, \tag{7}$$

where $g_k > 0$ is the weight of node k, which can be any omics type (gene expression, CNV or methylation). Note that for methylation, 1-methylation values were used since methylation is likely to be negatively correlated with gene expression.

The matrix p is a stochastic matrix, i.e., the state probability matrix from the current time step to the next, as follows:

$$\pi^{t+1} = \pi^t p, \tag{8}$$

where π^t is the distribution at time step t. In our setting, after a finite number of time steps, the initial distribution will converge to a stationary (invariant) distribution π such that

$$\pi = \pi p. \tag{9}$$

The stationary distribution has a closed form solution:

$$\pi_i = \frac{1}{Z} g_i \sum_{k \in N(i)} g_k, \tag{10}$$

where Z is the normalization factor to be a probability distribution.

This Markov process on the gene network mimics the interactions among genes and the stationary distribution gives a distribution that represents the information each gene has which includes not only its own value but the interactions with its neighbors. The Markov process was performed for each sample in individual omics types, separately, yielding invariant measures I_{ijk} for sample i, omics type j, and gene k.

Clustering based on the vector-valued Wasserstein distance

With the graph structure determined, the vector-valued Wasserstein distance was computed for each pair of samples, using the invariant measures of each omics type. Note that the network for METABRIC or TCGA breast cancer data consisted of 2 and 3 layers, respectively. That is, we fitted the multi-omics data into the vector-valued optimal transport model. The resulting distance matrix was then input to standard hierarchical clustering to identify clusters of tumors. Kaplan-Meier survival analysis with log-rank test was performed to assess the difference of 5-year survival rates among the clusters identified. Further, CIBERSORT scores were compared among the clusters to investigate the difference in immune cell types [29, 30]. This analysis was performed for METABRIC or TCGA breast cancer data, separately. vWCluster was implemented in MATLAB and the code is available on https://github.com/MSK-MOI/vWCluster.

Results

METABRIC data analysis

The vector-valued Wasserstein distance was computed on gene expression and CNV data for METABRIC data. As described above, the resulting distance matrix was input to standard hierarchical clustering. The clustering results are shown in Fig 2.

Based on the dendrogram and the number of intrinsic molecular subtypes in breast cancer, four clusters were chosen for further analysis. Kaplan-Meier analysis with log-rank test (without NA samples) resulted in a statistically significant survival difference among clusters with a log-rank $p < 0.0001$ (Fig 3).

The clustering results were compared with PAM50 and Claudin-low subtypes [31], and the associations were assessed using a chi-squared test, resulting in $p < 0.0001$ as shown in

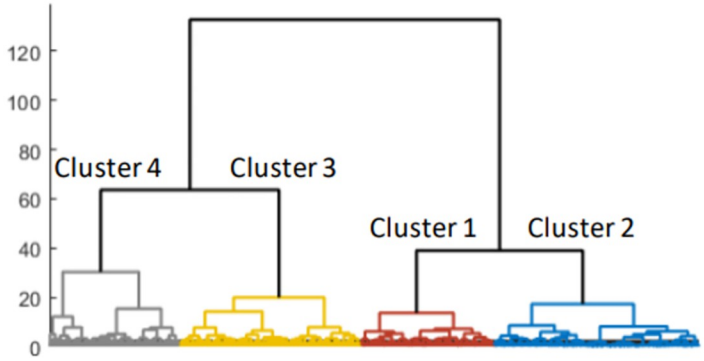

Fig 2. Clustering results employing the resultant vector-valued Wasserstein distance on METABRIC data.

https://doi.org/10.1371/journal.pone.0265150.g002

Table 1. Clusters 1 and 2 were enriched for Luminal A subtype. Cluster 3 was enriched for Luminal A and B subtypes, and cluster 4 was more enriched for basal subtype.

For the four clusters identified, 22 CIBERSORT immune cell types were compared using one-way analysis of variance (ANOVA) test. The twelve immune cell types were statistically significantly different among the four clusters. The top two significant immune cell types were M0 and M1 macrophages, for which Cluster 4 had the highest values with mean = 0.158 (standard deviation [SD] = 0.094) and 0.103 (0.051), respectively (Table 2).

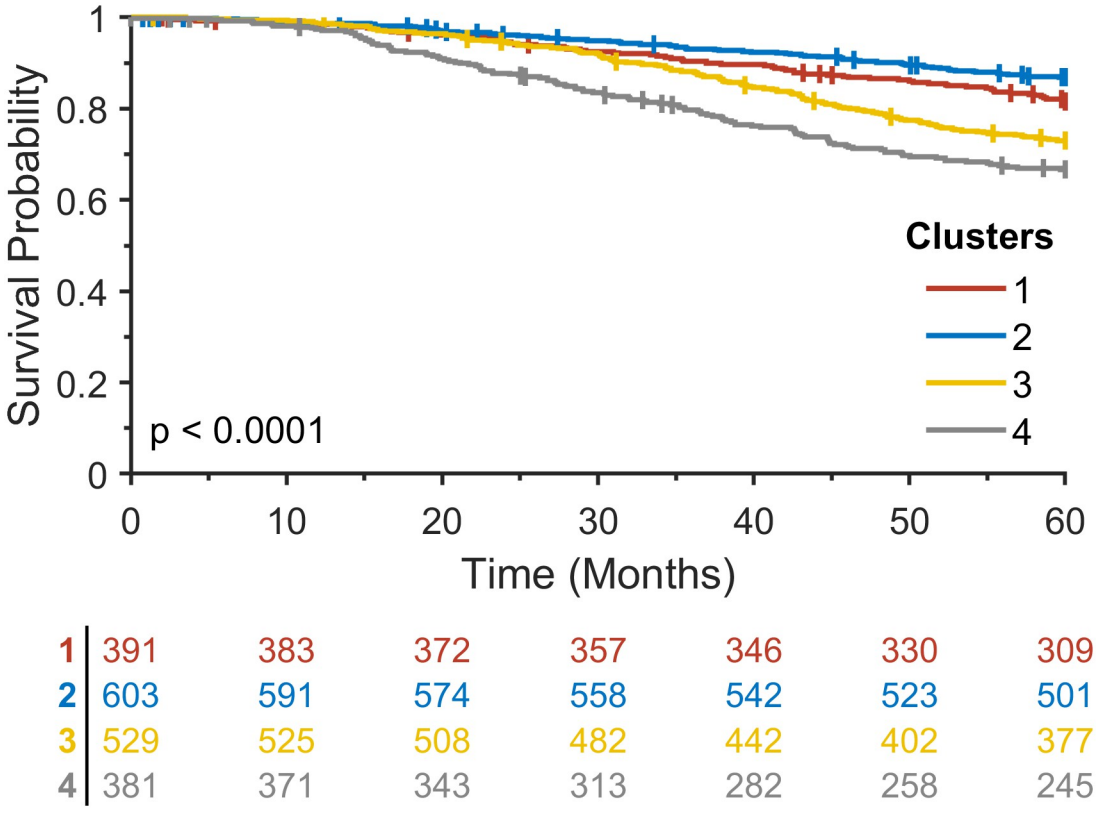

Fig 3. Kaplan-Meier analysis for four clusters that resulted from a hierarchical clustering method on the vector-valued Wasserstein distance matrix in the METABRIC study.

https://doi.org/10.1371/journal.pone.0265150.g003

Table 1. Comparison between PAM50 along with Claudin-low subtypes and clusters identified by the proposed method.

Cluster	Luminal A	Luminal B	Her2	Basal	Claudin-low	Normal-like	NA
1	189	94	33	16	24	32	3
2	316	78	32	18	95	62	2
3	147	205	85	31	27	33	1
4	27	84	70	134	53	13	0

NA: not available.

https://doi.org/10.1371/journal.pone.0265150.t001

TCGA data analysis

To validate the proposed method, we further analyzed multi-omics data in the TCGA breast cancer study, including gene expression, CNV, and methylation data. The clustering results are shown in Fig 4. Similar to the METABRIC analysis, four clusters were chosen for further analysis. Kaplan-Meier analysis with log-rank test resulted in a statistically significant survival difference among clusters with a log-rank p = 0.0088 (Fig 5).

The clustering results were compared with PAM50 subtypes. The associations were assessed using a chi-squared test, resulting in $p < 0.0001$ as shown in Table 3. Clusters 1 and 2 were enriched for Luminal A and B subtypes. Cluster 4 was more enriched for basal subtype.

For the four clusters identified, 22 CIBERSORT immune cell types were compared using the one-way ANOVA test. Fifteen immune cell types were statistically significantly different among the four clusters. The most significant immune cell type was M0 macrophages with p = 1.51E-08 (Table 4).

Table 2. Comparison of 22 CIBERSORT immune cell types among the four clusters identified in METABRIC, showing mean (standard deviation) values.

Immune cell types	Cluster 1	Cluster 2	Cluster 3	Cluster 4	P-value
B cells naive	0.008 (0.017)	0.008 (0.018)	0.008 (0.019)	0.007 (0.017)	0.8240
B cells memory	0.03 (0.036)	0.035 (0.047)	0.025 (0.031)	0.026 (0.031)	**3.60E-05**
Plasma cells	0.17 (0.098)	0.169 (0.097)	0.172 (0.093)	0.165 (0.089)	0.7405
T cells CD8	0.045 (0.05)	0.044 (0.048)	0.041 (0.049)	0.042 (0.047)	0.5181
T cells CD4 naive	0.014 (0.034)	0.017 (0.039)	0.013 (0.028)	0.009 (0.025)	**0.0028**
T cells CD4 memory resting	0.055 (0.06)	0.057 (0.058)	0.052 (0.057)	0.044 (0.053)	**0.0068**
T cells CD4 memory activated	0.001 (0.005)	0.001 (0.007)	0.001 (0.005)	0.001 (0.007)	0.6263
T cells follicular helper	0.057 (0.034)	0.053 (0.035)	0.055 (0.035)	0.068 (0.036)	**3.86E-10**
T cells regulatory (Tregs)	0.016 (0.02)	0.014 (0.02)	0.017 (0.021)	0.017 (0.021)	0.0771
T cells gamma delta	0.057 (0.045)	0.056 (0.044)	0.059 (0.044)	0.063 (0.044)	0.1604
NK cells resting	0.003 (0.013)	0.005 (0.016)	0.003 (0.011)	0.003 (0.013)	0.1057
NK cells activated	0.029 (0.026)	0.025 (0.026)	0.029 (0.025)	0.031 (0.028)	**0.0028**
Monocytes	0.017 (0.024)	0.021 (0.027)	0.015 (0.022)	0.019 (0.029)	**0.0018**
Macrophages M0	0.106 (0.095)	0.104 (0.092)	0.13 (0.098)	0.158 (0.094)	**3.50E-19**
Macrophages M1	0.076 (0.042)	0.069 (0.044)	0.081 (0.043)	0.103 (0.051)	**1.69E-28**
Macrophages M2	0.159 (0.089)	0.162 (0.095)	0.147 (0.077)	0.126 (0.067)	**6.74E-11**
Dendritic cells resting	0.003 (0.01)	0.005 (0.013)	0.003 (0.008)	0.005 (0.014)	**0.0323**
Dendritic cells activated	0.003 (0.01)	0.004 (0.013)	0.005 (0.017)	0.009 (0.023)	**3.89E-07**
Mast cells resting	0.147 (0.104)	0.148 (0.11)	0.144 (0.1)	0.101 (0.082)	**4.26E-13**
Mast cells activated	0.001 (0.004)	0.001 (0.005)	0.001 (0.004)	0.001 (0.008)	0.1828
Eosinophils	0 (0)	0 (0.001)	0 (0.001)	0 (0)	0.9508
Neutrophils	0.001 (0.005)	0.001 (0.004)	0.001 (0.006)	0.001 (0.002)	0.0684

https://doi.org/10.1371/journal.pone.0265150.t002

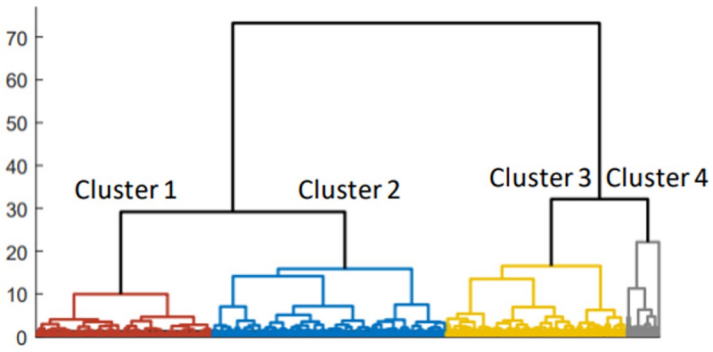

Fig 4. Clustering results employing the resultant vector-valued Wasserstein distance on TCGA data.

https://doi.org/10.1371/journal.pone.0265150.g004

Nine immune cell types were significantly different in both METABRIC and TCGA studies: memory B cells, resting memory CD4 T cells, follicular helper T cells, monocytes, M0 macrophages, M1 Macrophages, resting dendritic cells, activated dendritic cells, and resting mast cells (Fig 6). Three immune cell types, including naive CD4 T cells, activated NK cells, and M2 macrophages, showed statistical significance in METABRIC alone, whereas six immune cell types, including naive B cells, Plasma cells, CD8 T cells, activated memory CD4 T cells, regulatory T cells (Tregs), and resting NK cells, showed statistical significance in TCGA alone.

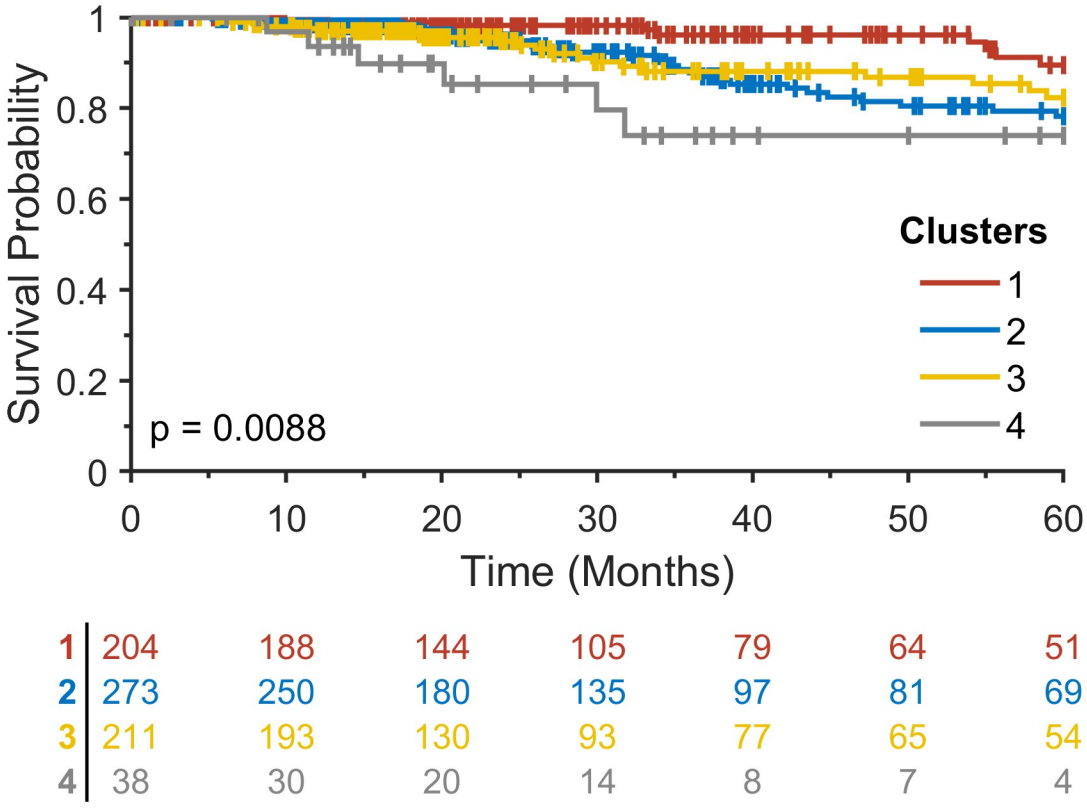

Fig 5. Kaplan-Meier analysis for four clusters that resulted from a hierarchical clustering method on the vector-valued Wasserstein distance matrix in the TCGA breast cancer study.

https://doi.org/10.1371/journal.pone.0265150.g005

Table 3. Comparison between PAM50 subtypes and clusters identified by the proposed method.

Cluster	Luminal A	Luminal B	Her2	Basal	Normal-like	NA
1	184	9	0	0	11	0
2	131	46	22	56	15	3
3	66	73	19	45	4	4
4	2	7	2	25	1	1

NA: not available.

https://doi.org/10.1371/journal.pone.0265150.t003

Comparison with SNF

The clustering performance of vWCluster was compared with that of SNF using gene expression and CNV data for METABRIC and gene expression, CNV, and methylation data for TCGA breast cancer. Kaplan-Meier analysis for the resulting four clusters from SNF yielded statistical significance for METABRIC with a log-rank $p < 0.0001$, but for TCGA breast cancer, the log-rank test was statistically insignificant with $p = 0.13$ (S1 Fig).

We also compared vWCluster with SNF for five clusters. For METABRIC, both methods resulted in significant log-rank p-values with $p < 0.0001$. For TCGA breast cancer, vWCluster resulted in statistical significance with $p = 0.013$, which was slightly worse than that of four clusters, whereas the p-value of SNF for five clusters remained statistically insignificant with $p = 0.0884$.

Table 4. Comparison of 22 CIBERSORT immune cell types among the four clusters identified in TCGA, showing mean (standard deviation) values.

Immune cell types	Cluster 1	Cluster 2	Cluster 3	Cluster 4	P-value
B cells naive	0.068 (0.045)	0.05 (0.046)	0.045 (0.046)	0.033 (0.037)	**2.29E-07**
B cells memory	0.007 (0.017)	0.015 (0.031)	0.01 (0.021)	0.017 (0.025)	**0.0005**
Plasma cells	0.049 (0.052)	0.037 (0.045)	0.046 (0.052)	0.032 (0.038)	**0.0341**
T cells CD8	0.106 (0.057)	0.108 (0.064)	0.093 (0.061)	0.094 (0.068)	**0.0331**
T cells CD4 naive	0 (0.004)	0 (0.001)	0.002 (0.011)	0 (0.002)	0.0662
T cells CD4 memory resting	0.135 (0.076)	0.122 (0.074)	0.098 (0.071)	0.068 (0.069)	**1.23E-06**
T cells CD4 memory activated	0 (0.002)	0.004 (0.012)	0.003 (0.011)	0.003 (0.009)	**0.0014**
T cells follicular helper	0.063 (0.039)	0.073 (0.04)	0.07 (0.045)	0.093 (0.068)	**0.0003**
T cells regulatory (Tregs)	0.014 (0.02)	0.026 (0.03)	0.02 (0.024)	0.015 (0.02)	**8.15E-06**
T cells gamma delta	0.003 (0.01)	0.003 (0.01)	0.002 (0.007)	0.004 (0.012)	0.4291
NK cells resting	0.003 (0.01)	0.006 (0.013)	0.005 (0.011)	0.012 (0.019)	**7.92E-05**
NK cells activated	0.02 (0.024)	0.019 (0.023)	0.021 (0.025)	0.017 (0.023)	0.8629
Monocytes	0.02 (0.023)	0.016 (0.019)	0.015 (0.026)	0.016 (0.018)	**0.0314**
Macrophages M0	0.056 (0.1)	0.087 (0.105)	0.116 (0.132)	0.177 (0.165)	**1.51E-08**
Macrophages M1	0.055 (0.032)	0.069 (0.045)	0.057 (0.04)	0.058 (0.054)	**0.0014**
Macrophages M2	0.271 (0.119)	0.273 (0.128)	0.301 (0.121)	0.286 (0.13)	0.1177
Dendritic cells resting	0.021 (0.031)	0.015 (0.025)	0.011 (0.024)	0.005 (0.012)	**0.0003**
Dendritic cells activated	0.002 (0.007)	0.004 (0.013)	0.009 (0.026)	0.021 (0.06)	**3.58E-07**
Mast cells resting	0.096 (0.071)	0.063 (0.062)	0.064 (0.075)	0.042 (0.044)	**1.41E-07**
Mast cells activated	0.007 (0.024)	0.005 (0.02)	0.011 (0.036)	0.004 (0.009)	0.4921
Eosinophils	0.001 (0.003)	0 (0.003)	0 (0.002)	0 (0)	0.8402
Neutrophils	0.003 (0.008)	0.002 (0.005)	0.004 (0.007)	0.003 (0.008)	0.1433

https://doi.org/10.1371/journal.pone.0265150.t004

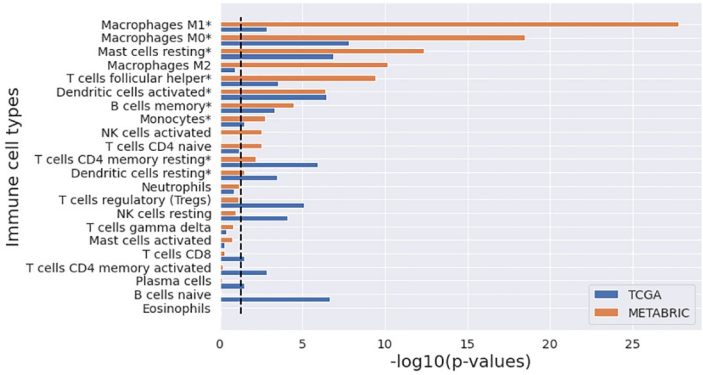

Fig 6. Comparison of 22 immune cell types in CIBERSORT among the four clusters identified in METABRIC and TCGA studies. The black dot line indicates -log10(p = 0.05).

https://doi.org/10.1371/journal.pone.0265150.g006

Discussion

The treatment of multi-omic biological data in a vector-valued manner may provide new insights for understanding the biological mechanisms of cancer biology, using complementary information offered by individual omics types. vWCluster is a data analysis methodology based on OMT theory, which enables the integration of multi-omics data in a vector-valued form, represented by multiple layers in a network. The Wasserstein distance computed on the vector-valued data was further employed to identify cancer subtypes. We applied this method to the two largest breast cancer studies, METABRIC and TCGA. The clusters identified showed significantly different survival rates in both studies.

vWCluster identified cluster 4 as a poor survival group in both METABRIC and TCGA breast cancer studies and cluster 2 and cluster 1 as a good survival group in METABRIC and TCGA breast cancer, respectively. In both studies, the poor survival group was enriched for basal subtype and the good survival group was enriched for Luminal A subtype. This is consistent with the clinical findings that in general, the triple-negative/basal-like subtype has a poor prognosis [32] while the Luminal A subtype has a better prognosis than other breast cancer subtypes [33].

CIBERSORT scores, consisting of 22 immune cell types, were further compared among the identified clusters. CIBERSORT employs gene expression profiles from a set of 547 genes to predict 22 immune cell types, using support vector regression [29]. ANOVA tests revealed that nine immune cell types were commonly statistically significant in both studies, indicating that the tumor immune microenvironment may differ among the identified clusters and this is associated with the difference in survival in breast cancer patients. Among the nine immune cell types, the poor survival group (cluster 4) had the lowest scores in memory resting CD4 T cells and resting mast cells, and the highest scores in follicular helper T cells, M0 macrophages, and activated dendritic cells in both METABRIC and TCGA breast cancer studies (Tables 2 and 4). By contrast, the good survival group (cluster 2 in METABRIC and cluster 1 in TCGA breast cancer) had the highest scores in memory resting CD4 T cells and resting mast cells, and the lowest scores in follicular helper T cells and M0 macrophages in both METABRIC and TCGA breast cancer studies. The score for activated dendritic cells was the lowest in TCGA breast cancer and the second lowest in METABRIC. A study revealed that M0 and M1 macrophages were significantly higher in the basal-like subtype compared to the Luminal A and B subtypes ($p < 0.001$) [34]. Recently, Gao *et al.* [35] investigated the difference of immune cells infiltration abundance between ER/PR-positive and triple-negative subtypes and reported that

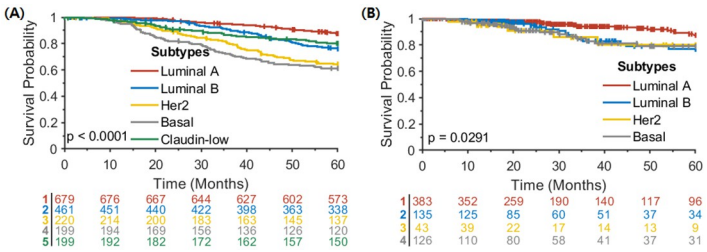

Fig 7. Kaplan-Meier analysis for intrinsic molecular subtypes in (A) METABRIC and (B) TCGA breast cancer studies with no normal-like samples.

https://doi.org/10.1371/journal.pone.0265150.g007

triple-negative tumors had significantly higher CIBERSORT scores for follicular helper T cells ($p < 0.001$) and lower CIBERSORT scores for resting memory CD4 T cells ($p = 0.002$) and resting mast cells ($p < 0.001$) compared to ER/PR-positive tumors. These results are consistent with our findings.

Kaplan-Meier analysis was performed for intrinsic molecular subtypes in the METABRIC and TCGA breast cancer studies (Fig 7). As in Kaplan-Meier analysis for the clusters identified by our method in METABRIC, an extremely significant survival difference was found among intrinsic subtypes with a log-rank $p < 0.0001$, showing the worst survival rate for basal subtype. By contrast, for the TCGA breast cancer cohort, our method resulted in much better statistical significance with a log-rank $p = 0.0088$ compared to marginal statistical significance with a log-rank $p = 0.0291$ among intrinsic subtypes in TCGA. It is worth noting that Kaplan-Meier survival curves for all four clusters in Fig 5 were separable, whereas in the intrinsic subtypes, only the Luminal A subtype was separated from others that had similar survival patterns, suggesting the potential of our proposed method to identify new subtypes in cancer and further stratify patients at high risk of mortality. Further investigation of the association between the tumor immune microenvironment and survival will be explored in future work.

Prior to this study, Chen *et al.* [8] and Ryu *et al.* [19] introduced vector-valued extensions of the Wasserstein distance metric. However, the current study is the first to employ the vector-valued Wasserstein distance methodology for the integration of multi-omics data and further to cluster samples.

Conclusion

We proposed a multi-omics data integration and clustering method, called vWCluster, based on the vector-valued Wasserstein distance. In this method, individual omics types represented as multiple layers in a network can be efficiently integrated, considering the biological interactions of biomarkers and providing complementary biological information. The formulation of vWCluster treats the data vectorially, which potentially minimizes information loss. vWCluster is flexible and applicable to the integration of multi-modal data including imaging and genomic data, which is a research direction we plan to explore in the future.

Supporting information

S1 Fig. Kaplan-Meier analysis for four clusters that resulted from SNF in (A) METABRIC and (B) TCGA breast cancer studies.
(PDF)

Acknowledgments

Code availability

vWCluster was implemented in MATLAB and the codes are available in https://github.com/MSK-MOI/vWCluster.

Author Contributions

Conceptualization: Allen R. Tannenbaum.

Data curation: Jiening Zhu, Jung Hun Oh.

Formal analysis: Jiening Zhu, Jung Hun Oh, Allen R. Tannenbaum.

Funding acquisition: Joseph O. Deasy, Allen R. Tannenbaum.

Investigation: Jiening Zhu, Jung Hun Oh, Allen R. Tannenbaum.

Methodology: Jiening Zhu, Jung Hun Oh, Allen R. Tannenbaum.

Project administration: Joseph O. Deasy, Allen R. Tannenbaum.

Resources: Joseph O. Deasy.

Software: Jiening Zhu.

Supervision: Jung Hun Oh, Joseph O. Deasy, Allen R. Tannenbaum.

Validation: Jiening Zhu.

Writing – original draft: Jiening Zhu.

Writing – review & editing: Jiening Zhu, Jung Hun Oh, Joseph O. Deasy, Allen R. Tannenbaum.

References

1. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Frontiers in Genetics. 2017; 8:84. https://doi.org/10.3389/fgene.2017.00084 PMID: 28670325
2. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009; 25 (22):2906–2912. https://doi.org/10.1093/bioinformatics/btp543 PMID: 19759197
3. Alkhateeb A, Zhou L, Tabl AA, Rueda L. Deep Learning Approach for Breast Cancer InClust 5 Prediction based on Multiomics Data Integration. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. ACM; 2020. Available from: https://doi.org/10.1145/3388440.3415992.
4. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods. 2014; 11(3):333–337. https://doi.org/10.1038/nmeth.2810 PMID: 24464287
5. Pouryahya M, Oh JH, Javanmard P, Mathews JC, Belkhatir Z, Deasy JO, et al. aWCluster: A Novel Integrative Network-based Clustering of Multiomics for Subtype Analysis of Cancer Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2020. https://doi.org/10.1109/TCBB.2020.3039511 PMID: 33226952
6. Villani C. Topics in Optimal Transportation. American Mathematical Soc.; 2003.
7. Villani C. Optimal Transport: Old and New. vol. 338. Springer Science & Business Media; 2008.
8. Chen Y, Georgiou TT, Tannenbaum A. Vector-valued optimal mass transport. SIAM Journal Applied Mathematics. 2018; 78(3):1682–1696. https://doi.org/10.1137/17M1130897
9. Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arxivorg. 2017;1701.07875.
10. Georgiou T, Michailovich O, Yogesh R, Malcolm J, Tannenbaum A. On the matrix Monge-Kantorovich problem. Linear Algebra and Its Applications. 2007; 425:663–672.

11. Carlen EA, Maas J. An analog of the 2-Wasserstein metric in non-commutative probability under which the Fermionic Fokker–Planck equation is gradient flow for the entropy. Communications in Mathematical Physics. 2014; 331(3):887–926. https://doi.org/10.1007/s00220-014-2124-8

12. Haker S, Zhu L, Tannenbaum A, Angenent S. Optimal mass transport for registration and warping. International Journal of Computer Vision. 2004; 60(3):225–240. https://doi.org/10.1023/B:VISI.0000036836.66311.97

13. Mittnenzweig M, Mielke A. An entropic gradient structure for Lindblad equations and coupling of quantum systems to macroscopic models. J Stat Physics. 2017; 167(2). https://doi.org/10.1007/s10955-017-1756-4

14. Rachev ST, Rüschendorf L. Mass Transportation Problems: Volumes I and II. Springer Science & Business Media; 1998.

15. Mathews JC, Nadeem S, Pouryahya M, Belkhatir Z, Deasy JO, Levine AJ, et al. Functional network analysis reveals an immune tolerance mechanism in cancer. Proceedings of the National Academy of Sciences. 2020; 117(28):16339–16345. https://doi.org/10.1073/pnas.2002179117 PMID: 32601217

16. Chen Y, Georgiou TT, Tannenbaum A. Matrix optimal mass transport: a quantum mechanical approach. IEEE Trans Automatic Control. 2018; 63(8):2612–2619. https://doi.org/10.1109/TAC.2017.2767707

17. Chen Y, Georgiou TT, Tannenbaum A. Interpolation of Density Matrices and Matrix-Valued Measures: The Unbalanced Case. Euro Jnl of Applied Mathematics. 2018; 30(3):458–480. https://doi.org/10.1017/S0956792518000219

18. Chen Y, Gangbo W, Georgiou T, Tannenbaum A. On the matrix Monge-Kantorovich problem. European J of Applied Mathematics. 2020; 31:574–600. https://doi.org/10.1017/S0956792519000172

19. Ryu EK, Chen Y, Li W, Osher S. Vector and Matrix Optimal Mass Transport: Theory, Algorithm, and Applications. SIAM Journal on Scientific Computing. 2018; 40(5):A3675–A3698. https://doi.org/10.1137/17M1163396

20. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486(7403):346–352. https://doi.org/10.1038/nature10983 PMID: 22522925

21. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61–70. https://doi.org/10.1038/nature11412

22. Evans LC, Gangbo W. Differential equations methods for the Monge-Kantorovich mass transfer problem. Memoirs of the American Mathematical Society. 1999; 137(653). https://doi.org/10.1090/memo/0653

23. Rachev ST, Rüschendorf L. Mass Transportation Problems: Volume I: Theory. Probability and its Applications. Berlin: Springer; 1998. Available from: http://link.springer.com/10.1007/b98894.

24. Kantorovich LV. On a problem of Monge. CR (Doklady) Acad Sci URSS (NS). 1948; 3:225–226.

25. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. Cancer Discovery. 2012; 2(5):401–404. https://doi.org/10.1158/2159-8290.CD-12-0095 PMID: 22588877

26. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Science Signaling. 2013; 6(269):pl1–pl1. https://doi.org/10.1126/scisignal.2004088 PMID: 23550210

27. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database–2009 update. Nucleic Acids Research. 2009; 37(Database):D767–D772. https://doi.org/10.1093/nar/gkn892

28. Chen Y, Cruz FD, Sandhu R, Kung AL, Mundi P, Deasy JO, et al. Pediatric Sarcoma Data Forms a Unique Cluster Measured via the Earth Mover's Distance. Scientific Reports. 2017; 7(1):7035. https://doi.org/10.1038/s41598-017-07551-8 PMID: 28765612

29. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods. 2015; 12(5):453–457. https://doi.org/10.1038/nmeth.3337 PMID: 25822800

30. Ali HR, Chlon L, Pharoah PDP, Markowetz F, Caldas C. Patterns of Immune Infiltration in Breast Cancer and Their Clinical Implications: A Gene-Expression-Based Retrospective Study. PLOS Medicine. 2016; 13(12):e1002194. https://doi.org/10.1371/journal.pmed.1002194 PMID: 27959923

31. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of Clinical Oncology. 2009; 27(8):1160–1167. https://doi.org/10.1200/JCO.2008.18.1370 PMID: 19204204

32. Fan M, Chen J, Gao J, Xue W, Wang Y, Li W, et al. Triggering a switch from basal- to luminal-like breast cancer subtype by the small-molecule diptoindonesin G via induction of GABARAPL1. Cell Death & Disease. 2020; 11(8). https://doi.org/10.1038/s41419-020-02878-z PMID: 32801338

33. Li Y, Ma L. Efficacy of chemotherapy for lymph node-positive luminal A subtype breast cancer patients: an updated meta-analysis. World Journal of Surgical Oncology. 2020; 18(1). https://doi.org/10.1186/s12957-020-02089-y PMID: 33267822

34. Hachim MY, Hachim IY, Talaat IM, Yakout NM, Hamoudi R. M1 Polarization Markers Are Upregulated in Basal-Like Breast Cancer Molecular Subtype and Associated With Favorable Patient Outcome. Frontiers in Immunology. 2020; 11. https://doi.org/10.3389/fimmu.2020.560074 PMID: 33304345

35. Gao C, Li H, Liu C, Xu X, Zhuang J, Zhou C, et al. Tumor Mutation Burden and Immune Invasion Characteristics in Triple Negative Breast Cancer: Genome High-Throughput Data Analysis. Frontiers in Immunology. 2021; 12. https://doi.org/10.3389/fimmu.2021.650491 PMID: 33968045