Integrating Omics Data with a Multiplex Network-based Approach for the Identification of Cancer Subtypes

Haiying Wang, *Member, IEEE*, Huiru Zheng, *Member, IEEE*, Jianxin Wang, Senior *Member, IEEE*, Chaoyang Wang, and, FangXiang Wu, Senior *Member, IEEE*

***Abstract*—Comprehensive characterization and identification of cancer subtypes have a number of applications and implications in life science and cancer research. Technologies centered on the integration of omics data hold great promise in this endeavor. This paper proposed a multiplex network-based approach for integrative analysis of heterogeneous omics data. It represents a useful alternative network-based solution to the problem and a significant step forward to the methods in which each type of data is treated independently. It has been tested on the identification of the subtypes of glioblastoma multiforme and breast invasive carcinoma from three omics data. The results obtained have shown that it has achieved the performance comparable to state-of-the-art techniques (Normalized Mutual Information > 0.8). In comparison to traditional systems biology tools, the proposed methodology has several significant advantages. It has the ability to correlate and integrate multiple data levels in a holistic manner which may be useful to facilitate our understanding of the pathogenesis of diseases and to capture the heterogeneity of biological processes and the complexity of phenotypes.**

***Index Terms*— Multiplex networks; omics data; cancer subtypes; data integration**

I. Introduction

COMPREHENSIVE characterization and identification of cancer subtypes associated with distinct molecular profiles and differential clinical outcomes has significant applications and implications in life science and cancer research since it may lead, for example, to a better understanding of cancer evolution, new treatment insights, optimal patient stratification and the design of new, effective therapeutic strategies [1], [2]. A breakthrough reclassification of pancreatic cancer has been published in Nature recently and a total of 4 key subtypes, i.e. Squamous, Oancreatic Progenitor, ADEX, and Immunogenic, have been identified, providing a basis to offer new insights into personalized therapeutic treatments [1]. A new approach to the classification of patients for therapeutic purposes based on the recognition of intrinsic biological subtypes within the breast cancer spectrum was adopted by the 12th St Gallen International Breast Cancer Conference Expert Panel [3]. It has been highlighted that Luminal A patients generally only receive endocrine therapy, while for most patients with Triple negative, chemotherapy is required.

However, due to its highly heterogeneous nature, different conclusions regarding the number of cancer subtypes in a tissue have been drawn depending on the types of data used and methodologies employed. In the context of the analysis of glioblastoma multiforme (GBM), for instance, Nigro et al. [4] identified two molecular subtypes with one group containing the most common copy number alteration, loss of chromosome 10. By applying consensus hierarchical clustering to the analysis of expression data from 200 GBM and 2 normal brain samples assayed on three gene expression platforms, Verhaak et al. [5] classified GBM into 4 subgroups, i.e. Proneural, Neural, Classical and Mesenchymal. Using the same datasets, i.e. DNA methylation, mRNA expression and miRNA from 215 patient samples with GBM and 105 samples with breast invasive carcinoma (BIC), Wang et al. [6] applied SNF and suggested 3 subtypes in GBM and 5 subtypes in BIC while Specicher and Pfeifer [7] identified 6 subgroups in GBM and 7 in BIC with multiple kernel learning.

Due to the ability to provide system-level measurements for nearly all biomolecules in the cell and opportunities to study biological systems at different levels, recent years have seen a growing trend toward the integration of diverse omics data for the identification of cancer subtypes. Recent examples include the identification of subtypes of pancreatic cancer associated with distinct histopathological characteristics and differential survival using a combination of the whole-genome and deep-

Manuscript received 29 Feburary, 2016.

H.Y. Wang is with the Computer Science Research Institute, University of Ulster, Jordanstown Campus, Shore Road, Newtownabbey BT37 0QB, United Kingdom.

H. Zheng is with the Computer Science Research Institute, University of Ulster, Jordanstown Campus, Shore Road, Newtownabbey BT37 0QB, United Kingdom.

J.X. Wang is with the School of Information Science and Engineering, Central South University, Changsha, 410083, P.R. China

C. Wang is with the College of Medicine and Veterinary Medicine, University of Edinburgh, and Faculty of Medicine at Imperial College London, UK. Email: chaoyang_wang@outlook.com

F. X. Wu is with Department of Mechanical Engineering and Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK S7N5A9, Canada.

*Corresponding author: e-mail: h.zheng@ulster.ac.uk, phone: 44-28-90366591

exome sequencing with gene copy number analysis [1].

While the growing availability of diverse omics data offers huge opportunities to generate a more thorough and comprehensive view of biological problems, mining such abundant information poses great challenges to research communities, requiring the development of advanced integrative analysis platform to capture the heterogeneity of biological processes and the complexity of phenotypes [6].

A. *Current effort on omics data integration: a brief overview*

The recognized significance of data integration in the era of omics has triggered intense efforts across the global. For example, as a large-scale, collaborative effort led by the National Institute of Health, The Cancer Genome Atlas (TCGA) has collected massive, high quality information generated from various molecular levels for over 30 types of human cancer derived from about 10,000 cases of tumor and matching normal tissues samples. By enabling to map molecular alternation at multiple levels, TCGA provides a valuable resource to accelerate our understanding of the molecular basis of human cancers [8]. A number of EU projects focusing on integrative analysis of diverse omics data have been funded under EU FP7-Health programme. Examples include the STATegra project (http://www.stategra.eu/), which involves 11 partners from different countries. Since 2007, the European Commission has actively participated several international large scale omics research initiatives including International Cancer Genome Consortium (https://icgc.org/) and International Human Epigenome Consortium (http://ihec-epigenomes.org/). Gomez-Cabrero et al. [9] characterized current efforts on data integration in the life science. A recent review on the emerging approaches for omics data integration to uncover genotype-phenotype interactions was provided by Ritchie et al. [10].

Over the past decades, a wide range of computational approaches have been proposed and developed. Using a model-based integration strategy, Akavia et al. [11] developed a computational framework that integrates chromosomal copy number and gene expression data for detecting aberrations that promote cancer progression. Relying on the use of kernel-based statistical learning methods, Lanckriet et al [12] introduced a computational framework for genomic data fusion, in which each type of data is represented via a kernel function that defines similarities between pairs of entities, such as genes or proteins. It has been shown that kernel functions derived from different types of omics data can be combined in a straightforward fashion. Kim et al. [13] introduced a graph-based approach for predicting clinical outcomes in brain cancer and ovarian cancer by integrating multi-omics data as a transformation-based integration. A graph-based semi-supervised learning was used as a classification algorithm. Integration of multi-level genomic data sources was achieved by finding an optimum value of the linear combination coefficient for the individual graphs derived from each type of data. Using a joint latent variable model for integrative clustering, the iCluster method [14] seeks to find a single common clustering structure for all omics data involved. The number of clusters needs to be estimated by heuristic approaches.

More recently, Wang et al. [6] introduced a novel network-based approach, i.e. Similarity Network Fusion (SNF), for aggregating data types on a genomic scale. It consists of two main steps: constructing a patient-similarity network for each available omics data and fusing all networks into a single similarity network with a nonlinear combination method to represent the full spectrum of underlying data. The approach has been applied to combine 3 omics data, i.e. mRNA expression, DNA methylationa, and miRNA expression for five cancer datasets including glioblastoma multiforme (GBM). It has been shown that SNF substantially outperforms single data type analysis and established integrative approaches.

B. *The objectives in this study*

In this study we proposed an alternative network-based data integration strategy, i.e., a multiplex network-based integrative approach for exploring large volumes of multivariate patient data based on the extension of our previous analysis [15]. Similar to SNF, for each type of data, a patient-similarity network is generated. After that, a multiplex network is formed by introducing a coupling strength that links each node in a network slice and its counterpart in each of the other network slices. To demonstrate its performance, the proposed method is applied to identify the subtypes of GBM and BIC. An empirical study of the impact of the selection of learning parameters on the performance is carried out.

The rest of paper is organized as follows. Section II briefly describes the methodology, datasets under study, and evaluation metrics used to assess the significance of results. The formation of multiplex networks and its implementation are provided. The results and discussions are presented in Section III. The conclusions, together with the discussion of limitations and future research, are given in Section IV.

II. METHODOLOGY

Inspired by the recent work published by Mucha et al. [16], a multiplex network(MN)-based clustering approach is proposed to explore large volumes of multivariate patient data. As illustrated in Fig. 1, for each given dataset, a network will be constructed, in which each node corresponds to a patient and each edge represents the similarity between a pair of patients derived from the given dataset. The whole multiplex networks can be represented using a 3rd-order tensor $A = (A_{ijs})_{n\times n\times k}$, where *n* is the number of patients and *k* is the number of datasets under consideration. Each element A_{ijs} is the non-negative value representing the weight associated with the link between a pair of patients in the network derived from dataset *s*.

A. *Cluster detection across multiscale networks*

Unlike the traditional approach, in which each network is treated independently, we propose a flexible framework for integrative clustering analysis of heterogeneous data based on the adaptation of the generalized modularity proposed by Mucha et al. [16]. The generalized modularity shown in (1) will be used as an objective function to optimize partitions across networks. As shown in (1), there are two parts in the representation, i.e. the first part is responsible for the modularity derived from each network [17] and the second part is to enforce a consensus in terms of cluster assignments. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2556640, IEEE Transactions on NanoBioscience

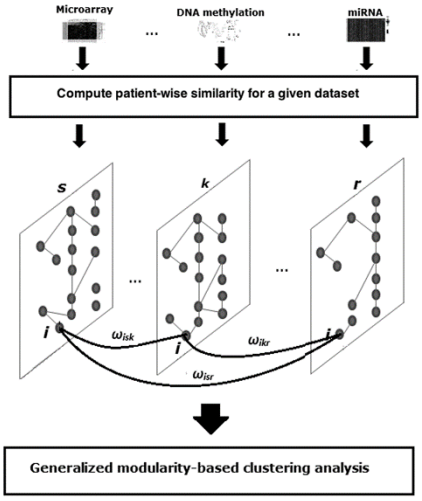

Fig. 1. A flexible, multiplex network-based framework for integrative clustering analysis. *s*, *k*, and *r* represent similarity networks constructed from the corresponding datasets. Each node in the networks is associated with a patient and each edge represents the similarity between a pair of patients derived from the given dataset. ω_{isr} represents the coupling strength between two slices, i.e. *s* and *r*, for node *i*.

optimal solution will be achieved when the same node (patient) across all the networks is assigned to the same cluster. The significance of the second part is determined by the coupling strength, ω, representing the relationship between two sets of datasets. When $\omega = 0$, the optimal partition is achieved from separate optimisation in each network. As ω is increased, the optimisation will gradually force the cluster assignment of a node to remain in the same partition across networks. This becomes more evident when similar patterns are observed across datasets. Such a feature lends itself naturally to providing a flexible framework for integrative clustering analysis of multiple heterogeneous data.

$$Q_m = \frac{1}{2\mu}\sum_{ijs}\left[\left(A_{ijs} - \gamma_s\frac{k_{is}k_{js}}{2m_s}\right)\delta(c_{is},c_{js})\right] + \frac{1}{2\mu}\sum_{isr}\omega_{isr}\delta(c_{is},c_{ir}) \quad (1)$$

where A_{ijs}, k_{is}, and m_s represent the adjacency matrix, the degree of node *i*, and the total number of links in network *s* respectively. ω_{isr} stands for the strength between networks constructed from datasets *s* and *r* for node *i* and $\mu = \frac{1}{2}\sum_{js}\left(\sum_i A_{ijs} + \sum_r \omega_{jsr}\right)$. For each network *s*, γ_s is the resolution parameter used to examine cluster structure at multiple scales and c_{is} represents cluster assignment of node *i* in network *s*. For simplicity, the inter-slice couplings between network *s* and *r*, ω_{isr}, take binary values {0, ω} indicating absence/presence of the inter-slice links. $\delta(c_{is}, c_{jr})$ is the Kronecker delta function which is equal to 1 when two nodes in a network or a node from two slices are assigned to the same community.

B. Implementation

The implementation was based on the generalized Louvain MATLAB code [18]. It implements a Louvain-like greedy community detection method that is based on modularity optimization [19]. As illustrated in Fig. 2, the algorithm consists of two main stages that are repeated iteratively. Starting with assigning a different cluster to each node in a network, the first phase is repeated by moving a node from its community and placing it in the community of its neighbours at a time to optimize the specified quality function until no further improvement can be achieved. The second phase is to build a new network whose nodes represent the communities found during the first stage.

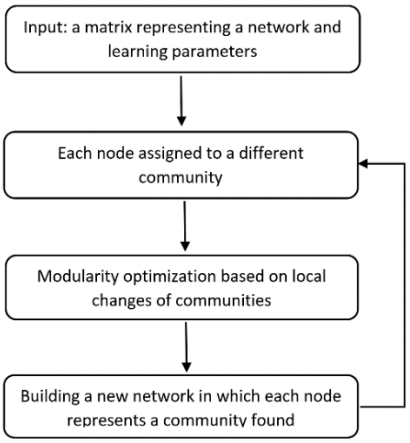

Fig. 2 An illustration of a Louvain-like greedy community detection method

The beauty of the generalized Louvain approach [18] is that it works directly with the modularity matrix and thus can be used with any quality function specified in terms of a modularity matrix. The corresponding multislice modularity matrix associated with the quality function defined in Eq.(1) can then be derived as illustrated in Fig. 3. In this study, we considered the type of interlayer connectivity as categorical, i.e. the interslice couplings connect an individual (patient in this study) in a network to himself or herself in each of remain networks as shown Fig. 1. The reader is referred to [18] for a detailed description of its implementation.

$$\begin{bmatrix} \begin{pmatrix} B_{001} & \cdots & B_{0n1} \\ \vdots & \ddots & \vdots \\ B_{n01} & \cdots & B_{nn1} \end{pmatrix} & \omega_{1s} & \omega_{1r} \\ \omega_{1s} & \begin{pmatrix} B_{00s} & \cdots & B_{0ns} \\ \vdots & \ddots & \vdots \\ B_{n0s} & \cdots & B_{nns} \end{pmatrix} & \omega_{sr} \\ \omega_{1r} & \omega_{sr} & \begin{pmatrix} B_{00r} & \cdots & B_{0nr} \\ \vdots & \ddots & \vdots \\ B_{n0r} & \cdots & B_{nnr} \end{pmatrix} \end{bmatrix}$$

Fig. 3 An illustration of a modularity matrix for categorical multislice networks. $B_{ijs} = A_{ijs} - \gamma_s\frac{k_{is}k_{js}}{2m_s}$ where A_{ijs} is the adjacency matrix for slice *s*. ω_{sr} represents the interslice coupling between slices *s* and *r*.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2556640, IEEE Transactions on NanoBioscience

C. Datasets under study

Three types of omics data available from the TCGA website preprocessed by Wang et al. [6] were used: mRNA expression, miRNA expression, and DNA methylation. The proposed method has been applied to the analysis of two cancer types, i.e. GBM with 215 samples in which 134 were male and 81 female and BIC with 104 female samples. The platforms used to generate the data and the details of data preprocessing including g normalization can be found in [6]

The formation of multiplex networks was based on patient-wise similarity matrices published by Wang et al. [6]. They were computed with a scaled exponential similarity kernel [6] as defined below.

$$P(i,j) = exp\left\{-\frac{d^2(\mathrm{i,j})}{\mu\epsilon_{i,j}}\right\} \qquad (2)$$

where μ is a hyperparameter and $\epsilon_{i,j}$ is used to avoid scaling problems. $P(i,j)$ represents similarity between two patients, i and j, and $d(i,j)$ is a distance function used to calculate the patient-wise distance for a given dataset. After that, a K nearest neighbours (KNN)-based method is used to estimate local affinity. The similarities between non-neighbouring patients are set to zero as illustrated below.

$$S(i,j) = \begin{cases} \dfrac{P(i,j)}{\sum_{m\in N_i} P(i,m)} & j \in N_i \\ 0 & otherwise \end{cases} \qquad (3)$$

where $S(i,j)$ represents the normalized similarity based on the K most similar patients (N_i) for each patient.

D. Evaluation metrics

To assess the significance of differences between GBM subtypes identified in terms of their survival profiles, the log rank test of the Cox regression model [21] was used. It is a nonparametric hypothesis test. The null hypothesis is two groups have identical survival functions. The p value estimated indicates how likely the observed differential survival profiles occur by chance. The Kaplan-Meier estimator [22] is utilized to estimate the survival function, $\hat{S}(t)$, i.e. the probability that a patient survives longer than time t.

In order to study whether certain type of proteins/genes are enriched in a GBM subtype, we adopted hypergeometric distribution function defined as follows.

$$p = 1 - \sum_{i=1}^{k-1} \binom{K}{i}\binom{N-K}{n-i}/\binom{N}{n} \qquad (4)$$

where N and K represent the sizes of population and the sample (subtype in our case) drawn from the population without replacement respectively, and n and k stand for the numbers of certain types of proteins/genes in the population and the sample respectively. The estimated p represents the probability of observing at least k members from a sample drawn from a population of size N having n members in total without replacement by chance.

III. Results and Discussion

E. GBM subtypes derived from clustering analysis of the multiplex networks

Two learning parameters need to be set in the clustering algorithm used in the study, i.e. γ (resolution parameter) and ω (coupling strength). Unless indicated otherwise, γ is set to 0.2 throughout this study. As expected, separate subtypes were generated with ω = 0 for each network with each patient was assigned to 3 separate subtypes. A total of 19 subtypes were produced when ω is set to zero: 6 for mRNA expression data, 5 for DNA methylation and 8 for miRNA expression. As ω was introduced, subtypes merged across networks quickly. This not only reduced the total number of subtypes but more importantly patients were gradually assigned to one subtype. When ω was increased to 0.3, a total of 3 subtypes were derived: 63 in G1, 23 in G2, and 131 in G3 as shown in Table I.

TABLE I THE CHARACTERISTICS OF 3 GBM SUBTYPES IDENTIFIED

Subtypes dentified	G1	G2	G3
Number of patients	61 (M: 38, F:23)	23 (M:11, F:12)	131 (M: 85, F:46)
Average age (years)	52.85	40.61	61.61
Average survival time (days)	657.56	1140.65	467.96

As illustrated in Fig. 4, similarity networks derived from 3 datasets exhibit very different patterns. DNA methylation appears to support connectivity in the medium sized cluster, i.e. Subtype G1 (Fig. 4(b)). While patterns shown in Fig.4(a) suggest relatively strong intercluster mRNA expression-based similarity, it would be hard to derive any convincing conclusion from the miRNA-based similarity network (Fig. 4(c)).

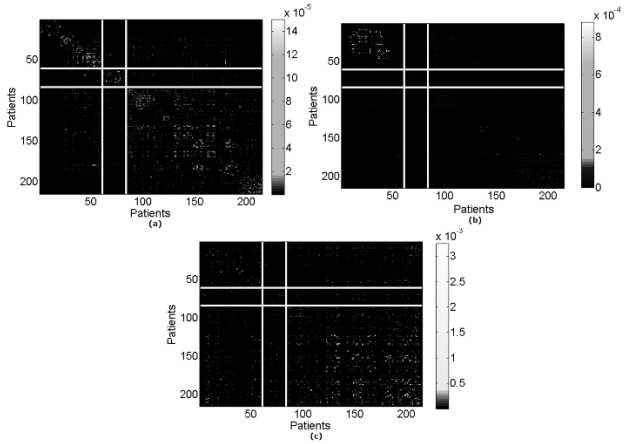

Fig. 4. Patient similarities in each subtype for each of the dataset: (a) mRNA expression data; (b) DNA methylation data; (c) miRNA expression data. The graph was drawn using MATLAB code released by Wang et al. [6]. The similarity value of each pair correlates with color intensity, black with the similarity level equal to zero.

F. Correlation with Clinical Variables

We first investigated the correlation between GBM subtypes identified and age, one of the most important

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2016.2556640, IEEE Transactions on NanoBioscience

prognostic factor in GBM [23]. A statistically significant difference in terms of the average age was observed across 3 subtypes (ANOVA test, $p < 0.0001$) with the smallest patient cluster (Subtype G2) being closely associated with younger patients (median age 34 years). Two post-hoc tests, namely *Bonferroni's method* and *Tukey's Honestly Significant Difference* (HSD) test, indicate that all pairs of subtype mean ages are significantly different ($p < 0.05$).

Next we studied survival profiles associated with each subtype, i.e. the number of days to the last follow-up where4 available [6]. As depicted in Fig. 5, survival times are significantly different among three GBM subtypes with patients in Subtype G2 having a more favorable prognosis (Average survival time 1140.65 days). The overall Cox log rank *p* value for 3 subtypes is 0.000251.

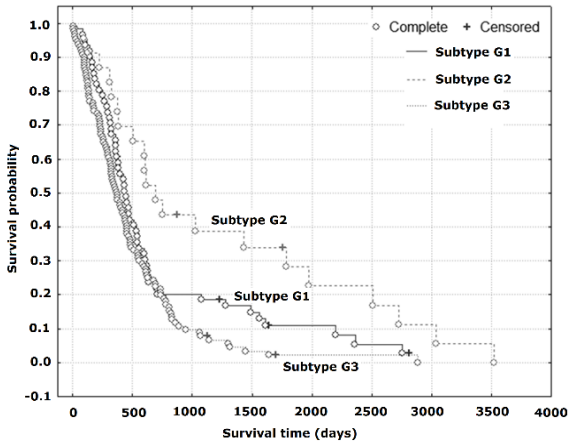

Fig. 5. Kaplan-Meier survival curves for three GBM subtypes as identified (overall Cox log rank *p*-value for 3 subtypes is 0.000251).

Finally, we examined patient response to treatment with temozolomide (TMZ), a chemotherapy drug used to treat certain types of brain tumors including GBM. As illustrated in Fig. 6, patients with GBM in Subtypes G1 and G3 had a significantly increased survival time (Cox log-rank test, $p < 0.005$), whereas for patients associated with Subtype G2, no significant difference in survival time was observed.

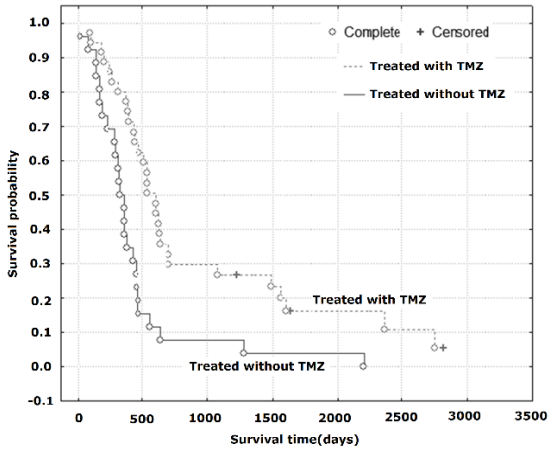

Fig 6. Survival analysis of GBM patients for treatments with TMZ in Subtype G1. Patients associated with Subtype G1 had a significantly increased survival time (Cox log-rank test, $p < 0.005$). Similar observation can be made when examining patients in Subtype G3.

G. Comparisons with state-of-the-art and established subtypes

We first compared our results with the study by Wang et al. [6] published in Nature Method in 2014. As summarized in Table II, a comparable result was obtained in our study. Patients assigned to Clusters 1, 2 and 3 by SNF are highly enriched in Subtypes G3, G1, and G2 respectively (hypergeometric test, $p < 0.0001$). The value of Normalized Mutual Information (NMI) between subtypes identified and cluster labels obtained by SNF (0.80) suggests a high concordance between two results.

TABLE II COMPARISONS WITH CLUSTERS IDENTIFIED USING SNF [6]

Subtypes identified	**Clusters identified by SNF**		
	Cluster 1	***Cluster 2***	***Cluster 3***
G1	1	59	1
G2	2	0	21
G3	126	5	0

The comparison with 4 established subtypes, i.e. Classical, Mesenchymal, Neural and Proneural, determined primarily by expression data [5] is summarized in Table III. Subtypes G1 and G2 are strongly enriched for the mesenchymal GBM (hypergeometric test, $p < 10^{-12}$) and the proneural type (hypergeometric test, $p < 10^{-13}$), respectively. Subtype G3 contains samples that belong to all 4 types of GBM, however, both classical and neural samples are over-represented in this subtype (hypergeometric test, $p < 0.01$). Given that 4 established subtypes were mainly determined based on the analysis of their expression data, the distribution of other omics data over these 4 subtypes deserves further investigation.

TABLE III COMPARISONS WITH 4 ESTABLISHED GBM SUBTYPES

Subtypes identified	**4 established subtypes [5]**			
	Classical	***Mesenchymal***	***Neural***	***Proneural***
G1	7	34	7	9
G2	1	0	1	20
G3	40	20	20	23

A recent study by Sturm et al. [24] identified an epigenetic subgroup of GBM with a distinct global methylation pattern characterized by a somatic mutation in IDH1. Interestingly we found that out of 15 patients with an IDH1 mutation, 13 belong to the Subtype 2 identified in this study.

H. Applying the MN approach to the analysis of breast cancer

To further evaluate the MN performance, we applied it to the analysis of BIC. The optimal number of subtypes identified is 3, which is in agreement with the numbers suggested by Wang et al. [6] based on the analysis of similarity networks using the two heuristics, i.e. eigengaps and rotation cost.

The characteristics of 3 BIC subtypes are shown in Table IV. Both patients in Subtypes B1 and B2 were diagnosed with infiltrating ductal carcinoma, which is the most common type of breast cancer. All 7 patients diagnosed with infiltrating lobular

carcinoma were found in the Subtype B1. However, no significant difference between ductal and lobular carcinomas was observed in terms of their survival profiles (Cox log-rank test, $p > 0.1$).

TABLE IV THE CHARACTERISTICS OF 3 BIC SUBTYPES IDENTIFIED

BIC Subtypes	B1	B2	B3
Number of patients	46	30	29
Average age (years)	56.43	51.94	60.38
Average survival time (days)	1310.98	939.17	733.41
Infiltrating Ductal Carcinoma	37	29	26
Infiltrating Lobular Carcinoma	7	0	0
ER+	45	8	27
PR+	45	5	20
Chemotherapy	21	23	16
Hormone therapy	21	4	14

*ER+: Estrogen-receptor-positive; PR+: Progesterone-receptor-positive;

There is marginally significant difference between the 3 subtypes in terms of their ages at initial pathologic diagnosis (Kruskal-Wallis test, $p = 0.05$) with the subtype B3 associated with elderly patients (mean rank: 61.55).

Turning to survival analysis, statistically significant difference in survival profiles between the subtypes were observed as depicted in Fig. 7 (Cox log-rank test, $p < 0.01$) with the largest subgroup (Subtype B1), in which 45 out of 46 patients are both estrogen receptor (ER) positive and progesterone receptor (PR) positive, having a more favorable prognosis (Average survival time 1310.98 days). This is consistent with the clinic observation that patients with both ER+ and PR+ have better clinical outcomes, which is supported by the recent study published in Nature [25].

According to latest 5-year survival rates for women of different ages with breast cancer in England from Cancer Research UK (https://www.breasthealthuk.com/about-breast-cancer/breast-cancer-survival-rates), women aged between 40 and 70 have better outcomes than younger women and women older than 70, especially for patients over 80 years of age whose survival rate is about 68.5%. However, no significant difference was found in survival between age groups across all three subgroups (Cox log-rank test, $p > 0.1$). This could be partially attributed to the lack of sufficient number of patients in some age groups. For example, only two subtype B1 patients belong to the groups younger than 40 and over 80, respectively.

A variety of drugs have been used to treat breast cancer. Among 76 patients which have drug information available, about 35 drugs have been used with *cyclophosphamide* being the commonly used one. While there are 11 drugs found to be used to treat all 3 subtype patients, some drugs are used to treat a particular subtype of patients. For example, drugs *bevacizumab*, *clodronic acid*, *doxorubicin*, *toremifene*, *gemcitabine*, *methotrexate*, and *Taxane* are used only to treat patients associated with Subtype B2.

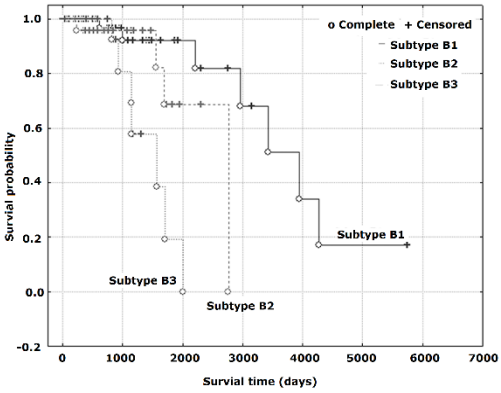

Fig. 7. Kaplan-Meier survival curves for three BIC subtypes as identified (overall Cox log rank *p*-value for 3 subtypes is less than 0.01).

Finally, we compared our results with state-of-the-art and known subtypes. The comparison with the 5 subtypes identified by Wang et al. [6] is shown in Table V. A high value of NMI (0.803) was obtained, indicating a high degree of concordance between two sets of clustering.

TABLE V COMPARISONS BIC SUBTYPES WITH CLUSTERS IDENTIFIED USING SNF [6]

Clusters identified by SNF	**BIC Subtypes identified**		
	B1	B2	B3
Cluster 1	0	7	0
Cluster 2	0	22	0
Cluster 3	0	0	10
Cluster 4	46	1	1
Cluster 5	0	0	18

TABLE VI COMPARISONS BIC SUBTYPES WITH 4 WELL KNOWN MOLECULAR SUBTYPES [25]

MOLECULAR SUBTYPES	**BIC Subtypes identified**		
	B1	B2	B3
Luminal A	40	2	9
Luminal B	3	0	15
Basal-like	0	23	0
HER2-enriched	2	5	5

It has been shown that each of four main breast cancers, i.e., Luminal A, Luminal B, Basal-like, and HER2-enriched exhibits significant molecular heterogeneity as highlighted in the study reported in [26]. Comparing with its results (Table VI), we found that luminal A cancers which are the most likely to retain activity of two major tumor suppressors, i.e. PB1 and TP53, are highly enriched in Subtype B1 that have the best prognosis as shown Fig. 7 (hypergeometric test, $p < 0.0001$). Luminal B tumours in which the TP53 pathway is often inactivated are highly over-represented in the more aggressive Subtype B3 patients. All 23 basal-like breast cancers which are more likely

to lose the function of TP53, RB1 and BRCA1 are found in the group of patients in Subtype B2, over 70% of which are triple negative, i.e. negative for ER, PR and HER2.

I. The impact of learning parameters

The construction of patient-wise similarity networks was based on the approach introduced in [6] in which the following two parameters were used: (1) k, the number of neighbours which is used to measure local affinity with K nearest neighbours (KNN); and μ, a hyperparameter used to determine similarity kernel. It was recommended setting μ in the range of [0.3, 0.8] and k less than 30. In this section we first examined the impact of these parameters on the performance. Without losing generality, the BIC dataset was used in this analysis. We assessed the performance based on the comparison with the SNF approach [6].

As shown in Table VII, the high level of concordance was achieved when k is set to a range between 7 and 10 which is consistent with Wang et al. study [6]. They suggested to set k equal to $N/10$ approximately (N is the number of subjects) where the knowledge of the number of clusters is not available. The performance is significantly deteriorated when k greater than 15, especially when k = 20, the model essentially fails to differentiate patients with all the patients grouped together.

TABLE VII THE IMPACT OF THE SELECTION OF THE NUMBER OF NEIGHBOURS (K) ON THE ANALYSIS

Number of neighbours (k)	**5**	**7**	**10**	**12**	**15**	**20**
The number of subtypes identified	5	3	3	2	2	1
NMI	0.667	0.812	0.803	0.618	0.701	0.000

The impact of the selection of the hyperparameter, i.e. μ, is depicted in Table VIII. The model appears to be sensitive to the variation of μ with the best performance was obtained when μ is set to the range between 0.45 and 0.50.

TABLE VIII THE IMPACT OF THE SELECTION OF THE HYPERPARAMETER, μ, ON THE ANALYSIS

μ	0.40	0.45	0.50	0.55	0.60	0.70
The number of subtypes identified	8	4	3	2	2	1
NMI	0.671	0.853	0.803	0.538	0.487	0.000

There are two learning parameters required for the multiplex network clustering algorithm used in our study, i.e. γ and ω. As expected the value of a resolution parameter γ has significant impact on the number of subtypes identified. The best performance was achieved when γ is set to the range of [0.2, 0.3]. Turning to the parameter ω representing to the couple strength between networks, we found that the system is robust to the selection of ω when γ is set to 0.2 and ω is greater than 0.1.

IV. CONCLUSIONS

It has been well recognized that comprehensive characterization and identification of cancer subtypes have a number of applications and implications in life science, for example, leading to a better understanding of heterogeneity of phenotypes and cellular organization at different levels. Technologies centered on the integration of omics data hold great promise in this endeavor. This paper proposed a multiplex networks-based approach for integrative analysis of heterogeneous omics data. It represents a useful alternative network-based solution and a significant step forward to the methods already in use in which each type of data is treated independently. It has been tested on the identification of GBM and BIC subtypes from three omics data, i.e. RNA expression, DNA methylation and miRNA expression. Results obtained have shown that a high level of concordance (NMI > 0.8) has been achieved in comparisons to state-of-the-art techniques. The proposed methodology has several useful features. For example, it allows researchers to compare the biological/clinical patterns observed in a patient against data from large numbers of other patients which may be from different ethnic groups and subject to different environmental and epigenetic influences. It provides a flexible platform to integrate different types of patient data, potentially from multiple sources, allowing discovering complex disease patterns with multiple facets. The proposed platform has the ability to correlate and integrate multiple data levels in a holistic manner to facilitate our understanding of the pathogenesis of disease.

This paper also provides an empirical analysis of the impact of the selection of some learning parameters on the analysis. It suggests that in general the results are not critically sensitive to the selection of k used to measure local affinity for a given patient. However, it appears that the system is quite sensitive to the variation of hyperparameter, i.e. μ. As expected, the values of the resolution parameter γ and the couple strength ω have impact on the number of subtypes identified although it appears to be robust to the selection of ω when γ = 0.2 and ω > 0.1 in the analysis of BIC data. However, there is no standard way to determine the optimal value of these learning parameters in advance. Currently the determination of the learning parameters including resolution and coupling strength was based on trial and error. How to automatically determine the best combination of learning parameters would be part of future research. Another future direction concerns the way in which the coupling strength is determined. For simplicity, in this paper we have specified the parameter to an equal value between networks. Clearly, a more desirable solution is to assign the strength between networks in the way which could reflect the characteristics of datasets under investigation.

The proposed method was applied to the identification of subtypes of GBM and BIC. We are extending our analysis to the study of other human cancers such as pancreatic cancer and colon adenocarcinoma.

REFERENCES

[1] P. Balley, D. K. Chang, K. Nones, A.L. Johns, A,Patch, M. Gingras, et al., "Genomic analyses identify molecular subtypes of pancreatic cancer," *Nature*, 2016, doi:10.1038/nature16965.

[2] A. A. Alizadeh, V. Aranda, A. Bardelli, C. Blanpain, C. Bock, C.Borowski, et al., "Toward understanding and exploiting tumor heterogeneity," *Nature Medicine* **21**, 2015, pp.846–853.

[3] A. Goldhirsch1, W. C. Wood, A. S. Coates, R. D. Gelber, B. Thürlimann, H.-J. Senn, et al., "Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011," *Annals of Oncology*, doi: 10.1093/annonc/mdr304.

[4] J. M. Nigro, A. Misra, L. Zhang, I. Smirnov, H. Colman, C. Griffin, et al., "Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma," *Cancer Res.* 65, 2005, pp.1678–1686.

[5] R. G. Verhaak, K.A. Hoadley, E. Purdom, V. Wang, Y. Qi, M.D. Wilkerson, et al.,. "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell* **17**,2010, pp. 98–110.

[6] B Wang, A Mezlini, F Demir, M Fiume, T Zu, M Brudno, B Haibe-Kains, A Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, doi:10.1038/nmeth.2810, Jan. 2014.

[7] N.K. Speicher and N. Pfeifer, "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery," *Bioinformatics*, 2015, 31 (12): i268-i275.

[8] The Cancer Genome Atlas Research Network, "Integrated genomic characterization of endometrial carcinoma," *Nature*, 2013, 497(7447), pp.67-73

[9] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, et al., "Data integration in the era of omics: current and future challenges," *BMC Systems Biology* 2014, 8(Suppl 2):I1.

[10] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, D. Kim, Pendergrass, and Dokyoon Kim, "Methods of integrating data to uncover genotype–phenotype interactions," *Nature Reviews Genetics* 16, 85–97 (2015)

[11] U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, et al., "An integrated approach to uncover drivers of cancer.," *Cell* 143, 2010, pp.1005–1017.

[12] G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. A.Noble, "Statistical framework for genomic data fusion," *Bioinformatics*, 20, 2004, pp.2626–2635.

[13] D. Kim, H. Shin, Y.S. Song, and J. H. Kim, "Synergistic effect of different levels of genomic data for cancer clinical outcome prediction.," *J. Biomed. Inform.* 45, 2012, pp.1191–1198.

[14] R. Shen, A. Olshen and M. Ladanyi, "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, pp. 2906–2912.

[15] HY.Wang and H. Zheng, "Integrating omic data for identifying disease subtypes: a multiple network-based approach," in the Proc. Of 2015 IEEE International Conferene on Bioinformatics and Biomedicine (BIBM), pp.581-586.

[16] P. Mucha, T. Richardson, K. Macon, M. Porter, J. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 208, pp. 876-878, 2010.

[17] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences of the United States of America*, 2006, **103** (23): 8577–8696.

[18] J. Inderjit, J. Lucas, and Mu. Peter, "A generalized Louvain method for community detection implemented in MATLAB," http://netwiki.amath.unc.edu/GenLouvain (2011-2014).

[19] V. Blondel, Jean-Loup Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, P10008 (2008).

[20] C.W. Brennan, R.G. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S.R. Salama, S. Zheng, et al. "The somatic genomic landscape of glioblastoma," *Cell*, 2013 Oct 10;155(2), pp.462-77.

[21] D. W. Hosmer, Jr., S. Lemeshow, and S. May. Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley, New York, 2011.

[22] J. M. Bland, D. G. Altman, "Survival probabilities. The Kaplan-Meier method, "BMJ 1998; 317: 1572.

[23] S. Bozdag, A. Li, G. Riddick G, Y. Kotliarov, M. Baysan, F.M. et al, "Age-specific signatures of glioblastoma at the genomic, genetic, and epigenetic levels," *PLoS* One, 2013 Apr 29;8(4):e62982.

[24] D. Sturm, H. Witt, V. Hovestadt, D.A. Khuong-Quang, D.T. Jones, C. Konermann, D. Sturm, et al., "Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma.," *Cancer Cell* 22, 2012, pp.425–437.

[25] H. Mohammed, I. Russell, R. Stark, O. Rueda, T. Hickey, G. Tarulli, et al. "Progesterone receptor modulates estrogen receptor-α action in breast cancer," *Nature*, 2015, 523, pp.313–317.

[26] The Cancer Genome Atlas Network. "Comprehensive Molecular Portraits of Human Breast Tumors." *Nature* 490.7418 (2012), p pp.61–70.

Haiying Wang (Member) received the PhD degree on artificial intelligence in biomedicine in 2004 and he is currently a Senior Lecturer in the School of Computing and Mathematics at Ulster University, Belfast, UK. His research interests lie broadly within the areas of artificial intelligence, complex network analysis, computational biology and bioinformatics. He has a particular research interest and expertise in the development of bio-inspired, self-adaptive and self-organization systems, advanced techniques for the extraction of functional modules from complex systems, and network-based approaches to the field of systems biology. Since 2004, he has published more than 110 peer-reviewed papers in refereed journals and conference proceedings.

Huiru Zheng (Member) received her PhD degree on data mining and bioinformatics from the University of Ulster, UK in 2003. Her re-search area lies on the broad area of healthcare informatics, including bioinformatics, medical informatics, data mining and artificial intelligence and their applications on systems biology, telecare and tele-medicine. She has published over 180 research papers in peer reviewed international journals and conferences. Dr. Zheng is currently a Reader with the School of Computing and Mathematics at the University of Ulster.

Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the vice dean and a professor in School of Information Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, Bioinformatics and computer network. He has published more than 150 papers in various International journals and refereed conferences. He is a senior member of the IEEE.

Chaoyang Wang started the MBChB degree at the University of Edinburgh, UK, in 2013. He is currently completing a BSc in Medical Sciences with Honours in Surgery and Anaesthesia with the Faculty of Medicine at Imperial College London, UK. He was with the Queen's Medical Research Institute, UK, from 2014 to 2015. His previous research focused on the physiological and pathological effects of sodium chloride intake in humans. His current research emphasis is the application of computer science in the field of biomedical science, with particular interest in the human cell cycle and cancer subtype identification.

Fang-Xiang Wu (M'06, SM'11) received the B. Sc. degree and the M. Sc. degree in Applied Mathematics, both from Dalian University of Technology, Dalian, China, in 1990 and 1993, respectively, the first Ph.D. in Control Theory and Its Applications from Northwestern Polytechnical University, Xi'an, China, in 1998, and the second Ph.D. in Biomedical Engineering from University of Saskatchewan (U of S), Saskatoon, Canada, in 2004. During 2004-2005, he worked as a postdoctoral fellow in the Laval University Medical Research Center (CHUL), Quebec City, Canada.

Dr. Wu is currently a full professor of Bioengineering in the Department of Mechanical Engineering and the Graduate Chair of the Division of Biomedical Engineering at the U of S. His current research interests include computational and systems biology, genomic and proteomic data analysis, biological system identification and parameter estimation, applications of control theory to biological systems. Dr. Wu has published more than 230 technical papers in refereed journals and conference proceedings. Dr. Wu is serving as the editorial board member of five international journals and as the guest editor of several international journals, and as the program committee chair or member of several international conferences. He has also reviewed papers for many international journals.