Methods 67 (2014) 344–353

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Incorporating inter-relationships between different levels of genomic data into cancer clinical outcome prediction

Dokyoon Kim [a,b], Hyunjung Shin [c], Kyung-Ah Sohn [d], Anurag Verma [b], Marylyn D. Ritchie [b], Ju Han Kim [a,e,*]

[a] Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110799, Republic of Korea
[b] Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA
[c] Department of Industrial & Information Systems Engineering, Ajou University, San 5, Wonchun-dong, Yeoungtong-gu, 443-749 Suwon, Republic of Korea
[d] Department of Information and Computer Engineering, Ajou University, Suwon 443-749, Republic of Korea
[e] Systems Biomedical Informatics Research Center, Seoul National University, Seoul 110799, Republic of Korea

ARTICLE INFO

Article history:
Available online 18 February 2014

Keywords:
Multi-omics data
Data integration
Inter-relationship
Clinical outcome prediction
TCGA
Ovarian cancer

ABSTRACT

In order to improve our understanding of cancer and develop multi-layered theoretical models for the underlying mechanism, it is essential to have enhanced understanding of the interactions between multiple levels of genomic data that contribute to tumor formation and progression. Although there exist recent approaches such as a graph-based framework that integrates multi-omics data including copy number alteration, methylation, gene expression, and miRNA data for cancer clinical outcome prediction, most of previous methods treat each genomic data as independent and the possible interplay between them is not explicitly incorporated to the model. However, cancer is dysregulated by multiple levels in the biological system through genomic, epigenomic, transcriptomic, and proteomic level. Thus, genomic features are likely to interact with other genomic features in the different genomic levels. In order to deepen our knowledge, it would be desirable to incorporate such inter-relationship information when integrating multi-omics data for cancer clinical outcome prediction. In this study, we propose a new graph-based framework that integrates not only multi-omics data but inter-relationship between them for better elucidating cancer clinical outcomes. In order to highlight the validity of the proposed framework, serous cystadenocarcinoma data from TCGA was adopted as a pilot task. The proposed model incorporating inter-relationship between different genomic features showed significantly improved performance compared to the model that does not consider inter-relationship when integrating multi-omics data. For the pair between miRNA and gene expression data, the model integrating miRNA, for example, gene expression, and inter-relationship between them with an AUC of 0.8476 (REI) outperformed the model combining miRNA and gene expression data with an AUC of 0.8404. Similar results were also obtained for other pairs between different levels of genomic data. Integration of different levels of data and inter-relationship between them can aid in extracting new biological knowledge by drawing an integrative conclusion from many pieces of information collected from diverse types of genomic data, eventually leading to more effective screening strategies and alternative therapies that may improve outcomes.

1. Introduction

Gene expression profiles have been widely used for predicting clinical outcomes for the diagnosis, treatment, or prognosis of cancer for several years [1–5]. In addition to gene expression at the transcriptome level, there have been many attempts at cancer clinical outcome prediction using different levels of genomic data such as copy number alteration (CNA) at the genomic level, miRNA expression, or DNA methylation at the epigenomic level [6–10]. Despite these efforts, explaining cancer clinical outcomes remains problematic since the cancer genome is neither simple nor independent but is complicated and dysregulated by multiple levels in the biological system through genomic, epigenomic, transcriptomic, proteomic level, etc [11,12]. In order to improve our understanding of cancer and develop multi-layered theoretical models, it will require an increased understanding of interactions between

* Corresponding author at: Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110799, Republic of Korea

E-mail addresses: duk27@psu.edu (D. Kim), shin@ajou.ac.kr (H. Shin), kasohn@ajou.ac.kr (K.-A. Sohn), anurag.verma@psu.edu (A. Verma), marylyn.ritchie@psu.edu (M.D. Ritchie), juhan@snu.ac.kr (J.H. Kim).

http://dx.doi.org/10.1016/j.ymeth.2014.02.003
1046-2023/

multiple levels of genomic data that contribute to tumor formation and progression [11].

For overcoming these problems in cancer research, the emerging multi-omics data and clinical information from a collaborative initiative such as the Cancer Genome Atlas (TCGA) have provided many opportunities to explore the complex multi-layered genomic basis of cancer for improving the ability to diagnose, treat, and prevent cancer. TCGA is a large-scale collaborative initiative to improve our understanding of multi-layered of molecular basis of cancer. In addition, the International Cancer Genome Consortium (ICGC) is another comprehensive collaborative initiative to characterize multi-omics data in 50 different cancer types [13]. While the TCGA and ICGC open unprecedented opportunities to deepen the novel knowledge of the molecular basis of cancer [13–22], integrative analysis of multi-omics data is still considered as one of important problems to better explain cancer phenotype, further providing an enhanced global view on interplays between different levels of genomic data.

Previously, we proposed a graph-based framework that integrates multi-omics data for predicting cancer clinical outcomes in glioblastoma multiforme and serous cystadenocarcinoma in an intermediate integration manner [23]. In addition, we have extended the previous framework to integrate genomic knowledge such as pathway or Gene Ontology [24]. The intermediate integration approach has an advantage that a model preserves data-specific properties by trying using optimally-weighted multiple graphs or kernel matrices transformed from multi-omics data as an intermediate level, compared to an early integration approach that combines input matrices before modelling. On the other hands, the late integration approach combines multiple predictive models by training multi-omics data individually in order to obtain the final model such as ensemble technique. The intermediate integration approach results into one prediction with one hypothesis, whereas the late integration approach has multiple independent hypotheses that have to be combined afterward. The strengths of our proposed framework as an intermediate integration approach were also highlighted as initiating its application using multiscale 'omics analytics, flexibility, and computation efficiency [25]. However, one of the disadvantages of intermediate integration approach is that it is difficult to consider inter-relationship between different levels of genomic data since each data is transformed before the integration as an individual intermediate level such as a graph.

There are possible multiple inter-relationships between different levels of genomic data such as 'copy number alteration region – genes located in the altered copy number region,' 'miRNA – its target genes,' and 'DNA methylation site – gene regulated by promoter regions,' etc (Fig. 1). In order to identify genes that are associated with gene dosage, many integrative analyses between copy number and gene expression have been conducted [26–29]. In addition, miRNA as one of the important regulators of gene expression can be integrated with gene expression to identify the selective inhibition of translation or selective degradation [30–32]. Furthermore, in terms of epigenetic regulation, histone modification or DNA methylation can serve to regulate gene expression in cancer [33–36]. The limitation of previous work was that we integrated multi-omics data for cancer clinical outcome prediction without considering inter-relationship between different levels of genomic features [23]. When integrating inter-relationship between different levels of genomic features, we assume that the prediction accuracy for cancer clinical outcome increases because of information fused over multiple genomic dataset and inter-relationship between them, providing an improved global view on interplays between different genomic levels in cancer mechanisms [11,37]. Thus, it will be desirable that a framework will be capable of containing the inter-relationship between different levels of genomic data when integrating multi-omics data.

In this study, we propose a new framework that integrates not only multi-omics data but inter-relationship between them in the intermediate integration manner for better elucidating cancer clinical outcomes. In order to highlight the validity of the proposed framework, serous cystadenocarcinoma data from TCGA was adopted as a pilot task. Serous cystadenocarcinoma is the most prevalent form of ovarian cancer, and is the 5th leading cause of cancer mortality in women in the United States [38]. Ovarian cancer patients are likely to be diagnosed with a late stage due to its asymptomatic nature, which are causing poor survival status [39]. Given multi-omics data, inter-relationships from one level to another may lead to some clues that help to uncover an unknown biological knowledge. Integrating multi-omics data and inter-relationship for predicting clinical outcomes will lead to better understand multi-layered genetic determinants of ovarian

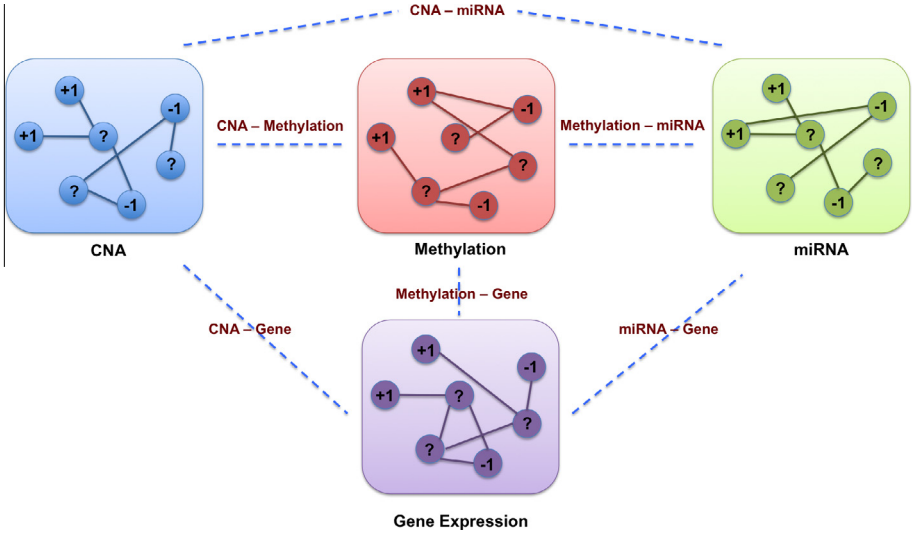

Fig. 1. Inter-relationships between different levels of genomic data. There are six possible inter-relationships between four types of genomic data including CNA, methylation, miRNA, and gene expression. Each genomic data can be converted into a graph where nodes represent patients and edges depict their similarities.

Table 1
Data description.

Data type	Platform	# Features
CNA	Agilent SurePrint G3 Human CGH Microarray Kit 1 × 1 M	54 cytobands
Methylation	Infinium Humanmethylation27 BeadChip	27,578 CpG loci
miRNA	Agilent Human miRNA Microarray Rel2.0	799 miRNAs
Gene expression	Affymetrix HT Human Genome U133 Array Plate Set	12,042 genes

cancer survival, further allowing for the possibility of leading alternative therapies that may improve outcomes.

2. Materials and methods

2.1. Data

Normalized multi-omics datasets in ovarian cancer were downloaded from TCGA data portal (http://tcga-data.nci.nih.gov/) (Table 1). Gene expression, miRNA, and methylation data contain 12,042 genes, 799 miRNAs, and 27,578 CpG loci, respectively. In order to use the results of altered regions of deletion or amplification across sets of patients, CNA data was retrieved from cBio Cancer Genomics Portal [40]. CNA data contains 54 significant cytoband regions. A binary classification of short-term and long-term survival was set for a pilot project. In the classification of *short-term or long-term survival*, 'short-term' represents the patients who survived less than 3 years, whereas 'long-term' indicates patients who survived longer than 3 years [41]. A total of 258 patients' records were available across the CNA, methylation, gene expression, and miRNA data sets ($N = 258$) with survival information, in which 110 were short-term survival and 148 were long-term survival.

2.2. Clinical outcome classification

We used a graph-based semi-supervised learning (SSL) as a classification algorithm, which is a halfway learning scheme between supervised and unsupervised learning [42–45]. One of the strengths of graph-based integration is its computational efficiency because of sparseness properties of input matrix while the accuracy remains comparable to the other methods such as kernel-based integration [46,47]. In addition, a graph-based SSL enjoys other advantages such as visualization, its relationship with kernel methods, solid mathematical background, and robust results in computational biology [48].

In this study, the common entity of each graph from multi-omics data is a patient (Fig. 1). If two patients were more closely related than to others, we assumed that clinical outcomes of those two patients would be more likely to be similar [4,49]. Thus, clinical outcome prediction can be conducted by considering similarities between patients based on their genomic profiles such as gene expression. Edges represent similarities between patients extracted from genomic profiles such as gene expression or methylation. An annotated patient is labeled either by '−1' or '1', indicating two possible clinical outcomes, either 'short-term survival' or 'long-term survival' (Fig. 2). In order to predict the label of the unannotated patient '?', the edges connected from/to the patient play an important role in influencing propagation between the patient and its neighbors. This idea can be easily formulated using graph-based SSL [45]. Technically, the data-setup of our experiment for the binary classification can be rephrased as $\{x_n, y_n\}_{n=1}^{N}$ where $x_n \in R^d$ (d is the number of features and N is the number of patients) and $y_n \in \{-1, 1\}$.

2.3. Graph-based SSL

Here, we present the formulation of the graph-based SSL. In the graph-based SSL, a patient x_i ($i = 1,\ldots,n$) is represented as a node i in a graph, and the relationship between patients is represented by an edge. The edge strength from node j to node i is encoded in element w_{ij} of a $n \times n$ symmetric weight matrix W. A Gaussian function of Euclidean distance between patients was used to state connection strength:

$$w_{ij} = \begin{cases} \exp\left(-\frac{(x_i - x_j)^T (x_i - x_j)}{\sigma^2}\right) & \text{if } i \sim j, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

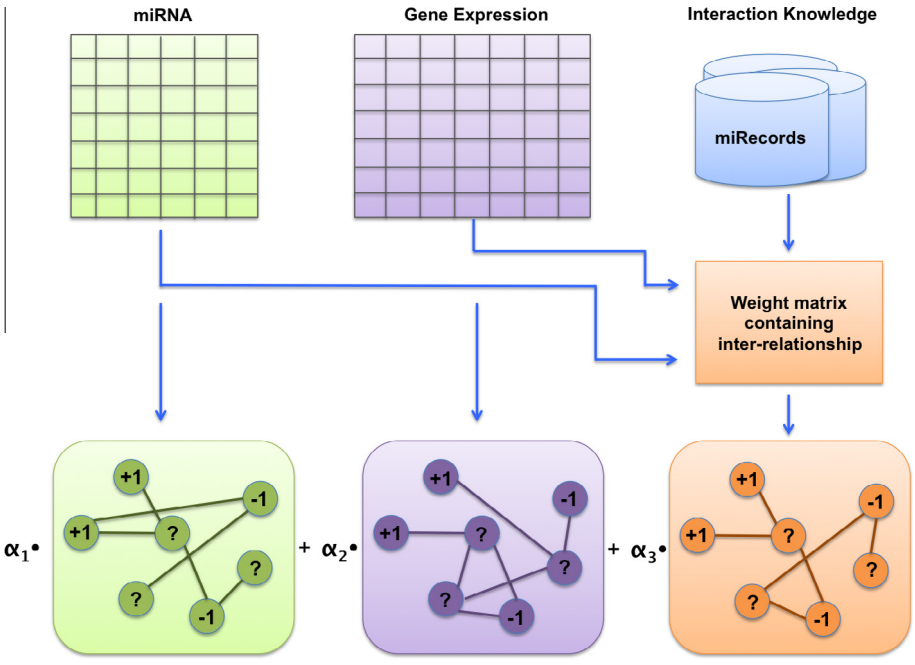

Fig. 2. Schematic overview of integrating different levels of genomic data and inter-relationship (e.g. miRNA and gene expression data).

Nodes i, j are connected by an edge if i is in j's k-nearest-neighborhood or vice versa. The labeled nodes have labels $y_l \in \{-1,1\}$, whereas the unlabeled nodes have zeros $y_u = 0$. An output of graph-based SSL is an n-dimensional real-valued vector $f = [f_l^T f_u^T]^T = (f_1,\ldots f_l, f_{l+1},\ldots f_{n=l+u})^T$, which can be thresholded to create label predictions on $f_l = f_1,\ldots f_n$ after learning. Graph-based SSL consists of two main conditions, which are loss condition and smoothness condition. It is assumed that f_i should be close to the given label y_i in labeled nodes as a loss condition, and overall, f_i should not be too different from the f_i of adjacent nodes as a smoothness condition. One can obtain f by minimizing the following quadratic functional [42,44,45]:

$$\min_f (f-y)^T(f-y) + \mu f^T L f, \tag{2}$$

where $y = (y_1,\ldots,y_l, 0,\ldots 0)^T$, and the matrix L, called the graph Laplacian matrix [50], is defined as $L = D - W$ where $D = \mathrm{diag}(d_i)$, $d_i = \sum_j w_{ij}$. The parameter μ trades off loss versus smoothness. The solution of this problem is obtained as

$$f = (I + \mu L)^{-1} y, \tag{3}$$

where I is the identity matrix.

2.4. Inter-relationship between different levels of genomic data

In order to overcome the disadvantage of intermediate integration, we create another graph that contains inter-relationship between different graphs (Fig. 2). In the Fig. 2 as an example of miRNA and gene expression, two graphs can be generated from miRNA and gene expression data, respectively. In addition, another graph containing inter-relationship between miRNA and its target genes can be generated as well. If information from inter-relationship graph is regarded as complementary to original graphs from miRNA or gene expression, the prediction accuracy through the integration three graphs will increase. Even though relations between miRNA and gene expression, between CNA and gene expression, or between methylation and gene expression have been investigated from several studies [26–36], to the best of our knowledge, there are few integrative studies between CNA and methylation, between CNA and miRNA, or between miRNA and methylation. In our study, nevertheless, we conducted all possible inter-relationships between four different genomic data, CNA, methylation, gene expression, and miRNA (Fig. 1).

(A) miRNA – Gene expression (RE)
(B) CNA – Gene expression (CE)
(C) Methylation – Gene expression (ME)
(D) CNA – Methylation (CM)
(E) CNA – miRNA (CR)
(F) Methylation – miRNA (MR)

In order to get inter-relationship between miRNA and gene expression (RE) from interaction knowledge, we used miRecords, which is integrated resources of miRNA that store target interactions produced by 11 established miRNA target prediction programs [51]. We created 10 variations for predicted target pairs between miRNA and its target genes by considering the number of positive voters from the included algorithms by miRecords in order to reduce false positives from 11 miRNA target prediction sources (Supplementary Table 1). Because most of the evaluation results from these variations were largely comparable, the most representative variation # 6 in Supplementary Table 1 was used to get inter-relationship between miRNA and its target genes for further study. For the inter-relationship between CNA and gene expression (CE), we used the chromosomal positional information of multiple genes in a specific CNA region since gene dosage from either deletion or duplication can affect gene expression. We considered a *cis* acting of methylation probe with respect to a given gene expression within 500 kb interval between methylation probe and genes for the inter-relationship between methylation and gene expression (ME) [52]. For the inter-relationships between CNA and methylation (CM) or between CNA and miRNA (CR), we also used the chromosomal positional information that the relation was set if miRNA or methylation probe are within CNA region. In addition, for the inter-relationship between methylation and miRNA (MR), an indirect mapping strategy through common genes was adopted because there is no known interaction knowledge between them. The relation between methylation and miRNA (MR) was set if a targeted gene from a specific miRNA is also shared by a specific methylation probe as a candidate cis-acting regulation.

2.5. Weight matrix incorporating inter-relationship

To calculate a weight matrix (W) containing inter-relationship between different levels of genomic data, we adopt a new measure that has been recently developed and re-validated in the previous study [53]. As an example of miRNA and gene expression data, let miRNA represent the miRNA data matrix of size N by N_{mi} and let gene denote the gene expression data matrix of size N by N_G, where N, N_{mi}, and N_G represent the number of patients, miRNAs, and gene expression traits, respectively. A new feature matrix X incorporating the miRNA-target gene information can be constructed by

$$X_{ij} = \sum_{m=1}^{N_G} \mathrm{miRNA}(i,j) \cdot \mathrm{gene}(i,m) \cdot \delta(j,m), \tag{4}$$

where $\delta(j,m) = 1$ if mth gene is targeted by jth miRNA, and 0 otherwise. After constructing the new matrix containing inter-relationship between two different types of genomic data, a Gaussian function of Euclidean distance between patients was used to calculate the final weight matrix using Eq. (1). Thus, nearby patients in Euclidean spaces are assigned large edge weights, which are likely to share similar inter-relationship pattern. This weight matrix containing inter-relationship can be used for graph-based SSL as an input, representing an inter-relationship graph. This approach was applied to other pairs including CE, ME, CM, CR, and MR.

2.6. Integrating multiple graphs

From different levels of genomic data and inter-relationship, multiple graphs can be generated (Fig. 2). However, clinical outcome prediction can benefit by integrating diverse graphs from multi-omics data and inter-relationship, rather than relying only on single level of genomic data that may have possible limitations, (i.e. incomplete information and noise). Information from each graph is regarded as partly independent from and partly complementary to others. When genomic data are presented as a graph form, integrating multi-omics data can be done by employing a graph integration method from finding optimum combination coefficients [23,46,54]. Based on the method, the integration of multiple graphs was conducted through finding an optimum coefficient of the linear combination for the individual graphs. This corresponds to finding the combination coefficients α for the individual Laplacians of the following mathematical formulation:

$$\min_{\alpha} y^T \left(I + \sum_{k=1}^{K} \alpha_k L_k \right)^{-1} y, \quad \sum_k \alpha_k \leqslant \mu, \tag{5}$$

where K is the number of graphs and L_k is the corresponding graph-Laplacian of graph G_k. Similar to the output prediction for single graphs, the solution is obtained by

$$f = \left(I + \sum_{k=1}^{K} \alpha_k L_k \right)^{-1} y. \quad (6)$$

3. Results and discussion

The receiver operating characteristic (ROC) curve plots sensitivity (true positive rate) as a function of 1-specificity (false positive rate) for a binary classifier system as its discrimination threshold is varied [55]. For each problem, we calculated area under the curve (AUC) of ROC as a performance measure. In order to avoid over-fitting, five-fold cross-validation was conducted. Because genomic data sources are normally high dimensional and noisy and contain many redundant features, which could incur computational difficulty and low accuracy, a t-test based feature selection method was used [56]. Even though there are many feature selection techniques such as filter, wrapper, and embedded method [57], a simple univariate feature selection method was used in order to emphasize not the effect of feature selection but the effect of integration with multi-omics data and inter-relationship between them in this study.

3.1. Selected features

In order to avoid the over-fitting, the feature selection was conducted using training dataset and repeated for five times. We set a 0.05 p-value threshold from t-test to fairly get selected features from different levels of genomic data. Among selected features from five times, we finally selected overlap features in order to make the weight matrix. The total numbers of final selected features of CNA, methylation, miRNA, and gene expression data are 2 CNA regions, 100 CpG loci, 25 miRNAs, and 99 genes, respectively. For constructing inter-relationship graph, we searched over possible pairs between selected features belonging to the different levels of genomic data based on interaction knowledge as described in the Method section. Due to the small number of selected CNA features, there were not enough pairs between CNA feature and other features. Thus, we searched over 54 CNA features rather than 2 features to have possible inter-relationship between CNA feature and other genomic features since 54 CNA regions were results of altered regions of amplification or deletion across sets of patients from GISTIC algorithm. The final numbers of pairs of CE, RE, ME, CM, CR, and MR are 13, 14, 202, 12, 2, and 22, respectively.

3.2. Integration with inter-relationship between different levels of genomic data

Fig. 3 shows the prediction performance on the classification of short-term and long-term survival in ovarian cancer for 6 types of pair, RE, CE, ME, CM, CR, and MR, between CNA, methylation, miRNA, and gene expression data. ROC curves can be found in the Supplementary information (Supplementary Fig. 1). For the pair between miRNA and gene expression data (RE), the SSL with miRNA data (R) and gene expression data (E) performed with AUCs of 0.6699 and 0.8191, respectively. In addition, the integration model with miRNA and gene expression data (RE) showed better performance with the AUC of 0.8404 than the model from the miRNA or gene expression data alone. In particular, the model integrating miRNA, gene expression, and inter-relationship between them with the AUC of 0.8476 (REI) even outperformed the model combining miRNA and gene expression data (RE) (Fig. 3(a)). Note that, similar results were also obtained for other pairs, CE, ME, CR, and MR except for CM (Fig. 3(b–f)). As we expected, ME and CE pairs also showed that the performance of model incorporating inter-relationship, MEI and CEI, increased compared to the one without inter-relationship, ME and CE, respectively (Fig. 3(b) and (c)). However, the performance of the integration model from CNA, methylation, and inter-relationship between them (CMI) showed worse than the model combining CNA and methylation data (CM) (Fig. 3(d)). Even though the integration model from CNA and miRNA (CR) showed the worse performance compared to the model with genomic data alone (R), the performance of the model incorporating inter-relationship (CRI) showed the best among any other models (Fig. 3(e)). Most interestingly, when incorporating inter-relationship between methylation and miRNA, the final integrated model (MRI) with the AUC of 0.7994 outperformed the integrated model (MR) with the AUC of 0.7865 (Fig. 3(f)). We found that the opportunity for success in prediction of clinical outcomes in ovarian cancerincreased when incorporating inter-relationship into the final SSL model.

3.3. Biological implication

On the bases of the results, the model incorporating inter-relationship between different genomic features showed the improvement compared to the model without inter-relationship when integrating multi-omics data. Taken together these results suggest that inter-relationship between difference levels of genomic data is complementary to the prediction power of explaining cancer clinical outcome. Through the proposed model, different levels of genomic features associated with survival in ovarian cancer were selected. In particular, selected features involved in inter-relationship might be regarded as acting important roles associated with survival in ovarian cancer. As different levels of genomic data including CNA, methylation, and miRNA might affect gene regulation through either specific or synergistic fashion, this approach will lead us to an enhanced global view on interplays between them [58,59].

In order to view the whole interplays from an inter-relationship network was generated using Cytoscape (Fig. 4) [60]. Nodes depict genomic features such as CNA, methylation, gene, or miRNA and edges represent inter-relationship between different genomic features based on interaction knowledge. Even though there are a few isolated networks that have one or two edges, many of genomic features are connected to each other as a big network (Fig. 4). This suggests that different levels of genomic features are not likely acting in isolation, but rather interact with other genomic features since cancer is dysregulated by multiple levels in the biological system through genomic, epigenomic, transcriptomic, proteomic level [11]. RE showed the largest number of edges from the network (Fig. 4). However, the number of edges is not biased to the prediction model with inter-relationship. MR, which has much smaller number of edges than the RE, showed the greatest improvement compared to the model without inter-relationship (Fig. 3). Among genes with a large number of edges, CHP gene is involved in MAPK signaling pathway that plays a critical role in the development and progression of ovarian cancer and other cancer types [61,62]. In addition, genetic and epigenetic regulation of the SLC22A3 gene associated with prostate cancer was well described [63] and the potential role of SLC22A3 as one of the members of FRA6E in ovarian cancer was also investigated [64]. EGF-like module containing mucin-like hormone receptor 2 (EMR2) is associated with survival in breast cancer [65]. Among miRNAs with a large number of edges, hsa-miR-146a is well known as a common mechanism of miRNA activity in cancer cell, which is trans-activated by the NF-kB pathway and negatively feeds back

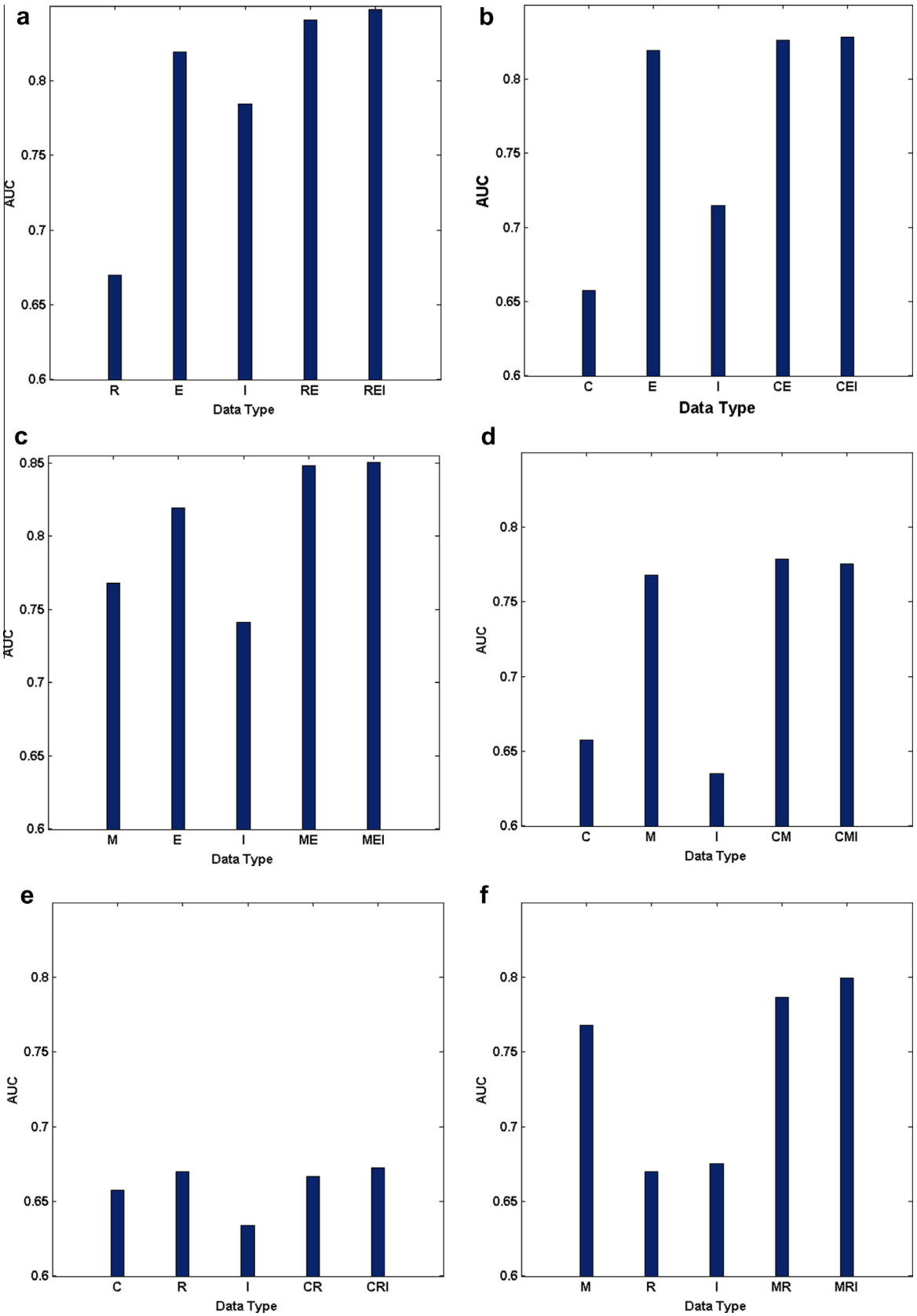

Fig. 3. Performance comparison between models: C stands for CNA, M for methylation, E for gene expression, R for miRNA, I for inter-relationship, RE for the integration model from miRNA and gene expression data, REI for the integration model from miRNA, gene expression, and inter-relationship (a) RE pair (b) CE pair (c) ME pair (d) CM pair (e) CR pair (f) MR pair.

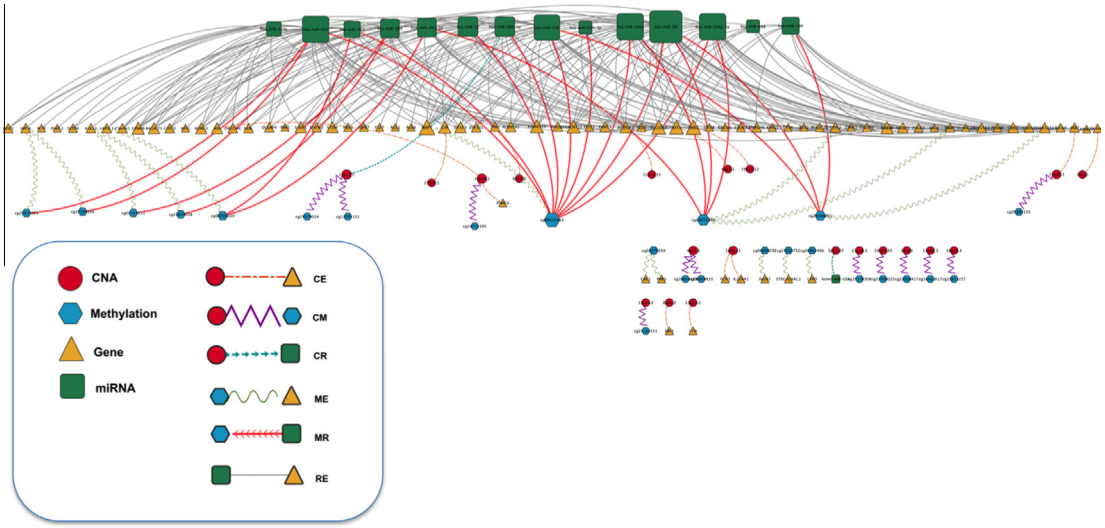

Fig. 4. Inter-relationship network from selected features involved in inter-relationship between different levels of genomic data. Nodes depict genomic features such as CNA, methylation, gene, or miRNA and edges represent inter-relationship between different genomic features based on interaction knowledge. Node size varies based on the degree of connectivity of the nodes. Edge shape is also different from the types of inter-relationship.

on this signaling cascade by targeting two upstream activators of the pathway, IRAK1 and TRAF6 [66].

Most interestingly, MR pair showed the great improvement by incorporating inter-relationship between miRNA and methylation through common genes (Fig. 3(f)). Thus, we created a network in order to see shared genes by methylation and miRNA (Fig. 5). Seven genes including RBBP7, CTPS2, KLK12, HPCA, UAP1L1, MAPT, and KIAA1033 were shared by miRNA and methylation. MAPT gene

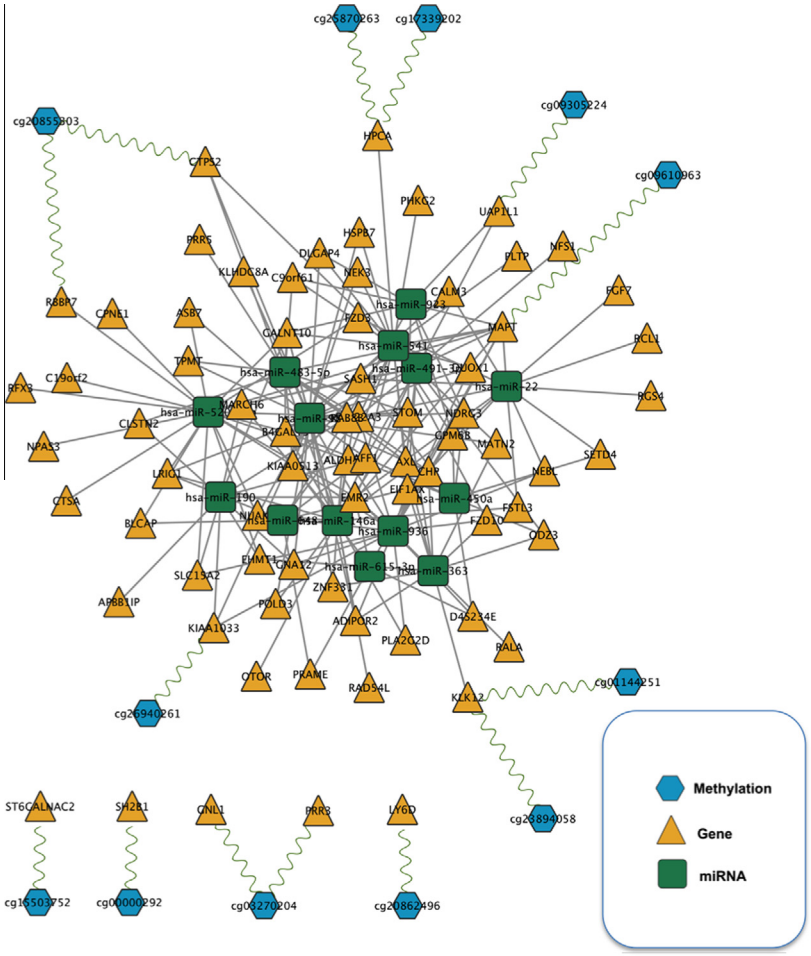

Fig. 5. Inter-relationship network from selected features for the inter-relationship between methylation and miRNA (MR) sharing common genes.

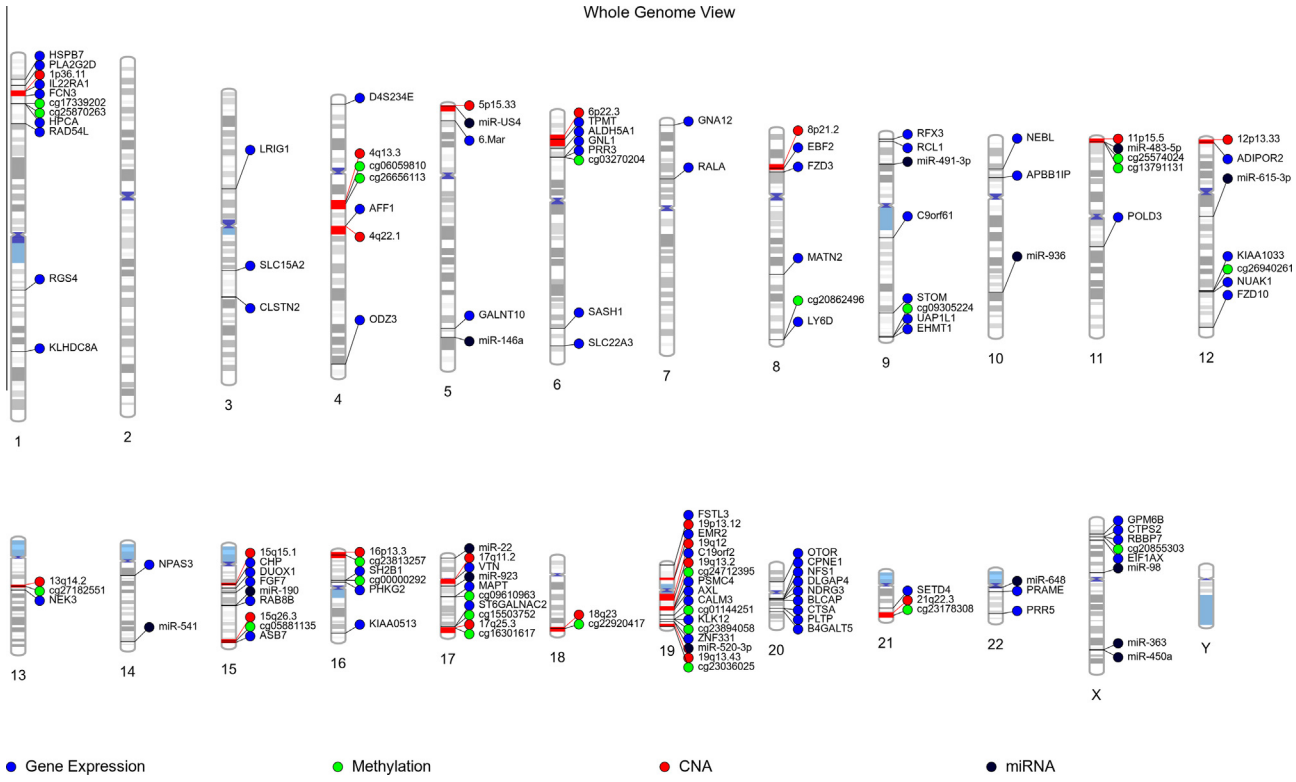

Fig. 6. Whole genome view of selected features using PhenoGram: blue circle stands for gene expression, green circle for methylation, red circle for CNA, and black circle for miRNA.

is also involved in MATK signaling pathway that is potential cancer therapy in ovarian cancer [67]. In addition, kallikrein gene 12 (KLK12) might be involved in the pathogenesis and progression of several cancer types and considered as a novel cancer biomarker [68].

Taken together these results suggest that there might be possible synergistic mechanism between methylation and miRNA regulation for the expression [69].

In order to provide the whole genome view, all selected features for the inter-relationship between four types of genomic data were plotted using PhenoGram visualization software (Fig. 6) [70]. In particular, chromosome 19 shows many of the features of genomic instability from different levels of genomic data, which was reported in the previous study [19]. One interesting possibility suggests that the same genomic loci might contribute clinical information in more than one domain – the same genes that change in their copy number, miRNA patterns, and methylation patterns also present predictive powers based on gene expression levels.

4. Conclusions

In this study, we addressed the issue of integrating inter-relationship between different levels of genomic data in the intermediated integration manner. We proposed the new framework that combines not only multi-omics data but inter-relationship between them in order to better predict cancer clinical outcomes. For demonstrating the validity of the proposed framework, ovarian cancer data from TCGA was adopted for classifying short-term and long-term survival as a pilot project.

On the bases of the results of our computational experiments, the model incorporating inter-relationship between different genomic features showed the modest improvement compared to the model without inter-relationship when integrating multi-omics data. We found that not only RE, CE, and ME pairs but CR and MR pairs showed the positive effect of integration with inter-relationship. These results suggest that inter-relationship between genomic features is complementary to the prediction power of explaining cancer phenotype because of the possible mechanisms that genomic features are likely to operate together in cancer. In addition, we could investigate the interplays associated with survival in ovarian cancer between different levels of genomic features through incorporating inter-relationship. Taken together these results suggest that proposed framework will lead us to an improved global view on interplays since different levels of genomic data might affect the cancer clinical outcome through either partly independent or partly complementary fashion. Notably, when combining inter-relationship between methylation and miRNA into the model, it showed the greatest improvement. It suggests that there might be possible synergistic regulatory mechanism between methylation and miRNA for gene expression. Thus, integration of different levels of data and inter-relationship between them can aid in extracting new biological knowledge by drawing an integrative conclusion from many pieces of information collected from diverse types of genomic data.

One of the limitations in the current study is that we only used the limited interactions knowledge such as ‘miRNA – its target genes’, ‘CNA regions – genes in the altered region’, ‘methylation – candidate genes targeted by methylation as cis acting regulation’, etc. We expect that the model integrating multi-omics data and inter-relationship will improve as long as the quality of interaction knowledge increases in the future. Moreover, another interesting direction for further works would be the incorporation of inter-relationship based on not interaction knowledge but interaction profiles as a data-driven approach. For instance, the direction of correlation between CNA and expression of genes in CNA region usually shows positive. On the other hands, the direction of correlations between gene expression and miRNA or between gene expression and methylation normally show negative. Even though

the current study is limited in predicting of short-term and long-term survival in ovarian cancer as a pilot task, the proposed framework can be applied to other clinical outcomes such as grade, stage, metastasis, recurrence, etc. Moreover, this framework can be applied to other cancer types as well for the future works. As multi-omics data from about 25 cancer types has exploded in use, our proposed framework will be valuable for explaining the underlying tumorigenesis, eventually leading to more effective screening strategies and therapeutic targets in many types of cancer. The Matlab code for graph-based semi-supervised learning will be available upon request.

Competing interests

All authors declared that there is no conflict of interest in this research.

Authors' contributions

DK and JHK designed and developed the study and wrote the manuscript. HS, KS, AV, and MDR provided the experimental results and interpreted the results. HS, MDR, and JHK provided intellectual guidance and mentorship. All authors read and approved the final manuscript.

Acknowledgements

This research was supported by a grant of the Korean Health Technology R&D Project, Ministry of Health and Welfare (H13C2164) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2010-0028631). MDR would like to gratefully acknowledge support by NIH grant 5R01 LM010040 and NHLBI grant 2U01 HL065962. In addition, we gratefully acknowledge the TCGA Consortium and all its members for the TCGA Project initiative, for providing samples, tissues, data processing and making data and results available.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2014.02.003.

References

[1] A. Berchuck, E.S. Iversen, J.M. Lancaster, J. Pittman, J. Luo, P. Lee, S. Murphy, H.K. Dressman, P.G. Febbo, M. West, et al., Clin. Cancer Res. 11 (10) (2005) 3686–3696.
[2] E. Huang, S.H. Cheng, H. Dressman, J. Pittman, M.H. Tsou, C.F. Horng, A. Bild, E.S. Iversen, M. Liao, C.M. Chen, et al., Lancet 361 (9369) (2003) 1590–1596.
[3] P. Roepman, L.F. Wessels, N. Kettelarij, P. Kemmeren, A.J. Miles, P. Lijnzaad, M.G. Tilanus, R. Koole, G.J. Hordijk, P.C. van der Vliet, et al., Nat. Genet. 37 (2) (2005) 182–186.
[4] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, et al., Nature 415 (6871) (2002) 530–536.
[5] X. Fan, L. Shi, H. Fang, Y. Cheng, R. Perkins, W. Tong, Clin. Cancer Res. 16 (2) (2010) 629–636.
[6] L.D. Wood, D.W. Parsons, S. Jones, J. Lin, T. Sjoblom, R.J. Leary, D. Shen, S.M. Boca, T. Barber, J. Ptak, et al., Science 318 (5853) (2007) 1108–1113.
[7] S. Myllykangas, J. Tikka, T. Bohling, S. Knuutila, J. Hollmen, BMC Med. Genomics 1 (2008) 15.
[8] J. Lu, G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando, et al., Nature 435 (7043) (2005) 834–838.
[9] M. Boeri, C. Verri, D. Conte, L. Roz, P. Modena, F. Facchinetti, E. Calabro, C.M. Croce, U. Pastorino, G. Sozzi, Proc. Natl. Acad. Sci. USA 108 (9) (2011) 3713–3718.
[10] R. Beroukhim, C.H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J.S. Boehm, J. Dobson, M. Urashima, et al., Nature 463 (7283) (2010) 899–905.
[11] S. Hanash, Nat. Rev. Cancer 4 (8) (2004) 638–644.
[12] L. Chin, J.W. Gray, Nature 452 (7187) (2008) 553–563.
[13] International Cancer Genome C, T.J. Hudson, W. Anderson, A. Artez, A.D. Barker, C. Bell, R.R. Bernabe, M.K. Bhan, F. Calvo, I. Eerola, et al., Nature 464 (7291) (2010) 993–998.
[14] Cancer Genome Atlas Research N, New Engl. J. Med. 368 (22) (2013) 2059–2074.
[15] Cancer Genome Atlas Research N, C. Kandoth, N. Schultz, A.D. Cherniack, R. Akbani, Y. Liu, H. Shen, A.G. Robertson, I. Pashtan, R. Shen, et al., Nature 497 (7447) (2013) 67–73.
[16] TCGA Network, Nature 487 (7407) (2012) 330–337.
[17] TCGA Network, Nature 489 (7417) (2012) 519–525.
[18] TCGA Network, Nature 490 (7418) (2012) 61–70.
[19] TCGA Network, Nature 474 (7353) (2011) 609–615.
[20] TCGA Network, Nature 455 (7216) (2008) 1061–1068.
[21] H. Noushmehr, D.J. Weisenberger, K. Diefes, H.S. Phillips, K. Pujara, B.P. Berman, F. Pan, C.E. Pelloski, E.P. Sulman, K.P. Bhat, et al., Cancer Cell 17 (5) (2010) 510–522.
[22] S. Srinivasan, I.R. Patric, K. Somasundaram, PLoS One 6 (3) (2011) e17438.
[23] D. Kim, H. Shin, Y.S. Song, J.H. Kim, J. Biomed. Inform. 45 (6) (2012) 1191–1198.
[24] D. Kim, J.G. Joung, K.A. Sohn, H. Shin, M.D. Ritchie, J.H. Kim, JAMIA (2013). in press.
[25] Y.A. Lussier, H. Li, J. Biomed. Inform. 45 (6) (2012) 1199–1201.
[26] R. Williams, J.E. Lim, B. Harr, C. Wang, R. Walters, M.G. Distler, M. Teschke, C.L. Wu, T. Wiltshire, A.I. Su, et al., PLoS One 4 (3) (2009).
[27] B.E. Stranger, M.S. Forrest, M. Dunning, C.E. Ingle, C. Beazley, N. Thorne, R. Redon, C.P. Bird, A. de Grassi, C. Lee, et al., Science 315 (5813) (2007) 848–853.
[28] L.D. Orozco, S.J. Cokus, A. Ghazalpour, L. Ingram-Drake, S. Wang, A. van Nas, N. Che, J.A. Araujo, M. Pellegrini, A.J. Lusis, Hum. Mol. Genet. 18 (21) (2009) 4118–4129.
[29] P. Cahan, Y. Li, M. Izumi, T.A. Graubert, Nat. Genet. 41 (4) (2009) 430–437.
[30] M. Rantalainen, B.M. Herrera, G. Nicholson, R. Bowden, Q.F. Wills, J.L. Min, M.J. Neville, A. Barrett, M. Allen, N.W. Rayner, et al., PLoS One 6 (11) (2011).
[31] J. Lu, A.G. Clark, Genome Res. 22 (7) (2012) 1243–1254.
[32] C. Borel, S. Deutsch, A. Letourneau, E. Migliavacca, S.B. Montgomery, A.S. Dimas, C.E. Vejnar, H. Attar, M. Gagnebin, C. Gehrig, et al., Genome Res. 21 (1) (2011) 68–73.
[33] E. Dudziec, A. Gogol-Doring, V. Cookson, W. Chen, J. Catto, PLoS One 7 (3) (2012) e32750.
[34] M. Li, C. Balch, J.S. Montgomery, M. Jeong, J.H. Chung, P. Yan, T.H. Huang, S. Kim, K.P. Nephew, BMC Med. Genomics 2 (2009) 34.
[35] J.R. Gibbs, M.P. van der Brug, D.G. Hernandez, B.J. Traynor, M.A. Nalls, S.L. Lai, S. Arepalli, A. Dillman, I.P. Rafferty, J. Troncoso, et al., PLoS Genet. 6 (5) (2010).
[36] J.T. Bell, A.A. Pai, J.K. Pickrell, D.J. Gaffney, R. Pique-Regi, J.F. Degner, Y. Gilad, J.K. Pritchard, Genome Biol. 12 (6) (2011).
[37] C.M. Croce, New Engl. J. Med. 358 (5) (2008) 502–511.
[38] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, M.J. Thun, CA Cancer J. Clin. 59 (4) (2009) 225–249.
[39] S.A. Cannistra, New Engl. J. Med. 351 (24) (2004) 2519–2529.
[40] E. Cerami, J. Gao, U. Dogrusoz, B.E. Gross, S.O. Sumer, B.A. Aksoy, A. Jacobsen, C.J. Byrne, M.L. Heuer, E. Larsson, et al., Cancer Discov. 2 (5) (2012) 401–404.
[41] A.H. Bild, G. Yao, J.T. Chang, Q. Wang, A. Potti, D. Chasse, M.B. Joshi, D. Harpole, J.M. Lancaster, A. Berchuck, et al., Nature 439 (7074) (2006) 353–357.
[42] O. Chapelle, J. Weston, B. Scholkopf, Adv. Neural Inform. Process. Syst. (NIPS) 15 (15) (2003) 585–592.
[43] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in: Proceedings of the Twenty-first International Conference on Machine Learning (ICML), AAAI Press, Washington, DC, 2003, pp. 912–919.
[44] M. Belkin, in: Proceedings of the 17th Annual Conference on Learning Theory (COLT) 3120 Lecture Notes in Computer Science, 2004, pp. 624–638.
[45] D. Zhou, O. Bousquet, J. Weston, B. Scholkopf, Adv. Neural Inform. Process. Syst. (NIPS) 16 (2004) 321–328.
[46] K. Tsuda, H. Shin, B. Scholkopf, Bioinformatics 21 (Suppl. 2) (2005) ii59–65.
[47] H. Shin, K. Tsuda, in: Olivier Chapelle, Bernhard Schölkopf, Alexander Zien (eds.), Semi-Supervised Learning, MIT press, 2006, pp. 339–352 (Chapter 20).
[48] T. Aittokallio, B. Schwikowski, Brief. Bioinform. 7 (3) (2006) 243–255.
[49] A. Gottlieb, G.Y. Stein, E. Ruppin, R.B. Altman, R. Sharan, BMC Med. 11 (2013) 194.
[50] F.R.K. Chung, Spectral Graph Theory, Number 92 in Regional Conference Series in Mathematics, 1997.
[51] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, T. Li, Nucleic Acids Res. 37 (Database issue) (2009) D105–110.
[52] K.R. van Eijk, S. de Jong, M.P. Boks, T. Langeveld, F. Colas, J.H. Veldink, C.G. de Kovel, E. Janson, E. Strengman, P. Langfelder, et al., BMC Genomics 13 (2012) 636.
[53] D. Kim, H. Shin, J.G. Joung, S.Y. Lee, J.H. Kim, BMC Syst. Biol. (2013), http://dx.doi.org/10.1186/1752-0509-7-S3-S8.
[54] H. Shin, A.M. Lisewski, O. Lichtarge, Bioinformatics 23 (23) (2007) 3217–3224.
[55] M. Gribskov, N.L. Robinson, Comput. Chem. 20 (1) (1996) 25–33.
[56] P. Jafari, F. Azuaje, BMC Med. Inform. Decis. Mak. 6 (2006) 27.
[57] Y. Saeys, I. Inza, P. Larranaga, Bioinformatics 23 (19) (2007) 2507–2517.
[58] K.A. Sohn, D. Kim, J. Lim, J.H. Kim, BMC Syst. Biol. 7 (Suppl. 6) (2013) S9.

[59] R. Louhimo, S. Hautaniemi, Bioinformatics 27 (6) (2011) 887–888.
[60] R. Saito, M.E. Smoot, K. Ono, J. Ruscheinski, P.L. Wang, S. Lotia, A.R. Pico, G.D. Bader, T. Ideker, Nat. Methods 9 (11) (2012) 1069–1076.
[61] Z. Cao, L.Z. Liu, D.A. Dixon, J.Z. Zheng, B. Chandran, B.H. Jiang, Cell. Signal. 19 (7) (2007) 1542–1553.
[62] A.S. Dhillon, S. Hagan, O. Rath, W. Kolch, Oncogene 26 (22) (2007) 3279–3290.
[63] L. Chen, C. Hong, E.C. Chen, S.W. Yee, L. Xu, E.U. Almof, C. Wen, K. Fujii, S.J. Johns, D. Stryke, et al., Pharmacogenomics J. 13 (2) (2013) 110–120.
[64] S.R. Denison, G. Callahan, N.A. Becker, L.A. Phillips, D.I. Smith, Genes Chromosom. Cancer 38 (1) (2003) 40–52.
[65] J.Q. Davies, H.H. Lin, M. Stacey, S. Yona, G.W. Chang, S. Gordon, J. Hamann, L. Campo, C. Han, P. Chan, et al., Oncol. Rep. 25 (3) (2011) 619–627.
[66] K.D. Taganov, M.P. Boldin, K.J. Chang, D. Baltimore, Proc. Natl. Acad. Sci. USA 103 (33) (2006) 12481–12486.
[67] L. Santarpia, S.M. Lippman, A.K. El-Naggar, Expert Opin. Ther. Targets 16 (1) (2012) 103–119.
[68] G.M. Yousef, A. Magklara, E.P. Diamandis, Genomics 69 (3) (2000) 331–341.
[69] Y.H. Taguchi, BioData Min. 6 (1) (2013) 11.
[70] D. Wolfe, S. Dudek, M.D. Ritchie, S.A. Pendergrass, BioData Min. 6 (1) (2013) 18.