Stage-Dependent Gene Expression Profiling in Colorectal Cancer

Man-Sun Kim, Dongsan Kim and Jeong-Rae Kim

Abstract— Temporal gene expression profiles have been widely considered to uncover the mechanism of cancer development and progression. Gene expression patterns, however, have been analyzed for limited stages with small samples, without proper data pre-processing, in many cases. With those approaches, it is difficult to unveil the mechanism of cancer development over time. In this study, we analyzed gene expression profiles of two independent colorectal cancer sample datasets, each of which contains 556 and 566 samples, respectively. To find specific gene expression changes according to cancer stage, we applied the linear mixed-effect regression model (LMER) that controls other clinical variables. Based on this methodology, we found two types of gene expression patterns: continuously increasing and decreasing genes as cancer develops. We found that continuously increasing genes are related to the nervous and developmental system, whereas the others are related to the cell cycle and metabolic processes. We further analyzed connected sub-networks related to the two types of genes. From these results, we suggest that the gene expression profile analysis can be used to understand underlying the mechanisms of cancer development such as cancer growth and metastasis. Furthermore, our approach can provide a good guideline for advancing our understanding of cancer developmental processes.

Index Terms— Protein-protein interaction network; TCGA data; gene expression profile; cancer developmental process; disease free survival; linear mixed-effect regression model

1 Introduction

Cancer is a complex disease that involves a sequence of gene-environment interactions in a progressive process that is derived from dysfunction in multiple systems, including DNA repair and immune functions [1]. Among them, colorectal cancer is a leading cause of cancer-related death in many countries [2, 3]. Colorectal cancer is a progressively developing disease with well-defined mutational sequences from adenomatous polyps to malignant transformations. Therefore, it is crucial to analysis on the changes of genomic programs according to cancer stages to understand the cancer development [4-6].

Previous work has focused on the correlation between cancer stage and individual gene [7-9]. For instance, Kasem *et al*. found that *FAM134B* expression was negatively correlated with cancer stage [7]. They found that *FAM134B* was also involved in patient age, cancer recurrence and clinical outcome of colorectal cancer. Some studies also investigated the relationship between microRNA expression and cancer stage. They found that various microRNAs were associated with progression of colorectal cancer [8]. Fang Yao, *et al*. identified correlation between gene expression levels and histological grades of breast cancer [9]. However, there was no such systematic comparison between gene expression profiles and cancer stages by considering various confounding factors that might affect gene expression profiles such as age or gender.

In this study, we investigated the correlation between cancer stages and gene expression profiles from TCGA in order to identify how gene expression changes as cancer evolves. In this step, it is necessary to minimize the effect of other factors than cancer stage, in order to specifically investigate the change of genes solely depending on the stage. To this end, we used the linear mixed-effect regression model (LMER) [10] and compared the model to simple linear model (LM). As a result, we found that continually increasing genes along the colorectal cancer development are related to the nervous and developmental system whereas continually decreasing genes are related to the cell cycle and metabolism. We further applied the same procedure to the independent gene expression profiles [11] and validated the findings from TCGA. In conclusion, the continually increasing or decreasing genes are robustly observed from the two independent datasets and they were found to play an important role in the process where the degree of malignancy is increased.

Based on LMER, we propose that the temporal gene expression profile analysis can be effectively used to understand the underlying mechanisms of cancer development. Ultimately, we also suggest that understanding the signaling aberration during cancer development can be applied to cancer therapeutics and provide insights into the progression from tumor malignancy.

2 Materials and Methods

A conceptual illustration for our methodology is given in Fig. 1.

- *M.-S. Kim. is with the Department of Horticulture, College of Agriculture and Life Sciences, Chungnam National University, Daejeon, 34134, Republic of Korea. E-mail: kms77@cnu.ac.kr.*
- *D. Kim is with the Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea. E-mail: dongsan@kaist.ac.kr.*
- *J.-R. Kim is with the Department of Mathematics, University of Seoul, 163 Siripdaero, Dongdaemun-gu, Seoul 130-743, Republic of Korea. E-mail: jrkim@uos.ac.kr.*

Manuscript received xx mmm. 2017; revised xx mmm 2017; accepted xx mmme 2017. Date of publication xx mmm. 2017; date of current version xx mmm. 2017. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below Digital Object Identifier no. 10.1109/TCBB.2017.0000000

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

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/TCBB.2018.2814043, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Firstly, we collected gene expression profiles of colorectal cancer patients from the TCGA database and performed statistical analysis based on the data. We compared three clinical variables (stage, gender and age) with gene expression profiles and found that stage was the most influential variables on gene expression (Fig. 2a and Supplementary Fig. 1a) although the other clinical variables also impacted on gene expressions. Therefore, we applied the LMER to the data to remove the influence of the other factors. After then, we identify two types of gene expression patterns: continually increasing and decreasing genes according to cancer stage. Finally, we analyzed their functional and structural properties by using the Gene Ontology enrichment test [12] and KEGG pathway enrichment test [13] and by reconstructing a relevant molecular sub-network of the protein-protein interaction network [14].

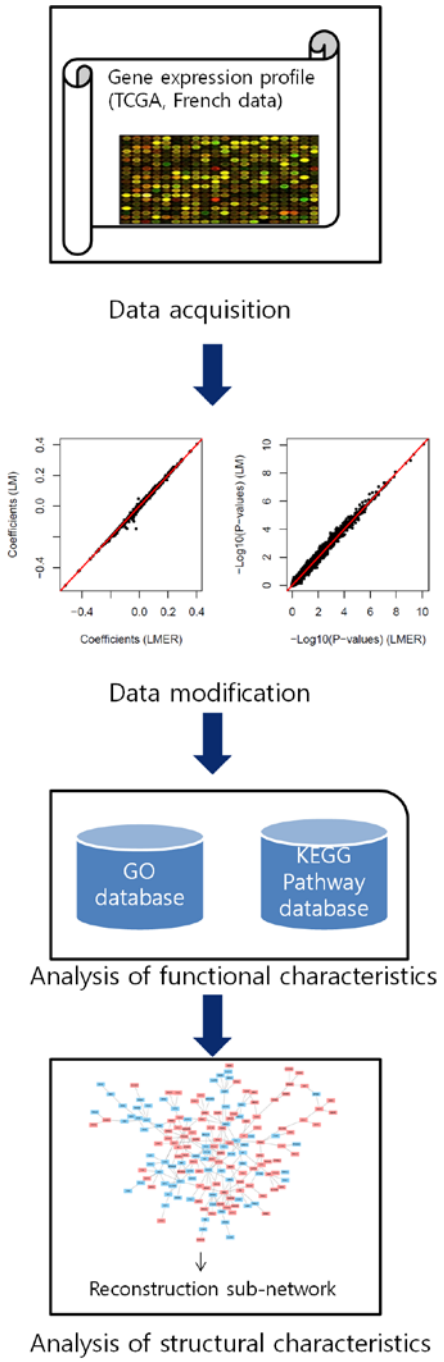

Fig. 1. The conceptual illustration of our methodology. After we obtained the P-values based on LMER, we analyzed the functional enrichment of the genes whose expression levels are positively or negatively correlated with cancer stage using GO and KEGG databases. Finally, we identified sub-networks which are composed of the (positively or negatively, respectively) stage-dependent genes using the PPI network and analyzed the structural characteristics of the networks.

2.1 Gene expression profiles and clinical variables of colorectal cancer patients.

We obtained an RNA sequencing dataset of colorectal cancer patients from TCGA. From the dataset, we removed normal samples and samples without any clinical variables. We obtained fragments per kilobase million (FPKM) normalized data and finally obtained log2 transformed gene expression profiles. Ensemble genes were collapsed into unique HUGO gene symbols based on collapseRows in R. Finally, we analyzed 24,504 genes and 554 samples in total (hereafter referred to as 'TCGA'). To validate the findings in TCGA, we further obtained robust multi-array average normalized gene expression profiles based on Affymetrix U133plus2 platform from the French Ligue Nationale Contre le Cancer (GSE39582) [11]. We also removed normal samples. Probes were collapsed into unique HUGO gene symbols based on collapseRows in R. This dataset is composed of 20,514 genes and 566 samples in total (hereafter referred to as 'French').

2.2 Statistical methods

All analyses were undertaken using the R platform for statistical computing (version 3.1.0) [15]. We fitted the LMER in R with the *lmer* function in the lme4 package [16]. The coefficients between cancer stages and gene expression levels were computed and the significance was estimated based on t-statistics. We computed P-values adjusted for multiple testing by controlling the false discovery rate (FDR) with the Benjamini-Hochberg procedure [17] in R and used a threshold of 0.05.

2.3 Gene Ontology (GO) and KEGG pathway enrichment analyses

To investigate the functional properties of the selected genes, we utilized the two databases (GO and KEGG). First, the Gene Ontology (Gene Ontology Consortium 2004) [18] consists of three types of information: cellular component, molecular function and biological process. GO annotation provides an indication of the trait of a group of genes [19]. GO functional annotations in this study were obtained from the GO database. We evaluated the statistical significance of the overlap between the selected genes through GO terms enrichment using clusterProfiler in R [12]. P-values were computed based on a hypergeometric distribution.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) consists of the information about molecular interaction networks, genes, proteins, chemical compounds, and chemical reactions [13]. We also evaluated the statistical significance of the overlap between the selected genes and the genes in each pathway using clusterProfiler in R [20]. P-values were computed based on a hypergeometric distribution.

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/TCBB.2018.2814043, IEEE/ACM Transactions on Computational Biology and Bioinformatics

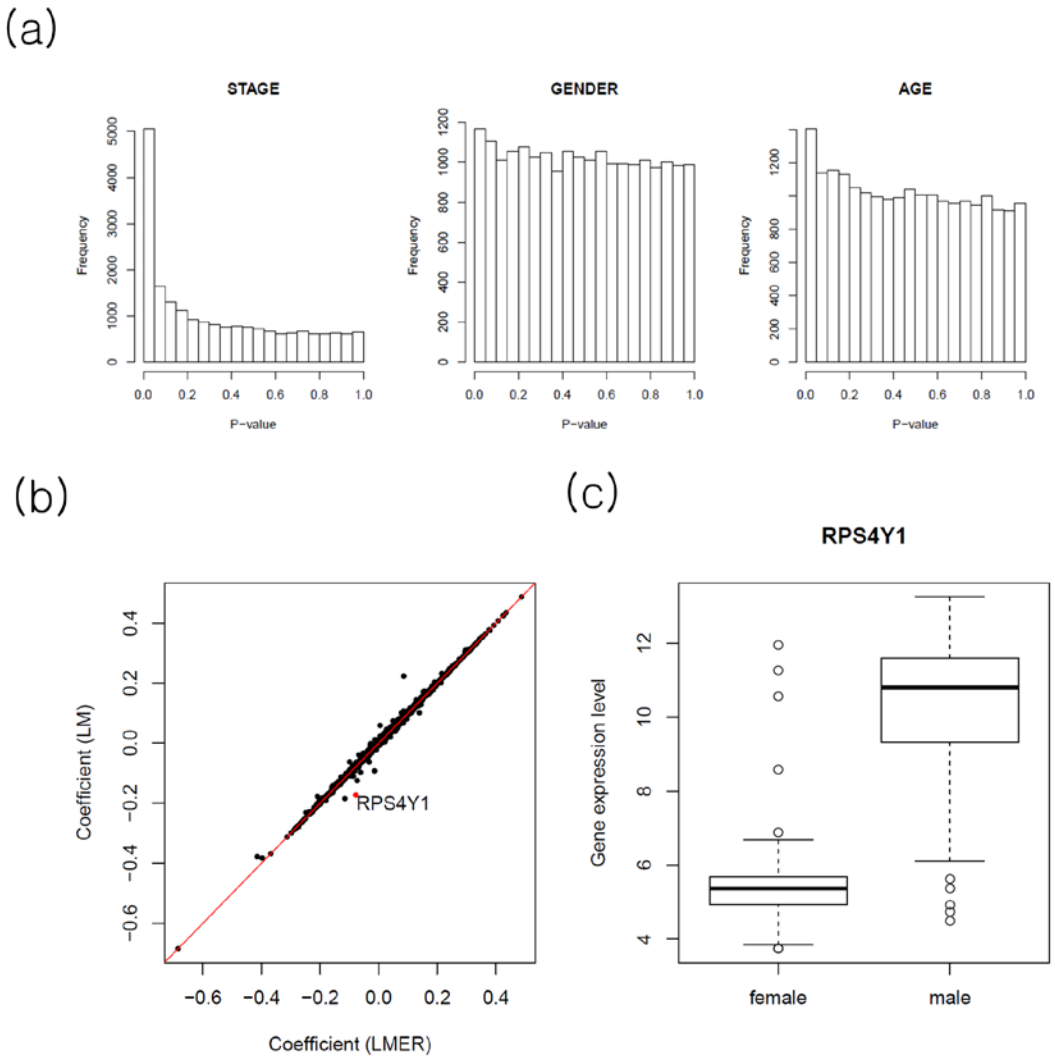

Fig. 2. (a) Distributions of correlation coefficients between gene expressions and each of three clinical factors. Here we used P-values corresponding to correlation coefficients. This result shows that the stage factor (I, II, III, and IV) of colorectal cancer has the most significant effect on monotone (increasing or decreasing) gene expression patterns. (b) Coefficients between cancer stages and gene expression levels based on linear models (LMs) and linear mixed-effect regression models (LMERs). (c) A representative gene (RPS4Y1) that shows different coefficient between LM and LMER. This gene has significantly different gene expression patterns according to gender among colorectal cancer patients. This gene has significantly different gene expression patterns according to gender among colorectal

2.4 Protein-protein Interaction (PPI) network

In order to reconstruct the PPI network of the selected genes, we used HPRD (Human Protein Reference Database, Release 9, 2010) [14]. After excluding self-interactions, there remained 37,080 binary PPIs involving 9,465 genes. In order to reconstruct a largest connected sub-network corresponding to the two gene sets, we used the following rules: if two genes connected by an edge in the PPI network are also included in the gene set, then both of the genes and the connecting link remained from the PPI network, and finally we obtained the largest connected sub-network among the remained PPI network.

3 Results and Discussion

3.1 Computing association between cancer stage and gene expression level using linear-mixed effect regression model

An examination of gene expression changes depending on the stage of colorectal cancer can significantly contribute to understandings of incidence and developmental process of cancer [4, 6]. However, the difference in gene expressions among patients with cancers may be associated with not only the stage of a cancer, but also many other factors such as gender, age, etc. and the factors' effects on each kind of gene expression level are

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/TCBB.2018.2814043, IEEE/ACM Transactions on Computational Biology and Bioinformatics

varied. For instance, there is a finding that the mortalities of male patients with breast cancer are considerably lower than those of female patients with the same cancer [2]. In addition, the occurrence frequencies of cancer are different according to age: it has been reported that each incidence of leukemia, thyroid carcinoma and gastric cancer would be increased at young, middle and old age, respectively [21]
.

Table 1. The GO and KEGG analyses results for INC and DEC type genes.

Type	ID	Description	P-value (TCGA)	P-value (French)
INC	GO:001525	angiogenesis	0.000214013	2.95E-12
	GO:001704	formation of primary germ layer	0.000743754	1.56E-07
	GO:010975	regulation of neuron projection development	1.33E-06	2.12E-06
	GO:022604	regulation of cell morphogenesis	0.000158596	3.05E-06
	GO:048514	blood vessel morphogenesis	0.000245984	1.56E-13
	GO:051216	cartilage development	0.000777562	1.61E-09
	GO:061448	connective tissue development	0.000814961	3.37E-10
	GO:061564	axon development	2.13E-06	7.05E-07
	hsa 05414	Dilated cardiomyopathy (DCM)	0.000119914	4.21E-07
	hsa 05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.000345959	0.000371123
	hsa 05410	Hypertrophic cardiomyopathy (HCM)	0.00108833	1.99E-05
	hsa 04360	Axon guidance	0.00253445	0.015690297
	hsa 04261	Adrenergic signaling in cardiomyocytes	0.002902014	0.034460936
DEC	GO:0033014	tetrapyrrole biosynthetic process	2.458794e-07	1.295534e-05
	GO:0017004	cytochrome complex assembly	3.464272e-06	1.681904e-05
	GO:0009062	fatty acid catabolic process	1.615264e-06	1.728045e-05
	GO:0008535	respiratory chain complex IV assembly	5.060928e-04	2.200430e-05
	GO:0072329	monocarboxylic acid catabolic process	6.175800e-07	2.200571e-05
	GO:0046935	carboxylic acid catabolic process	2.592615e-07	3.375507e-05
	GO:0044282	small molecule catabolic process	2.705056e-09	4.071183e-05
	GO:0016054	organic acid catabolic process	2.884503e-07	5.484548e-05
	hsa 04146	Peroxisome	0.002989731	0.004334261
	hsa 00860	Porphyrin and chlorophyll metabolism	0.004012654	0.002844474
	hsa 00970	Aminoacyl-tRNA biosynthesis	0.005079422	1.169021e-09
	hsa 00480	Glutathione metabolism	0.012560292	0.000626897
	hsa 00071	Fatty acid degradation	0.021762159	0.004983422
	hsa 00280	Valine, leucine and isoleucine degradation	0.03050471	0.000720611
	hsa 01212	Fatty acid metabolism	0.03050471	0.004983422
	hsa 00670	One carbon pool by folate	0.035236276	0.017158322
	hsa 04110	Cell cycle	0.038489352	1.68E-06

In order to quantitatively examine the effects of clinical factors on gene expression levels, each correlation between three crucial clinical factors (stage, age and gender) and individual gene expression is separately analyzed (Fig. 2a and Supplementary Fig. 1a). These results showed that the stage of colorectal cancer has the most significant effect on gene expressions among the three factors. The tendency to change the gene expression levels depending on age was relatively weaker than stage. The effect on gene expression changes by gender occurs only for several genes located at the sex chromosomes (Fig. 2b and Supplementary Fig. 1b). These findings coincide with previous literature studies that there were no gender difference between colorectal cancer patients [22, 23]. From the findings, we can conclude that analyzing the differentially expressed genes according to the cancer stage is the most important in understanding the molecular differences between patients. However, the other factors also affect the gene expressions even though the effects are relatively low. Hence it is necessary to consider the effect of other factors when we explore gene expression differences depending on the stage. To this end, we applied LMER [10] and compared the results with those based on LM. From the comparison, there are small number of genes that show large difference between LMER and LM. Among them, RPS4Y1 gene showed the largest difference between the two models. Based on LM, RPS4Y1 gene seems to be negatively associated with cancer stage, however, such association disappeared based on LMER (Fig. 2b). We note that the negative association between RPS4Y1 and cancer stage might due to gender bias of the gene expression level (Fig. 2c and Supplementary Fig. 1c). In conclusion, we could identify genes that are dependent only on cancer stage by correction of age and gender using the LMER.

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/TCBB.2018.2814043, IEEE/ACM Transactions on Computational Biology and Bioinformatics

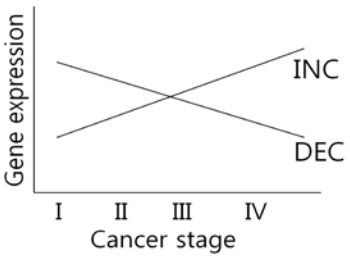

Fig. 3. An illustration of INC and DEC type genes.

3.2 Functional characteristics of the stage-dependent genes

Based on the LMER analysis, we found that the computed coefficients between gene expression levels and cancer stages based on TCGA dataset consistently hold for those based on French dataset (Supplementary Fig. 2). From the result, we could consider two types of genes whose expression levels are linearly dependent on stage (Fig. 3). One (INC type, 817 and 907 genes from TCGA and French dataset, respectively) is a continually increasing gene expression pattern and the other (DEC type, 579 and 1,350 genes from TCGA and French dataset, respectively) is a continually decreasing gene expression pattern (Supplementary Table S1). Since genes with similar expression pattern, in general, share functional characteristics, we examined the functional characteristics of the two types of genes based on the GO and KEGG pathway analysis. The analysis results from each dataset were listed in Supplementary Table S2. Among the findings, the significant GO terms or KEGG pathways from both datasets were listed in Table 1.

In order to further understand the roles of the two types of genes, we obtained gene families from Molecular signatures database (MSigDB) [24]. The gene families are categorized into eight groups: oncogenes, tumor suppressors, translocated cancer genes, transcription factors, protein kinases, homeodomain proteins, cell differentiation markers and cytokines/growth factors.

3.2.1 INC type genes

We found that the INC type genes are highly related to various developmental processes and are likely to activate the nerve system (Table 1). From the GO term enrichment analysis, we found the neuron related GO terms (GO 48667, GO 10975, GO 97485, GO 7411, GO 10977, GO 10976 and GO 50768), developmental system-related terms (GO 61564, GO 51961, GO 10721, GO 60485, GO 51216, GO 61448 and GO 16358), differentiation-related terms (GO 48667, GO 10769 and GO 10771) are enriched in INC type genes. The KEGG pathway analysis result includes many cardiomyopathy related pathways (hsa 05414, hsa 05412, hsa 05410 and hsa 04261). We suggest that these pathway activities are associated with the increment of cancer associated fibroblast or other fibrous connective tissue development in colorectal cancer samples [25, 26] .

Among the INC type genes, NGF (Nerve growth factor) related developmental processes were the most prominent. Why is the NGF expression increasing while the cancer is developing? We suggest the reason behind the expression of genes related to neuron increased along with cancer stage might be originated from genomic instability. The genomic instability can activate nerve system-related molecular pathways through NGF that should be turned off from normal colon cells. Finally, such genomic instability can contribute to the gene expression heterogeneity of cancer cells. Surprisingly, recent studies identified that generation of neurons called neurogenesis is observed from colorectal cancer patients [27]. They also found that patients with high degree of neurogenesis tend to obtain later cancer stage and worse clinical outcome. In addition, another study found that colorectal cancer cell can differentiated into neurons [28]. These findings support our result that colorectal cancer samples contain neuronal gene expression programs as cancer develops.

By comparing between the eight gene families and INC type genes, we found that transcription factors are enriched in the INC type genes. One of these genes, CDK8, which encodes a member of the mediator complex, is located at 13q12.13, a region of recurrent copy number gain in a substantial fraction of colon cancers. It is also well known that CDK8 is a colorectal cancer oncogene that regulates β-catenin activity [29]. In addition, SOX13 was known as a potential biomarker of colorectal cancer metastasis [30].

3.2.2 DEC type genes

We found that the DEC type genes are related with malfunctions of cell cycle and metabolic processes (Table 1). From the GO term enrichment analysis, most of the identified terms are catabolism related GO terms (GO 48667, GO 10975, GO 97485, GO 7411, GO 10977, GO 10976 and GO 50768) are enriched in DEC type genes. The KEGG pathway analysis result includes many metabolic pathways (hsa 00860, hsa 05412, hsa 05410 and hsa 04261) and cell cycle pathway (hsa 04110).

In addition, some genes related with the metabolic function showed the most significantly negative correlations with cancer stages (AGPAT5, FECH, LAP3, NAT1, and TYMS). Main pathways and genes related to those include cytokine receptor families-related IL2, IL7, IL12, oxidative phosphorylation-related ATP5A1 and cytokine receptor interaction-related CXCL (1, 2, 3, 9, 10, 11 and 13), and TNF Receptor Superfamily Member (TNFRSF 9, 10, 11 and 18). Here, TNF (tumor necrosis factor) is known to worsen inflammation and contribute to increase of cancer cell's oxidative stress. The reduction of such gene expressions may imply the degradation of their functions. In addition, the NOS2 which is known to be related to a very important expression control was identified. It serves to suppress the angiogenesis in the colorectal cancer. In summary, we found that various metabolic activities are weakened as colorectal cancer develops.

By comparing between the eight gene families and DEC type genes, we found cell differentiation markers are enriched in the DEC type genes. In particular, most of the cluster of differentiation (CD) genes were in the T cell receptor signaling pathway, which suggests that they can effectively mediate the interactions between cancer cells and immune system. We further found TAP1 gene, which was related to antigen processing. This finding suggests that colorectal cancer cells might lose immunological responsiveness, and therefore lose antitumor inflammatory response according to cancer development [31].

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/TCBB.2018.2814043, IEEE/ACM Transactions on Computational Biology and Bioinformatics

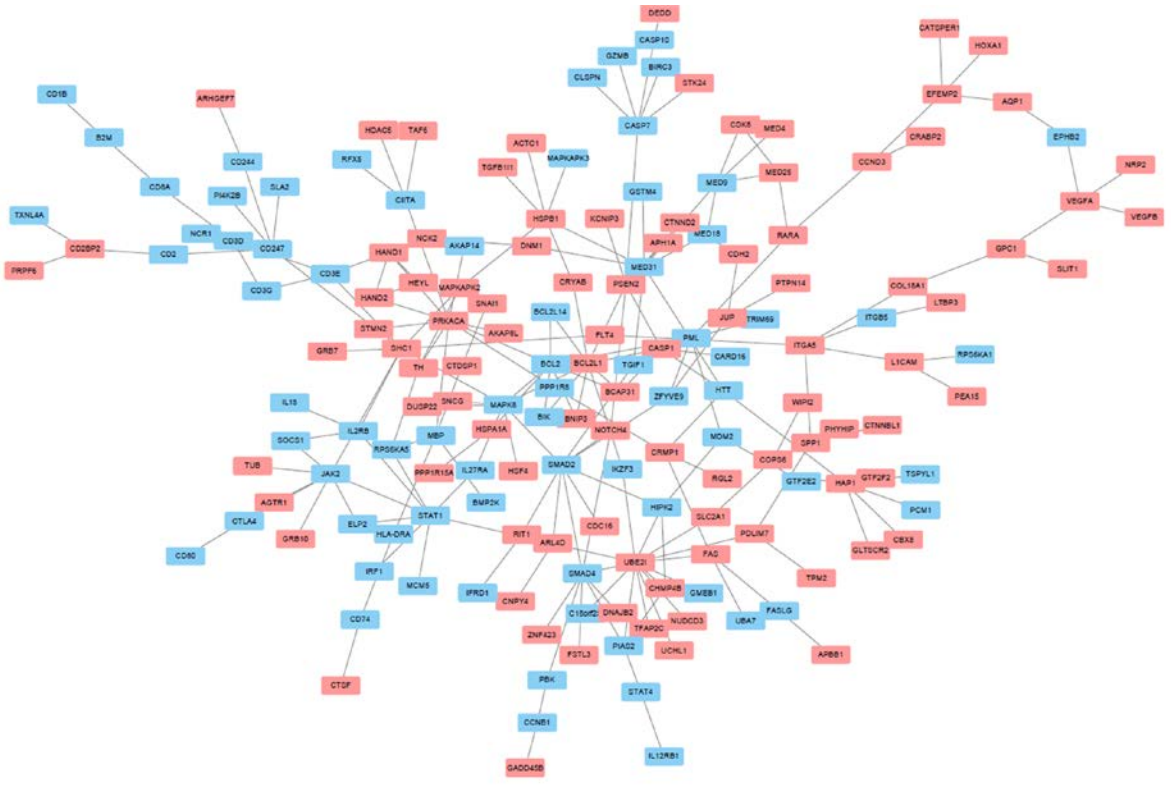

Fig.4. The network consists of 170 nodes and 210 links. Red (INC type genes) and blue (DEC type genes) rectangles denote the nodes in largest connected sub-network.

3.3 Topological characteristics of the stage-dependent genes

The network visualization is useful to understand the entire association between group of genes more clearly and intuitively[19, 32-34]. The nodes in a network can be genes (or proteins), and they can be connected through the links when they are related to each other [35-37]. In order to investigate how the two types (INC and DEC) of genes are closely located with each other in terms of molecular network, we reconstructed the largest connected sub-network by using the HPRD network Based on the INC and DEC type genes (Fig. 4 and Methods). The network consists of 170 nodes and 210 links. Based on the result, we found that both types of genes perform quite different functional roles, however, they are actually closely associated with each other in the network.

3.4 Clinical characteristics of the stage-dependent genes

The disease-free survival (DFS) analysis [38, 39] was conducted to further examine the relationship between stage-dependent gene expression and clinical outcomes. The DFS denotes the survival period without a cancer relapse. For each gene, we classified the patients into two subgroups based on the gene expression level. Patients with gene expression levels higher (lower) than the median expression levels of the given gene were classified into 'high' ('low'), respectively. Then, we performed survival analysis based on the Kaplan-Meier method for all the genes in TCGA dataset. Among the genes, we found two genes that show significant survival curve difference between two sub-groups (Fig. 5). Previous reports showed that MPP2, one of INC type genes, (MPP2 known as HNF-3, HFH-11, MPP2) promotes tumorigenesis and is widely overexpressed in a multitude of human solid tumors [40]. From the univariate Cox proportional-hazards model analysis, we found that the

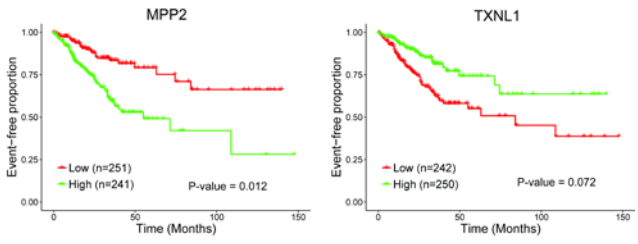

Fig. 5. The Kaplan–Meier estimates of cancer relapse for 'high' or 'low' expression group of each of two genes MPP2 and TXNL1. The MPP2 (a representative gene among the INC type genes) high-expressed group has higher risk of cancer relapse than MPP2 low-expressed group. On the other hand, the TXNL1 (a representative gene among the DEC type genes) low-expressed group has higher risk of cancer relapse than TXNL1 high-expressed group. The p-values were corrected by Benjamini-Hochberg procedure. The number of samples (n) were denoted on each group label.

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/TCBB.2018.2814043, IEEE/ACM Transactions on Computational Biology and Bioinformatics

MPP2-low group showed better prognosis than the MPP2-high group. The hazard ratio (HR) was 0.3889 and the 95% confidence interval (CI) was 0.2579 to 0.5863 (P value = 2.8E – 06). We also obtained consistent results after we consider age, gender and cancer stage based on multivariate analysis (HR = 0.4223, 95% CI, 0.2776 – 0.6424, P = 5.6E – 05). On the other hand, TXNL1, one of DEC type genes, encodes a protein that consists of thioredoxin family and can maintain genomic instability [41]. Therefore, the reduction of TXNL1 might be associated with genomic instability and aneuploidy of colorectal cancer [41]. From the univariate analysis, we found that the TXNL1-low group showed worse prognosis than the TXNL1-high group. The HR was 2.133 and the 95% CI was 1.421 to 3.2 (P = 2.6E – 4). We also obtained consistent results based on multivariate analysis (HR = 1.8823, 95% CI, 1.2509 – 2.833, P = 0.002). This result indicates that the gene expression changes along the cancer stage critically influences on the patient survival. In conclusion, the two gene expression patterns (INC and DEC types) can be used as important indicators in the survival analysis.

4. Conclusion

In this study, we investigated the association between gene expression level of TCGA dataset and cancer stage. We further applied the same procedure and validated our findings using independent gene expression profiles. An examination of gene expressions changed depending on the stage of colorectal cancer can significantly contribute to understandings of incidence and developmental process. However, the difference in gene expressions among patients with cancers may be caused by not only the stage of a cancer, but also many other factors such as gender, age, etc. and the factors' effects on each kind of cancer are varied. In order to specifically investigate the change of genes solely depending on the stage, we used LMER.

As a result, we found that continually increasing genes according to cancer stage are related to the nervous and developmental system whereas continually decreasing genes according to cancer stage are related to the cell cycle and metabolism. Also, we suggested that the crosstalk between ERK and SMAD signaling pathways might be crucial that regulate both INC and DEC type genes (Fig. 4). Based on the above procedure, we propose that the temporal gene expression profile analysis using our method can be effectively used to understand the underlying mechanisms of cancer development such as cancer growth, and weakening the cell cycle and metabolic functions.

Why do we choose LMER than simple LM? In fact, such genes that show large difference between LMER and LM are limited from our result; however, we still could find several genes that showed large difference between the two models. Furthermore, we note that the LMER will be a more essential model as we could obtain more samples and clinical variables together, for finding true association between gene expression levels and clinical variables by correcting other confounding factors.

Investigation of causative relationship between cancer stage and gene expression level based on correlative analysis is one of limitations of our study. However, based on the two independent gene expression datasets, we could consistently obtain similar results between the two datasets. We also note that our results coincide with recent findings that neurogenesis and fibroblast increment can occur for later-staged colorectal cancer patients [27]. Furthermore, a recent study showed that LMER can be applied to identify causal relationship between age and gene expression level [10]. Thus, we note that our findings can reflect gene expression changes along colorectal cancer progression.

Ultimately, the proposed approach can be applied to diagnosis of the progression of tumor malignancy and cancer therapeutics and comparison of expression profiles between the benign and malignant cases may be useful if we can obtain the expression profiles of colorectal adenomas.

Acknowledgment

The authors would like to thank Dr. Kyungoh Choi for valuable comments and technical support on the manuscript. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1A2B4012942).

References

[1] S. S. Knox, "From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer," *Cancer Cell Int,* vol. 10, pp. 11, 2010.

[2] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun, "Cancer statistics, 2008," *CA Cancer J Clin,* vol. 58, no. 2, pp. 71-96, Mar-Apr, 2008.

[3] S. Rathore, M. Hussain, and A. Khan, "GECC: Gene Expression Based Ensemble Classification of Colon Samples," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 11, no. 6, pp. 1131-45, Nov-Dec, 2014.

[4] R. Leake, "The cell cycle and regulation of cancer cell growth," *Ann N Y Acad Sci,* vol. 784, pp. 252-62, Apr 30, 1996.

[5] T. Jacks, and R. A. Weinberg, "The expanding role of cell cycle regulators," *Science,* vol. 280, no. 5366, pp. 1035-6, May 15, 1998.

[6] L. H. Hartwell, and M. B. Kastan, "Cell cycle control and cancer," *Science,* vol. 266, no. 5192, pp. 1821-8, Dec 16, 1994.

[7] K. Kasem, V. Gopalan, A. Salajegheh, C. T. Lu, R. A. Smith, and A. K. Lam, "The roles of JK-1 (FAM134B) expressions in colorectal cancer," *Exp Cell Res,* vol. 326, no. 1, pp. 166-73, Aug 01, 2014.

[8] M. M. Vickers, J. Bar, I. Gorn-Hondermann, N. Yarom, M. Daneshmand, J. E. Hanson, C. L. Addison, T. R. Asmis, D. J. Jonker, J. Maroun, I. A. Lorimer, G. D. Goss, and J. Dimitroulakos, "Stage-dependent differential expression of microRNAs in colorectal cancer: potential role as markers of metastatic disease," *Clin Exp Metastasis,* vol. 29, no. 2, pp. 123-32, Feb, 2012.

[9] F. Yao, C. Zhang, W. Du, C. Liu, and Y. Xu, "Identification of Gene-Expression Signatures and Protein Markers for Breast Cancer Grading and Staging," *PLoS One,* vol. 10, no. 9, pp. e0138213, 2015.

[10] D. Glass, A. Vinuela, M. N. Davies, A. Ramasamy, L. Parts, D. Knowles, A. A. Brown, A. K. Hedman, K. S. Small, A. Buil, E. Grundberg, A. C. Nica, P. Di Meglio, F. O. Nestle, M.

Ryten, U. K. B. E. consortium, T. c. Mu, R. Durbin, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, M. E. Weale, V. Bataille, and T. D. Spector, "Gene expression changes with age in skin, adipose tissue, blood and brain," *Genome Biol*, vol. 14, no. 7, pp. R75, Jul 26, 2013.

[11] L. Marisa, A. de Reynies, A. Duval, J. Selves, M. P. Gaub, L. Vescovo, M. C. Etienne-Grimaldi, R. Schiappa, D. Guenot, M. Ayadi, S. Kirzin, M. Chazal, J. F. Flejou, D. Benchimol, A. Berger, A. Lagarde, E. Pencreach, F. Piard, D. Elias, Y. Parc, S. Olschwang, G. Milano, P. Laurent-Puig, and V. Boige, "Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value," *PLoS Med*, vol. 10, no. 5, pp. e1001453, 2013.

[12] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, G. O. H. Ami, and G. Web Presence Working, "AmiGO: online access to ontology and annotation data," *Bioinformatics*, vol. 25, no. 2, pp. 288-9, Jan 15, 2009.

[13] M. Kanehisa, and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res*, vol. 28, no. 1, pp. 27-30, Jan 1, 2000.

[14] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human Protein Reference Database--2009 update," *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D767-72, Jan, 2009.

[15] A. Doi, K. Ishikawa, N. Shibata, E. Ito, J. Fujimoto, M. Yamamoto, H. Shiga, H. Mochizuki, Y. Kawamura, N. Goshima, K. Semba, and S. Watanabe, "Enhanced expression of retinoic acid receptor alpha (RARA) induces epithelial-to-mesenchymal transition and disruption of mammary acinar structures," *Mol Oncol*, vol. 9, no. 2, pp. 355-64, Feb, 2015.

[16] D. Bates, M. M¨achler, and B. Bolker, "lme4: Linear mixed-effects models using s4 classes http://cran.R-project.org/package=lme4. R package version 0.999375-42.," 2011.

[17] K. Dobbin, J. H. Shih, and R. Simon, "Statistical design of reverse dye microarrays," *Bioinformatics*, vol. 19, no. 7, pp. 803-10, May 1, 2003.

[18] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, no. 1, pp. 25-9, May, 2000.

[19] C. H. Seo, J. R. Kim, M. S. Kim, and K. H. Cho, "Hub genes with positive feedbacks function as master switches in developmental gene regulatory networks," *Bioinformatics*, vol. 25, no. 15, pp. 1898-904, Aug 01, 2009.

[20] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, "clusterProfiler: an R package for comparing biological themes among gene clusters," *OMICS*, vol. 16, no. 5, pp. 284-7, May, 2012.

[21] T. H. Keegan, L. A. Ries, R. D. Barr, A. M. Geiger, D. V. Dahlke, B. H. Pollock, W. A. Bleyer, A. National Cancer Institute Next Steps for, and G. Young Adult Oncology Epidemiology Working, "Comparison of cancer survival trends in the United States of adolescents and young adults with those in children and older adults," *Cancer*, vol. 122, no. 7, pp. 1009-16, Apr 1, 2016.

[22] M. A. Gordon, W. Zhang, D. Yang, S. Iqbal, A. El-Khouiery, F. Nagashima, G. Lurje, M. Labonte, P. Wilson, A. Sherrod, R. D. Ladner, and H. J. Lenz, "Gender-specific genomic profiling in metastatic colorectal cancer patients treated with 5-fluorouracil and oxaliplatin," *Pharmacogenomics*, vol. 12, no. 1, pp. 27-39, Jan, 2011.

[23] M. Martinelli, L. Scapoli, F. Cura, M. T. Rodia, G. Ugolini, I. Montroni, and R. Solmi, "Colorectal cancer susceptibility: apparent gender-related modulation by ABCB1 gene polymorphisms," *J Biomed Sci*, vol. 21, pp. 89, Sep 04, 2014.

[24] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdottir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739-40, Jun 15, 2011.

[25] H. Ueno, A. M. Jones, K. H. Wilkinson, J. R. Jass, and I. C. Talbot, "Histological categorisation of fibrotic cancer stroma in advanced rectal cancer," *Gut*, vol. 53, no. 4, pp. 581-6, Apr, 2004.

[26] A. Jacobson, and J. L. Cunningham, "Connective tissue growth factor in tumor pathogenesis," *Fibrogenesis Tissue Repair*, vol. 5, no. Suppl 1, pp. S8, 2012.

[27] D. Albo, C. L. Akay, C. L. Marshall, J. A. Wilks, G. Verstovsek, H. Liu, N. Agarwal, D. H. Berger, and G. E. Ayala, "Neurogenesis in colorectal cancer is a marker of aggressive tumor behavior and poor outcomes," *Cancer*, vol. 117, no. 21, pp. 4834-45, Nov 01, 2011.

[28] C. F. Ran Lu, Wenqi Shangguan, Yuan Liu, Yu Li, Yanna Shang, Dongqin Yin, Shengliang Zhang…, Qiaorong Huang, Xue Li, Wentong Meng, Hong Xu, Zongguang Zhou, Jiankun Hu, Weimin Li, Lunxu Liu, Xianming Mo, and "Neurons generated from carcinoma stem cells support cancer progression " *Signal Transduction and Targeted Therapy*, 2017.

[29] R. Firestein, A. J. Bass, S. Y. Kim, I. F. Dunn, S. J. Silver, I. Guney, E. Freed, A. H. Ligon, N. Vena, S. Ogino, M. G. Chheda, P. Tamayo, S. Finn, Y. Shrestha, J. S. Boehm, S. Jain, E. Bojarski, C. Mermel, J. Barretina, J. A. Chan, J. Baselga, J. Tabernero, D. E. Root, C. S. Fuchs, M. Loda, R. A. Shivdasani, M. Meyerson, and W. C. Hahn, "CDK8 is a colorectal cancer oncogene that regulates beta-catenin activity," *Nature*, vol. 455, no. 7212, pp. 547-51, Sep 25, 2008.

[30] P. Zhang, Y. Ma, F. Wang, J. Yang, Z. Liu, J. Peng, and H. Qin, "Comprehensive gene and microRNA expression profiling reveals the crucial role of hsa-let-7i and its target genes in colorectal cancer metastasis," *Mol Biol Rep*, vol. 39, no. 2, pp. 1471-8, Feb, 2012.

[31] A. Kasajima, C. Sers, H. Sasano, K. Johrens, A. Stenzinger, A. Noske, A. C. Buckendahl, S. Darb-Esfahani, B. M. Muller, J. Budczies, A. Lehman, M. Dietel, C. Denkert, and W. Weichert, "Down-regulation of the antigen processing machinery is linked to a loss of inflammatory response in colorectal cancer," *Hum Pathol*, vol. 41, no. 12, pp. 1758-69,

Dec, 2010.

[32] M. S. Kim, J. R. Kim, D. Kim, A. D. Lander, and K. H. Cho, "Spatiotemporal network motif reveals the biological traits of developmental gene regulatory networks in Drosophila melanogaster," *BMC Syst Biol,* vol. 6, pp. 31, May 01, 2012.

[33] M. S. Kim, J. R. Kim, and K. H. Cho, "Dynamic network rewiring determines temporal regulatory functions in Drosophila melanogaster development processes," *Bioessays,* vol. 32, no. 6, pp. 505-13, Jun, 2010.

[34] D. Kim, M. S. Kim, and K. H. Cho, "The core regulation module of stress-responsive regulatory networks in yeast," *Nucleic Acids Res,* vol. 40, no. 18, pp. 8793-802, Oct, 2012.

[35] F. Zhang, M. Wu, X. J. Li, X. L. Li, C. K. Kwoh, and J. Zheng, "Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates," *J Bioinform Comput Biol,* vol. 13, no. 3, pp. 1541002, Jun, 2015.

[36] J. R. Kim, S. M. Choo, H. S. Choi, and K. H. Cho, "Identification of Gene Networks with Time Delayed Regulation Based on Temporal Expression Profiles," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 12, no. 5, pp. 1161-8, Sep-Oct, 2015.

[37] S. Gao, I. Karakira, S. Afra, G. Naji, R. Alhajj, J. Zeng, and D. Demetrick, "Evaluating predictive performance of network biomarkers with network structures," *J Bioinform Comput Biol,* vol. 12, no. 5, pp. 1450025, Oct, 2014.

[38] M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-Meier estimate," *Int J Ayurveda Res,* vol. 1, no. 4, pp. 274-8, Oct, 2010.

[39] A. Yamada, T. Ishikawa, I. Ota, M. Kimura, D. Shimizu, M. Tanabe, T. Chishima, T. Sasaki, Y. Ichikawa, S. Morita, K. Yoshiura, K. Takabe, and I. Endo, "High expression of ATP-binding cassette transporter ABCC11 in breast tumors is associated with aggressive subtypes and low disease-free survival," *Breast Cancer Res Treat,* vol. 137, no. 3, pp. 773-82, Feb, 2013.

[40] X. Luo, J. Yao, P. Nie, Z. Yang, H. Feng, P. Chen, X. Shi, and Z. Zou, "FOXM1 promotes invasion and migration of colorectal cancer cells partially dependent on HSPA5 transactivation," *Oncotarget,* vol. 7, no. 18, pp. 26480-95, May 03, 2016.

[41] T. Gemoll, U. J. Roblick, S. Szymczak, T. Braunschweig, S. Becker, B. W. Igl, H. P. Bruch, A. Ziegler, U. Hellman, M. J. Difilippantonio, T. Ried, H. Jornvall, G. Auer, and J. K. Habermann, "HDAC2 and TXNL1 distinguish aneuploid from diploid colorectal cancers," *Cell Mol Life Sci,* vol. 68, no. 19, pp. 3261-74, Oct, 2011.

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/TCBB.2018.2814043, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Man-Sun Kim received her Ph.D. degree in Computer Engineering from Gongju National University, South Korea in 2005. From 2008 to 2011, she worked as a post doctor at the Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST). She is currently a research professor at Graduate School of New Drug Discovery and Development, Chungnam National University since 2016. Her research interests include systems biology, bioinformatics and cancer biology for OMICS data analysis in genome-wide scale and disease prediction model.

Dongsan Kim received his Ph.D. degree in Bio and Brain Engineering in Korea Advanced Institute of Science and Technology (KAIST), South Korea in 2014. He is a post doctor at the same department since 2014. His research interests include systems biology, bioinformatics and cancer biology based on big data analysis.

Jeong-Rae Kim received the M.S. degree and the Ph.D. degree in mathematics from Seoul National University, South Korea in 1997 and 2004, respectively. From 2005 to 2007, he was with the Bio-Max Institute in Seoul National University as a Senior Researcher. From 2007 to 2010, he was with the Department of Bio and Brain Engineering in Korea Advanced Institute of Science and Technology (KAIST) as a research professor. Since 2010, he has been a professor of the Department of Mathematics at the University of Seoul. His research interests include systems biology, bioinformatics and numerical analysis.