GMR Genetics and Molecular Research
ISSN 16765680

Detecting the potential cancer association or metastasis by multi-omics data analysis

L. Hua[1,2]*, WY. Zheng[1,2]*, H. Xia[1,2] and P. Zhou[1,2]

[1] School of Biomedical Engineering, Capital Medical University, Beijing, China
[2] Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, Capital Medical University, Beijing, China

*These authors contributed equally to this study.
Corresponding author: L. Hua
E-mail: hualin7750@139.com

Genet. Mol. Res. 15 (3): gmr.15038987
Received July 20, 2016
Accepted July 20, 2016
Published August 18, 2016
DOI http://dx.doi.org/10.4238/gmr.15038987

ABSTRACT. Comprehensive multi-omics data analyses have become an important means for understanding cancer incidence and progression largely driven by the availability of high-throughput sequencing technologies for genomes, proteomes, and transcriptomes. However, how tumor cells from the site of origin of the cancer begin to grow in other sites of the body is very poorly understood. In order to examine potential connections between different cancers and to gain an insight into the metastatic process, we conducted a multi-omics data analysis using data deposited in The Cancer Genome Atlas database. By combining somatic mutation data along with DNA methylation level and gene expression level data, we applied a Bayesian network analysis to detect the potential association among four distinct cancer types namely, Head and neck squamous cell carcinoma (Hnsc), Lung adenocarcinoma (Luad), Lung squamous cell carcinoma (Lusc), and Skin cutaneous melanoma (Skcm). Further validation based on the ‘identification of somatic signatures’ and the ‘association rules analysis’ confirmed these associations. Previous investigations have suggested that common risk factors and molecular abnormalities in cell-cycle regulation and signal transduction predominate among these cancers. This evidence indicates that our study provides a rational analysis and hopefully will help shed light on the links between different cancers and metastasis as a whole.

Key words: Somatic mutation; DNA methylation; TCGA;
Gene expression

INTRODUCTION

Current large-scale cancer sequencing projects have identified a large number of somatic mutations derived from an increasing number of different cancer tissues and patients. These studies have enabled comprehensive characterization of somatic mutations in a large number of tumor samples, and provide valuable information to aid in increasing the understanding of cancer incidence and progression (Watson et al., 2013). For example, lung cancers from smokers have ten times as many somatic mutations as those from non-smokers (Vogelstein et al., 2013). Experimental evidence has suggested that mitochondrial dysfunction, particularly due to mitochondrial DNA somatic mutation, could be a factor determining a cancer cells’ susceptibility to anti-cancer drugs that target energy metabolism (Kim, 2014). A recent study has found that genes with the highest frequency of somatic mutations can be detected in high-grade gliomas, T-cell lineage acute lymphoblastic leukemia and medulloblastoma (Huether et al., 2014). In addition, some of the most frequently mutated genes have been proven to be tractable targets for new anti-cancer drugs. For example, in several solid human tumors, recent studies have observed a high frequency of somatic mutation in the gene encoding phosphoinositide-3-kinase catalytic alpha (PIK3CA). A specific kinase inhibitor of PIK3CA was found to be a potentially effective therapeutic reagent against head and neck squamous cell carcinoma (Hnsc) (Qiu et al., 2006). On the other hand, it is well known that the inactivation of certain tumor-suppressor genes occurs because of hypermethylation within promoter regions. In fact, numerous studies have demonstrated that a broad range of genes are silenced by DNA methylation in different cancer types (Kulis and Esteller, 2010). Researchers have identified epigenetic patterns that are relevant to carcinogenesis, by analyzing the increasing amount of DNA methylation data. In addition, analyses of somatic alterations in some cancers support the hypothesis that one alteration can predispose to a subsequent specific alteration (Sweeney et al., 2009). Most frequently mutated genes in tumor samples may exhibit either DNA hypermethylation or hypomethylation. For example, analyses for several independent data sets are in agreement that mutations of the BRAF gene are much more frequent among tumors exhibiting CpG island methylation phenotype (CIMP) than in tumors without CIMP (Li et al., 2006).

Determining the functional impact of genes harboring the more frequent somatic mutations is crucial to understanding tumorigenesis and metastasis. Interestingly, it has been reported that, in some cases, somatic mutations of certain genes are found to be shared among different cancer types. For example, by mapping missense somatic mutations to protein domains, Yang et al. found, for twenty-one cancer types, that the vast majority of within-domain mutational hotspots shared by multiple cancer types occurred at functional sites (Yang et al., 2015). Recent sequencing of DNA derived from several cancers has also provided a comprehensive analysis of somatic mutations across entire genomes. For example, somatic mutations in the 3' untranslated regions (3'UTRs) of genes identified in four cancers have been reported in a new study, and this study computationally predicted how they may alter miRNA targeting, potentially resulting in dysregulation of the expression of the genes harboring these mutations (Ziebarth et al., 2012). Therefore, the integration of information about multiple features of DNA (such as DNA copy number, allelic status, sequence mutations, and DNA methylation) with gene expression patterns (Chari et al., 2010) has dramatically improved our ability to predict the risk of cancer association or metastasis. In this study, we have conducted a multi-omics data analysis to detect the potential risk of cancer association or metastasis for four distinct cancers: Head and neck squamous cell carcinoma (Hnsc), Lung adenocarcinoma (Luad), Lung squamous cell carcinoma (Lusc) and Skin cutaneous melanoma (Skcm). First, we identified the common genes harboring the most somatic mutations shared by the four cancers, and then we analyzed the DNA methylation status of these frequently mutated genes. Next, we constructed three Bayesian networks, based on DNA methylation levels, gene expression levels, and the observed correlation coefficients between DNA methylation levels, and the gene expression levels of genes found to be in common (genes-in-common) across these four cancers in order to explore potential associations between these four cancers. Finally, further validation, based on the ‘identification of somatic signature’ and ‘association rules analysis’ confirmed the association between these cancers. Previous investigations have reported common risk factors and that molecular abnormalities in cell-cycle regulation and signal transduction predominate among these cancers. The new evidence presented here also support that our study provides a rationale analysis to help shed light on links between cancers and metastasis as a whole.

MATERIAL AND METHODS

Data sources

Somatic mutation data

We searched The Cancer Genome Atlas (TCGA) database (http://cancergenome.nih.gov) in which somatic mutation data are available. For the purposes of the present study, we selected somatic mutation data for four cancers: Head and neck squamous cell carcinoma (Hnsc), Lung adenocarcinoma (Luad), Lung squamous cell carcinoma (Lusc) and Skin cutaneous melanoma (Skcm). In total, 67,125 somatic mutations across 319 Hnsc cancer patients, 208,724 somatic mutations across 519 Luad cancer patients, 61,485 somatic mutations across 178 Lusc cancer patients and 200,589 somatic mutations across 264 Skcm patients were used. Based on the Illumina GAIIx platform, these mutations were initially captured by whole-exome sequencing performed on tumors. All the categories of somatic mutation, including, Frame_Shift_Deletion, Frame_Shift_Insertion, In_Frame_Deletion, In_Frame_Insertion, Missense_Mutation, Nonsense_Mutation, Nonstop_Mutation, Silent, Splice_Site, Translation_Start_Site, RNA, 3’untranslated region (3’UTR), 5’Flank, 5’UTR and Intron were all put into the analysis. For each cancer, we used the Somatic Cancer Alterations package of R software (http://www.r-project.org) to compute the somatic mutation frequency along with the somatic mutation types for genes. Specifically, we selected those genes-in-common (i.e.

those that were shared by the four selected cancers and harbored the most somatic mutations) for further analysis.

DNA methylation levels and gene expression levels

For those genes-in-common, we used the MethHC database (http://MethHC.mbc.nctu.edu.tw) (Huang et al., 2015) to obtain their promoter region DNA methylation levels as well as the gene expression levels in tumor and normal samples respectively. MethHC currently consists of 6,548 DNA methylation data generated using the Illumina HumanMethylation450K BeadChip, which includes more than 480,000 CpG sites and 12,567 mRNA/microRNA expression data calculated by RNAseq/microRNA-seq. In this paper, the DNA methylation levels and gene expression levels were processed by MethHC.

Construction of Bayesian networks

To explore the potential association or metastasis among four cancers, for the genes-in-common we constructed three Bayesian networks based on 1) the DNA methylation levels; 2) the gene expression levels of in tumor samples versus normal samples and 3) the relationship between DNA methylation levels and the gene expression levels. In the present study, Pearson correlation coefficients were computed for the latter analysis. Briefly, the Bayesian network construction process was performed as follows: under the assumption of parameter independence, an initial Bayesian network structure S was learned from the training data. From this initial network, a greedy search algorithm with random restarts was performed to obtain the highest score posterior network to avoid local maxima. Finally, an optimized Bayesian network that maximizes the Bayesian factor is obtained using a heuristic search of the network space in a specified domain. The conditional likelihood of the variables given their parents is represented in a Bayesian network by using Gaussian conditional densities. In this study, we used the BNarray package (Chen et al., 2006) of R software (http://www.r-project.org) to construct the Bayesian networks.

Further validation

To further validate the association between the different cancer types obtained from the Bayesian networks, we used another two methods, namely ‘identification of somatic signatures’ (Gehring et al., 2015) and ‘association rules analysis’ (Wright et al., 2013), to explore the relationships between the four cancer types. As part of this validation process, an additional four cancer types were included: Gliobastoma multiforme (Gbm), Kidney renal clear cell carcinoma (Kirc), Ovarian cancer (Ov) and Thyroid carcinoma (Thca). These tumors had the following numbers of somatic mutations; 19,938 somatic mutations across 291 Gbm patients, 178,142 somatic mutations across 293 Kirc patients, 5,872 somatic mutations across 142 Ov patients and 6,716 somatic mutations across 403 Thca patients were included. The processes used for identification of somatic signatures and association rules analysis are described briefly below:

Identification of somatic signatures

The process for identification of somatic signatures was divided into two steps. In the first step, each somatic mutation was described in relation of the sequence context in which it occurred. Then a matrix, M, was used to represent the frequency of motifs across multiple samples. In the second step, the matrix M was numerically decomposed as M=WH. Where, W is the composition of each signature in term of somatic motifs and H indicates the contribution of the signature to the alterations for each sample. Principal component analysis (PCA) was used to employ the eigenvalue decomposition of M (Stacklies et al., 2007). Finally, cancer links were obtained by clustering the somatic signatures of the eight cancer types. We used the SomaticSignatures package (Watson et al., 2013) of R software (http://www.r-project.org) to perform this analysis.

Association rules analysis

Association rules analysis is a data mining method, and the Apriori algorithm is often used to discover the association rules. An association rule is expressed in the form X⇒Y. The strength of an association rule in the Apriori algorithm is often determined by its support and confidence. A high support indicates that a rule does not simply occur by chance, and thus this means that this rule has a higher reliability. Confidence determines how often items in Y appear in records that contain X. The higher the confidence, the more likely it is for Y to be present in transactions that contain X. In order to perform association rules analysis, we constructed a gene-cancer association matrix. For 16,383 genes and 8 types of cancer, the element of the matrix a_{ij} is defined as 1 if the ith gene has somatic mutations in the jth cancer whereas it is defined as 0 if the ith gene has no somatic mutation in the jth cancer. To obtain more effective rules, we set 0.8 as both the support and the confidence threshold. We used the arulesViz package (http://lyle.smu.edu/~mhahsler) of R software (http://www.r-project.org) to implement the analysis and the cancers associations were visualized using the obtained association rules. An overview of our study is shown in Figure 1.

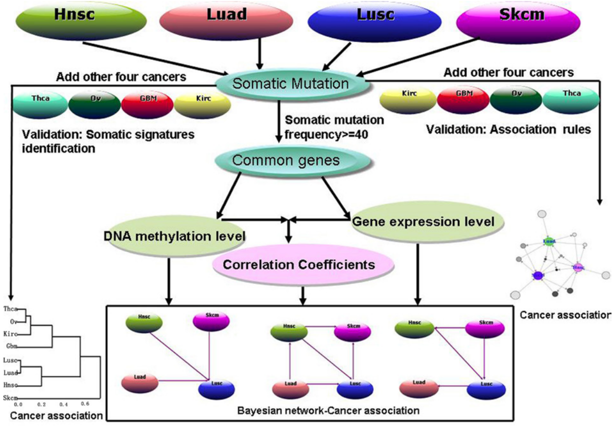

Figure 1. Study overview. First, genes harboring the most somatic mutations were identified and those genes that were common amongst the initial four different cancer types were selected. Following this, the individual gene's DNA methylation levels and expression levels were identified and a Pearson correlation coefficient calculated to reflect the association between the level of DNA methylation level and gene expression. Next, three Bayesian networks were constructed based on DNA methylation levels, gene expression levels, and the Pearson correlation coefficients, respectively. After then adding in an additional four cancer types, the 'identification of somatic signatures' and the 'association rules analysis' were used to validate the cancer associations obtained from the constructed Bayesian networks.

RESULTS

Identification of the genes-in-common harboring the most mutations shared by four cancers

After computing the mutation frequency of individual genes in the four separate cancer types, we then identified the most frequently mutated genes. There were five genes showing the most somatic mutations across all four of the different cancer types and these were: TTN, MUC16, CSMD3, LRP1B and RYR2 (see Table 1). TTN was found to harbor the most mutations across all of the genes examined and this could possibly be explained by the fact that larger genes have a greater chance of harboring mutation based on the assumption that mutations occur randomly across the genome (Jia and Zhao, 2014). As shown in Table 1, we noted especially that both TTN and MUC16 harbored a greater number of mutatios in Skcm than was observed in the other three cancers: In fact, 83 and 24% of Skcm patients have mutations in TTN and MUC16 respectively. Previous studies have identified TTN somatic mutations in multiple cancer types including melanoma, glioblastoma and pancreatic carcinoma, suggesting that the cellular functions of these molecules could possibly be related to a common tumor progression mechanism (Balakrishnan et al., 2007). MUC16 performs a number of important biological roles in cancer cell signaling, metastasis, regulation of immune responses and in anti-cancer therapeutic strategies (Felder et al., 2014). Also of note, we observed that CSMD3 harbored more mutations in Luad than in other types of cancer; in fact, 44% of Luad patients were found to have mutations in the CSMD3 region. Interestingly, a recent study has found that CSMD3 is the second most frequently mutated gene (next to TP53) in lung cancer. This study demonstrated that loss of CSMD3 might be causative for increased proliferation of airway epithelial cells (Liu et al., 2012). The finding of a significant number of somatic mutations in the LRP1B gene is not simply a result of of its long coding region; recently published cancer genome studies have also found its potential association with glioblastoma (GBM) and lung adenocarcinoma (Lawrence et al., 2013).

Table 1. Five genes having the highest somatic mutation frequency commonly Found in four distinct cancer types.

Gene	Somatic mutation frequency in four cancer types				
	Hnsc (Fraction)	Luad (Fraction)	Lusc (Fraction)	Skcm (Fraction)	Sum
TTN	401 (0.52)	945 (0.54)	441 (0.79)	1609 (0.83)	3396
MUC16	155 (0.07)	643 (0.10)	200 (0.12)	1158 (0.24)	2156
CSMD3	130 (0.25)	540 (0.44)	145 (0.51)	176 (0.35)	991
LRP1B	115 (0.24)	450 (0.39)	120 (0.42)	273 (0.44)	958
RYR2	78 (0.15)	507 (0.44)	142 (0.46)	137 (0.32)	864

Fraction refers to the proportion of patients that have mutations in these gene regions.

For the five most mutated genes, we calculated the frequency of their individual major somatic mutation types across the four cancer types. The frequency distribution of somatic mutation types is shown in Figure 2. From this, it can be seen that, missense mutations account for the highest proportion among all the different somatic mutation types. In fact, many studies have shown that missense mutations might play an important role in carcinogenesis. For example, missense mutations in oncogenes and tumor suppressors can cause structural effects or cause changes in function (Stehr et al., 2011). Moreover, cancer missense mutations can alter the binding properties of proteins and their cellular interaction profile (Nishi et al., 2013).

In addition, we observed that silent mutations were observed for each gene and that their frequency for each gene varied across the four different cancer types. Recent studies have shown that silent mutations frequently contribute to human cancer (Supek et al., 2014).

In order to filter the appropriate number of genes-in-common, we determined the appropriate cut-off by analyzing the relationship between somatic mutation frequency and the number of genes in common. As the number of genes-in-common increased, the somatic mutation frequency decreased rapidly (See Figure S1). Finally, we selected a somatic mutation frequency of 40 as the cut-off to filter genes-in-common across the four cancer types. According to this criterion, thirty-two genes with a somatic mutation frequency ≥40 (See Table S1) shared by four cancers were filtered for the Bayesian network construction.

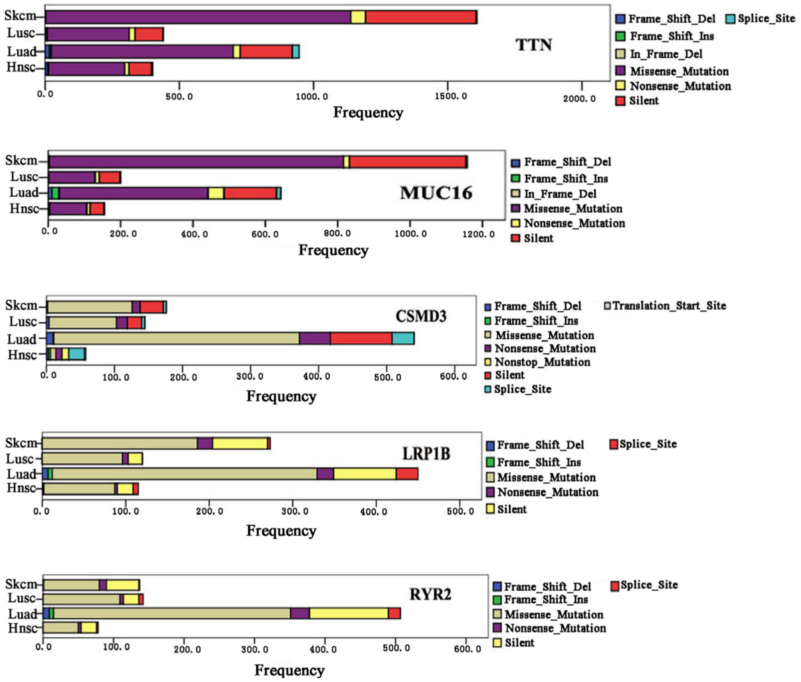

Figure 2. Frequency of different major somatic mutation types across four different cancer types, and summation thereof, for the top five genes-in-common identified in this study (namely, TTN, MUC16, CSMD3, LRP1B and RYR2). The major somatic mutation types are: Frame_Shift_Deletion, Frame_Shift_Insert, Missense_Mutation, Nonsense_Mutation, Silent, and Splice_Site.

Construction of Bayesian networks

DNA methylation levels of genes-in-common across four different cancer types

For the thirty-two genes-in common, we used the MethHC database to obtain information regarding their DNA methylation levels specifically in the promoter regions as well as their respective gene expression levels in tumor and normal samples. Assessing the average DNA methylation levels of the top five genes-in-common, the DNA methylation levels of TTN, MUC16, CSMD3 and RYR2 were all significantly different between normal tissue and tissue from the four different tumor types (P < 0.05) (See Figure 3). LRP1B did not show significant differences in DNA methylation level with the exception of normal versus Luad tissue. We observed that TTN displayed the lowest methylation levels in tumor samples compared to normal samples (P < 0.005). With the exception of Luad tumor samples, MUC16 also displayed the lowest methylation levels in the three other tumor samples compared to normal samples. In contrast, with the exception of Skcm tumor samples, CSMD3 and RYR2 exhibited the highest methylation levels in the three other tumor samples versus normal samples. It is known that aberrant DNA methylation is strongly associated with human cancer. Therefore, the aberrant methylation of these genes in these four distinct cancer types suggests the possible association or metastasis among these four cancers.

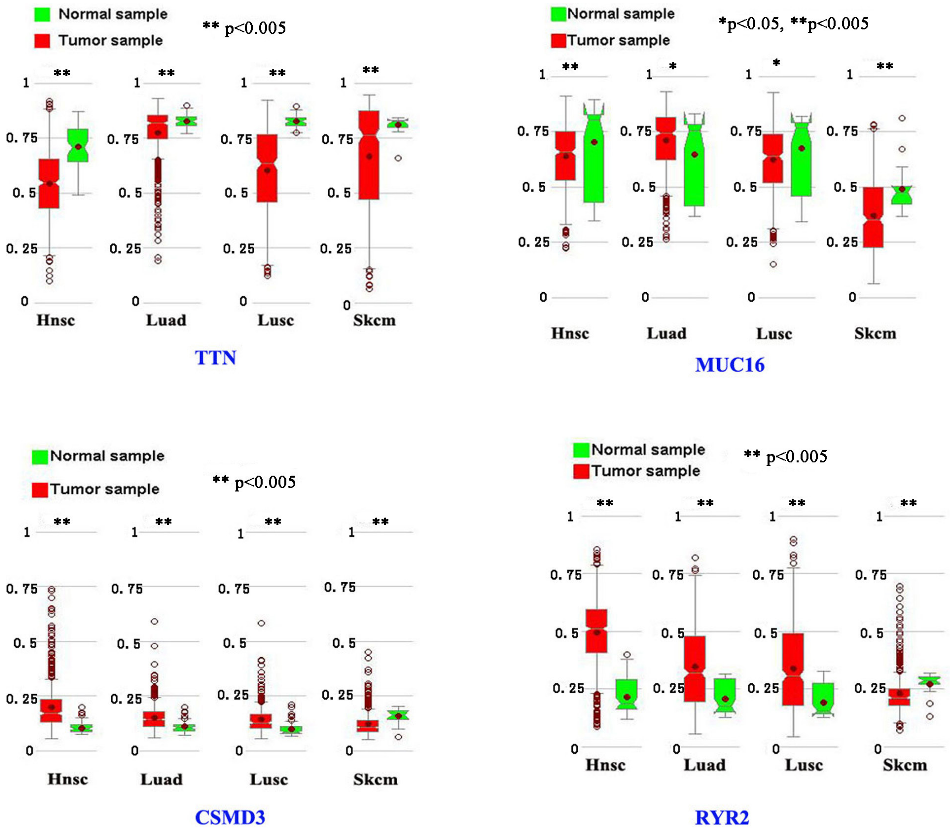

Figure 3. Average DNA methylation levels for the four genes, harboring the most somatic mutations; a comparison in normal versus four different tumor types (Hnsc, Luad, Lusc, and Skcm).

Construction of Bayesian networks

Based on DNA methylation levels, gene expression levels and the Pearson correlation coefficients between DNA methylation levels and gene expression levels for these thirty-two genes-in-common we constructed three Bayesian networks respectively (See Figures 4, 5 and 6). From Figure 4, we can see that DNA methylation levels of these thirty-two genes-in-common show no significant differences between normal and tumor samples ($P > 0.05$).

Hnsc, Luad and Skcm are direct causes of Lusc; the correlation coefficients of any two cancers based on DNA methylation levels were all ≥0.9 ($P < 0.001$). It is reported that at least 75 percent of head and neck cancers are caused by tobacco and alcohol use. Therefore, people who use tobacco and alcohol are at greater risk of Hnsc, and have an increased chance of developing new cancers, such as lung cancer (Do et al., 2003). New evidence suggests that melanoma metastasis to the lung is not uncommon and carries a poor prognosis (Seitelman et al., 2011).

From Figure 5, we observed that the gene expression levels of thirty-two genes-in-common all displayed significant differences between normal and tumor samples ($P < 0.01$). The average gene expression level of these thirty-two genes in tumor samples were all found to be higher than in normal samples. It is known that driver mutations can affect gene expression by means of aberrant transcription, epigenetic regulation, cell signaling and gene dosage effects (Gerstung et al., 2015). Although exactly how driver mutations interfere with the transcriptomic state and affect gene expressions is not well known, our results suggest that these genes harboring the most somatic mutations shared by four cancers also have the highest expression in the tumor samples. In addition, Figure 5 also shows the potential association among Hnsc, Luad and Skcm; the correlation coefficients of any two cancers based on the gene expression levels were all ≥0.6 ($P < 0.001$). In particular, as shown in Figure 5, we found an association between Hnsc and Skcm. A previous study has suggested that the INK4a/p16 germline mutation associated with familial atypical multiple mole melanoma syndrome can also be associated with familial head and neck squamous cell carcinoma syndrome. Young Hnsc patients, with a family history, may have a germline p16 defect that could predispose them to develop other cancers, including melanoma and pancreatic cancer (Vinarsky et al., 2009).

Based on the Pearson correlation coefficients comparing DNA methylation levels and and gene expression levels for these thirty-two genes in common, it is apparent from Figure 6 that similar results were obtained as for the gene expression analysis data shown in Figure 5. There was no significant difference in the Pearson correlation coefficients between normal and tumor samples ($P > 0.05$). The potential associations among Hnsc, Luad and Lusc were also observed; the correlation coefficients of any two cancers based on the Pearson correlation coefficients between the DNA methylation levels and the gene expression levels were all >0.4 ($P < 0.05$). Newly published evidence supports these results; performing a review of thirty-four lung cancer patients and twenty five autopsies of lung cancer with skin metastasis, a previous study concluded that the incidence of cutaneous metastasis is high for large-cell carcinoma and low for squamous and small-cell carcinoma (Terashima and Kanazawa, 1994). A recent Lung Screening Study from Pittsburgh observed that subjects with incidence of head and neck squamous cell carcinoma are at high risk of lung cancer. This study provided a rationale for offering head and neck cancer screening along with computed tomography screening for lung cancer. Randomized controlled trials that assess the effectiveness of adding the examination of the head and neck area to lung cancer screening programs are therefore warranted (Dixit et al., 2015).

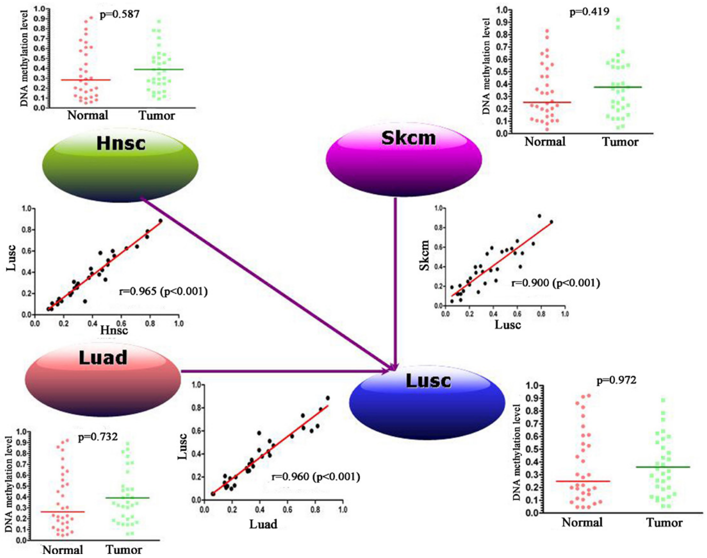

Figure 4. Constructed Bayesian network based on the DNA methylation levels of the thirty-two genes-in-common identified in tumor samples. The scatter dot plots show the average DNA methylation levels of the thirty-two genes-in-common. In each scatter dot plot, the red color indicates the normal sample and the green color indicates the tumor sample. The horizontal lines indicate the median DNA methylation levels. The scatter plots display the correlation, based on DNA methylation levels in tumor samples, between two cancers along with the linear regression lines.

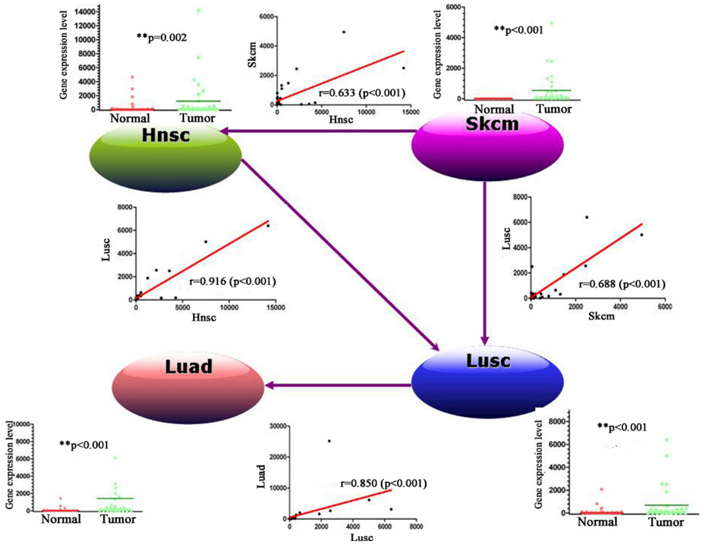

Figure 5. Constructed Bayesian network based on the gene expression levels of the thirty-two genes-in-common from tumor samples. The scatter dot plots show the average gene expression levels for thirty-two genes-in-common. In each scatter dot plot, the red color indicates the normal sample and the green color indicates the tumor sample. The horizontal lines indicate the median gene expression levels. The scatter plots display the correlation, based on tumor sample gene expression levels, between two cancers along with the linear regression lines.

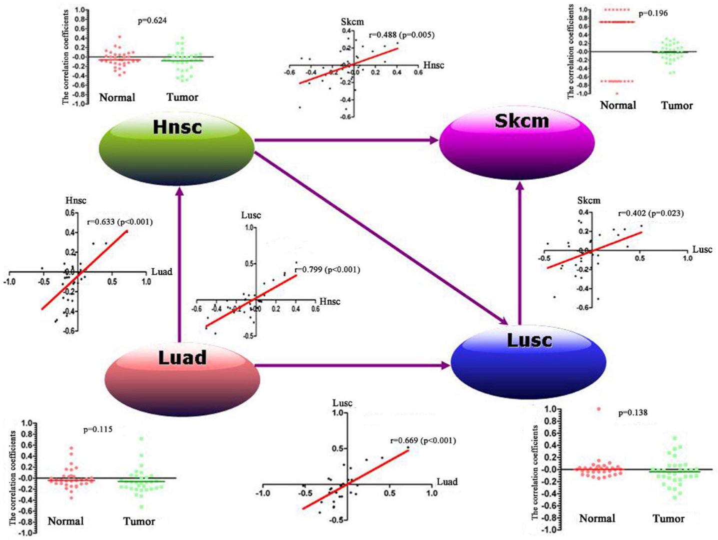

Figure 6. Constructed Bayesian network based on the Pearson correlation coefficients between DNA methylation levels and gene expression levels for thirty-two genes-in-common identified in tumor samples. The scatter dot plots show the Pearson correlation coefficients for the thirty-two common genes. In each scatter dot plot, the red color indicates the normal sample and the green color indicates the tumor sample. The horizontal lines indicate the median Pearson correlation coefficients. The scatter plots display the correlation, based on the Pearson correlation coefficients in tumor samples, between two cancers along with the linear regression lines.

Further validation

Further validation based on the identification of somatic signatures

By identifying the somatic signatures of eight distinct cancer types, there was an obvious difference in distribution of the somatic motifs between them. We noted that the contribution of C>T was higher in Gbm and Skcm than in other cancers. After the somatic signatures were clustered, we observed that Thca, Ov, Kirc and Gbm were clustered into one group whereas Lusc, Luad, Hnsc and Skcm were clustered into another separate group (See Figure 7A). This result supports our above conclusion in which the associations exist among these selected four cancers.

Further validation based on the association rules analysis

Applying the Apriori algorithm, with a cut-off of support of 0.8 and confidence of 0.8, we obtained twelve association rules. By analyzing these association rules, we found that Hnsc, Skcm and Luad are potentially associated. We used a graph-based technique (http://lyle.smu.edu/~mhahsler) which offers a very clear representation to visualize the twelve association rules we obtained (See Figure 7B). In this graph, the vertices represent items or item sets and the edges indicate the relationship in rules. From Figure 7B, we can see that Luad, Skcm and Hnsc are linked indirectly by sharing some common items. This result therefore also supports our above conclusion.

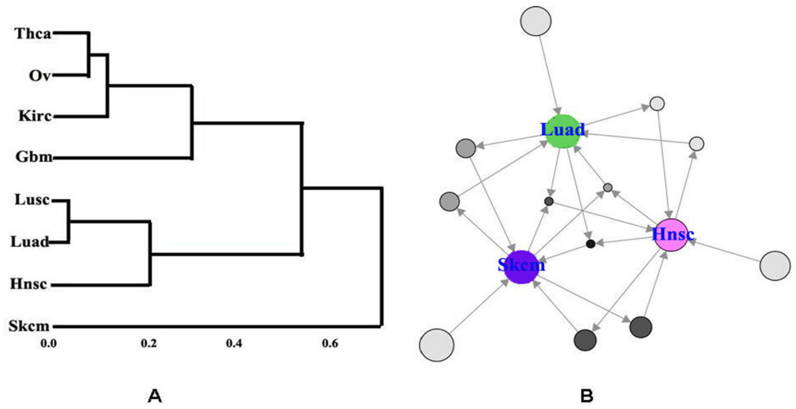

Figure 7. Further validation based on two different methodologies. **A.** Further validation based on the identification of somatic signatures. Following clustering of the somatic signatures, Thca, Ov, Kirc, and Gbm were clustered into one group whereas Lusc, Luad, Hnsc, and Skcm were clustered into another separate group. **B.** Further validation based on the association rules analysis. This graph visualizes the twelve association rules. The vertices represent items or item sets and the edges indicate the relationship in rules. Luad, Skcm and Hnsc are linked indirectly by sharing common items.

DISCUSSION

Cancer metastasis results from several interconnected processes such as cell proliferation, cell adhesion, migration, and invasion into the surrounding tissue. Metastasis is the leading reason for the resulting mortality of cancer patients (Khan and Mukhtar, 2010). Therefore, it is important to study cancer associations or metastasis from multiple molecular biology levels. The identification of clinically relevant biomarkers will help achieve a more effective diagnosis and prognosis contributing to personalized or precision cancer therapy. In practice, many studies involved in the somatic mutation analysis of cancers have provided valuable information for understanding cancer development or metastasis. However, characterizing somatic mutations and their functional consequences in tumor tissues remains a challenge. With the rapid technological advances in acquiring data from diverse platforms in cancer research, numerous large scale datasets have become available, providing high resolution views and multi-faceted descriptions of biological systems (Kong et al., 2011). Accordingly, multilevel -omics data integration approaches will help researchers to uncover further systemic information about cancer associations and metastasis. In the current study, we conducted a multi-omics data analysis to detect the potential risk of association or metastasis for four cancers. We identified thirty-two genes-in-common that harbored the greatest number of somatic mutations and these genes were shared by four distinct cancer types. Data was obtained on these genes-in-common concerning their DNA methylation status and levels of gene expression and a Pearson correlation coefficient calculated to assess the correlation between DNA methylation level and gene expression. Based on these three parameters, we constructed three individual Bayesian networks, and from these networks, we observed that there was a significant association betwen these four cancers. Further validation, based on the ‘identification of somatic signatures’ and the ‘association rules analysis’ confirmed these associations. Our analysis will not only help understand the potential links between different cancers as a whole, but can also help prioritize candidate cancer-causing mutations or genes and to elucidate potential cancer-type-dependent functional effects.

The limitations of our study should however be addressed. Among the thirty-two gene-in-common, TP53 was excluded based on its lower mutation frequency in Skcm cancer. A previous study has reported that p53 mutations in human cutaneous melanoma correlate with sun exposure but do not contribute to melanomagenesis (Zerp et al., 1999). However, it is known that somatic mutations in the TP53 gene are one of the most frequent alterations in human cancers (Olivier et al., 2010). In our analysis, TP53 displayed the greatest mutation frequency in Hnsc (Mutation frequency = 323, ranked 2nd in the mutation frequency list), as well as in Luad (Mutation frequency = 361, ranked 9th in the mutation frequency list) and in Lusc (Mutation frequency = 154, ranked 3rd in the mutation frequency list). Therefore, the use of a cut-off to eliminate genes such as TP53, which are not be shared all four cancer types, shows that important genes may be inadvertently eliminated from the analysis. Another limitation is that although our studies are supported by previous studies, this may not be sufficient supporting evidence. The integration of other available data, such as somatic copy number alterations (SCNAs) which affect a larger fraction of the genome in cancers than do any other type of somatic mutation, will be needed to validate these results. In addition, both network context and pathway information will help improve the power of data integration analysis and aid in finding those driver mutations which can confer metastatic potential. However, our current studies did not include any of this type of information. In the future we expect that the integration of more available data types along with biological context will allow a greater ability to detect and to find the potential cancer associations or predict metastasis in future studies

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the Beijing Natural Science Foundation (Grant #7142015) and the Excellent Talent Cultivation Project of Beijing (#2012D005018000002). Research also supported by the foundation-Clinical Cooperation Project of Capital Medical University (#14JL43).

REFERENCES

Balakrishnan A, Bleeker FE, Lamba S, Rodolfo M, et al. (2007). Novel somatic and germline mutations in cancer candidate genes in glioblastoma, melanoma, and pancreatic carcinoma. *Cancer Res.* 67: 3545-3550. http://dx.doi.org/10.1158/0008-5472.CAN-07-0065

Chari R, Thu KL, Wilson IM, Lockwood WW, et al. (2010). Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer. *Cancer Metastasis Rev.* 29: 73-93. http://dx.doi.org/10.1007/s10555-010-9199-2

Chen X, Chen M and Ning K (2006). BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics* 22: 2952-2954. http://dx.doi.org/10.1093/bioinformatics/btl491

Dixit R, Weissfeld JL, Wilson DO, Balogh P, et al. (2015). Incidence of head and neck squamous cell carcinoma among subjects at high risk of lung cancer: results from the Pittsburgh Lung Screening Study. *Cancer* 121: 1431-1435. http://dx.doi.org/10.1002/cncr.29189

Do KA, Johnson MM, Doherty DA, Lee JJ, et al. (2003). Second primary tumors in patients with upper aerodigestive tract cancers: joint effects of smoking and alcohol (United States). *Cancer Causes Control* 14: 131-138. http://dx.doi.org/10.1023/A:1023060315781

Felder M, Kapur A, Gonzalez-Bosquet J, Horibata S, et al. (2014). MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. *Mol. Cancer* 13: 129. http://dx.doi.org/10.1186/1476-4598-13-129

Gehring JS, Fischer B, Lawrence M and Huber W (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* 31: 3673-3675 10.1101/010686.

Gerstung M, Pellagatti A, Malcovati L, Giagounidis A, et al. (2015). Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat. Commun.* 6: 5901. http://dx.doi.org/10.1038/ncomms6901

Huang WY, Hsu SD, Huang HY, Sun YM, et al. (2015). MethHC: a database of DNA methylation and gene expression in human cancer. *Nucleic Acids Res.* 43: D856-D861. http://dx.doi.org/10.1093/nar/gku1151

Huether R, Dong L, Chen X, Wu G, et al. (2014). The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. *Nat. Commun.* 5: 3630. http://dx.doi.org/10.1038/ncomms4630

Jia P and Zhao Z (2014). VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLOS Comput. Biol.* 10: e1003460. http://dx.doi.org/10.1371/journal.pcbi.1003460

Khan N and Mukhtar H (2010). Cancer and metastasis: prevention and treatment by green tea. *Cancer Metastasis Rev.* 29: 435-445. http://dx.doi.org/10.1007/s10555-010-9236-1

Kim A (2014). Mitochondrial DNA somatic mutation in cancer. *Toxicol. Res.* 30: 235-242. http://dx.doi.org/10.5487/TR.2014.30.4.235

Kong J, Cooper LAD, Wang F, Gutman DA, et al. (2011). Integrative, multimodal analysis of glioblastoma using TCGA molecular data, pathology images, and clinical outcomes. *IEEE Trans. Biomed. Eng.* 58: 3469-3474. http://dx.doi.org/10.1109/TBME.2011.2169256

Kulis M and Esteller M (2010). DNA methylation and cancer. *Adv. Genet.* 70: 27-56. http://dx.doi.org/10.1016/B978-0-12-380866-0.60002-2

Lawrence MS, Stojanov P, Polak P, Kryukov GV, et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499: 214-218. http://dx.doi.org/10.1038/nature12213

Li WQ, Kawakami K, Ruszkiewicz A, Bennett G, et al. (2006). BRAF mutations are associated with distinctive clinical, pathological and molecular features of colorectal cancer independently of microsatellite instability status. *Mol. Cancer* 5: 2. http://dx.doi.org/10.1186/1476-4598-5-2

Liu P, Morrison C, Wang L, Xiong D, et al. (2012). Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis* 33: 1270-1276. http://dx.doi.org/10.1093/carcin/bgs148

Nishi H, Tyagi M, Teng S, Shoemaker BA, et al. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. *PLoS One* 8: e66273. http://dx.doi.org/10.1371/journal.pone.0066273

Olivier M, Hollstein M and Hainaut P (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* 2: a001008. http://dx.doi.org/10.1101/cshperspect.a001008

Qiu W, Schönleben F, Li X, Ho DJ, et al. (2006). PIK3CA mutations in head and neck squamous cell carcinoma. *Clin. Cancer Res.* 12: 1441-1446. http://dx.doi.org/10.1158/1078-0432.CCR-05-2173

Seitelman E, Donenfeld P, Kay K, Takabe K, et al. (2011). Successful treatment of primary pulmonary melanoma. *J. Thorac. Dis.* 3: 207-208.

Stacklies W, Redestig H, Scholz M, Walther D, et al. (2007). pcaMethods - a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 23: 1164-1167. http://dx.doi.org/10.1093/bioinformatics/btm069

Stehr H, Jang SHJ, Duarte JM, Wierling C, et al. (2011). The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol. Cancer* 10: 54. http://dx.doi.org/10.1186/1476-4598-10-54

Supek F, Miñana B, Valcárcel J, Gabaldón T, et al. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156: 1324-1335. http://dx.doi.org/10.1016/j.cell.2014.01.051

Sweeney C, Boucher KM, Samowitz WS, Wolff RK, et al. (2009). Oncogenetic tree model of somatic mutations and DNA methylation in colon tumors. *Genes Chromosomes Cancer* 48: 1-9. http://dx.doi.org/10.1002/gcc.20614

Terashima T and Kanazawa M (1994). Lung cancer with skin metastasis. *Chest* 106: 1448-1450. http://dx.doi.org/10.1378/chest.106.5.1448

Vinarsky V, Fine RL, Assaad A, Qian Y, et al. (2009). Head and neck squamous cell carcinoma in FAMMM syndrome. *Head Neck* 31: 1524-1527. http://dx.doi.org/10.1002/hed.21050

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, et al. (2013). Cancer genome landscapes. *Science* 339: 1546-1558. http://dx.doi.org/10.1126/science.1235122

Watson IR, Takahashi K, Futreal PA and Chin L (2013). Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* 14: 703-718. http://dx.doi.org/10.1038/nrg3539

Wright A, McCoy A, Henkin S, Flaherty M, et al. (2013). Validation of an association rule mining-based method to infer associations between medications and problems. *Appl. Clin. Inform.* 4: 100-109. http://dx.doi.org/10.4338/ACI-2012-12-RA-0051

Yang F, Petsalaki E, Rolland T, Hill DE, et al. (2015). Protein domain-level landscape of cancer-type-specific somatic mutations. *PLoS Comput. Biol.* 11: e1004147. http://dx.doi.org/10.1371/journal.pcbi.1004147

Zerp S, Elsas Av and Schrier LPaP (1999). p53 mutations in human cutaneous melanoma correlate with sun exposure but are not always involved in melanomagenesis. *Brit. J. Cancer* 79: 921-926.

Ziebarth JD, Bhattacharya A and Cui Y (2012). Integrative analysis of somatic mutations altering microRNA targeting in cancer genomes. *PLoS One* 7: e47137. http://dx.doi.org/10.1371/journal.pone.0047137

Supplementary material

Table S1. Thirty-two genes-in-common harboring the most somatic mutations across four different cancer types.

Figure S1. Determination of the cut off for somatic mutation frequency. In order to filter the appropriate number of genes in common across the four different cancer types based on somatic mutation frequency, the cut off was determined by analyzing the relationship between somatic mutation frequency and the number of genes in common. As the number of genes in common increased, the somatic mutation frequency decreased rapidly. The somatic mutation frequency of 40 was taken as the cut off to filter genes in common across the four different cancer types.