Exploration of the cancers association based on somatic data in TCGA

Hong Xia[1,2,a], Lin Hua[1,2,b,*], WeiYing Zheng[1,2,c] and Ping Zhou[1,2,d]

[1] School of Biomedical Engineering, Capital Medical University, Beijing, 100069, China

[2] Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, Capital Medical University, Beijing, 100069, China

[a]merina.xia@gmail.com, [b]hualin7750@139.com, [c]zhengwy2013@sina.com, [d]wjzpwyz@163.com

* corresponding author

Keywords: Somatic Data, Cancers, Association, TCGA

Abstract. As the widely development of Next-generation sequencing (NGS), many studies conducted the somatic data analysis of cancers, which provides the valuable information for understanding cancer incidence and progression. In this paper, we conducted a somatic data analysis for eight cancers, they are: Glioblastoma Multiforme, Head and Neck squamous cell carcinoma, Kidney renal clear cell carcinoma (Kirc), Lung Adenocarcinoma (Luad), Lung squamous cell carcinoma (Lusc), Ovarian Cancer (Ov), Skin Cutaneous Melanoma (Skcm) and Thyroid carcinoma (Thca). We explored the potential association between these cancers based on the somatic signatures identification and the association rules analysis. We found that TTN, TP53, CSMD3, MUC16 and PCDHGC5 were genes haboring the highest mutations ratio for eight cancers. Furthermore, we found the potential association among Hnsc, Skcm and Luad using association rules method. Some evidences have approved the common risk factors and molecular abnormalities in cell-cycle regulation and signal transduction predominate among these three cancers. Our analysis might help shed light on the links between different cancers as a whole.

1. Introduction

Recently, as the widely development of Next-generation sequencing (NGS), many studies conducted the somatic mutations analysis of cancers, which provides the valuable information for understanding cancer incidence and progression. Some of mutated cancer genes have been proven to be tractable targets for new anticancer drug [1]. For example, Wang et al. found that alterations of the PTEN gene occur in glioblastoma multiforme, and PTEN thus is the major target of inactivation on chromosome 10q in glioblastoma multiforme [2]. Recent studies have reported high frequencies of somatic mutations in the phosphoinositide-3-kinase catalytic alpha (PIK3CA) gene in several human solid tumors. A specific kinase inhibitor to PIK3CA was found to be potentially effective therapeutic reagent against head and neck squamous cell carcinoma [3]. In addition, some studies also provide the evidences of somatic mutations in primary lung adenocarcinoma for several tumour suppressor genes involved in other cancers [4]. For example, previous studies support that a tumour suppressor mechanism for BRCA1; somatic mutations and loss of heterozygosity (LOH) may result in inactivation of BRCA1 in at least a small number of ovarian cancers [5].

Interestingly, we found that some somatic mutations of genes are shared by different cancers. For example, TP53 somatic mutation are frequent in most human cancers and germline TP53 mutations predispose to a wide spectrum of early-onset cancers (Li-Fraumeni (LFS) and Li-Fraumeni-like syndromes (LFL)) [6]. Specially, we found the special metastatic sites of some cancers. For example, a recent case report presented a patient who developed metastatic thyroid lesions of a primary small cell lung cancer [7]. Therefore, our study aims to explore the potential association between different cancers based on the somatic data analysis and somatic signatures identification.

Here, we conducted a somatic data analysis for eight cancers involved in The Cancer Genome Atlas (TCGA) database (http://cancergenome.nih.gov), and these eight cancers are: Glioblastoma Multiforme (Gbm), Head and Neck squamous cell carcinoma (Hnsc), Kidney renal clear cell

carcinoma (Kirc), Lung Adenocarcinoma (Luad), Lung squamous cell carcinoma (Lusc), Ovarian Cancer (Ov), Skin Cutaneous Melanoma (Skcm) and Thyroid carcinoma (Thca). We explored the association between these cancers based on the somatic signatures identification and the association rules analysis. The results showed the potential association between Hnsc, Skcm and Luad. In fact, some previous evidences have approved the common risk factors and molecular abnormalities in cell-cycle regulation and signal transduction predominate among these three cancers. Our analysis might help shed light on the links between different cancers as a whole.

2. Methods

2.1. Data source

In the present study, we used The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov) which includes a variety of data types such as gene expression data, DNA methylation data, copy number changes, somatic mutations and so on. We selected the somatic data of eight cancers: Gbm, Hnsc, Kirc, Luad, Lusc, Ov, Skcm and Thca. The number of patients for Gbm, Hnsc, Kirc, Luad, Lusc, Ov, Skcm and Thca are 291, 319, 293, 519, 178, 142, 264 and 403, respectively. The Sequencer is Illumina GAIIx. The number of somatic variants for Gbm, Hnsc, Kirc, Luad, Lusc, Ov, Skcm and Thca are 19,938, 67,125, 24,158, 208,724, 61,485, 5,872, 200,589 and 6,716 respectively. We used SomaticCancerAlterations package which includes the data of these eight cancers [8] of R software (http://www.r-project.org) to compute the somatic mutations frequency of genes for each cancer. Specially, we recorded those genes haboring the most somatic mutations of each cancer.

2.2. Identification of somatic signatures

In the current study, we used the somatic signatures identification method described by Gehring et al. to implement our analysis [9]. The basic idea of this method is divided into two steps. In the first step, each somatic mutation is described in relation of the sequence context in which it occurs. Then a matrix Mij is used to represent the frequency of motifs across multiple samples. Where i indicate the number of motifs and j indicates the number of samples. In the second step, the matrix M is numerically decomposed into two matrices W and H. The formula is as following: $M_{ij} = \sum_{i=1}^{r} W_{ik} H_{kj}$.

Where r indicates the fixed number of signatures. W indicates the composition of each signature in term of somatic motifs. For each sample, H shows the contribution of the signature to the alterations [9]. The Principal component analysis (PCA) was used to employ the eigenvalue decomposition of M [10].We used SomaticSignatures package [9] of R software (http://www.r-project.org) to implement this analysis.

2.3. Exploration of the potential association between cancers using the association rules analysis

In order to explore the potential association between eight cancers, we used association rules method [11] to perform the analysis. Association rule is one of data mining methods, and the Apriori algorithm is often used to discover association rules. Association rule is expressed with the form X⇒Y. Where X is called antecedent (left-hand-side or LHS) and Y is called consequent (right-hand-side or RHS) of the rule. The strength of an association rule in the Apriori algorithm is often determined by its support and confidence. The support of a rule is calculated as: $Support(X \Rightarrow Y) = P(X \bigcup Y)$.

The support determines how often a rule is observed in the data. Therefore, a low support indicates that a rule has simply occurred by chance, and thus means that the rule might have a lower reliability. The confidence of a rule is defined as: $Conf(X \Rightarrow Y) = Support(X \bigcup Y) / Support(X)$.

The confidence determines how often items in Y appear in records that contain X. The confidence is the certainty of a rule. Another popular measure for association rules used is lift, which is defined as the following: $Lift(X \Rightarrow Y) = Support(X \bigcup Y) / (Support(X).Support(Y))$.

Lift can be interpreted as the deviation of the support of the whole rule from the support expected under independence given the supports of both sides of the rule. Greater lift values indicate the stronger associations.

Here, we construct a gene-cancers association matrix. For 16,383 genes and 8 cancers, the element of the matrix mij is defined as 1 if the ith gene has mutations in jth cancer whereas is defined as 0 if the ith gene has no any mutation in jth cancer. In order to get more effective rules, we set 0.8 as the support threshold and the confidence threshold respectively. In the present study, we used arulesViz package (http://lyle.smu.edu/~mhahsler) of R software (http://www.r-project.org) to implement the analysis and visualize the obtained association rules.

3. Results

3.1. Mutation gene frequency analysis and mapping the specially gene regions

We used SomaticCancerAlterations package of R software (http://www.r-project.org) to compute the mutations frequency of genes for each cancer. In Table 1, we listed the top five genes showing the most mutations sum of eight cancers. From Table 1, we can see that TTN, MUC16, TP53, CSMD3 and PCDHGC5 were the most frequently mutated genes identified in this study.

Table 1 The top five genes showing the most mutations sum of eight cancers

Gene	Eight Cancers								
	Gbm	Hnsc	Kirc	Luad	Lusc	Ov	Skcm	Thca	Sum
TTN	121	401	125	945	441	30	1609	14	3686
MUC16	68	155	46	643	200	12	1158	22	2304
TP53	101	323	8	361	154	118	44	3	1112
CSMD3	11	130	24	540	145	8	176	5	1039
PCDHGC5	62	3	4	735	165	9	12	9	999

We graphed the mutation ratio of eight cancers for these top five genes (See Figure 1). We found that TTN and TP53 were genes harboring the highest mutations ratio across eight cancers. From Figure 1, we can see that TTN showed more mutations in Lusc and Skcm than other cancers. However, although TTN displayed the most frequently mutation, it is presently hard to determine whether mutations in TTN act as drivers or are only passengers in lung cancer [12]. As the second gene which has the highest mutation ratio, TP53 showed more mutations in Lusc and Ov than other cancers. We know that mutations of TP53 gene are common genetic change in the malignant progression of many human cancers. Previous findings approved that alterations of TP53 play a major role in ovarian cancer, and suggest a possible mechanism for somatic mutations leading to ovarian cancer [13]. In addition, we also found that CSMD3 showed more mutations in Luad, Lusc and Skcm than other cancers. 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 cause the increased proliferation of airway epithelial cells [14]. However, from Figure 1, we can find that MUC16 showed the lower mutations, and PCDHGC5 showed fewer mutations across eight cancers.

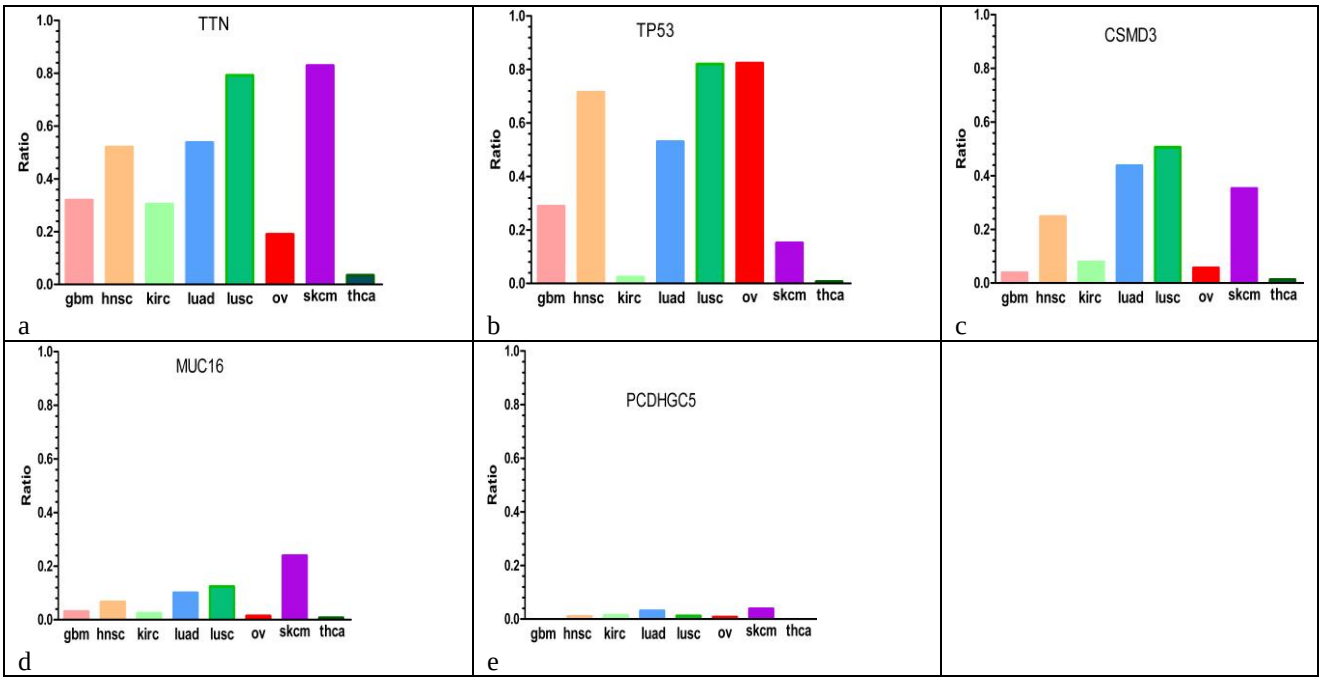

Fig.1 The top five genes showing the highest mutation ratio for eight cancers

3.2. Identification of somatic signatures

In this analysis, we used SomaticSignatures package which includes the data of these eight cancers [9] of R software to identify the somatic signatures. First, the observed occurrences of the somatic motifs were visualized across cancers as the following graph (See Figure 2). From Figure 2, we can see the obvious distribution difference of the motifs between the eight cancers. For example, the contribution of C>T was higher in Gbm and Skcm than in other cancers.

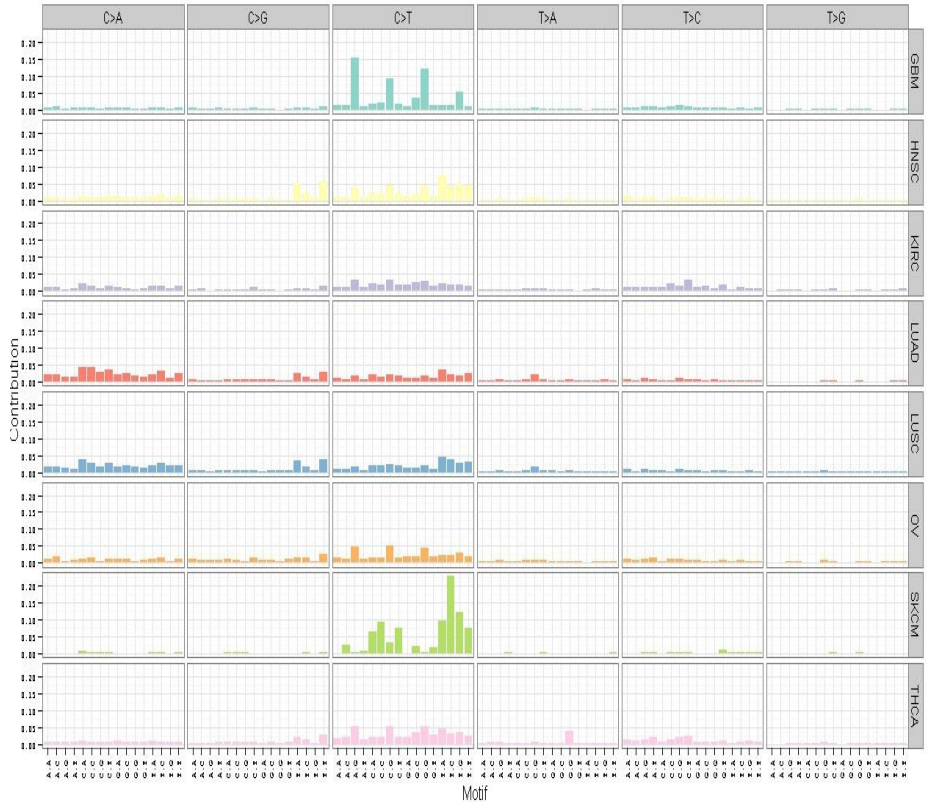

Fig.2 The obvious distribution difference of the motifs between eight cancers

Consider decomposition into smaller number signatures is nevertheless meaningful, and decompositions with more signatures naturally approximate the data better, we therefore select to decompose the data into 8 signatures in this analysis [8, 9]. We selected the Principal component analysis (PCA) to employ the eigenvalue decomposition of matrix. The eight somatic signatures (named S1 to S8) were visualized as a heatmap which was shown in Figure 3. In Figure 3, each signature represents different properties of the somatic spectrum observed in the data. We can see that signature S1 showed the most alterations, which are followed by signature S2 and signature S3. Signature S4-S8 showed the similarity distribution across the motifs.

From Figure 4, we can see that signature S1 is specially associated with Kirc and Thca. Signature S3 showed the strong association with Skcm. Signature S4 showed the strong association with Kirc. Signature S5 showed the association with Gbm and Thca. Signature S6 showed the strong association with Thca. Signature S7 showed the strong association with Ov. Signature S8 showed the strong association with Lusc. Furthermore, we clustered the somatic motifs, and the cluster graph was shown in Figure 5.

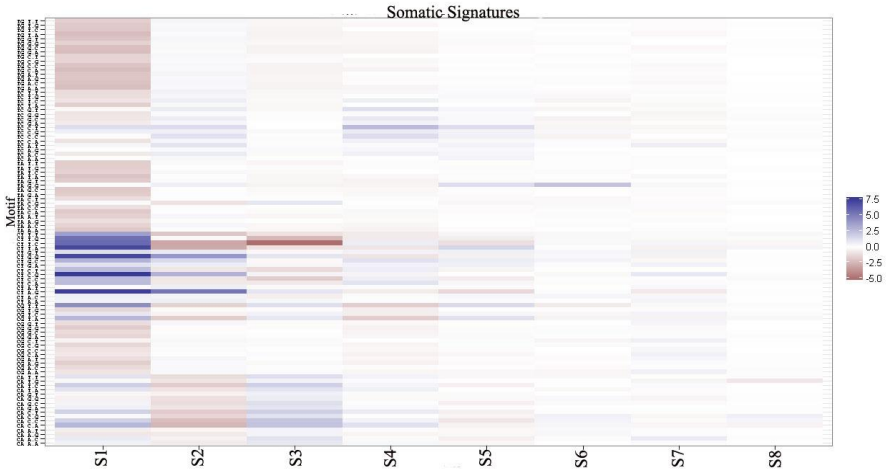

Fig.3 The heatmap of eight somatic signatures (S1-S8)

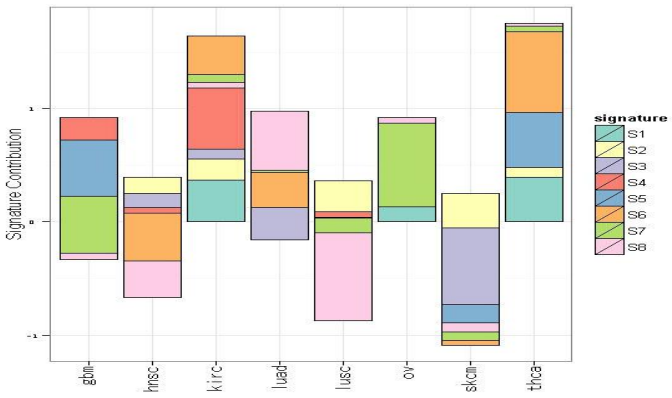

Fig.4 The association between eight somatic signatures (S1-S8) and cancers

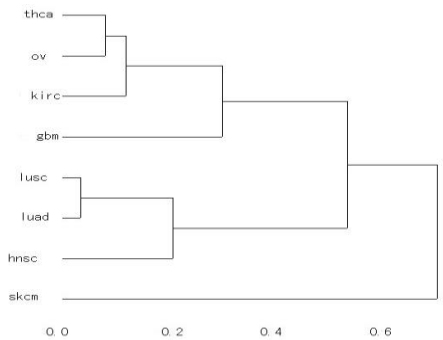

Fig.5 The cluster graph based on the somatic motifs

From Figure 5, we can see that Thca, Ov and Kirc and Gbm were clustered to one group whereas Lusc, Luad, Hnsc and Skcm were clustered to another group. Next, we will use association rules analysis to explore the association between these eight cancers.

3.3. Exploration the cancers association based on the association rules analysis

In this analysis, we used association rules method to explore the potential association between eightcancers. Applying Apriori algorithm, with the cut off of support of 0.8 and confidence of 0.8, we obtained 12 association rules. The matrix graphs of 12 rules (2D and 3D) along with the lift values were shown in Figure 6a and Figure 6b, respectively.

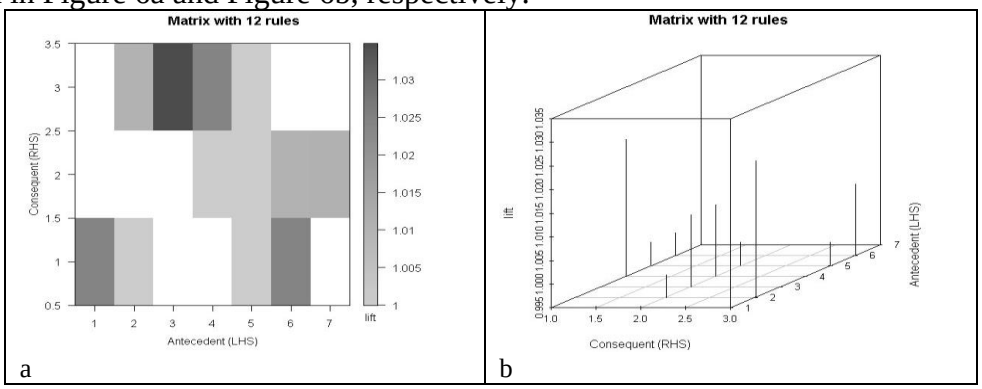

Fig.6 The matrix graphs (2D and 3D) of 12 association rules

Excluding three rules which are no sense, the rest 9 association rules along with their support, confidence and lift values were shown in Table 2.

Table 2 9 association rules obtained from Apriori algorithm

Rules	Support	Confidence	Lift
{Hnsc}⇒{Luad}	0.843	0.918	1.000
{Luad}⇒{Hnsc}	0.843	0.918	1.000
{Hnsc}⇒{Skcm}	0.879	0.958	1.024
{Skcm}⇒{Hnsc}	0.879	0.940	1.024
{Luad}⇒{Skcm}	0.867	0.944	1.010
{Skcm}⇒{Luad}	0.867	0.927	1.010
{Hnsc,Luad}⇒{Skcm}	0.815	0.968	1.035
{Hnsc,Skcm}⇒{Luad}	0.815	0.928	1.010
{Luad,Skcm}⇒{Hnsc}	0.815	0.940	1.024

From the results of association rules, we can see that Hnsc, Skcm and Luad have potential association. Here we used a new visualization technique [15] that enhances matrix-based visualization using grouping of rules via clustering to handle a larger number of rules. The arulesViz package (http://lyle.smu.edu/~mhahsler) of R software was used to make this graph. A balloon plot with antecedent groups as columns and consequents as rows is used to visualize the grouped matrix (See Figure 7). The colors of the balloons represent the aggregated interest measure in the group with a certain consequent and the size of the balloon shows the aggregated support. From Figure 7, we observed that there are one rule which contain “Hnsc" and up to 1 item in the antecedent and the consequent is “Skcm”. We also observed another rule which contain “Luad” and up to 1 item in the antecedent and the consequent is “Hnsc”.

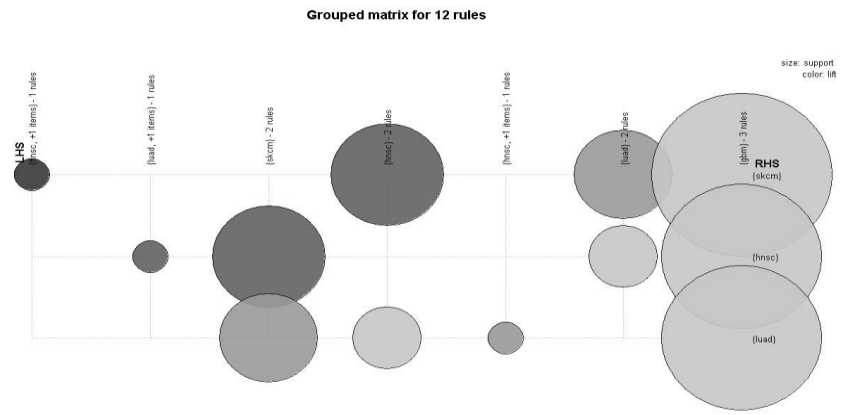

Fig.7 Visualize the grouped matrix of the obtained 12 association rules

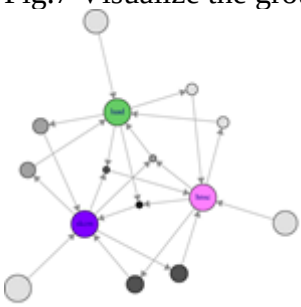

Fig.8 Graph-based visualization of the obtained association rules

In addition, we used another graph-based technique (http://lyle.smu.edu/~mhahsler) which offers a very clear representation of rules to visualize the obtained 12 association rules (See Figure 8). In this graph, the vertices represent items or item sets and the edges indicate the relationship in rules. From Figure 8, we can see that Luad, Skcm and Hnsc are linked indirectly by sharing some items.

In summary, the potential cancers association was found between Luad, Skcm and Hnsc by sharing some items. This result is completely consistent with the cluster analysis results. That is, Luad, Hnsc and Skcm were clustered to one group. Some previous evidences have approved the common risk factors and molecular abnormalities in cell-cycle regulation and signal transduction predominate among these three cancers. For example, for those patients with head and neck squamous cell carcinoma, the development of squamous cell carcinoma in the lung may signal a new primary or the onset of metastatic dissemination [16]. Previous study 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 [17]. These evidences support our analysis results.

4. Discussions

Currently, Next-generation sequencing (NGS) has been used to characterize the overall genomic landscape of human cancers. During carcinogenesis, the somatic alterations are often existed in tumor suppressor genes or oncogenes. Therefore, the identification of genomic regions with recurrent copy number alterations in tumor genomes is an efficient way to detect cancer genes [18]. In the practice, many studies involved in the somatic mutations analysis of cancers have provided the valuable information for understanding cancer progression.

Here, we conducted a somatic data analysis for eight cancers, and we explored the association between these cancers based on the somatic signatures identification and association rules methods. Our studies found the potential association between Head and Neck squamous cell carcinoma (Hnsc), Skin Cutaneous Melanoma (Skcm) and lung adenocarcinoma (Luad). Some evidences have approved the common risk factors and molecular abnormalities in cell-cycle regulation and signal transduction predominate among these three cancers. Our analysis might help understand the potential links between different cancers as a whole.

Certainly, it should be pointed out the limitation of our study. For example, we only used the somatic mutation data of eight cancers. In the practice, different data types should be used in the data integration analysis, such as SNP genotype data, gene expression data, DNA methylation data, and so on. In addition, the current biology network context and pathway information will help improve the power of data analysis results. Thus, more available data and biology context will be used to valid the results in future studies.

Acknowledgement

We would like to thank Dengke Niu for his helpful suggestions on this paper. This work is supported by Beijing Natural Science Foundation (Grant No. 7142015), National Natural Science Foundation of China (Grant Nos. 31100905) and the Science Technology Development Project of Beijing Municipal Commission of Education (SQKM201210025008). This study is also funded by the excellent talent cultivation project of Beijing (2012D005018000002) and the young backbone teacher's cultivation project of Beijing Municipal Commission of Education, supported by the foundation-clinical cooperation project of capital medical university (14JL43) , and supported by the natural science project of capital medical university (2014ZR21). We would like to thank Dengke Niu for his helpful suggestions on this paper.

References

[1] Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF. Statistical Analysis of Pathogenicity of Somatic Mutations in Cancer. Genetics. 2006 Aug; 173(4): 2187–2198.
[2] Wang SI, Puc J, Li J, Bruce JN, Cairns P, Sidransky D, Parsons R. Somatic mutations of PTEN in glioblastoma multiforme. See comment in PubMed Commons below Cancer Res. 1997 Oct 1; 57(19): 4183-4186.
[3] Qiu W, Schönleben F, Li X, Ho DJ, Close LG, Manolidis S, Bennett BP, Su GH. PIK3CA mutations in head and neck squamous cell carcinoma. Clin Cancer Res. 2006 Mar 1; 12(5):1441-1446.
[4] Ding L, Getz G, Wheeler DA, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature 2008; 455:1069-1075.
[5] Merajver SD, Pham TM, Caduff RF, Chen M, Poy EL, Cooney KA, Weber BL, Collins FS, Johnston C, Frank TS. Somatic mutations in the BRCA1 gene in sporadic ovarian tumours. Nat Genet. 1995 Apr; 9(4):439-443.
[6] Hainaut P, Hollstein M. p53 and human cancer: the first ten thousand mutations. Adv Cancer Res. 2000; 77: 81–137.
[7] Stamatis Katsenos, Stavros Archondakis, Michalis Vaias, and Nikolaos Skoulikaris, Thyroid gland metastasis from small cell lung cancer: an unusual site of metastatic spread. J Thorac Dis. 2013 Apr; 5(2): E21–E24.
[8] Gehring J. SomaticCancerAlterations: Somatic Cancer Alterations. R package version 1.3.2. 2014.
[9] Gehring J, Fischer B, Lawrence M and Huber W. SomaticSignatures: Inferring Mutational Signatures from Single Nucleotide Variants. bioRxiv. 2015, 1-6. http://dx.doi.org/10.1101/010686.
[10] Wolfram Stacklies, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. pcaMethodsa bioconductor package providing PCA methods for incomplete data Bioinformatics 2007, 23 (9): 1164-1167.
[11] Wright A, McCoy A, Henkin S, Flaherty M, Sittig D. Validation of an association rule mining-based method to infer associations between medications and problems. Appl Clin Inform. 2013 Mar 6; 4(1): 100-109.
[12] Yin S, Yang J, Lin B, et al. Exome sequencing identifies frequent mutation of MLL2 in non-small cell lung carcinoma from Chinese patients. Sci Rep. 2014, 4:6036.

[13] J Kupryjańczyk, A D Thor, R Beauchamp, et al. p53 gene mutations and protein accumulation in human ovarian cancer. Proc Natl Acad Sci U S A. 1993 Jun 1; 90(11): 4961–4965.
[14] Liu P, Morrison C, Wang L, Xiong D, Vedell P et al. Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. Carcinogenesis. 2012 Jul; 33(7): 1270-1276.
[15] Hahsler M, Chelluboina S. Visualizing Association Rules in Hierarchical Groups. In 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms (Interface 2011). The Interface Foundation of North America. 2011.
[16] Bishop JA, Ogawa T, Chang X, Illei PB, Gabrielson E, Pai SI, Westra WH. HPV analysis in distinguishing second primary tumors from lung metastases in patients with head and neck squamous cell carcinoma. Am J Surg Pathol. 2012 Jan; 36(1):142-148.
[17] Vinarsky V, Fine RL, Assaad A, Qian Y, Chabot JA, Su GH, Frucht H. Head and neck squamous cell carcinoma in FAMMM syndrome. Head Neck. 2009 Nov; 31(11):1524-1527.
[18] Chen M, Gunel M, Zhao H. SomatiCA: identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data. PLoS One. 2013 Nov 12; 8(11): e78143.