G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2015;13(4):132-136
http://dx.doi.org/10.5808/GI.2015.13.4.132

ORIGINAL ARTICLE

Expressional Subpopulation of Cancers Determined by G64, a Co-regulated Module

Jae-Woong Min and Sun Shim Choi*

Department of Medical Biotechnology, College of Biomedical Science, and Institute of Bioscience & Biotechnology, Chuncheon 24341, Korea

Studies of cancer heterogeneity have received considerable attention recently, because the presence or absence of resistant sub-clones may determine whether or not certain therapeutic treatments are effective. Previously, we have reported G64, a co-regulated gene module composed of 64 different genes, can differentiate tumor intra- or inter-subpopulations in lung adenocarcinomas (LADCs). Here, we investigated whether the G64 module genes were also expressed distinctively in different subpopulations of other cancers. RNA sequencing-based transcriptome data derived from 22 cancers, except LADC, were downloaded from The Cancer Genome Atlas (TCGA). Interestingly, the 22 cancers also expressed the G64 genes in a correlated manner, as observed previously in an LADC study. Considering that gene expression levels were continuous among different tumor samples, tumor subpopulations were investigated using extreme expressional ranges of G64—i.e., tumor subpopulation with the lowest 15% of G64 expression, tumor subpopulation with the highest 15% of G64 expression, and tumor subpopulation with intermediate expression. In each of the 22 cancers, we examined whether patient survival was different among the three different subgroups and found that G64 could differentiate tumor subpopulations in six other cancers, including sarcoma, kidney, brain, liver, and esophageal cancers.

Keywords: differentially expressed genes, lung adenocarcinoma, single cell analysis, survival analyses, tumor heterogeneity

Introduction

Different tumors or cancers in different patients have distinct genetic and cellular profiles, including kinds of genetic mutations, patterns of gene expression, and metastatic potential, which is often called cancer heterogeneity. The heterogeneity occurs both within and between tumors and leads to intra-tumor heterogeneity and inter-tumor heterogeneity, respectively [1-3]. Understanding cancer heterogeneity has recently been one of the important research subjects, because it is the base of difficulty in developing effective cancer treatments.

Tumor heterogeneity is basically due to the different origins of tumor cells or tissues. Various biological factors, such as smoking, gender, age, or hormonal status, can influence cancer initiation or progression, as well. Fundamentally, there are genetic variations among the hosts, even for the same cancer [4-6]. Astonishingly rapid development of massively parallel sequencing, alternatively called next-generation sequencing technologies, has recently elucidated the extent of tumor heterogeneity, and several hundred somatic mutations and structural variants that drive cancers have been identified [7-9].

Previously, we reported the characterization of the heterogeneity of human lung adenocarcinomas (LADCs) from patient-derived xenografts via single-cell transcriptome sequencing [10]. These cells were categorized into two separate subpopulations based on their expression of a module gene named G64. We found that G64 up-regulation/down-regulation was also present in patient tissue samples obtained from the Cancer Genome Atlas (TCGA), with G64 up-regulation corresponding to poor survival and associated with multiple clinical variables, such as smoking status (which exhibited the highest correlation) and tumor stage.

In the present study, we investigated if G64 can differentiate tumor subpopulations in other cancers, as well.

Received October 13, 2015; Revised December10, 2015; Accepted December 12, 2015

***Corresponding author:** Tel: +82-33-250-6487, Fax: +82-33-241-6480, E-mail: schoi@kangwon.ac.kr

Methods

Websites for downloading RNA sequencing (RNA-seq) data for 22 cancer datasets

The TCGA website (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp) was used for downloading RNA-seq-based transcriptome data and patient survival information for 22 cancers. The RNA-seq data downloaded for each cancer are listed in Table 1 and Supplementary Table 1. As described in the previous report [10], expression values of $0<\text{FPKM}$ (fragments per kilobase of exon per million fragments mapped) <0.1 were all converted to 0.1 to avoid the infinity problem. Each FPKM value of each gene was divided by the average FPKM estimated for the total patient samples where the gene is expressed and was log2-transformed, for which heat map analysis combined with hierarchical clustering was performed with the 'hclust' function of R package (https://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html).

Statistical tests

We used R (v 3.1.3) for all statistical tests (https://cran.r-project.org/). Using the factoMineR (http://factominer.free.fr) and rgl (https://r-forge.r-project.org/projects/rgl/) packages, principal component analysis (PCA) and visualization were performed [11]. The survival packages were used (http://r-forge.r-project.org) to compare patient survival rates and draw Kaplan-Meier plots. The coxph function in the survival packages was used for performing the Cox regression analysis [12, 13].

Results

Selection of two extreme patient groups and one intermediate group determined by G64 expression

The G64 genes were originally identified by their co-regulated expression pattern in single cells derived from a single LADC tumor region [10]. Interestingly, we have reported that 488 LADC patients samples downloaded from the TCGA were also separated into two distinct groups by the G64 genes. In the present work, we examined whether G64 could be a classifier for other cancers, as well. For this analysis, 22 RNA-seq-based transcriptome data with a sample size $\geq$150 were retrieved from the TCGA. The threshold of 150 was selected, because it was considered to be the minimum number of samples for statistical tests between

Table 1. List of the 22 cancers

Types of cancers	Total	Extreme	P1	P2	P3
Kidney renal clear cell carcinoma	531	160	0.000*	0.146	0.000*
Brain lower-grade glioma	514	154	0.001*	0.220	0.000*
Liver hepatocellular carcinoma	351	106	0.001*	0.291	0.041*
Kidney renal papillary cell carcinoma	289	86	0.007*	0.488	0.000*
Esophageal carcinoma	184	56	0.018*	0.731	0.027*
Sarcoma	257	78	0.036*	0.019*	0.604
Pancreatic adenocarcinoma	178	54	0.053	0.920	0.931
Stomach adenocarcinoma	373	112	0.116	0.999	0.114
Cervical squamous cell carcinoma and endocervical adenocarcinoma	302	90	0.157	0.355	0.469
Head and neck squamous cell carcinoma	514	154	0.335	0.427	0.258
Colon adenocarcinoma	192	58	0.412	0.188	0.017
Acute myeloid leukemia	173	52	0.457	0.731	0.027
Thyroid carcinoma	505	152	0.461	0.028	0.640
Skin cutaneous melanoma	467	140	0.628	0.844	0.822
Lung squamous cell carcinoma	501	150	0.679	0.720	0.724
Breast invasive carcinoma	1092	328	0.818	0.009	0.016
Ovarian serous cystadenocarcinoma	263	78	0.878	0.863	0.172
Glioblastoma multiforme	161	48	0.883	0.760	0.134
Bladder urothelial carcinoma	407	122	0.912	0.115	0.131
Prostate adenocarcinoma	486	146	0.978	0.488	0.000
Pheochromocytoma and paraganglioma	179	54	0.999	0.999	0.999
Testicular germ cell tumors	150	44	NA	0.728	0.835

P1, P2, and P3 indicate three different comparisons: the highest 15% versus the lowest 15%, the lowest 15% versus the intermediate, and the highest 15% versus the intermediate, respectively.
*Significant comparisons of $p<0.05$.

different groups (Table 1, Supplementary Table 1).

As done in the previous LADC study [10], heat map analysis was performed for each cancer type to see how patient samples were separated by G64 (Supplementary Fig. 1). As shown in Supplementary Fig. 1, the co-regulated patterns of G64 expression were confirmed in all types of cancers we tested. However, distinct two-group separations were not as evident as the case for LADC. Therefore, we classifies patient samples further into three different categories by using a 15% extreme threshold of G64 expression (Supplementary Fig. 2): the highest 15% group, the lowest 15% group, and an intermediate group. For this purpose, the patients were sorted out by the intensity of the average FPKM values of the G64 genes expressed in each patient. As a result, for instance, among a total of 531 kidney renal clear cell carcinoma (KIRC) samples, 160 samples were assigned into the two extreme groups—the highest 15% and the lowest 15%—whereas the remaining 371 samples were classified into the intermediate group.

G64 up-regulating cancers tend to have a poor prognosis in six different cancers

We next investigated whether patient survival was significantly different in these three groups classified by the average intensity of G64 expression. Patient survival rates were compared among the three patient groups for each of the 22 cancers. As done in the LADC study, we performed Kaplan-Meier analysis using the patients' survival data obtained from the TCGA in each type of cancer. Survival rates were compared between the lowest 15% versus highest 15% groups (i.e., P1 comparison in Table 1), the lowest 15% versus intermediate groups (i.e., P2 comparison in Table 1), and the intermediate versus highest 15% groups (i.e., P3 comparison in Table 1). PCA confirmed the groups' separations by G64 expression (Fig. 1, Supplementary Fig. 3–7). As summarized in Table 1, only six cancers were shown to have statistically significant differences between groups, including KIRC, brain lower grade glioma, liver hepatocellular carcinoma, kidney renal papilloma cell carcinoma, esophageal carcinoma, and sarcoma. In all comparisons

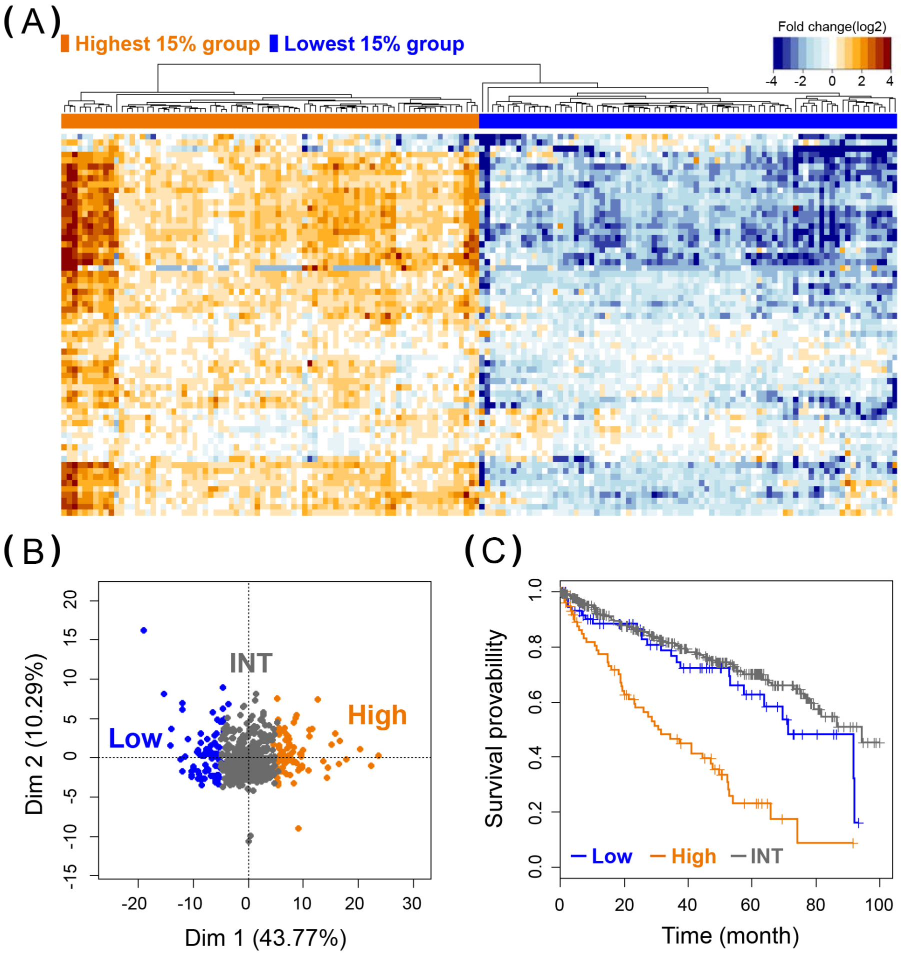


Fig. 1. Heat map, principal component analysis (PCA), and Kaplan Meier (KM) plot analysis of G64 in kidney renal clear cell carcinoma (KIRC). (A) Heat map was plotted using the 160 samples corresponding to the highest 15% and lowest 15% of G64 expression among the total of 531 KIRC samples. (B) A PCA plot of G64 expression was performed using the 531 KIRC samples. Low, the lowest 15% group; INT, intermediate group; High, the highest 15% group. (C) KM analysis of the 531 KIRC samples based on the average G64 expression. Cox regression analysis was used to investigate whether the survival duration of the two different groups was significantly different ($p<0.001$), as shown in Table 1.

between groups in these 6 cancers, G64 up-regulation was consistently observed to be associated with worse patient prognosis in the Cox regression analysis (Fig. 1, Supplementary Fig. 3–8).

The 22 cancers were subsequently divided into two batches: i.e., six survival-differentiating cancers and 16 survival-non-differentiating cancers. After the RNA-seq samples from the 22 cancers were collected as separate batches, we examined the aspect of survival difference by G64 expression in each batch. Consequently, the survival-differentiating batch, mixing up six different cancer samples (531 + 514 + 351 + 289 + 184 + 257 = 2,126 samples), clearly maintained the survival differentiation among the lowest 15%, the highest 15%, and the intermediate groups (Fig. 2), whereas the survival-non-differentiating batch, mixing up 16 different cancer samples (178 + 373 + 302 + ⋯. + 179 + 150 = 5943 samples), showed no survival difference among the three different groups determined by G64 expression levels (Fig. 2). Taken together, the G64 genes can differentiate cancer samples by their survival rate differences in cancers other than LADC.

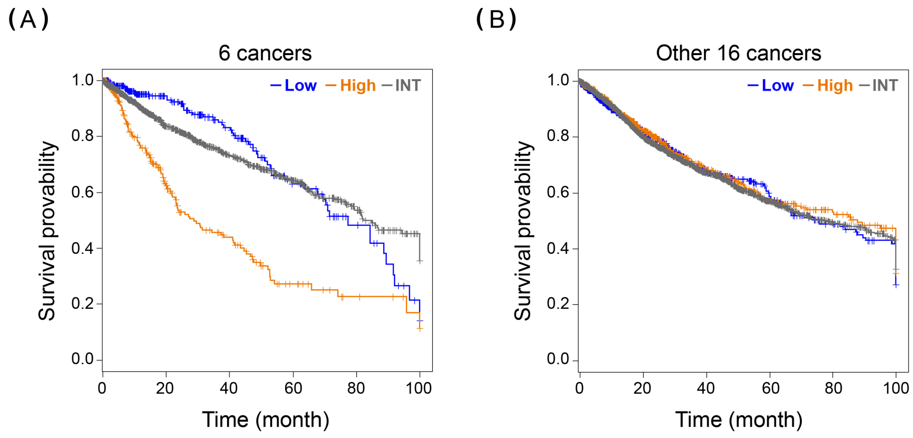
Discussion

Basically, it was an interesting finding that genes, named the G64 module in our previous study (i.e., a highly co-regulated gene group), can classify not only single cells derived from a single tumor but also tumors that have originated from several different cancers. Various biomarkers have been identified and developed for molecular diagnostics in cancers. In fact, numerous genes included in the list of the G64 module have already been identified as cancer diagnostic or prognostic markers by several independent researchers [14-16]. For instance, CDCA5 and NCAPH have been characterized to be highly expressed in lung cancers [17, 18]. Our previous study showed that cell cycle genes were a main functional GO category in the G64 module. Consistently, dysregulation of the cell cycle has long been proven to be a critical process causing cancers [19-22]. However, to our knowledge, it was the first finding ever that G64 genes were co-regulated in various cancers, and patient prognosis can be differentiated by these genes. Here, we clearly showed that the G64 module can be a good biomarker, predicting cancer prognosis at least for six different cancers—i.e., seven different cancers if LADC is included.

It is unclear what common molecular characteristics the seven cancers (i.e., 6 cancers in the present study and LADC) share together other than co-regulated G64 expression. The seven cancers must be initiated and progress by different genetic or environmental causes and pathways, and there is no proper explanation for the common subpopulation differentiation carried through G64 expression. It is also hard to explain at this moment why the 64 genes are co-regulated on an inter-tumoral level and intra-tumoral level and why up-regulating patients tend to have a poor prognosis. One possibility we can think of is that expression of these genes may be associated with drug metabolism, contributing to drug sensitivity or drug resistance. It will be interesting to study further how expression of G64 is changed during the recurrence of cancers. Taken together, G64 genes would be a good candidate of developing prognostic markers for multiple cancers.

Supplementary materials

Supplementary data, including one table and eight figures,



Groups	p value	*H.R	*95% C.I
Low vs High	0.000	3.00	2.18 ~ 4.13
INT vs Low	0.561	0.92	0.69 ~ 1.22
INT vs High	0.000	2.61	2.09 ~ 3.25

Groups	p value	*H.R	*95% C.I
Low vs High	0.523	0.94	0.76 ~ 1.15
INT vs Low	0.806	1.02	0.87 ~ 1.20
INT vs High	0.553	0.95	0.81 ~ 1.12

Fig. 2. KM analysis of combined samples of cancers on expression of G64. (A) KM plot of the combined six samples of the survival-differentiating cancers by G64, including kidney renal clear cell carcinoma, brain lower-grade glioma, liver hepatocellular carcinoma, kidney renal papilloma cell carcinoma, esophageal carcinoma, and sarcoma (see Table 1, upper six cancers). (B) KM plot of the 16 remaining samples of the survival-non-differentiating cancers (see Table 1, lower 16 cancers). Low, the lowest 15% group; INT, intermediate group; High, the highest 15% group; HR, hazard ratio; 95% CI, 95% confidence interval.

can be found with this article online at http://www.genominfo.org/src/sm/gni-13-132-s001.pdf.

Acknowledgments

This research was supported by a 2015 research grant from Kangwon National University (No. 520150328) to S.S.C.

References

1. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. *Nature* 2013;501:355-364.
2. Cusnir M, Cavalcante L. Inter-tumor heterogeneity. *Hum Vaccin Immunother* 2012;8:1143-1145.
3. Lyng H, Vorren AO, Sundfor K, Taksdal I, Lien HH, Kaalhus O, *et al*. Intra- and intertumor heterogeneity in blood perfusion of human cervical cancer before treatment and after radiotherapy. *Int J Cancer* 2001;96:182-190.
4. Jiang YH, Yuen RK, Jin X, Wang M, Chen N, Wu X, *et al*. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am J Hum Genet* 2013;93:249-263.
5. Pinto N, Dolan ME. Clinically relevant genetic variations in drug metabolizing enzymes. *Curr Drug Metab* 2011;12:487-497.
6. Yasuda SU, Zhang L, Huang SM. The role of ethnicity in variability in response to drugs: focus on clinical pharmacology studies. *Clin Pharmacol Ther* 2008;84:417-423.
7. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, *et al*. Mutational landscape and significance across 12 major cancer types. *Nature* 2013;502:333-339.
8. Roukos D, Batsis C, Baltogiannis G. Assessing tumor heterogeneity and emergence mutations using next-generation sequencing for overcoming cancer drugs resistance. *Expert Rev Anticancer Ther* 2012;12:1245-1248.
9. Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, Boerno ST, *et al*. Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. *PLoS One* 2010;5:e15661.
10. Min JW, Kim WJ, Han JA, Jung YJ, Kim KT, Park WY, *et al*. Identification of Distinct Tumor Subpopulations in Lung Adenocarcinoma via Single-Cell RNA-seq. *PLoS One* 2015; 10:e0135817.
11. Miyagi A, Takahashi H, Takahara K, Hirabayashi T, Nishimura Y, Tezuka T, *et al*. Principal component and hierarchical clustering analysis of metabolites in destructive weeds; polygonaceous plants. *Metabolomics* 2010;6:146-155.
12. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Methodol* 1972;34:187-220.
13. Efron B. Logistic regression, survival analysis, and the Kaplan-Meier curve. *J Am Stat Assoc* 1988;83:414-425.
14. Hayama S, Daigo Y, Yamabuki T, Hirata D, Kato T, Miyamoto M, *et al*. Phosphorylation and activation of cell division cycle associated 8 by aurora kinase B plays a significant role in human lung carcinogenesis. *Cancer Res* 2007;67:4113-4122.
15. Soria JC, Jang SJ, Khuri FR, Hassan K, Liu D, Hong WK, *et al*. Overexpression of cyclin B1 in early-stage non-small cell lung cancer and its clinical implication. *Cancer Res* 2000;60:4000-4004.
16. Zhong X, Guan X, Dong Q, Yang S, Liu W, Zhang L. Examining Nek2 as a better proliferation marker in non-small cell lung cancer prognosis. *Tumour Biol* 2014;35:7155-7162.
17. Beer D, Taylor J, Chen G, Kim S. Lung cancer signature. United States patent US 20120295803 A1. 2012 Nov 22.
18. Nguyen MH, Koinuma J, Ueda K, Ito T, Tsuchiya E, Nakamura Y, *et al*. Phosphorylation and activation of cell division cycle associated 5 by mitogen-activated protein kinase play a crucial role in human lung carcinogenesis. *Cancer Res* 2010;70: 5337-5347.
19. Goodwin G, Johns E, Lehn D, Landsman D, Wright J, Ferrari S, *et al*. A cell cycle regulator potentially involved in genesis of many tumor types. *Biol Chem* 1986;261: 2274.
20. Pan H, Griep AE. Altered cell cycle regulation in the lens of HPV-16 E6 or E7 transgenic mice: implications for tumor suppressor gene function in development. *Genes Dev* 1994;8: 1285-1299.
21. Sherr CJ. Cancer cell cycles. *Science* 1996;274:1672-1677.
22. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, *et al*. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* 2002;13:1977-2000.