Biochemical and Biophysical Research Communications xxx (2017) 1–6

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Patient sample-oriented analysis of gene expression highlights extracellular signatures in breast cancer progression

Yourae Hong, Nayoung Kim, Chao Li, Euna Jeong, Sukjoon Yoon*

Center for Advanced Bioinformatics & Systems Medicine, Department of Biological Sciences, Sookmyung Women's University, Hyochangwon-gil 52, Yongsan-gu, Seoul, 140-742, Republic of Korea

ARTICLE INFO

Article history:
Received 5 April 2017
Accepted 11 April 2017
Available online xxx

Keywords:
Cancer transcriptome
Gene ontology
Collagen
Breast invasive carcinoma (BRCA)

ABSTRACT

Although a large collection of cancer cell lines are useful surrogates for patient samples, the physiological relevance of observed molecular phenotypes in cell lines remains controversial. Because transcriptome data are a representative set of molecular phenotypes in cancers, we systematically analyzed the discrepancy of global gene expression profiles between patient samples and cell lines in breast cancers. While the majority of genes exhibited general consistency between patient samples and cell lines, the expression of genes in the categories of extracellular matrix, collagen trimers, receptor activity, catalytic activity and transporter activity were significantly up-regulated only in tissue samples. Genes in the extracellular matrix, particularly collagen trimers, showed a wide variation of expression in tissue, but minimal expression and variation in cell lines. Further analysis of tissue samples exclusively revealed that collagen genes exhibited a cancer stage-dependent expressional variation based on their supramolecular structure. Prognostic collagen biomarkers associated with survival rate were also readily predicted from tissue-oriented transcriptome analysis. This study presents the limitations of cell lines and the exclusive features of tissue samples in terms of functional categories of the cancer transcriptome.

1. Introduction

The availability of large amount of cancer omics data is expected to contribute to a better understanding of cancer heterogeneity and the development of patient-oriented therapies. Although sequence-based genotyping provides extensive information on cancer-specific somatic mutations, only a limited number of alterations within a few oncogenes, such as RAS/RAF, EGFR and PI3K, have shown clinical efficacy as drug targets and/or biomarkers for patient selection [1,2]. Considering the varied nature of a mutant gene in allele frequency, substitution type and mutation position among cancer samples, molecular phenotypic signatures derived from cancer transcriptome or proteome data may provide alternatives for better prediction of the drug response or patient selection [3–6]. Particularly, the massive accumulation of transcriptome data from both patient samples and cell lines represents an attractive resource for identifying non-genetic biomarkers in cancer therapies [7]. Representatively, the Cancer Genome Atlas (TCGA) released multi-level omics datasets generated from thousands of patients' tissue samples covering 12 cancer types [8]. Furthermore, transcriptome data of approximately 1000 well-defined cancer cell lines covering approximately 30 cancer types have been widely used as a surrogate for studying the association with cancer progression and drug responses [9]. However, because there are genomic differences between cell lines and tissue samples from an identical origin [10,11], it is necessary to systematically investigate the consistency or discrepancy of transcriptomes between patient samples and cell lines in terms of the functional diversity of genes within the transcriptome.

In this study, we retrieved whole transcriptome data of breast cancer (BRCA) patients' samples and cell lines from TCGA and the Cancer Cell Line Encyclopedia (CCLE), respectively. For global comparative transcriptome analysis, we categorized thousands of genes based on their functional annotation in Gene Ontology (GO). Although patient samples and cell lines exhibited general consistency in lineage-specific transcriptome signatures, we expected that differences in microenvironment might lead to different transcriptional phenotypes. Thus we attempted to quantitatively describe the exclusive signatures in patient samples in terms of functional categories of genes. Particularly, intercellular interactions and the heterogenic microenvironment were unique

* Corresponding author.
E-mail address: yoonsj@sookmyung.ac.kr (S. Yoon).

http://dx.doi.org/10.1016/j.bbrc.2017.04.055
0006-291X/

features in patient tumors but were minimally represented by in vitro cell lines. Therefore, it is thus necessary to comparatively investigate transcriptional activities related to receptors, transporters and extracellular matrices between patient samples and cell lines. The extracellular matrix (ECM) is a major scaffold for the reconstruction of tissues [12]. This matrix contains structural and functional proteins, such as collagen, laminins, fibronectin, and linker proteins, constituting the basement membrane and the interstitial matrix. The ECM is important in cancer progression, where it shows abnormalities and remodeling [13,14]. In particular, collagen is the most abundant component of the ECM, implying its critical role in cancer progression that is not appropriately captured in cell line samples [15]. In this study, we further analyzed the exclusive transcriptional profiles of the collagen family in BRCA patient samples. Through the identification of discrepant transcriptomes between tissue samples and cell lines, this study will contribute to more effective utilization of cell line and tissue data in cancer studies.

2. Materials and methods

2.1. Data acquisition

The gene expression datasets of the cell lines and tissue samples were obtained from the CCLE database and TCGA data portal, respectively [8,9]. The gene expression dataset for the cell lines obtained from the CCLE database was created using Affymetrix U133 plus 2.0 array chips and normalized with the robust multichip average (RMA) method. For the publicly available tissue panel dataset updated in 2013, the RNASeq version 2 dataset was evaluated using the Illumina HiSeq 2000 and Illumina Genome Analyzer (GA) platforms. The expression signal of each gene was normalized based on the RNA-Seq by expectation maximization (RSEM) count estimates method and converted to the log scale. To compare the same cancer type in the cell lines and tissue panel, the CCLE cell lines were re-annotated using the Genomics of Drug Sensitivity in Cancer (GDSC) data for cancer types (Supplemental Table 1) [16]. To compare genes between TCGA and CCLE, we first selected common genes in the 2 datasets. There were 17,673 selected genes with 37,148 gene probes.

For the tissue samples, clinical data were downloaded from the TCGA data portal. This database describes specific metrics, such as ‘pathologic stage’ information, allowing for the classification of the cancer samples into four cancer stages, as well as survival times from diagnosis to ‘days to death’ and ‘days to last follow-up.’

2.2. Selection of gene ontology (GO) terms using gene set enrichment analysis (GSEA)

GO terms were used to identify the function of the gene or protein. Specifically, GO terms selectively enriched in tissue samples were analyzed using GSEA [17]. Additionally, selected GO terms were downloaded from the Gene Expression Omnibus (GEO), GPL570.

2.3. Heatmap analysis of collagen family genes

To compare collagen gene expression during cancer progression (cancer stage I, II, III, and IV), collagen genes were analyzed using the hierarchical clustering software, QCanvas, which can cluster and visualize the results [18]. This software can be freely downloaded from the Sookmyung CBiS (http://compbio.sookmyung.ac.kr/~qcanvas/), Center for Advanced Bioinformatics & Systems Medicine.

2.4. Acquisition and classification of collagen supramolecular structure

To more precisely understand its function, we classified collagen according to its structure and types of domains. The domain information and structure information were obtained from published papers [19]. In the TCGA RNA sequencing dataset, 43 collagen genes were detected and divided into 8 structures, anchor fibril, fibril-forming, hexagonal network, transmembrane, multiplexin, network-forming, beaded filament and FACITs (fibril-associated collagens with interrupted triple helices) collagens.

2.5. Survival analysis

The overall survival of stage II/III breast cancer invasive carcinoma (BRCA) patients was analyzed using the Kaplan-Meier method and the log-rank test via R. The patients were divided into two groups, based on low (>2-fold the median) or high (<-2-fold the median) expression of particular genes in the stage II/III samples.

3. Results and discussion

3.1. Comparison of gene expression between tissues and cell lines in BRCA

In our previous study, we found that the expression levels of two ECM genes, cartilage oligomeric matrix protein (COMP) and collagen X (COL10A1), were specifically associated with breast cancer tissues [20]. In the present study, we investigated the expression profiles of these genes in 14 cancer types using both TCGA patient samples and CCLE cell line panels. The significant over-expression of COMP and COL10A1 was confirmed in BRCA patient tissue samples, while the expression of these genes was greatly decreased in most other cancer types (Fig. 1A). Interestingly, this selective gene expression pattern of two proteins in BRCA was not reproduced in the analysis of the cell line panel of same lineage diversity (Fig. 1A), implying that the transcriptome profile of the cell lines differed in relation to ECM genes.

We thus systematically compared the global transcriptome profile of BRCA between 5594 tissue samples and 432 cell lines. A total of 16,344 genes were assigned to at least one gene set in GO, and comparative GSEA analysis of gene expression was then performed between tissue data and cell line data. The majority of gene sets showed no significant difference ($P < 0.01$) in their GSEA score (i.e., expression level) between the two groups. However, five categories of gene sets including ECM, collagen trimers, receptor activity, catalytic activity and transporter activity, revealed significant over-enrichment in tissue samples (Fig. 1C). Most of the over-enriched categories were related to functions in intercellular interactions and the microenvironment. Although advanced 3D and co-cultures provide physiologically relevant conditions for cell line cultures, the present analysis indicates that reference transcriptome data obtained under monolayer culture conditions might not capture the signatures of patient tumors, particularly those related to extracellular activities.

Three of the over-enriched gene sets belonged to the category of molecular function in GO (Fig. 2A). The genes in these three sets exhibited a high level of expression with small variation in tissue samples, whereas their expression was widely varied in the cell lines (Fig. 2A). This finding implies that the gene expression levels in these categories undergo significant down-regulation in cell lines compared to the expression level in tissue samples. Consistently, previous reports have shown that ECM-receptor interaction gene sets are down-regulated in BRCA cell lines [21]. The other

Please cite this article in press as: Y. Hong, et al., Patient sample-oriented analysis of gene expression highlights extracellular signatures in breast cancer progression, Biochemical and Biophysical Research Communications (2017), http://dx.doi.org/10.1016/j.bbrc.2017.04.055