SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
MICROARRAYS
DATA MINING

Received
30 June 2014

Accepted
8 August 2014

Published
13 October 2014

Correspondence and requests for materials should be addressed to S.J. (jiaosc@vip.sina.com)

* These authors contributed equally to this work.

Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications

Weihong Zhao[1]*, Jiancheng Luo[2]* & Shunchang Jiao[1]

[1]Department of Medical Oncology, the General Hospital of the People's Liberation Army, Beijing, China, [2]Solomonbrothers Medical Institute, Wilmington, New Castle DE, USA.

Long non-coding RNAs (lncRNAs) are a kind of RNAs with regulation that participate fundamental cellular processes via diverse mechanisms. Despite the potential importance of lncRNAs in multiple kinds of cancer has been well studied, no comprehensive survey of cancer subtype associated lncRNAs. Here, we performed an array-based transcriptional survey of lncRNAs across 150 lung cancer samples comprising both adenocarcinoma and squamous cell carcinoma, and 306 breast cancer patients with clear clinical information. In lung cancer, 72 lncRNAs are identified to be associated with tumor subtypes and their functions as well as the associated proteins are predicted by constructing coding-non-coding co-expression network. The results suggest that they are mostly related with epidermis development, cell adhesion and response to stimulus. The validation results show the high concordance and confirmed the robust of the identification results. In breast cancer, we found 3 lncRNA genes are associated with estrogen receptor α (ER) positive and ER negative subtypes and tumor histologic grade. Survival (Kaplan-Meier) analysis results suggest that the expression pattern of the 3 lncRNAs is significantly correlated with clinical outcomes. The current study provides the first large-scale survey of lncRNAs within cancer subtypes, and may offer new targets for their diagnosis, therapy and prognosis.

Lung cancers and breast cancers can be classifiled into various subtypes on the basis of molecular, histological and clinical characteristics[1,2]. Different subgroups are associated with different clinical outcomes, suggesting a biologic basis behind the clinical heterogeneity of these cancers. Lung adenocarcinoma and squamous cell carcinoma are currently the most common types of non-small-cell lung carcinoma (NSCLC), and account for the majority of lung cancer deaths worldwide[3]. Comparably, adenocarcinoma was more often seen peripherally in the lungs and more common in never smokers, whereas squamous cell carcinoma tended to be more often centrally located and closely correlated with a history of tobacco smoking[4]. Furthermore, transition from squamous cell carcinoma to adenocarcinoma is also observed[5]. These results suggest complex dynamic biological process between these two types of lung cancer, which may include multiple steps of transcriptom alterations including aberrations in expression of both protein-coding and noncoding RNAs[6]. Breast cancer could be classified into ER positive and ER negative subtypes according to ER status. Approximately two-thirds of all breast cancer patients are ER positive at the time of diagnosis[7]. Previous studies demonstrated that ER positive and ER negative tumors display remarkably different gene expression patterns not solely explained by differences in estrogen responsiveness[7,8]. Furthermore, different ER status contributes to different clinical outcomes[8,9]. Thus, a detailed survey of the transcriptom difference between those tumor subtypes is essential, which may contribute to the diagnosis and treatment of each subtype of these cancer. Recently, accumulating evidence has shown that long non-coding RNAs (lncRNAs) may play critical roles in multiple cancers and may provide new insights into the molecular basis underlying the cancer subtypes[10,11].

LncRNAs, whose transcript length is more than 200 nt, have been found to be pervasively transcribed in the mammalian genome[12]. Functional mechanisms of lncRNAs include chromatin modification and gene expression regulation in a *cis* or *trans* manner[13]. Several lncRNAs were found as oncogenic or tumor-suppressor genes. For example, a lncRNA gene, *HOTAIR (Hox transcript antisense intergenic RNA)* is significantly highly expressed in NSCLC tissues and cell lines, and regulates NSCLC cell invasion *in vitro* and cell metastasis *in vivo*[10]. Moreover,

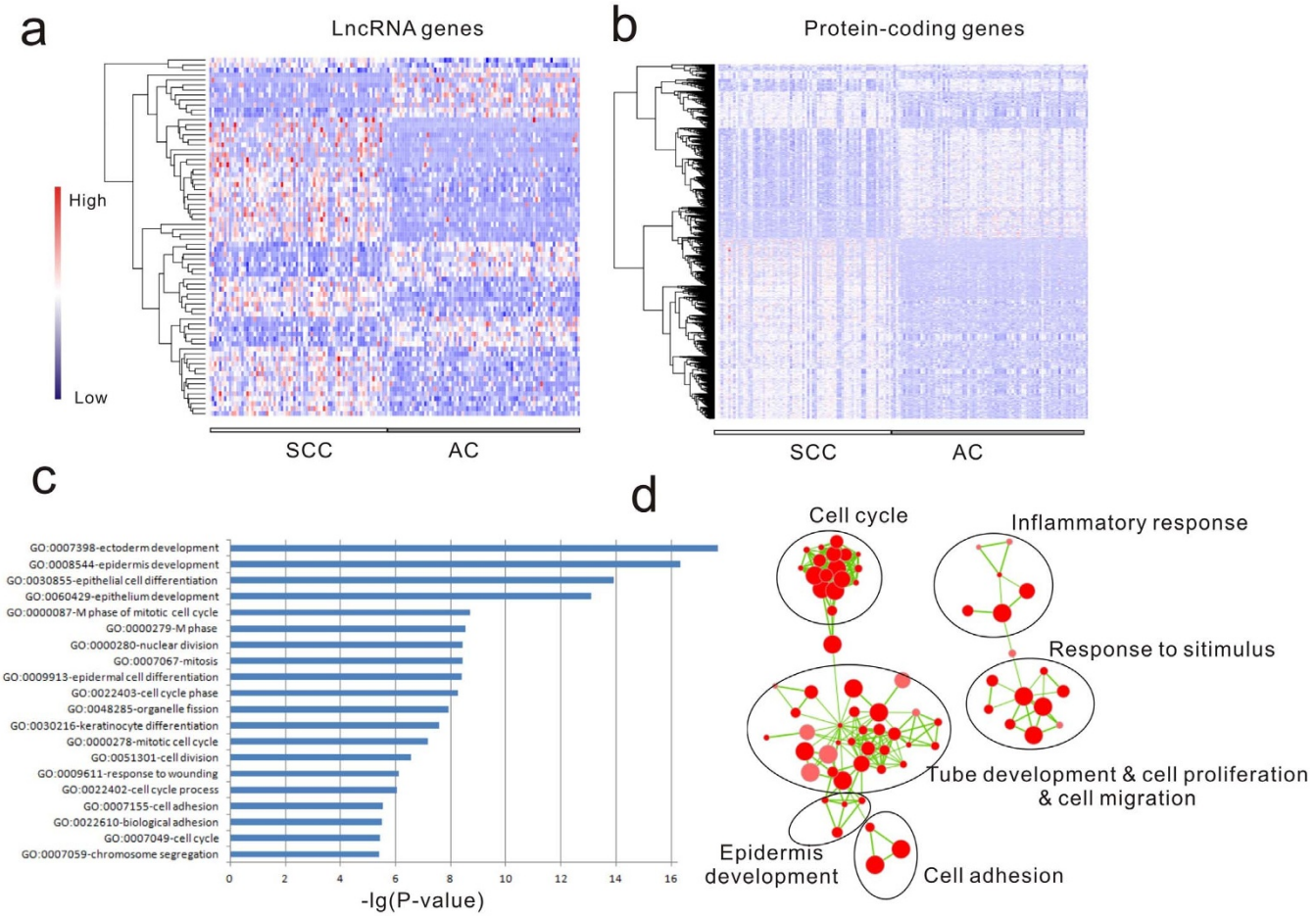

Figure 1 | Expression profile and functional enrichment of cancer subtype associated genes. (a) and (b) denote the expression profiles of lncRNA and protein-coding genes respectively. SCC: squamous cell carcinoma; AC: adenocarcinoma; (c) The functional enrichment results of differentially expressed protein-coding genes. (d) The functional enrichment map of GO terms with each node represents an GO term and an edge represents existing genes shared between connecting GO terms. Node size denotes the number of gene in the GO term. The main functional processes are marked for each group of GO terms.

HOTAIR is a predictor of breast cancer patient survival, and increased *HOTAIR* expression in patients correlated with enhanced breast cancer metastasis[14]. Another example is *MALAT-1* (*Metastasis-Associated-in-Lung-Adenocarcinoma-Transcript-1*), which is highly expressed in several human NSCLC cell lines. Highly *MALAT-1* gene expression in lung squamous cell carcinoma was associated with a poor prognosis[11]. The dysregulated lncRNAs in cancers suggest they are key components in biological network and may participate in tumorigenesis and metastasis. Dynamic changes in lncRNA expression have been observed across different cancer lines and different stages of cancer development[15]. However, our understanding of the lncRNAs biological contributions to histological subtypes of cancer is still unclear.

In order to study the functional significance of lncRNAs in cancer subtypes, we carry out a comprehensive study of the lncRNAs across 150 lung cancer samples comprising both adenocarcinoma and squamous cell carcinoma, and 306 breast cancer patients with clear clinical information. To provide more insights into tumor subtype associated lncRNAs, we investigate the whole-transcriptomic landscape of co-expressed relationships between lncRNAs and protein-coding genes and accordingly predicted hundreds of lncRNA functions. Furthermore, in breast cancer, by integrating clinical information we performed survival analysis and shown that the expression pattern of identified lncRNAs are associated with clinical outcomes. To the best of our knowledge, the current study represents the first exploration of lncRNAs and their functional as well as clinical significance within histological subtypes of cancer.

Results

Transcriptomic landscape of lung adenocarcinoma and squamous cell carcinoma. To obtain the globe expression profiles of lncRNA genes, we re-annotated the entire collection of probe sets for human Affymetrix microarrays (HGU133plus2.0) using ncFANs utility[16], which enable us to profile 2,812 lncRNA and 17,282 protein-coding genes simultaneously. Then, we examined the whole transcriptomic pattern across 150 non-small cell lung cancer samples including stage I of 41 adenocarcinoma, 36 stage II of adenocarcinoma, 34 stage I of squamous cell carcinoma, and 39 stage II of squamous cell carcinoma samples. To gain a detailed understanding of the biological significance of these transcripts in cancer subtypes, we further performed gene expression difference analysis and totally identified thousands of lncRNAs and protein-coding genes that are significantly differentially expressed between lung adenocarcinoma and squamous cell carcinoma samples (detail described in the Materials and Methods) (Figure 1a and 1b). Specifically, there are 72 lncRNAs and 1,191 coding genes that represent about 2.6% and 6.9% of the corresponding total gene numbers in microarrays (Supplementary File 1). Gene Ontology (GO) biological process enrichment analysis of all differentially expressed protein-coding genes demonstrates that ectoderm development, epidermis development, epithelial cell differentiation are the three most significantly enriched functions (Figure 1c). Those consist with the previous study and reflect histological difference between adenocarcinoma and squamous cell carcinoma[17,18]. Other

Figure 2 | Genomic context of *NKX2-1-AS1* (a) and *DSCAM-AS1* (b). Evolutionary conservation status as shown under the gene is measured by multiple alignments of 100 and 46 vertebrate species respectively.

enrichment results include cell cycle process, cell proliferation, cell adhesion and response to stimulus, which may reflect the different pathogenesis and prognosis of the two cancer subtypes (Figure 1d, Supplementary File 2).

Characterization of cancer subtype associated lncRNAs. We next sought to characterize lncRNAs that showed significant expression differences between lung adenocarcinoma and squamous cell carcinoma samples and found 23 upregulated and 49 down-regulated lncRNAs (P value < 0.01, fold change ≥ 1.5) (Figure 1a, Supplementary File 1). For example, expression of *Xist* (*inactive X chromosome-specific transcripts*) lncRNA has been demonstrated to be associated with human cancers and correlated with cancer outcomes[19,20], and it is also significantly increased in our analysis of lung adenocarcinoma samples. In addition, we found two upregulated antisense lncRNA genes, *NKX2-1-AS1* and *DSCAM-AS1*, whose host protein-coding genes are also associated with lung cancer (Figure 2). Specifically, *NKX2-1-AS1* is 1,775 nt with 2 exons, whose complementary gene is *TTF-1* (thyroid transcription factor-1; also known as *Nkx2-1*) (Figure 2a). *TTF-1* is a nuclear transcription factor that is expressed in lung and thyroid tissues, and is effective in distinguishing lung adenocarcinoma from pleural mesothelioma, that may enable it as a highly specific marker for lung adenocarcinoma in body cavity fluids[21,22]. Another is *DSCAM* (*Down Syndrome Cell Adhesion Molecule*) antisense lncRNA (*DSCAM-AS1*) that include four isoforms (Figure 2b). *DSCAM* is a member of the immunoglobulin superfamily of cell adhesion molecules, whose polymorphisms could influence overall survival in treatment of advanced NSCLC patients[23]. These differentially expressed antisense lncRNAs (such as *NKX2-1-AS1* and *DSCAM-AS1*) might interact together with their host genes to fulfill functions in different subtypes of lung cancer.

LncRNAs functional predictions and their co-expressed protein-coding genes. In order to study the functions and interactions of interesting lncRNAs, we constructed a coding-non-coding gene co-expression network (also called "two-color" network)[24] using the expression profiles of all 150 samples. The resulting network include 245 lncRNA genes and 4,567 coding genes with 65,417 connections, which comprising 61,669 coding-coding edges, 3,631 coding-noncoding edges and 117 noncoding-noncoding edges (Supplementary File 3). Next, we used hub-based subnetworks, in which lncRNAs as the hubs and surrounded by protein-coding genes, to predict the functions of lncRNA genes. To achieve this, we parsed the whole co-expression network into different hub-based subnetworks according to the topology. Among these subnetworks, 57 were lncRNA-centered with at least 10 co-expressed protein-coding genes and have at least one significantly enriched Gene Ontology term which include biological process (BP), molecular function (MF) and cellular component (CC) (Supplementary File 4). That means the functions of totally 57 lncRNAs could be predicted through such methods.

Next, we focused on the functions and mechanisms of differentially expressed lncRNAs between lung adenocarcinoma and squamous cell carcinoma. There are five differentially expressed lncRNA genes whose functions are predicted (Supplementary File 4). Consistent with enriched functions of differentially expressed protein-coding genes, these lncRNA genes are mostly related epidermis development, cell adhesion and response to stimulus (Supplementary File 4). Specifically, lncRNA gene, *SFTA1P* (*Surfactant Associated 1*), that is significantly up-regulated in adenocarcinoma, connected with 50 protein-coding genes (Figure 3). One of the most significant enriched functions of *SFTA1P* is surfactant homeostasis (Figure 3). Interestingly, in non-small-cell lung cancer, the expression of surfactant-associated protein (*SP-A*) occurs predominantly in lung adenocarcinomas[25].

Validation of differential expressed lncRNAs using an independent cohort of samples. To further validate our results, we performed the similar analysis pipeline using an entirely

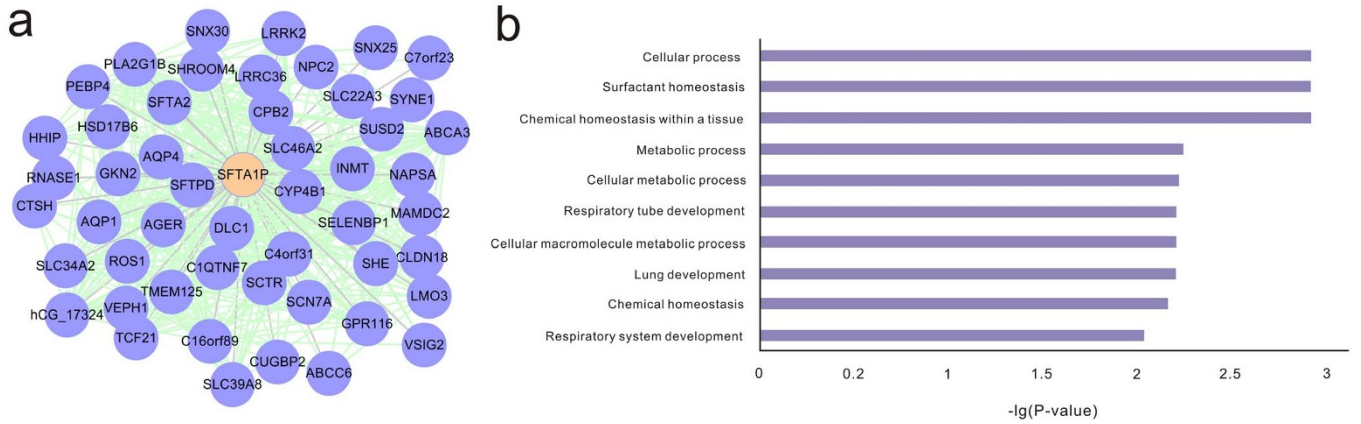

Figure 3 | The *SFTA1P*-centered subnetwork and the functional enrichment results of its neighboring coding genes. Red edges represent the connections between *SFTA1P* and coding genes, blue edges represent the connections between coding genes.

independent validation cohort of patients which was composed of 40 lung adenocarcinoma and 18 squamous cell carcinoma samples[1]. There are totally 1455 coding genes and 93 lncRNAs that were identified as cancer subtype associated genes using the same threshold. We observed 85% and 90% of original differently expressed gene set are concordances with the independent set, which suggest that the cancer subtype associated lncRNA gene set we identified in this study is robust (Figure 4a). Then, we performed principal-component analysis (PCA) using cancer subtype associated genes to test the ability to discriminate the lung adenocarcinoma from squamous cell carcinoma samples. The distinct expression patterns between cancer subtypes are clear shown in Figure 4b.

Breast tumor subtype and grade associated lncRNAs significantly correlate with clinical outcomes. We next carried out the similar analysis for breast cancer to identify lncRNA genes associated with ER positive and ER negative subtypes (Figure 5a). A total of 306 patients (712 microarray data, as described in the recent publications[2,26]) with clear clinical information were selected, including 39 ER negative, 261 ER positive and 6 status-unknown patients (Supplementary File 5). We used 119 samples without tamoxifen treatment to identify genes that were differentially expressed between ER (+) and ER (−) subtypes, so that the gene list identifiled was not affected by drug treatment. As a result, we obtained 307 protein-coding and 20 lncRNA genes that were significantly correlated with breast cancer subtypes. Previous studies have demonstrated that different subtypes of breast cancer could contribute to different prognosis[8,9]. Accordingly, we further examined the cancer subtype associated genes that were related with tumor grade (Figure 5a). Among 306 patients (266 of them have tumor grade information, see Supplementary File 5), 65 and 59 were assigned histologic grade 1 and 3 status corresponding to low and high risk of recurrence respectively, and 142 were classifiled as histologic grade 2 which denote intermediate risk of recurrence. For both tamoxifen-treatment and un-treatment samples, we carried out differential expression analysis between histologic grade 1 and 3 tumors independently. Totally, 291 protein-coding and 14 lncRNA genes were identified. By comparison with cancer subtype associated genes, 89 protein-coding and 3 lncRNA genes (*LINC00324*, *PTPRG-AS1* (*protein tyrosine phosphatase, receptor type, G, antisense*) and *SNHG17* (*small nucleolar RNA host gene 17*)) were in common (Supplementary File 6). We then performed functional enrichment of the 89 genes using GO biological process terms. The enrichment

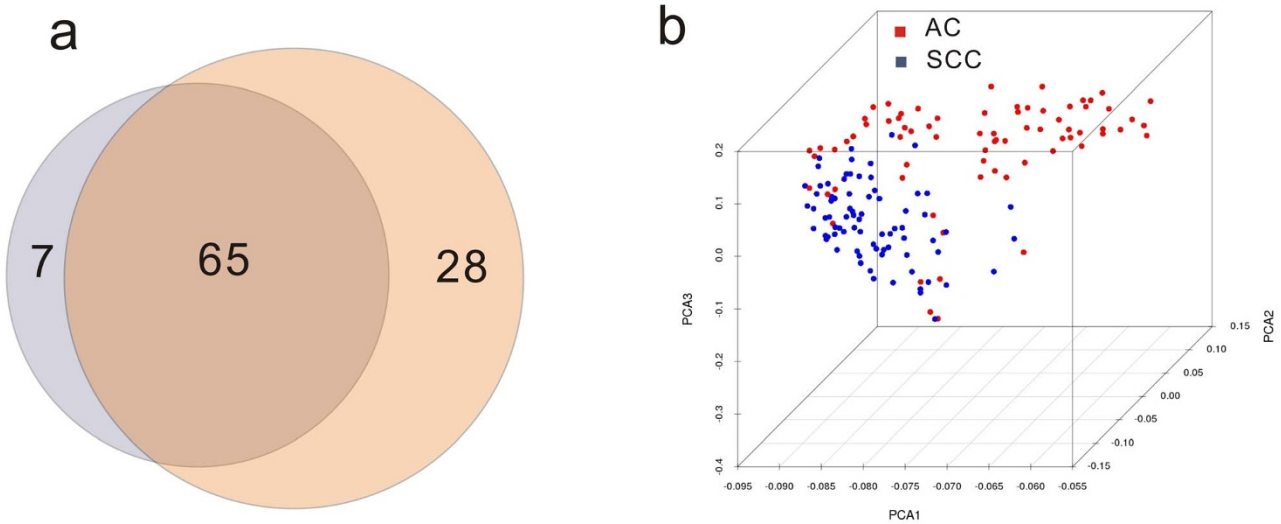

Figure 4 | Validation study of cancer subtype associated lncRNAs. (a). The number of cancer subtype associated lncRNAs (intersection) confirmed by an independent validation set (pink). (b). Principal-component analysis (PCA) of cancer subtype associated lncRNAs for lung adenocarcinoma and squamous cell carcinoma samples. PCA1, PCA2 and PCA3 represent the top three dimensions of cancer subtype associated lncRNA genes.

results suggested these genes were associated with cell cycle progression and proliferation process (Figure 5b), that showed the similar results with previous studies[2,26].

To investigate whether the three lncRNA genes identified by integrating ER status with histologic grade analysis were of clinical importance, unsupervised clustering and Kaplan-Meier survival analyses were performed (see Materials and Methods). First, 306 patients were classified into two groups (171 and 135 corresponding to group 1 and group 2, respectively) using k-means clustering methods based on patterns of expression levels of the three lncRNAs. Then, Kaplan-Meier survival analyses were performed comparing the two groups according to clustering results. As shown in Figure 5c, Kaplan–Meier curves showed a highly significant difference in relapse-free survival between the two groups (P value = 1.17e − 06). The result suggested the expression profiling of the three lncRNAs was able to stratify all patients into low- and high-risk groups. Specifically, the patients with statistically high expression of *LINC00324* gene and low expression of *PTPRG-AS1* and *SNHG17* gene (referred to group 1) associated with the long survival times. No significant difference was observed when performing the same clustering analyses on randomly selected lncRNA genes for ten times. Noteworthy, *PTPRG-AS1* is an antisense lncRNA, whose complementary gene is *PTPRG*. *PTPRG* is a member of the protein tyrosine phosphatase family that are known to regulate a variety of cellular processes, such as cell growth, mitotic cycle. Furthermore, *PTPRG* has been implicated as a tumor suppressor gene in breast, kidney and lung cancers[27,28]. Detail description of the lncRNAs can be found in discussion.

Discussion

The genomewide expression patterns for both coding and noncoding genes are a representation of the biology of the tumors; the difference in patterns reflects biological and histological diversity[8,9]. Thus, relating gene expression patterns to tumor subtypes is a key issue in understanding the molecular basis of tumorigenesis. The cellular and molecular heterogeneity in lung and breast tumors and the functional importance of lncRNA involved in controlling cell development emphasize the significance of studying tumor subtype associated lncRNAs in concert[13]. Several microarray studies carried out systematic investigation of the correlation between expression patterns of protein-coding genes and specific features of phenotypic variation, followed by the identification of predictive gene signatures for important clinical parameters such as relapse or overall survival[2,7–9,26]. As lncRNAs do not encode proteins and can function on the RNA molecular level, their functions are closely associated with their expressions. Although RNA sequencing (RNA-seq) is an effective way to study transcript abundance of lncRNA[29], the high cost and limited number of public datasets hindered its application in large-scale samples. In comparison, there are huge numbers of microarray data sets covering various biological and clinical conditions. Although lncRNAs are not considered in the original array design, a portion of microarray probes could perfectly match known lncRNAs, suggesting microarray probes can be reannotated for measuring lncRNA expression[24]. Here, we used this method to analyze large microarray datasets of lung and breast cancer, and obtained the expression profiles of both protein-coding and lncRNA genes simul-

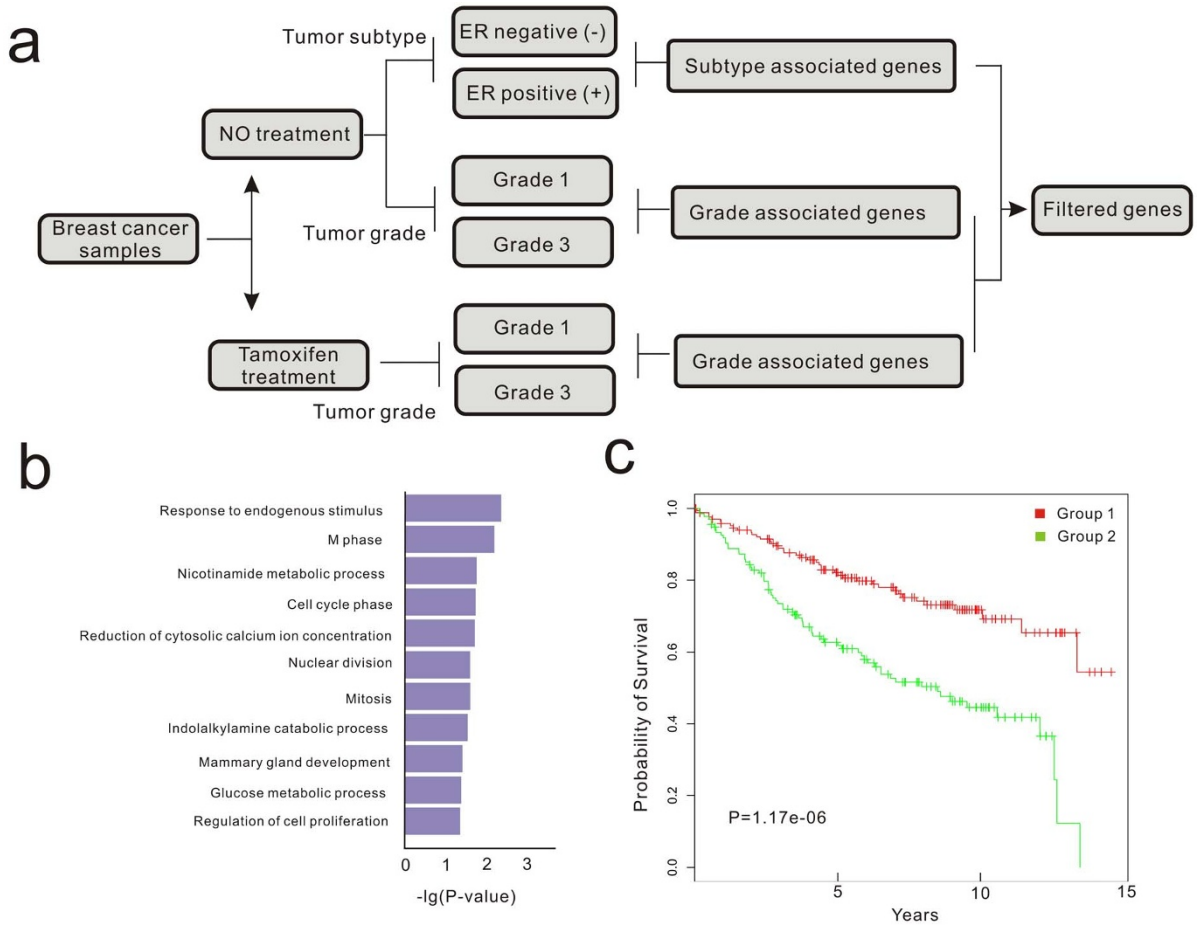

Figure 5 | Characterization of breast cancer subtype associated genes. (a). The overview of selecting lncRNA genes based on tumor subtype and grade. (b). The functional enrichment results of filtered protein-coding genes. (c). Kaplan-Meier survival curves for relapse free survival of two groups that clustered based on lncRNA expression pattern.

taneously. Due to the low technical variation and good detection sensitivity especially for low-abundance transcripts as well as in strand-specific manner, microarray probe reannotation method has been used in many other fields[16,24,30].

Lung cancer is generally heterogeneous and can be divided into multiple histological subtypes, each of which possesses specific biological and clinical behaviors[1]. Lung adenocarcinoma and squamous cell carcinoma are the most common types of NSCLC and have distinct pathogenesis and diagnosis. LncRNAs with subtype-specific expression may have important functions in certain histological subtypes, but the mechanisms still unclear mainly due to the lack of information to carry out the detail experimental study. Recent study have shown that acquired chromosomal aberrations play an important in common epithelial malignancies[31]. One such alteration in lung cancer is amplifilcation of 14q13.3, which contains the *TTF1* gene locus including both *TTF1* and its antisense lncRNA gene (*NKX2-1-AS1*) (Figure 2a). Increased *TTF1* expression is signifilcantly associated with multiple clinicopathological factors including tumour stage and overall survival among lung adenocarcinoma patients. In contrast, squamous cell carcinoma patients shown a decreased *TTF1* expression level, and no signifilcant association was observed between *TTF1* expression and clinicopathological factors[31]. Nevertheless, although amplification of *TTF1* gene locus portends a remarkable event in lung adenocarcinoma, the mechanism accounting for their association remains unknown. In current study, we found an interesting lncRNA, *NKX2-1-AS1*, which located in the opposite strand of *TTF1* and shown significant difference in expression level between lung adenocarcinoma and squamous cell carcinoma. The amplification of 14q13.3 could influence the expression of both *TTF1* and *NKX2-1-AS1* gene, and the dysregulation of the sense-antisense pair could be an initiative event in the development of lung adenocarcinoma. Further study will be necessary to determine whether *NKX2-1-AS1* expression directly and independently accounts for the aberration of *TTF1* expression in lung adenocarcinoma, and which proteins or pathways are effected by the dysregulation of this sense-antisense gene pair.

Other lung cancer subtype associated lncRNAs was also observed for cell-cell adhesions. As we known, dysfunction of genes in cell junction, a form of cell adhesions structure, can lead to tumorigenesis, tumor development and metastasis[1]. And two types of genes involved in cell junctions have been observed to be associated with lung cancer subtypes. Specifically, the genes account for desmosomes and gap junctions were highly expressed and tight junction genes were generally less expressed in squamous cell carcinoma compared to adenocarcinoma[1]. However, how these genes interact together to regulate cell-cell adhesions and further contribute to the different types of lung cancer remains unsettled. In this study, we identified several lung cancer subtype associated lncRNAs that may play potential roles in cell adhesions. For example, *DSCAM-AS1* (Figure 2b) is transcribed from the antisense strand of *DSCAM* and is overexpressed in lung adenocarcinoma. According to previous studies, such lncRNA-mRNA pairs generally characterized by inverse regulation in mammals, but the regulational roles of lncRNAs in cell adhesions has not been well studied.

For the ER positive and ER negative subtypes of breast cancer, we identified three subtype associated lncRNAs that were also related with tumor grade, suggesting that a higher complex order of coordination exists during tumorigenesis. Moreover, the expression pattern of the three lncRNAs were significantly correlated with clinical outcomes. Interestingly, the three lncRNAs corresponding to different lncRNA categories, such as antisense, intergenic and small nucleolar RNA host gene. In general, sense-antisense gene pairs regulated each other in inverse manner[13]. Accordingly, high expression of *PTPRG-AS1* may down-regulated expression level of *PTPRG* gene, a tumor suppressor gene in breast cancer. This observation is in agreement with the finding that low expression of *PTPRG-AS1* associated with the long survival times. The other gene, *LINC00324*, is a long intergenic noncoding RNA (lincRNA), which located in the desert region of genome (intergenic region). The regulation mechanism of lincRNA gene is poorly understand. Some evidences suggested that lincRNAs could *in cis* regulate the expression of their neighboring genes[13,32]. We found the *LINC00324* is located in 774 bp downstream of the 3′UTR of *CTC1* gene, which participates in DNA replication process and plays an essential role in protecting telomeres from degradation[33]. Furthermore, previous study suggested that the variation in *TERT* (*telomerase reverse transcriptase*), a protein-coding gene that has important role in regulating telomerase and protect telomeres, is associated with ER negative breast cancer[34]. Nevertheless, the relationship between *LINC00324* and *CTC1* has not be elucidated. Overall, the current results would provide the new clues and valuable information for further experimental study to investigate the mechanisms of these pathways in tumorigenesis and identify specific biomarkers in diagnosis of cancer subtypes.

Methods

Data set description and array data processing. The microarray data of 150 lung cancer and 306 breast cancer samples used in this study were obtained from the Gene Expression Omnibus (GEO) database[35] and can be directly downloaded from the website (accession number: GSE43580 for lung cancer and GSE6532 for breast cancer). The validation dataset of 58 surgically treated patients with non-small cell lung cancer were obtain under accession number GSE10245. By means of re-annotation of microarray probes, ncFANs[16] could calculate the expression level of both protein-coding genes and lncRNA genes simultaneously. Firstly, all the microarray raw data (CEL format) are submitted to ncFANs as input files. Then, Robust Multichip Average (RMA) method are performed to calculate the expression values (log2-transformed) of both protein-coding and lncRNA genes.

Differentially expression analysis. For the microarray dataset, many biological replicates are performed for each sample. Student's t-Test analysis and Benjamini Hochberg (BH) FDR correction in R is used to identify both protein-coding and lncRNA genes with statistically significant differential expression. The genes with fold change cut-offs of >1.5 and BH FDR-adjusted P values < 0.01 are selected as differentially expressed genes.

Gene Ontology enrichment analysis. We estimated the functional enrichment of differentially expressed protein-coding genes using the DAVID Bioinformatics Tool[36] and reported the results for the Gene Ontology (GO)-FAT biological process (BP) terms[37]. The P value should be lower than 0.01. GO-FAT is a subset of the GO annotation set derived by eliminating broad GO terms that are high in the GO term tree hierarchy to avoid the redundancy of annotation sets and overshadowing of the broad terms when applying multiple testing corrections. Then, the GO enrichment results are visualized using the Enrichment Map[38] plugin in Cytoscape[39] to group the GO terms with similar functions.

Co-expression network construction. The expression profiles of 150 lung cancer data sets are used as input file of ncFANs to construct the coding-non-coding gene co-expression network ("two-color" co-expression network)[16]. Multiple criteria are used to filter insignificant edges in network. Specifically, genes (for both coding and non-coding genes) with expressional variance ranked in the top 75th percentile of are retained. Then, P value of Pearson correlation coefficient (Pcc) for each gene pair (including coding-coding, coding-lnRNA and lncRNA-lncRNA gene pairs) is estimated using Fisher's asymptotic test and adjusted using the Bonferroni multiple test correction. Only gene pairs with a P value no more than 0.01 and with a Pcc value ranked in the top or bottom 0.5 percentile for each gene are retained and regarded as co-expressed gene pairs.

LncRNA gene function prediction. Hub-based method embedded in ncFANs is used to obtain the functional characteristics of lncRNAs[16,24]. The main process involved in ncFANs is as follows. First, the co-expression network is parsed into many hub-based subnetworks, each of which consist of a central lncRNA gene and its directly connected protein-coding genes. Only lncRNA genes with ten or more immediate protein-coding neighbors with gene ontology (GO) annotations including biological process (BP), molecular function (MF) and cellular component (CC), were considered. For all neighboring coding genes of each lncRNA gene, GO enrichment analysis was performed. The P value of the functional enrichment (<0.01) is used as parameter in the function prediction of lncRNAs genes.

Survival analysis of breast cancer subtype associated lncRNA genes. The expression values of the lncRNA genes in all patients are extracted. The patients are clustered by k-means clustering method in R (with cluster parameter is set to 2). Survival curves are visualized using R and compared using log-rank tests ("survdiff" function in R) using the relapse-free survival times and events.

1. Kuner, R. *et al.* Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer* **63**, 32–38 (2009).
2. Loi, S. *et al.* Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol* **25**, 1239–1246 (2007).
3. Yildiz, O. *et al.* Facial nerve palsy: an unusual presenting feature of small cell lung cancer. *Case Rep Oncol* **4**, 35–38 (2011).
4. Sun, S., Schiller, J. H. & Gazdar, A. F. Lung cancer in never smokers–a different disease. *Nat Rev Cancer* **7**, 778–790 (2007).
5. Kanazawa, H. *et al.* Transition from squamous cell carcinoma to adenocarcinoma in adenosquamous carcinoma of the lung. *Am J Pathol* **156**, 1289–1298 (2000).
6. Hamamoto, J. *et al.* Identification of microRNAs differentially expressed between lung squamous cell carcinoma and lung adenocarcinoma. *Mol Med Rep* **8**, 456–462 (2013).
7. Gruvberger, S. *et al.* Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* **61**, 5979–5984 (2001).
8. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869–10874 (2001).
9. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418–8423 (2003).
10. Liu, X. H. *et al.* The long non-coding RNA HOTAIR indicates a poor prognosis and promotes metastasis in non-small cell lung cancer. *BMC Cancer* **13**, 464 (2013).
11. Schmidt, L. H. *et al.* The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. *J Thorac Oncol* **6**, 1984–1992 (2011).
12. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775–1789 (2012).
13. Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43**, 904–914 (2011).
14. Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
15. Ozgur, E. *et al.* Differential expression of long non-coding RNAs during genotoxic stress-induced apoptosis in HeLa and MCF-7 cells. *Clin Exp Med* **13**, 119–126 (2013).
16. Liao, Q. *et al.* ncFANs: a web server for functional annotation of long non-coding RNAs. *Nucleic Acids Res* **39**, W118–124 (2011).
17. McDoniels-Silvers, A. L., Nimri, C. F., Stoner, G. D., Lubet, R. A. & You, M. Differential gene expression in human lung adenocarcinomas and squamous cell carcinomas. *Clin Cancer Res* **8**, 1127–1138 (2002).
18. Hofmann, H. S. *et al.* Identification and classification of differentially expressed genes in non-small cell lung cancer by expression profiling on a global human 59.620-element oligonucleotide array. *Oncol Rep* **16**, 587–595 (2006).
19. Yildirim, E. *et al.* Xist RNA is a potent suppressor of hematologic cancer in mice. *Cell* **152**, 727–742 (2013).
20. Huang, K. C. *et al.* Relationship of XIST expression and responses of ovarian cancer to chemotherapy. *Mol Cancer Ther* **1**, 769–776 (2002).
21. Bakir, K., Kocer, N. E., Deniz, H. & Guldur, M. E. TTF-1 and surfactant-B as co-adjuvants in the diagnosis of lung adenocarcinoma and pleural mesothelioma. *Ann Diagn Pathol* **8**, 337–341 (2004).
22. Gomez-Fernandez, C., Jorda, M., Delgado, P. I. & Ganjei-Azar, P. Thyroid transcription factor 1: a marker for lung adenoarinoma in body cavity fluids. *Cancer* **96**, 289–293 (2002).
23. Wu, X. *et al.* Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy. *J Natl Cancer Inst* **103**, 817–825 (2011).
24. Liao, Q. *et al.* Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res* **39**, 3864–3878 (2011).
25. Linnoila, R. I., Mulshine, J. L., Steinberg, S. M. & Gazdar, A. F. Expression of surfactant-associated protein in non-small-cell lung cancer: a discriminant between biologic subsets. *J Natl Cancer Inst Monogr*, 61–66 (1992).
26. Li, J. *et al.* Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat Commun* **1**, 34 (2010).
27. Liu, S., Sugimoto, Y., Sorio, C., Tecchio, C. & Lin, Y. C. Function analysis of estrogenically regulated protein tyrosine phosphatase gamma (PTPgamma) in human breast cancer cell line MCF-7. *Oncogene* **23**, 1256–1262 (2004).
28. Zheng, J. *et al.* 17 beta-estradiol-regulated expression of protein tyrosine phosphatase gamma gene in cultured human normal breast and breast cancer cells. *Anticancer Res* **20**, 11–19 (2000).
29. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915–1927 (2011).
30. Du, Z. *et al.* Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol* **20**, 908–913 (2013).
31. Perner, S. *et al.* TTF1 expression in non-small cell lung carcinoma: association with TTF1 gene amplification and improved survival. *J Pathol* **217**, 65–72 (2009).
32. Dimitrova, N. *et al.* LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint. *Mol Cell* **54**, 777–790 (2014).
33. Chen, L. Y., Redon, S. & Lingner, J. The human CST complex is a terminator of telomerase activity. *Nature* **488**, 540–544 (2012).
34. Haiman, C. A. *et al.* A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet* **43**, 1210–1214 (2011).
35. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210 (2002).
36. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1–13 (2009).
37. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).
38. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).
39. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).

Acknowledgments

This work is funded by Beijing Municipal Science & Technology Commission. Grant Number: Z111107067311018.

Author contributions

S.J. designed experiments. W.Z. and J.L. analyzed data and wrote this manuscript. All authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Zhao, W., Luo, J. & Jiao, S. Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Sci. Rep.* **4**, 6591; DOI:10.1038/srep06591 (2014).