A proteogenomic profile of early lung adenocarcinomas by protein co-expression network and genomic alteration analysis

Toshihide Nishimura1,2,9, Haruhiko Nakamura1,2,9, Kien Thiam Tan3,9, De-Wei Zhuo3,9, Kiyonaga Fujii1,2,9, Hirotaka Koizumi4, Saeko Naruki4,9, Masayuki Takagi4, Naoki Furuya5, Yasufumi Kato5, Shu-Jen Chen3, Harubumi Kato7,8 & Hisashi Saji2

The tumourigenesis of early lung adenocarcinomas, including adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and lepidic predominant invasive adenocarcinoma (LPA), remains unclear. This study aimed to capture disease-related molecular networks characterising each subtype and tumorigenesis by assessing 14 lung adenocarcinomas (AIS, five; MIA, five; LPA, four). Protein–protein interaction networks significant to the three subtypes were elucidated by weighted gene co-expression network analysis and pairwise G-statistics based analysis. Pathway enrichment analysis for AIS involved extracellular matrix proteoglycans and neutrophil degranulation pathway relating to tumour growth and angiogenesis. Whereas no direct networks were found for MIA, proteins significant to MIA were involved in oncogenic transformation, epithelial-mesenchymal transition, and detoxification in the lung. LPA was associated with pathways of HSF1-mediated heat shock response regulation, DNA damage repair, cell cycle regulation, and mitosis. Genomic alteration analysis suggested that LPA had both somatic mutations with loss of function and copy number gains more frequent than MIA. Oncogenic drivers were detected in both MIA and LPA, and also LPA had a higher degree of copy number loss than MIA. Our findings may help identifying potential therapeutic targets and developing therapeutic strategies to improve patient outcomes.

Lung adenocarcinoma is the most common histological subtype of non-small-cell lung cancer (NSCLC), and it accounts for the highest prevalence rate1. Lung adenocarcinoma is thought to develop from cells in the distal bronchial epithelia, terminal bronchioles, and alveoli including Clara cells and alveolar type I or type II pneumocytes2. About 60% of lung adenocarcinoma subtypes have unique protein markers as oncogenic driver mutations such as *EGFR*, *ALK*, *ROS1*, *HER2*, *KRAS*, and *BRAF*. These are crucial biomarkers in molecular targeted therapy. However, around 40% of lung adenocarcinoma subtypes do not show oncogenic driver mutations, and the prognoses of these patients are relatively poor.

Recent advances in chest high-resolution computed tomography have helped to detect small adenocarcinoma nodules at earlier stages4,5. In 2011, the International Association for the Study of Lung Cancer, the American Thoracic Society and the European Respiratory Society proposed a new pathological classification of lung adenocarcinoma. This classification introduced the concepts of adenocarcinoma in situ (AIS) and minimally invasive adenocarcinoma (MIA), resulting in the elimination of the term “bronchioloalveolar carcinoma”6. While both AIS

---

1Department of Translational Medicine Informatics, St. Marianna University School of Medicine, Kawasaki, Kanagawa 216-8511, Japan. 2Department of Chest Surgery, St. Marianna University School of Medicine, Kawasaki, Kanagawa 216-8511, Japan. 3ACT Genomics Co., LTD., Taipei 114, Taiwan. 4Department of Pathology, St. Marianna University Hospital, Kawasaki, Kanagawa 216-8511, Japan. 5Division of Respiratory Medicine, Department of Internal Medicine, St. Marianna University School of Medicine, Kawasaki, Kanagawa 216-8511, Japan. 6Department of Thoracic Surgery, Kanto Central Hospital, Tokyo 158-8531, Japan. 7Tokyo Medical University, Tokyo 160-0023, Japan. 8International University of Health and Welfare, Tokyo 107-8402, Japan. 9These authors contributed equally: Toshihide Nishimura, Haruhiko Nakamura, Kien Thiam Tan, De-Wei Zhuo, Kiyonaga Fujii and Saeko Naruki. *email: t-nisimura@marianna-u.ac.jp
and MIA included tumours of 3 cm or smaller in size, more specifically, AIS was defined as a pre-invasive lesion showing pure lepidic growth without invasion, whereas MIA was specifically defined as a solitary tumour that also exhibited lepidic predominant growth but showing invasion of 5 mm or less. Conversely, lepidic predominant invasive adenocarcinoma (LPA) with more than 5 mm of invasion is categorised as an invasive adenocarcinoma according to the predominant histological pattern, which can either be acinar, papillary (PAP), micropapillary (MP), solid SLD, or a combination of the four variants.

Lepidic-type adenocarcinomas have been thought to encompass a stepwise progression from AIS to MIA to LPA. Generally, complete AIS or MIA resection leads to 100% recurrence-free 5-year survival. However, some previous studies have reported that LPA might recur after complete resection. AIS plus MIA and LPA have different prognoses after resection, and differential protein expressions are thought to be associated with cancer cell invasiveness in each subtype and play important roles in local recurrence and prognosis.

The proteogenomic analysis may be effective in understanding the carcinogenesis and tumorigenesis of each lung adenocarcinoma subtype and identifying specific protein markers to drive differential diagnosis and subtype-specific treatment. Therefore, it is crucial to understand the molecular and tumourigenesis profiles of early lung adenocarcinomas from the perspective of interactions and alterations of disease-related molecular networks.

Recent advances in mass spectrometry (MS) have made MS-based shotgun sequencing and quantitative analysis highly capable of identifying a large number of disease-related proteins expressed in clinical specimens. The use of laser microdissection (LMD) enables the collection of a certain type of target cells from formalin-fixed, paraffin-embedded (FFPE) cancer tissue sections. Remarkable advances have been made in next-generation sequencing (NGS), thus enabling complete massive parallel sequencing using only a small number of FFPE specimens to collect comprehensive information regarding copy number variations (CNVs) and mutations in cancer-related genes.

Weighted gene co-expression network analysis (WGCNA) is an extensively applied, unsupervised gene-clustering method designed based on the correlation network of gene expressions. We have recently identified the key protein modules that characterise small-cell lung carcinoma and large-cell neuroendocrine lung carcinoma using WGCNA and the systematic network analysis of clinical tissue proteome datasets.

Therefore, the present study aimed to identify key protein–protein interaction (PPI) networks that characterise each lung adenocarcinoma subtype and to attain genomic alteration information, which would reflect the disease nature and treatment outcomes in patients with lung adenocarcinoma.

**Results**

**Proteome datasets of lung adenocarcinoma.** MS-based proteomic analysis was conducted on 14 FFPE tissue specimens of lung adenocarcinomas: five AIS, five MIA, and four LPA. These specimens were selected for their preserved condition, tumour area, and well-clarified pathological diagnosis (Table 1). Presurgical treatment was not performed for any of these lung adenocarcinomas.

| Sample no | Histological type | Age (years) | Sex | Location | Tumour size on CT (mm) | Clinical TNM classification | Clinical stage | EGF mutation |
|-----------|------------------|-------------|-----|----------|-----------------------|-----------------------------|----------------|-------------|
| A. AIS (n = 5) | | | | | | | | |
| AIS_T53 | AIS | 53 | F | RS8 | 11 | cT1a | cN0 | cM0 | cIA | EGF(R) (+) |
| AIS_T54 | AIS | 74 | M | RS6 | 29 | cT1b | cN0 | cM0 | cIA | EGF(R) (−) |
| AIS_T56 | AIS | 69 | F | RS6 | 16 | cT1a | cN0 | cM0 | cIA | EGF(R) (−) |
| AIS_T58 | AIS | 78 | M | LS1 + 2 | 28 | cT1b | cN0 | cM0 | cIA | EGF(R) (−) |
| AIS_T59 | AIS | 68 | F | RS1 | 13 | cT1a | cN0 | cM0 | cIA | EGF(R) (+) |
| Average ± SD | | 68.4 ± 8.5 | | | | 19.4 ± 7.6 | | | |
| B. MIA (n = 5) | | | | | | | | |
| MIA_T73 | MIA | 58 | M | RS9 | 20 | cT1a | cN0 | cM0 | cIA | EGF(R) (+) |
| MIA_T74 | MIA | 67 | F | LS3 | 19 | cT1a | cN0 | cM0 | cIA | EGF(R) (+) |
| MIA_T75 | MIA | 77 | F | RS8 | 20 | cT1a | cN0 | cM0 | cIA | EGF(R) (−) |
| MIA_T79 | MIA | 61 | M | LS6 | 10 | cT1a | cN0 | cM0 | cIA | EGF(R) (+) |
| MIA_T80 | MIA | 63 | F | RS3 | 12 | cT1a | cN0 | cM0 | cIA | EGF(R) (−) |
| Average ± SD | | 65.2 ± 6.6 | | | | 16.2 ± 4.3 | | | |
| C. LPA (n = 4) | | | | | | | | |
| LPA_T85 | LPA | 68 | M | RS1 | 30 | cT1b | cN0 | cM0 | cIA | EGF(R) (−) |
| LPA_T87 | LPA | 73 | F | RS3 | 28 | cT1b | cN0 | cM0 | cIA | EGF(R) (+) |
| LPA_T88 | LPA | 59 | M | LS1 + 2 | 30 | cT1b | cN0 | cM0 | cIA | EGF(R) (−) |
| LPA_T89 | LPA | 67 | F | RS6 | 20 | cT1a | cN0 | cM0 | cIA | EGF(R) (−) |
| Average ± SD | | 66.8 ± 5.0 | | | | 27.0 ± 4.1 | | | |

**Table 1.** Clinicopathological information of patients with early lung adenocarcinoma. AIS, adenocarcinoma in situ; MIA, minimally invasive adenocarcinoma; LPA, lepidic predominant invasive adenocarcinoma; ADC, adenocarcinoma; ANOVA, analysis of variance.
A total of 2,023 proteins were identified; among these, about 48.6% were commonly expressed in the cancer cells of the three aforementioned subtypes of lung adenocarcinoma (Fig. 1). Overall, < 1% of unique to AIS and MIA each, whereas ca. 34% of the identified proteins were unique to LPA (Fig. 1A).

When a protein is expressed with its spectral count ($SpC \geq 0$ and in more than three samples of a subtype, the protein can be referred to be characteristic to the lung adenocarcinoma subtype. The spectral count ($SpC$) is the number of tandem MS/MS spectra assigned to each protein. Overall, 760, 673, and 1,357 proteins were characteristic to AIS, MIA, and LPA, respectively.

The gene ontology (GO) analysis was performed using the Protein Analysis Through Evolutionary Relationships (PANTHER) version 14.1 software program (Paul D. Thomas, University of Southern California, Los Angeles, CA, USA)\textsuperscript{20}, and presented its results quite similar among the three subtypes (Fig. 1B). Commonly to all three subtypes, proteins were abundantly associated with cellular process, localization, biological regulation, response to stimulus and metabolic process in biological process (GO), binding, structural molecule activity and catalytic activity in molecular function (GO), and cytoskeletal protein, transferase, oxidoreductase, nucleic acid binding, enzyme modulator, and hydrolase in protein class (GO) (Fig. 1B).
Identification of key protein modules by WGCNA. We constructed a weighted gene co-expression network and clustered all the identified proteins, and identified 49 protein modules (Fig. 2A). In the WGCNA, a soft threshold power of 10 was selected to define the adjacency matrix according to the criteria of approximate scale-free topology, with a minimum module size of 10 and a module detection sensitivity deepSplit of 4. The clinical traits for patients were set according to the lung adenocarcinoma subtype—AIS, MIA, or LPA. The correlations between the resultant modules and clinical traits were determined to identify the protein modules whose expressions were upregulated or downregulated in the subtype samples. A heatmap of the proteome abundance of eigen-proteins and samples and pairwise correlations between the modules in the heatmap of eigen-protein expressions were presented (Fig. 2B, C).

We identified several modules that showed high and/or moderate correlations with clinical traits (correlation: \(|r|>0.5\)) (Fig. 3). We performed multiple testing correction by the Benjamini–Hochberg method. Finally, the eight modules—WM26, WM27, WM29, WM30, WM31, WM32, WM33, and WM35 (indicated by the red dashed squares) associated with LPA were found to be statistically significant whereas none of the modules associated with AIS or MIA remained significant.

Identification of the PPI networks associated with AIS and MIA by pairwise G-statistical analysis of identified proteins. Trait analysis in WGCNA trends to overlook important modules for investigating differential disease mechanisms whereas WGCNA is a powerful tool in identifying the co-expression of molecular modules. Indeed, we could not identify key co-expression modules associated with AIS and MIA with high statistical confidence. We adopted the pairwise G-statistics approach\(^{21}\) that can identify individual proteins with significant differences in spectral counting (SpCs)-based proteome abundance among different patient groups, where William’s correction for continuity was applied to the \(2 \times 2\) tables\(^{13}\). The adjusted G-statistical calculation enabled us to handle the data containing small spectral counts including zero. A protein was defined to be significant to AIS or MIA when the protein was expressed in \(\geq 60\%\) samples of the group and has its relative abundance > 50% and \(q\) values < 0.05 that were corrected by the Benjamini–Hochberg method for pairwise
G-statistics p-values among the three groups. We thus identified thirty and eight proteins significant to AIS and MIA, respectively (Fig. 4).

Functional enrichment analysis of the PPI networks. To characterise the key protein networks, the biological association among the proteins in each network was analysed by mapping the network proteins in the human protein–protein interaction (PPI) network and among the biological pathways by pathway enrichment analysis (Figs. 5 and 6).

The PPI networks generated using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database22 were reconstructed using the CYTOSCAPE version 3.7.2 software program (Institute for Systems Biology, Seattle, WA, USA) for the protein networks of AIS and MIA identified by the pairwise G-statistics and the eight WGCNA network modules associated with LPA (WM26, WM27, WM29, WM30, WM31, WM32, WM33, and WM35). Top hub proteins were identified according to maximal clique centrality (MCC) by using the CYTOHUBBA PLUGIN23. Figure 5 shows the protein networks of AIS, MIA, and the top three WGCNA modules.
modules—WM 29, WM30 and WM31 highly correlated with LPA, where eigen- and/or hub-proteins in their networks were indicated by blue and red dotted circles, respectively. Figures S1 and S2 in Supplementary Information File 1 represent core protein networks and their pathways enriched for WM26, WM27, WM32, WM33, and WM35.

The enriched pathways of the protein networks associated with AIS (Fig. 5A) included (1) biological process (GO): extracellular matrix organisation, regulation of vascular endothelial growth factor (VEGFR) signalling pathway, and cell adhesion; (2) Reactome pathways: extracellular matrix organization, ECM proteoglycans, molecules associated with elastic fibres, and non-integrin membrane-ECM interactions. The top MCC-hub protein heparan sulfate proteoglycan 2 (HSPG2) (also known as Perlecan) is an integral component of basement membranes and interacts with other ECM components as captured in the PPI networks associated with AIS involving laminin (LAMB2 and LAMC1) and prolargin (PRELP) (Fig. 5A). Perlecan binds growth factors via heparan sulfate chains and interacts with vascular endothelial growth factor receptors 2 (VGFR2) which plays a major role in tumour angiogenesis. Dynamic remodelling of tissue architecture takes place during tumour growth and angiogenesis, where perlecan is one of the ECM constituents of the tumour microenvironment. Whereas the intact protein perlecan is known to possess pro-angiogenic properties, its C-terminal fragment, which is released by proteolysis during cancer remodelling and known as endorepellin, has the opposite functions of anti-angiogenic activity.

Perlecan interacts with elastin (ELN), vitronectin (VTN), LTBP2, TGFBI, elastin microfibril interfacer 1 (EMILIN1), collagen alpha-2(VI) chain (COL6A2), peristin (POSTN), and lysosome C (LZY) (as seen in Fig. 5A) which are involved in ECM organization and cell adhesion. The subnetwork of STOM, LAMP1, ARL8B, and RAB27A, and LYZ (Fig. 5A) suggests an involvement of the neutrophil degranulation pathway. Recently, the role of the neutrophil degranulation has been discussed how the release of neutrophil granule proteins is associated with cancer development and tumour progression via neutrophil-mediated transport of cancer cells leading to different cellular phenotypes and into different tissues. The subnetwork consisting of ACADSB and ECHDC1 may relate to the carboxylic acid catabolic process.

No protein networks were obtained in-network depth of 0 interactions mainly because of the limited number of proteins significant to MIA by which no hub proteins were identified (Fig. 5B). The q values by multiple testing correction using the Benjamini–Hochberg method are presented.

Figure 4. Proteins significant to AIS and MIA identified by pairwise G-statistics analysis. The q values by multiple testing correction using the Benjamini–Hochberg method are presented.

B. Minimally invasive adenocarcinoma (MIA)
was obtained as interferon-gamma signalling in Reactome pathways (Fig. 6), which included MHC class II HLA-DR beta 1 chain (HLA-DRB1) and protein tyrosine phosphatase non-receptor type 1 (PTPN1). PTPN1 (also known as PTP1B) belongs to protein tyrosine phosphatases (PTPs) family. PTP1B has been known to have the two faces in tumorigenesis that PTP1B promotes tumour progression in some cancers but functions as a tumour suppressor in other cancers26, whereas the role of PTP1B in NSCLC has been unknown. It has been reported that the high expression of PTP1B in NSCLC tissues was associated with the stage and overall survival of NSCLC patients26. PTP1B inhibitors are currently considered a promising anti-cancer therapy due to its involvement in the progression of numerous types of cancers via oncogenic transformation27. Hematopoietic pre-B-cell leuke-emia transcription factor (PBX)-interacting protein (PBXIP1), also known as HPIP, is a corepressor of PBX. Shi et al.28 demonstrated that HPIP silencing suppressed TGF-β1-induced epithelial-mesenchymal transition (EMT) in A549 lung cancer cells in vitro. DCXR encodes diacetyl/l-xylulose reductase, a multifunctional enzyme in glucose metabolism. Reactive carbonyls are known to cause severe respiratory diseases, which are detoxified by carbonyl reductases in the lung, in particular, DCXR that mediates chemical redox cycling29.

The eight WGCNA modules significant to LPA (Fig. 5C–E and Fig. S1A–E in Supplementary Information File 1) were subjected to further bioinformatics analysis. The top WGCNA modules were WM31, WM30, and WM29 in the order of correlation values: 0.995, 0.904, and 0.889, respectively. The enriched pathways of the WM31 (Fig. 5C) included (1) biological process (GO): spliceosomal snRNP assembly and termination of RNA polymerase II transcription; (2) Reactome pathways: snRNP assembly, mRNA splicing—minor pathway, histone stem-loop-binding protein (SLBP) independent processing of histone pre-mRNAs, cleavage of a growing transcript in the termination region, and citric acid cycle (TCA cycle). The eigengene 

\[ \text{PPP1CB} \]

encodes serine/threonine-protein phosphatase 1 (PP1)-beta catalytic subunit (PP-1B) which is one of the catalytic subunits of PP1 involved in regulating cell division. Low levels of PP1s and their key regulatory subunit PPP1R9B (known as Spinophilin) are related to a poor prognosis in numerous cancers including lung cancers, which was more involved in squamous cell lung carcinoma than lung adenocarcinoma26. It might be noteworthy that a malignant glioma of infancy was recently found to harbour a novel PPP1CB-ALK fusion protein31, to be compared with the EML4-ALK translocation mutation present in ca. 6% of NSCLC. The main PPI network (1) including PP-1B, damage specific DNA binding protein 1 (DDB1) and nuclear mitotic apparatus protein 1 (NUMA1) (Fig. 5C) is
involved in mesenchyme migration, cell cycle, mitotic/recruitment of NuMA to mitotic centrosomes and regulation of Polo-like kinase 1 (PLK1) activity at G2/M Transition. The subnetwork (2) including the hub protein small nuclear ribonucleoprotein Sm D2 (SM-D2) encoded by SNRPD2 belongs to the spliceosomal snRNP assembly, and the subnetwork (3) including KARS, IARS2, TIMM50, and ACO2 is involved in the tRNA aminoacylation for protein translation and carboxylic acid metabolic process.

The enriched pathways of the WM30 module included (1) biological process (GO): chaperone-mediated protein folding, response to endoplasmic reticulum stress, and protein de-ubiquitination; (2) Reactome pathways: AU-rich element RNA-binding protein 1 (AUF1) (also known as hnRNP D0) binding and destabilisation of mRNA, activation of nuclear factor kappa B (NF-κB) in B-cells, and oxygen-dependent proline hydroxylation of hypoxia-inducible factor-alpha (HIF-alpha) (Figs. 5D and 6). The eigengene $PSMC4$ encodes 26S proteasome regulatory subunit 6B, a component of the 26S proteasome, playing a key role in the maintenance of protein homeostasis by removing misfolded or damaged proteins. The translation elongation factor eukaryotic elongation factor 1-alpha1 (EEF1A1) was reported to participate in the entire heat shock response process from transcription through translation. Depending on the type of stimulus, EEF1A1 can be phosphorylated and/or methylated and can interact with trans-acting factors including AUF1, which possibly determines the stability of specific mRNAs and its interaction with 70-kd heat shock protein (HSP70s) mRNA. Stress-induced phosphoprotein 1 (STIP1), also known as HOP/P60/STI1, is a chaperone protein that comprises three tetratricopeptide repeat domains, which can simultaneously bind HSP70s and HSP90s. STIP1 is a tumour-associated antigen (TAA) and its overexpression has been identified in numerous cancers, including colorectal carcinoma, pancreatic cancer, cholangiocellular carcinoma and ovarian cancer, and it is possibly associated with poor survival outcomes in patients with cancer. STI3 encodes STI3 HSP70s-interacting protein (HIP)/putative tumour suppressor STI3 suppressing of tumorigenicity 13 protein, which is also involved in the regulation of heat shock factor 1 (HSP1)-mediated heat shock response and which mediates the association of HSP70s and HSP90s. In an in vitro study on pancreatic ductal adenocarcinoma, Ma et al. reported that signal recognition particle receptor subunit alpha (SRP54) was downregulated in pancreatic ductal adenocarcinoma cells compared to normal pancreatic ductal cells. This downregulation was associated with increased cell proliferation and decreased cell apoptosis, suggesting a potential role for SRP54 in the development and progression of pancreatic ductal adenocarcinoma.
(SRPRB) plays a central role in the interaction between proteins and stress-associated endoplasmic reticulum (ER) protein 1 (SERP1), which is responsible for the accumulation of unfolded protein in ER stress. They also noted that the downregulation of SERP1 significantly increased SRPRB expression, leading to cell apoptosis through NF-κB activation. Hypoxia upregulated protein 1 (HYOU1) has a significant degree of interaction with SRPRB that well correlated with both the upregulated expressions of SRPRB and HYOU1 observed in our protein expression data. Yamaguchi et al.40 found that SERP1 expression was enhanced in vitro by hypoxia, which was associated with the accumulation of unfolded protein in ER stress. The subnetwork (4) is involved in the pathway of carbohydrate metabolic process.

A genomic alteration analysis for MIA and LPA. Protein expression profiles were found to be highly differentiated between MIA and LPA (Figs. 1A and 2B). We then conducted NGS to obtain genomic alteration profiles in FFPE tissue specimens, which were obtained from the same FFPE blocks used for the proteomic analyses. NGS was performed to identify genomic alterations such as gene mutations including single-nucleotide variations, insertion/deletion variations, and gene CNVs, using the ACTONCO panel (ACT Genomics, Taipei, Taiwan), which comprises of 440 cancer-related genes. Oncoprints were prepared to visualise somatic mutations and CNVs by heatmap together with the TCGA lung adenocarcinoma datasets for comparison (Fig. 7A, B). TCGA lung ADC datasets (M stage: M0, Mx, NA; N stage: N0, NA; sample number: n = 471) were obtained from the cBioPortal Pan-Lung cancer (TCGA, Nat Genet 2016) (https://www.cbioportal.org/). LPA was suggested to accompany both somatic mutations with loss of function and copy number gains more frequent than the MIA group. Given the small sample size, it is hard to perform statistical analysis or to find any enriched biomarkers in MIA or LPA. However, it was shown that all both MIAs and LPAs harboured at least one oncogenic driver with EGFR oncogenic mutations being most prevalent. Other oncogenic events detected include ERBB2 p. Gly776delinValVal in the sample MIA_T75, and EGFR copy number gain in the samples of LPA_T87 and
MIA_T80. Oncogenic drivers identified in our study were quite similar to those from the large-scale genomic studies. Aside from the tumour which exhibited the APOBEC-like mutation signature, in our cohort, a high percentage of MIA and LPA were EGFR mutant, which may explain why the number of somatic mutations is low and no difference in tumour mutational burden was found between two groups.

Discussion

The tumourigenesis of early lung adenocarcinoma is thought to progress in a stepwise manner in the order of AIS, MIA, and LPA; however, this has not been completely elucidated to date. Therefore, understanding of disease-related molecular mechanisms and profiles in early lung adenocarcinomas would be markedly useful for selecting treatment strategies and could improve the treatment outcomes of individual patients. Protein-based networks by both WGCNA and pairwise G-statistics analyses identified several protein modules and networks that were potentially associated with disease mechanisms driven by distinct early lung adenocarcinomas.

Most activities in ECM organization and integrin-mediated cell–cell adhesion characterised by AIS were consistent with its growth pattern restricted to neoplastic cells with pre-existing alveolar structures without stroma, vascular, or pleural invasion. MIA might be associated with oncogenic transformation and detoxification in lung cell-extracellular matrix interactions although no direct networks were identified. This seems to coincide with the observation of tumour histology defined as MIA presenting a limited size of invasion with a myofibroblastic stroma-associated invasive tumour. Pathways of cell surface–receptor signalling, and HSF1-mediated heat shock response regulation, cell cycle regulation were characteristic to LPA.

Assuming that the tumourigenesis proceeds along with the AIS–MIA–LPA axis, the key protein networks identified in this study appear to reflect the nature of the cancer cells of these three subtypes. Generally, the following two possible scenarios can be considered: either LPA cells emerge via transformation of MIA cells or from a different cell origin. The former scenario requires LPA cells to have the same gene mutation profiles as MIA cells and transformed LPA cells to possibly not grow so rapidly but instead co-exist with MIA cells. Conversely, the latter scenario instead requires both MIA and LPA cells to have distinctly different gene mutation profiles from one another. Cancer stem cells (CSCs) are known to emerge via the dysfunctional organogenesis of organ stem cells resulting from the loss of gene regulatory control. In this process, new aberrant cells emerge and acquire a self-renewing capability via the neoplastic transformation of CSCs. Such aberrant cells would emerge as a sub-population of tumour cells owing to genetic intra-tumour heterogeneity, and their rapid growth would disrupt tumour environment and result in predominant cell survival. The orthogonal partial least square-discriminant analysis (OPLS-DA) performed for mutant proteins identified in the previous study revealed the profound differences in distance between the LPA group and the AIS plus MIA group.

The mutational landscape shown in the larger study conducted by Qian et al. suggests that a majority of patients in their study have an ethnic background different from ours. Our cohort has a higher rate of oncogenic EGFR mutations (57% vs 29%) and a lower rate of KRAS mutations (0% vs 25%). Moreover, in the study performed by Chen et al., oncogenic drivers were not detected in up to half of their samples. Although our study sample size is limited and ethnic background is different, we found similarities between two studies of which oncogenic drivers were detected in MIA and LPA (Fig. 7A, B). Besides, we also found that LPA had a higher degree of copy number loss (one or two copy deletion) than MIA (Fig. 7C), which feature was also seen in the genomic study.

In conclusion, our results could identify disease-related protein networks that are possibly associated with distinct early lung adenocarcinomas—AIS, MIA, and LPA. The large genomic studies suggested mutation signature profiling did not vary significantly throughout pre-invasive lung ADCs—AIS and MIA and invasive lung ADCs. However, together with our protein network-based profiles and the previous OPLS discriminant analysis of mutant proteomes, it remains disputable that LPA cells emerge via a direct transformation from AIS or MIA, whereas no evidence was attained by the genomic alteration analysis performed in this study. It should be also noted that their studies grouped all 11 subtypes including abundant papillary and acinar ADCs as invasive lung ADCs. A large-scale genomic alteration study using tissue specimens histologically well-defined as LPA would be needed for further investigation. Network-based investigations regarding tumourigenesis will further provide clinically important information about proteogenomic landscapes in lung adenocarcinoma.

Methods

FFPE tissue specimens and sample preparation. Among 974 patients who underwent surgical lung cancer resection at St. Marianna University Hospital between 2000 and 2014, only 674 (69.3%) had tumours that were histologically confirmed adenocarcinomas. The pathological specimens were independently reviewed by two pathologists (H. N. and M. T.) to confirm that they satisfied the 2015 World Health Organization classification criteria of lung tumours (histological criteria). FFPE tumour tissue blocks from 14 surgical specimens histologically well-judged as AIS, MIA, and LPA were obtained without patient identifiers from the St. Marianna University School of Medicine Hospital. Informed consent was obtained from all participating subjects, and the protocol was approved by the institutional review board of St. Marianna University School of Medicine (approval no. 1461) and was conducted in accordance with the Helsinki Declaration. For tissue microdissection, 10-μm-thick sections from the FFPE tumour blocks were cut and placed on DIRECTOR slides (OncoPlex Diagnostics Inc., Rockville, MD, USA). The sections were deparaffinised and stained with haematoxylin using standard histological methods prior to dissection. Microdissection was performed using a Leica LMD7 microdissection microscope (Leica, Wetzlar, Germany). A total area of 4 mm² with approximately 15,000 tumour cells was directly transferred from the FFPE sections via laser dissection into the cap of a 200-μL low-binding tube. Proteins were extracted and digested with trypsin using LIQUID TISSUE MS Protein Prep kits (Onco-
A total of eight tumour samples (four MIA and four LPA) selected from the specimens used for proteomic analyses, were subjected to NGS. The RECOVER ALL Total Nucleic Acid Isolation kit (Qiagen, Hilden, Germany) was used to isolate genomic DNA from FFPE tumour samples. The DNA concentration and integrity were analysed using the QUBIT-IT dsDNA HS assay (Invitrogen, Carlsbad, CA, USA) and FRAGMENT ANALYSER (Advanced Analytical Technologies, Ankey, IA, USA), respectively.

**Liquid chromatography-tandem MS (LC–MS/MS)-based proteomic analysis.** A label-free quantitation approach using spectral counting by LC–MS/MS was adopted for the global proteomic analysis. The digested samples (5 μL for a single run) were analysed in triplicate by LC–MS/MS using a reverse-phase LC system interfaced with a Q EXACTIVE ORBITRAP mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA) via a nano-electrospray ionisation device (AMR Inc., Tokyo, Japan). LC–MS/MS analyses were performed using the Q-Exactive-Orbitrap mass spectrometer (Thermo Fisher Scientific) with the nEASY-LC-2.0 nanoelectrospray flow system (Advion Biotechnologies, Ithaca, NY, USA). Liquid chromatography-tandem mass spectrometry (LC-MS/MS) analyses were conducted using the Q-Exactive-Orbitrap mass spectrometer (Thermo Fisher Scientific) with the nEASY-LC-2.0 nanoelectrospray flow system (Advion Biotechnologies, Ithaca, NY, USA). The raw data were processed using the MaxQuant software (version 1.5.2.2) with the Andromeda search engine (version 1.5.1.1) (Thermo Fisher Scientific). The search parameters included the following: trypsin/digested, fixed modifications of oxidation of cysteine and methionine oxidation, variable modifications of deamidation of asparagine and glutamine, and peptide mass tolerance of 15 ppm. Protein identification was based on at least two unique peptides. The protein quantification was performed using the spectral counting method. The fold change of protein expression between the two groups was calculated and proteins with a fold change of ≥ 2 were considered to be significantly differentially expressed.

**WGCNA.** The similarity among protein expression patterns for all protein pairs was calculated according to their pairwise Pearson's correlation coefficient, i.e. the similarity between proteins i and j was defined as \((1 - r_{ij})/2\), where \(r_{ij}\) is the correlation of the protein expression patterns between the two proteins i and j. An adjacency matrix was then computed by increasing the similarity matrix up to the power of 10 to generate a co-expression network with scale-free properties. Subsequently, from the resultant scale-free co-expression network, we generated a topological overlap matrix (TOM) that considers topological similarities between a pair of proteins in the network. Using the dissimilarity according to TOM (1-TOM), we conducted hierarchical clustering to generate a tree that clustered proteins in its branches. Dynamic tree cutting was used to trim the branches to identify protein modules. A protein module was summarised by the top hub protein (referred to as eigen-protein) with the highest connectivity in the module. To identify the protein modules associated with clinical traits, we calculated the correlation coefficients between the eigen-proteins and clinical traits. WGCNA was conducted using a GARUDA PLATFORM GADGET (The Systems Biology Institute, Tokyo, Japan) that implemented the WGCNA pipeline based on the WGCNA R-package.

**Protein–protein interaction network construction.** To construct a protein interaction network for a protein module, we used the STRING database (version 11.0) (32), which accumulates information on protein–protein interactions from various other databases such as IntAct, Reactome, DIP, BioGRID, MINP, KEGG, NCI/Nature PID, The Interactive Fly, and BioCyc. STRING networks were constructed under the criteria for linkage only with experiments, databases, text mining, and co-expression using the default settings, i.e. a medium confidence score of 0.400, a network depth of 0, or 50 interactions. Subsequently, protein networks imported from the STRING database were visualised using CYTOSCAPE version 3.7.2. Functional enrichment results were obtained for canonical pathways considering \(p < 0.05\) to be statistically significant.

**Data availability**

The unfiltered MS datasets generated and analysed in this study have been deposited in the PRIDE archive (http://www.ebi.ac.uk/pride/archive/) via the PRIDE partner repository and jPOST with the dataset identifiers PXD017334 and JPST000741, respectively.

Received: 22 February 2020; Accepted: 24 July 2020

**References**

1. American Society of Clinical Oncology (ASCO). Cancer Net. 2019. https://www.cancer.net/cancer-types/lung-cancer-non-small-cell/statistics (2019).
2. Hanna, J. M. & Onaitis, M. W. Cell of origin of lung cancer. J. Carcinog. 12, 6. https://doi.org/10.4103/1477-3163.109033 (2013).
3. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550 (2014).
4. Nakamura, H. & Saij, H. Worldwide trend of increasing primary adenocarcinoma of the lung. Surg. Today. 44, 1004–1012 (2014).
12. Kato, Y.
11. Nishimura, T.
10. Gu, J.
8. Yoshizawa, A.
5. Wu, K., Huang, R. S., House, L. & Cho, W. C. Next-generation sequencing for lung cancer.
6. Travis, W. D.
16. Devarakonda, S., Masood, A. & Govindan, R. Next-generation sequencing of lung cancers: lessons learned and future directions.
19. Tang, Y., Ke, Z. P., Peng, Y. G. & Cai, P. T. Coexpression analysis reveals key gene modules and pathways of human coronary heart disease.
21. Sokal, R. R. & Rohlf, F. J.
31. Aghajan, Y., Levy, M. L., Malicki, D. M. & Crawford, J. R. Novel fusion protein in a high-grade glioma of infancy.
30. Verdugo-Sivianes, E. M.
23. Chin, C. H.
22. Szklarczyk, D.
24. Douglass, S., Goyal, A. & Iozzo, R. V. The role of perlecan and endorepellin in the control of tumor angiogenesis and endothelial cell autophagy.
25. Mollinedo, F. Neutrophil degranulation, plasticity, and cancer metastasis.
33. Xu, Y. W.
40. Yamaguchi, A.
41. Chen, H.
34. Kubota, H.
35. Walsh, N.
39. Ma, Q.
38. Yamaguchi, A., et al. Stress-associated endoplasmic reticulum protein 1 (SERP1)/Ribosome-associated membrane protein 4 (RAMP4) stabilizes membrane proteins during stress and facilitates subsequent glycosylation. J. Cell Biol. 147, 1195–1204 (1999).
37. Chen, H. et al. Genomic and immune profiling of pre-invasive lung adenocarcinoma. Nat. Commun. 10, 5472. https://doi.org/10.1038/s41467-019-13460-3 (2019).
36. Qian, J. et al. Genomic underpinnings of tumor behavior in in situ and early lung adenocarcinoma. Am. J. Respir. Crit. Care Med. 201, 697–706 (2020).
35. Willis, E. R. Targeted cancer therapy: vital oncogenes and a new molecular genetic paradigm for cancer initiation progression and treatment. Int. J. Mol. Sci. 17, E1552. https://doi.org/10.3390/ijms17091552 (2016).
34. Bylesjo, M. et al. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J. Chemom. 20, 341–351 (2006).
33. WHO classification of tumours-pathology and genetics of tumours of the lung, pleura, thymus and heart. 4th Edition, Vol. 7. (eds. Travis, W. D., Bambilla, E., Burke, A. P., Marx, A. & Nicholson, A. G.) (IARC Press, 2015).
32. Travis, W. D. et al. The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification. J. Thorac. Oncol. 10, 1243–1260 (2015).
47. Prieto, D. A. et al. Liquid tissue: proteomic profiling of formalin-fixed tissues. Biotechniques 38, S32–S35 (2005).

48. ZybaIov, B., Coleman, M. K., Florens, L. & Washburn, M. P. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labelling. Anal. Chem. 77, 6218–6224 (2005).

**Acknowledgements**

Financial support by Chugai Pharmaceutical Co., Ltd on this study is gratefully acknowledged. The authors would like to thank Dr. Hiroki Sakai for his assistance regarding the case survey of lung cancer resections in the Department of Chest Surgery, St. Marianna University Hospital.

**Author’s contributions**

T.N., H.N., H.S. Y.K., and Ha.K. conceptualised this study. T.N., H.N., Ha.K., and S-J. C. initiated and managed the collaboration. S.N. and Hi.K. selected the FFPE tissue blocks of lung adenocarcinoma and prepared the slides. H.N. and M.T. reviewed the pathological specimens. K.F. performed laser microdissection, sample preparation, and MS-based proteomic analysis. T.N. designed and performed bioinformatics analysis for protein expression data. W.C., K.T.T., and S-J. C. performed NGS and analysed the genomic alteration data. T.N. wrote the main manuscript text and prepared Table 1, Figs. 1, 2, 3, 4, 5, 6 and 7, and Supplementary Information File 1. T.N., H.N., N.F. Y.K. and H.S. wrote the first draft of the manuscript. All authors have reviewed the manuscript and agree with the submission.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-70578-x.

Correspondence and requests for materials should be addressed to T.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020