



OPEN

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

Toshihide Nishimura^{1,2,9}✉, Haruhiko Nakamura^{1,2,9}, Kien Thiam Tan^{3,9}, De-Wei Zhuo^{3,9}, Kiyonaga Fujii^{1,2,9}, Hirotaka Koizumi⁴, Saeko Naruki^{4,9}, Masayuki Takagi⁴, Naoki Furuya⁵, Yasufumi Kato⁶, Shu-Jen Chen³, Harubumi Kato^{7,8} & Hisashi Saji²

The tumorigenesis 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*³. 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 stages^{4,5}. 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-nishimura@marianna-u.ac.jp

Sample no	Histological type	Age (years)	Sex	Location	Tumour size on CT (mm)	Clinical TNM classification			Clinical stage	EGFR mutation
						c-T	c-N	c-M		
A. AIS (n=5)										
AIS_T53	AIS	53	F	RS8	11	cT1a	cN0	cM0	cIA	EGFR(+)
AIS_T54	AIS	74	M	RS6	29	cT1b	cN0	cM0	cIA	EGFR(-)
AIS_T56	AIS	69	F	RS6	16	cT1a	cN0	cM0	cIA	EGFR(-)
AIS_T58	AIS	78	M	LS1+2	28	cT1b	cN0	cM0	cIA	EGFR(-)
AIS_T59	AIS	68	F	RS1	13	cT1a	cN0	cM0	cIA	EGFR(+)
	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	EGFR(+)
MIA_T74	MIA	67	F	LS3	19	cT1a	cN0	cM0	cIA	EGFR(+)
MIA_T75	MIA	77	F	RS8	20	cT1a	cN0	cM0	cIA	EGFR(-)
MIA_T79	MIA	61	M	LS6	10	cT1a	cN0	cM0	cIA	EGFR(+)
MIA_T80	MIA	63	F	RS3	12	cT1a	cN0	cM0	cIA	EGFR(-)
	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	EGFR(-)
LPA_T87	LPA	73	F	RS3	28	cT1b	cN0	cM0	cIA	EGFR(+)
LPA_T88	LPA	59	M	LS1+2	30	cT1b	cN0	cM0	cIA	EGFR(-)
LPA_T89	LPA	67	F	RS6	20	cT1a	cN0	cM0	cIA	EGFR(-)
	Average ± SD	66.8 ± 5.0			27.0 ± 4.1					
<i>p</i> value_ANOVA		0.816			0.077					

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.

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 survival⁷. However, some previous studies have reported that LPA might recur after complete resection^{8–10}. 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 tumorigenesis 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^{11–14}. 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^{15,16}.

Weighted gene co-expression network analysis (WGCNA)¹⁷ is an extensively applied, unsupervised gene-clustering method designed based on the correlation network of gene expressions^{14,18,19}. 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 datasets¹⁴.

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). Pre-surgical treatment was not performed for any of these lung adenocarcinomas.

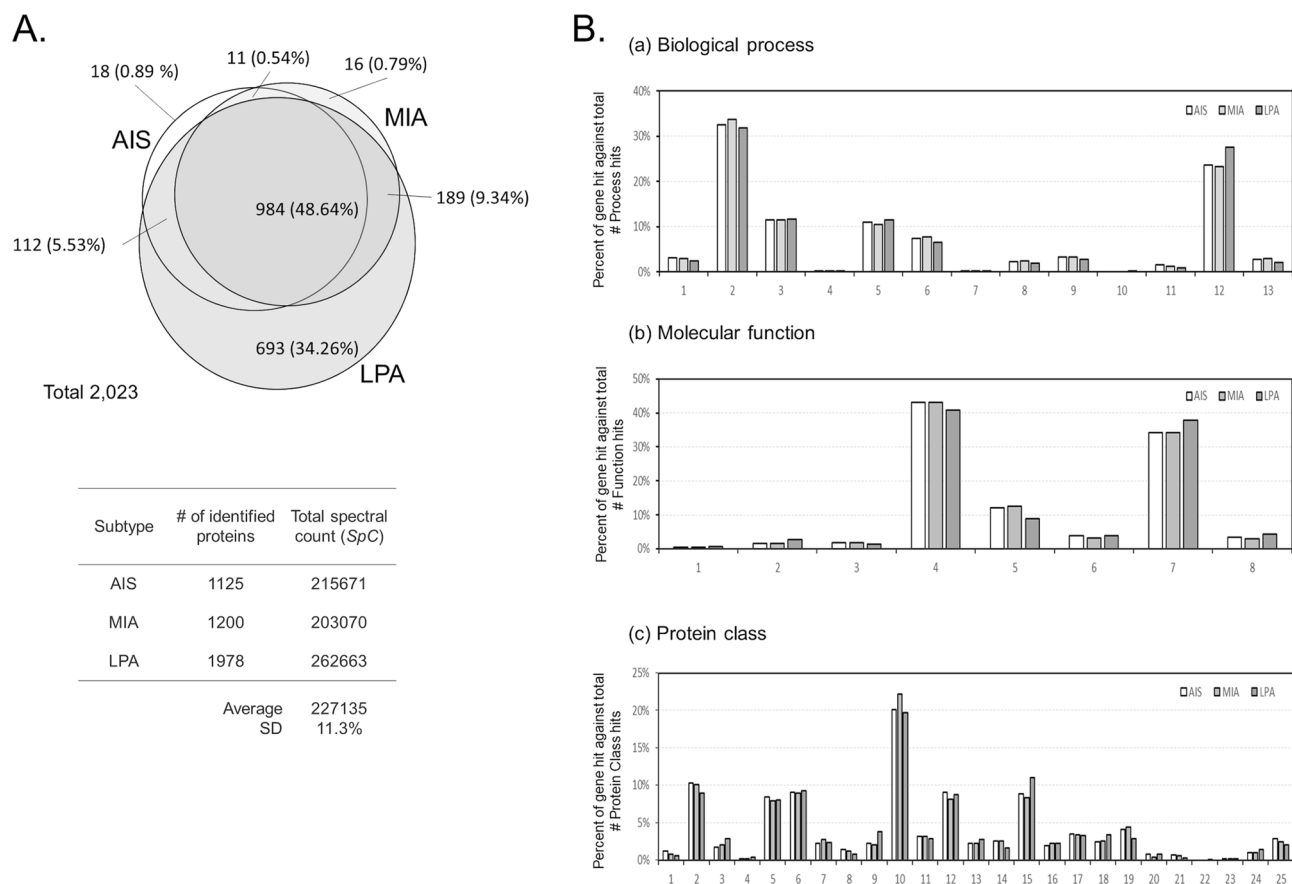


Figure 1. Venn map and hierarchical clustering of the identified proteins. **(A)** Venn map of the identified proteins. **(B)** Gene ontology (GO) analysis of the identified proteins to AIS, MIA and LPA. (a) Biological process. 1, cellular component organization or biogenesis (GO:0071840); 2, cellular process (GO:0009987); 3, localization (GO:0051179); 4, reproduction (GO:0000003); 5, biological regulation (GO:0065007); 6, response to stimulus (GO:0050896); 7, pigmentation (GO:0043473); 8, developmental process (GO:0032502); 9, multicellular organismal process (GO:0032501); 10, rhythmic process (GO:0048511); 11, biological adhesion (GO:0022610); 12, metabolic process (GO:0008152); 13, immune system process (GO:0002376). (b) Molecular function. 1, translation regulator activity (GO:0045182); 2, transcription regulator activity (GO:0140110); 3, molecular transducer activity (GO:0060089); 4, binding (GO:0005488); 5, structural molecule activity (GO:0005198); 6, molecular function regulator (GO:0098772); 7, catalytic activity (GO:0003824); 8, transporter activity (GO:0005215). (c) Protein class. 1, extracellular matrix protein (PC00102); 2, cytoskeletal protein (PC00085); 3, transporter (PC00227); 4, transmembrane receptor regulatory/adaptor protein (PC00226); 5, transferase (PC00220); 6, oxidoreductase (PC00176); 7, lyase (PC00144); 8, cell adhesion molecule (PC00069); 9, ligase (PC00142); 10, nucleic acid binding (PC00171); 11, signaling molecule (PC00207); 12, enzyme modulator (PC00095); 13, calcium-binding protein (PC00060); 14, defense/immunity protein (PC00090); 15, hydrolase (PC00121); 16, transfer/carrier protein (PC00219); 17, membrane traffic protein (PC00150); 18, transcription factor (PC00218); 19, chaperone (PC00072); 20, cell junction protein (PC00070); 21, surfactant (PC00212); 22, structural protein (PC00211); 23, storage protein (PC00210); 24, isomerase (PC00135); 25, receptor (PC00197).

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*) ≥ 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)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).

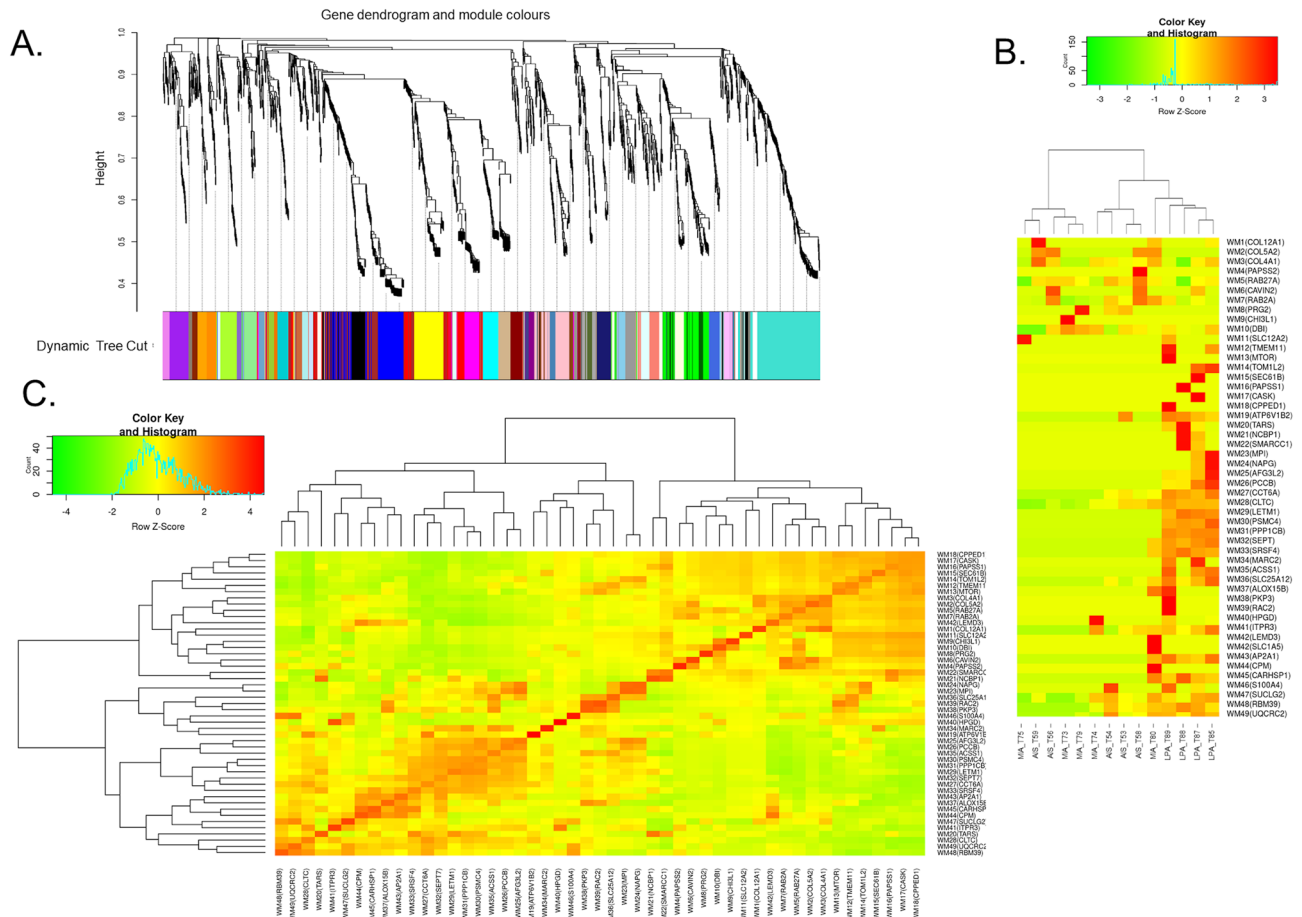


Figure 2. Gene modules identified by WGCNA. (A) Gene dendrogram obtained by clustering dissimilarity according to topological overlap with the corresponding module. The coloured rows correspond with the 49 modules identified by dissimilarity according to topological overlap. (B) A heatmap of the proteome abundance of eigen proteins in the 49 protein modules and samples. (C) Pairwise correlations between the modules in the heatmap of eigen-protein expressions.

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, WM 29, 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 tends 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 approach²¹ 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 2×2 tables¹³. 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

Module ID	Module Color	Module trait relationships									# of proteins in the module
		Adenocarcinoma <i>in-situ</i> (AIS)			Minimally invasive adenocarcinoma (MIA)			Lepidic predominant invasive adenocarcinoma (LPA)			
		Correlation	p-value	BH-corrected q-value	Correlation	p-value	BH-corrected q-value	Correlation	p-value	BH-corrected q-value	
WM1	darkturquoise	0.35	0.22	0.72	-0.129	0.66	0.899	-0.234	0.4204	0.572	33
WM2	darkorange2	0.628	0.016	0.395	-0.215	0.461	0.753	-0.439	0.1167	0.26	11
WM3	yellowgreen	0.548	0.042	0.345	0.091	0.756	0.904	-0.679	0.0076	0.027	18
WM4	orange	0.431	0.124	0.506	-0.111	0.705	0.864	-0.339	0.2353	0.412	28
WM5	sienna3	0.632	0.015	0.749	0.022	0.939	0.959	-0.694	0.0059	0.024	18
WM6	lightgreen	0.524	0.054	0.38	-0.242	0.405	0.709	-0.3	0.2978	0.471	39
WM7	lightsteelblue1	0.618	0.019	0.304	0.112	0.702	0.882	-0.774	0.0012	0.009	13
WM8	skyblue3	0.046	0.875	0.952	0.293	0.309	0.756	-0.361	0.2054	0.387	17
WM9	darkred	-0.168	0.566	0.894	0.406	0.149	0.732	-0.253	0.3833	0.537	35
WM10	paleturquoise	-0.009	0.976	0.996	0.426	0.129	0.902	-0.442	0.1131	0.277	21
WM11	floralwhite	-0.16	0.586	0.87	0.407	0.149	0.81	-0.262	0.3649	0.526	11
WM12	lightcyan1	0.063	0.829	0.945	0.248	0.392	0.768	-0.331	0.248	0.405	12
WM13	magenta	0.126	0.667	0.861	0.29	0.315	0.671	-0.441	0.1145	0.267	58
WM14	yellow	0.113	0.7	0.88	0.293	0.309	0.722	-0.431	0.1243	0.265	89
WM15	tan	0.133	0.649	0.884	0.299	0.299	0.772	-0.458	0.0992	0.256	50
WM16	brown	0.22	0.449	0.917	0.399	0.158	0.704	-0.656	0.0108	0.035	108
WM17	cyan	0.278	0.336	0.866	0.449	0.108	1.054	-0.771	0.0013	0.009	46
WM18	red	0.256	0.377	0.881	0.432	0.123	1.002	-0.73	0.003	0.017	84
WM19	darkorange	0.194	0.506	0.919	-0.343	0.23	0.752	0.158	0.5901	0.705	28
WM20	turquoise	-0.227	0.436	0.971	-0.214	0.463	0.731	0.467	0.092	0.251	209
WM21	black	-0.06	0.839	0.934	0.049	0.868	0.966	0.012	0.9687	0.969	80
WM22	blue	0.088	0.764	0.913	0.247	0.394	0.743	-0.356	0.2116	0.384	124
WM23	darkgreen	-0.128	0.663	0.879	-0.041	0.888	0.967	0.179	0.5396	0.661	34
WM24	lightyellow	-0.066	0.824	0.961	0.04	0.892	0.95	0.027	0.9264	0.946	35
WM25	green	-0.27	0.351	0.861	-0.243	0.403	0.732	0.543	0.0447	0.137	89
WM26	royalblue	-0.337	0.238	0.686	-0.326	0.256	0.738	0.703	0.005	0.022	35
WM27	plum1	-0.135	0.646	0.904	-0.626	0.017	0.405	0.808	4.78E-04	0.005	16
WM28	plum2	-0.044	0.881	0.939	-0.457	0.101	1.235	0.531	0.0507	0.146	10
WM29	salmon	-0.456	0.101	0.497	-0.382	0.177	0.669	0.889	2.13E-05	3.48E-04	47
WM30	lightcyan	-0.44	0.115	0.514	-0.412	0.143	0.877	0.904	9.40E-06	2.30E-04	40
WM31	skyblue	-0.479	0.083	0.452	-0.459	0.099	1.609	0.995	1.48E-13	7.26E-12	25
WM32	darkmagenta	-0.569	0.034	0.414	-0.254	0.381	0.777	0.873	4.63E-05	0.001	18
WM33	white	-0.555	0.04	0.388	-0.113	0.701	0.904	0.708	0.0046	0.023	27
WM34	midnightblue	-0.222	0.446	0.95	-0.17	0.561	0.785	0.416	0.1391	0.284	41
WM35	grey60	-0.362	0.203	0.711	-0.333	0.244	0.747	0.738	0.0026	0.016	39
WM36	ivory	-0.181	0.536	0.875	-0.08	0.785	0.894	0.277	0.3371	0.516	11
WM37	brown4	-0.294	0.307	0.836	0.121	0.682	0.903	0.184	0.5282	0.664	11
WM38	darkgrey	-0.187	0.523	0.884	-0.171	0.558	0.804	0.38	0.1805	0.354	31
WM39	darkolivegreen	-0.095	0.746	0.914	-0.022	0.941	0.941	0.124	0.6724	0.766	19
WM40	greenyellow	-0.207	0.478	0.937	0.382	0.178	0.623	-0.186	0.5254	0.677	51
WM41	mediumpurple3	-0.164	0.576	0.881	0.088	0.766	0.894	0.081	0.7831	0.816	14
WM42	steelblue	-0.006	0.985	0.985	0.658	0.01	0.512	-0.692	0.0061	0.023	21
WM43	darkslateblue	-0.52	0.057	0.348	0.2	0.492	0.731	0.339	0.2364	0.399	10
WM44	pink	-0.198	0.497	0.937	0.392	0.165	0.674	-0.206	0.4797	0.635	74
WM45	saddlebrown	-0.413	0.142	0.535	0.309	0.282	0.769	0.111	0.7066	0.769	24
WM46	purple	0.339	0.236	0.723	-0.212	0.468	0.717	-0.135	0.6457	0.753	56
WM47	violet	0.142	0.629	0.906	-0.035	0.905	0.944	-0.113	0.7	0.78	21
WM48	bisque4	0.19	0.514	0.9	-0.292	0.311	0.693	0.108	0.714	0.761	11
WM49	orangered4	-0.032	0.914	0.952	-0.219	0.453	0.765	0.286	0.3582	0.532	16

Figure 3. Relationship between module eigen-proteins and the clinical traits of subtypes AIS, MIA, and LPA. Each row in the embedded table represents weighted gene co-expression network analysis results for each module. The first and second columns in the table represent the module identification and colour name of the module, respectively. The twelfth column represents the number of proteins in each module. The p values of the correlation coefficients and q values by multiple testing correction using the Benjamini–Hochberg method are presented. The table is colour-coded by the correlation coefficient according to the colour legend on the right side of the figure. The intensity and direction of the correlations are indicated on the right side of the heatmap (red, positive correlation; blue, negative correlation). Columns with significant q values are highlighted in bright red background.

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, WM 29, WM30, WM31, WM32, WM33, and WM35). Top hub proteins were identified according to maximal clique centrality (MCC) by using in the CYTOHUBBA PLUGIN²³. Figure 5 shows the protein networks of AIS, MIA, and the top three WGCNA

A. Adenocarcinoma *in-situ* (AIS)

UniProtKB Acc.No.	Gene name	Description	Relative abundance (%)			Pairwise G-statistics						Relative expression (%)
			AIS	MIA	LPA	AIS vs. MIA			AIS vs. LPA			
						G-value	p-value	BH-corrected q-value ^{a)}	G-value	p-value	BH-corrected q-value ^{a)}	
P61626	LYZ	Lysozyme C	71.6	0.0	28.4	70.31	2.42E-18	9.07E-18	23.18	1.48E-06	1.92E-06	71.6 0.0
O15382	BCAT2	Branched-chain-amino-acid aminotransferase, mitochondrial	65.5	0.0	34.5	128.08	1.08E-29	1.08E-28	24.97	5.83E-07	7.99E-07	
P27105	STOM	Erythrocyte band 7 integral membrane protein	64.4	23.6	12.0	39.00	4.25E-10	7.96E-10	113.57	1.62E-26	4.85E-26	
Q6NZ12	CAVIN1	Caveolae-associated protein 1	64.3	19.2	16.5	51.54	7.00E-13	1.91E-12	90.21	2.14E-21	4.93E-21	
P55268	LAMB2	Laminin subunit beta-2	63.5	21.1	15.4	79.72	4.31E-19	1.85E-18	168.25	1.78E-38	1.07E-37	
P58107	EPK1	Epiplakin	62.6	19.0	18.4	234.89	5.12E-53	1.54E-51	374.80	1.69E-83	2.53E-82	
P04004	VTN	Vitronectin	61.4	21.4	17.2	60.20	8.56E-15	2.57E-14	120.65	4.55E-28	1.71E-27	
P11047	LAMC1	Laminin subunit gamma-1	58.8	25.0	16.2	17.78	2.48E-05	2.86E-05	51.26	8.11E-13	1.35E-12	
Q9NTX5	ECHDC1	Ethylmalonyl-CoA decarboxylase	58.3	0.0	41.7	73.66	9.28E-18	3.09E-17	6.73	9.47E-03	9.80E-03	
P10909	CLU	Clusterin	57.4	26.1	16.5	25.42	6.61E-07	6.58E-07	81.57	1.69E-19	3.38E-19	
P11279	LAMP1	Lysosome-associated membrane glycoprotein 1	57.3	12.2	30.5	23.70	1.13E-06	1.47E-06	11.79	5.97E-04	6.88E-04	
Q8NBQ5	HSD17B11	Estradiol 17-beta-dehydrogenase 11	57.2	13.0	29.7	37.63	8.54E-10	1.51E-09	20.82	5.05E-06	6.31E-06	
P00325	ADH1B	Alcohol dehydrogenase 1B	57.2	15.7	27.1	89.52	3.03E-21	1.51E-20	70.12	5.58E-17	9.85E-17	
Q15083	POSTN	Periostin	56.9	15.9	27.2	109.06	1.58E-25	1.18E-24	85.39	2.45E-20	5.25E-20	
Q9NZM1	MYOF	Myoferrin	56.6	29.1	15.2	26.31	2.91E-07	4.36E-07	128.17	2.82E-29	1.21E-28	
Q9NOC3	RTN4	Reticulon-4	55.3	12.8	31.9	24.42	7.77E-07	1.06E-06	11.04	8.90E-04	9.89E-04	
Q9Y6C2	EMILIN1	EMILIN-1	54.5	23.2	22.4	42.46	7.22E-11	1.44E-10	79.37	5.15E-19	9.65E-19	
P13645	KRT10	Keratin, type I cytoskeletal 10	53.9	18.1	27.9	104.68	1.45E-24	8.69E-24	90.96	1.47E-21	3.66E-21	
E7ER71	CD74	HLA class II histocompatibility antigen gamma chain	53.8	10.8	35.4	18.51	1.09E-05	2.03E-05	5.38	2.03E-02	2.03E-02	
P49554	ACADSB	Short/branched chain specific acyl-CoA dehydrogenase, mitochondrial	53.5	10.2	36.2	37.47	9.29E-10	1.55E-09	9.68	1.87E-03	2.00E-03	
Q9NVJ2	ARL8B	ADP-ribosylation factor-like protein 8B	53.5	31.6	14.9	5.07	2.43E-02	2.43E-02	35.55	2.49E-09	3.93E-09	
Q14767	LTBP2	Latent-transforming growth factor beta-binding protein 2	53.5	15.2	31.3	46.64	8.53E-12	1.83E-11	25.39	4.69E-07	6.70E-07	
Q18653	AOC3	Membrane primary amine oxidase	53.4	26.3	20.2	50.33	1.30E-12	3.25E-12	146.03	1.28E-33	6.39E-33	
P51888	PRELP	Prolargin	53.1	36.7	10.2	5.10	2.40E-02	2.48E-02	99.83	1.66E-23	4.53E-23	
P12110	COL6A2	Collagen alpha-2(VI) chain	52.8	32.8	14.4	31.42	2.08E-08	3.29E-08	273.60	1.86E-61	1.86E-60	
P15502	ELN	Elastin	52.2	29.9	17.9	47.13	6.64E-12	1.53E-11	242.99	8.75E-55	6.59E-54	
Q15582	TGFBI	Transforming growth factor-beta-induced protein ig-h3	51.4	32.9	15.7	13.14	2.90E-04	3.22E-04	120.24	5.61E-28	1.87E-27	
P62879	GNB2	Guanine nucleotide-binding protein G(I)/G(S)/G(T) subunit beta-2	50.7	18.1	31.2	19.47	1.02E-05	1.28E-05	12.88	3.32E-04	3.98E-04	
P51159	RAB27A	Ras-related protein Rab-27A	50.7	28.3	21.1	7.80	5.23E-03	5.60E-03	28.95	7.43E-08	1.11E-07	
P98160	HSPG2	Basement membrane-specific heparan sulfate proteoglycan core protein	50.0	27.0	23.0	189.41	4.29E-43	6.43E-42	545.65	1.12E-120	3.35E-119	

B. Minimally invasive adenocarcinoma (MIA)

UniProtKB Acc.No.	Gene name	Description	Relative abundance (%)			Pairwise G-statistics						Relative expression (%)
			AIS	MIA	LPA	AIS vs. MIA			MIA vs. LPA			
						G-value	p-value	BH-corrected q-value ^{a)}	G-value	p-value	BH-corrected q-value ^{a)}	
P00918	CA2	Carbonic anhydrase 2	9.6	68.3	22.1	55.16	1.11E-13	1.78E-13	39.41	3.44E-10	9.18E-10	68.3 3.3
Q96A06	PBXIP1	Pre-B-cell leukemia transcription factor-interacting protein 1	14.0	61.3	24.7	81.74	1.55E-19	6.22E-19	63.15	1.92E-15	7.67E-15	
P13760	HLA-DRB1	HLA class II histocompatibility antigen, DRB1-4 beta chain	21.9	60.2	17.9	95.75	1.30E-22	1.04E-21	165.59	6.80E-38	5.44E-37	
P04424	ASL	Argininosuccinate lyase	3.3	59.8	36.9	79.02	6.15E-19	1.64E-18	15.80	7.04E-05	1.13E-04	
Q724V1	DCXR	L-xylulose reductase	15.0	54.7	30.4	65.23	6.67E-16	1.33E-15	36.22	1.76E-09	3.52E-09	
P18031	PTPN1	Tyrosine-protein phosphatase non-receptor type 1	18.3	51.2	30.5	14.87	1.15E-04	1.15E-04	9.77	1.77E-03	2.03E-03	
Q8N1A0	KRT222	Keratin-like protein KRT222	13.9	50.9	35.2	26.59	2.52E-07	2.88E-07	8.99	2.72E-03	2.72E-03	
Q95833	CLIC3	Chloride intracellular channel protein 3	14.6	50.3	35.0	35.81	2.18E-09	2.90E-09	12.67	3.72E-04	4.95E-04	

^{a)} Multiple testing correction for pairwise G-statistics p-values by using the Benjamini-Hochberg method.

Figure 4. Proteins significant to AIS and MIA identified by pairwise G-statistics analysis²¹. The *q* values by multiple testing correction using the Benjamini–Hochberg method are presented.

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 organization, regulation of vascular endothelial growth factor receptor (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 (VEGFR2) 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 activity²⁴. Perlecan interacts with elastin (ELN), vitronectin (VTN), LTBP2, TGFBI, elastin microfibril interfacier 1 (EMILIN1), collagen alpha-2(VI) chain (COL6A2), periostin (POSTN), and lysozyme C (LYZ) (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 tissues²⁵. 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 only enriched pathway

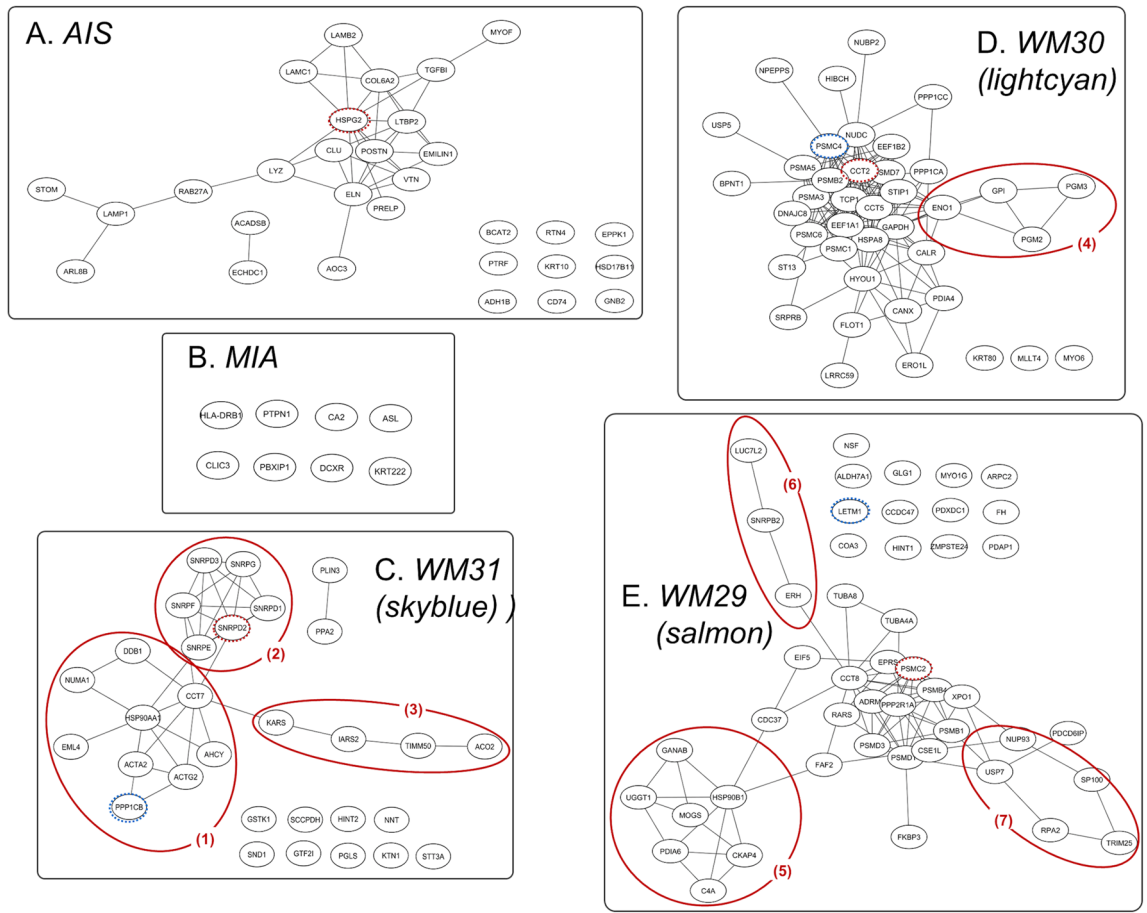


Figure 5. Protein–protein interaction networks identified for (A) AIS and (B) MIA, and the top three LPA protein networks: (C) WM31, (D) WM30, and (E) WM29 modules. Dotted circle nodes in blue and red represent eigen-proteins and hub proteins, respectively, for each module. Solid red circles with numbers represent subnetworks.

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 leukemia 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- β 1-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 *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 adenocarcinoma³⁰. 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

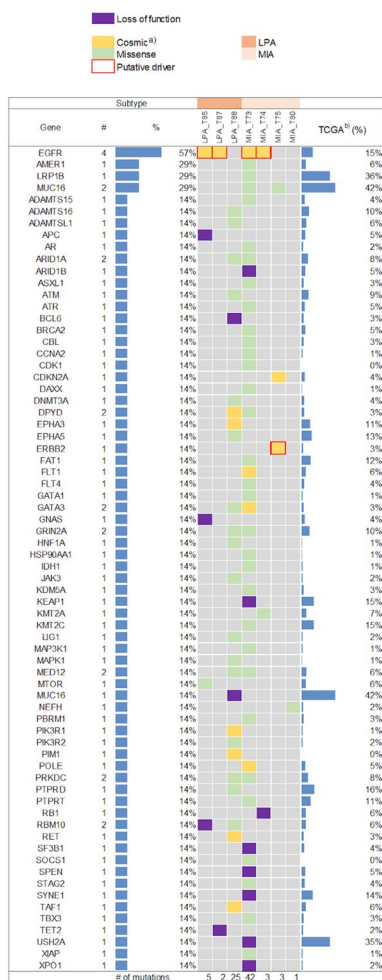
Subtype	WGCNA Module ID	Biological Process (GO)				Reactome Pathways			
		#term ID	Term description	Observed gene count	False discovery rate	#term ID	Term description	Observed gene count	False discovery rate
Adenocarcinoma in-situ (AIS)		GO:0030198	Extracellular matrix organization	9	0.00000688	HSA-1474244	Extracellular matrix organization	8	0.00000205
		GO:0030947	Regulation of vascular endothelial growth factor receptor signaling pathway	3	0.0073	HSA-3000178	ECM proteoglycans	5	0.00000959
		GO:0007155	Cell adhesion	8	0.0098	HSA-2129379	Molecules associated with elastic fibres	4	0.0000211
		GO:0051270	Regulation of cellular component movement	8	0.0098	HSA-3000171	Non-integrin membrane-ECM interactions	4	0.0000683
		GO:1905907	Negative regulation of amyloid fibril formation	2	0.0098	HSA-3000157	Laminin interactions	3	0.00042
		GO:0009653	Anatomical structure morphogenesis	11	0.016	HSA-1474228	Degradation of the extracellular matrix	4	0.0013
		GO:0009887	Cellular process	30	0.016	HSA-977225	Amyloid fiber formation	3	0.0047
		GO:0016043	Cellular component organization	18	0.016	HSA-216083	Integrin cell surface interactions	3	0.0049
		GO:0070206	Protein trimerization	3	0.016	HSA-70895	Branched-chain amino acid catabolism	2	0.0097
		GO:2000145	Regulation of cell motility	7	0.016	HSA-8874081	MET activates PTK2 signaling	2	0.0136
Minimally invasive adenocarcinoma (MIA)		N/a				HSA-877300	Interferon gamma signaling	2	0.0264
Lepidic predominant invasive adenocarcinoma (LPA)	WM31 (skyblue)	GO:0000387	Spliceosomal snmp assembly	6	4.35E-08	HSA-191859	snRNP Assembly	6	3.41E-08
		GO:0000245	Spliceosomal complex assembly	5	0.0000119	HSA-72165	mRNA Splicing - Minor pathway	6	3.41E-08
		GO:0022618	Ribonucleoprotein complex assembly	7	0.0000119	HSA-111367	SLBP independent processing of histone pre-mnas	4	0.00000181
		GO:0051170	Import into nucleus	6	0.0000119	HSA-77588	SLBP Dependent processing of replication-dependent histone pre-mRNAs	4	0.00000197
		GO:0008334	Histone mna metabolic process	4	0.0000146	HSA-72163	mRNA splicing - Major pathway	6	0.00000729
		GO:0006369	Termination of RNA polymerase II transcription	4	0.0000399	HSA-109688	Cleavage of growing transcript in the termination region	4	0.000059
		GO:0034622	Cellular protein-containing complex assembly	8	0.0017	HSA-379726	Mitochondrial tRNA aminoacylation	3	0.0000876
		GO:0090131	Mesenchyme migration	2	0.0027	HSA-71403	Citric acid cycle (TCA cycle)	2	0.0067
		GO:0006418	tRNA aminoacylation for protein translation	3	0.0033	HSA-163560	Triglyceride catabolism	2	0.0074
		GO:0051649	Establishment of localization in cell	10	0.0043	HSA-445355	Smooth muscle contraction	2	0.0119
	WM30 (lightcyan)	GO:0061077	Chaperone-mediated protein folding	9	2.18E-11	HSA-450408	AUF1 (hnRNP D0) binds and destabilizes mRNA	8	7.82E-11
		GO:0008457	Protein folding	11	3.7E-10	HSA-1169091	Activation of NF-kappaB in B cells	7	1.09E-09
		GO:0006458	'De novo' protein folding	6	0.00000015	HSA-1234176	Oxygen-dependent proline hydroxylation of Hypoxia-inducible Factor Alpha	7	1.09E-09
		GO:0034976	Response to endoplasmic reticulum stress	9	0.000000435	HSA-1236974	ER-Phagosome pathway	8	1.09E-09
		GO:0016579	Protein deubiquitination	8	0.0000212	HSA-1236978	Cross-presentation of soluble exogenous antigens (endosomes)	7	1.09E-09
		GO:0051084	'De novo' posttranslational protein folding	4	0.000037	HSA-174084	Autodegradation of Cdh1 by Cdh1:APC/C	7	1.09E-09
		GO:1901575	Organic substance catabolic process	15	0.0000461	HSA-174113	SCF-beta-TRCP mediated degradation of Emi1	7	1.09E-09
		GO:0044248	Cellular catabolic process	15	0.0000548	HSA-174154	APC/C:Cdc20 mediated degradation of Securin	7	1.09E-09
		GO:0031647	Regulation of protein stability	7	0.0000978	HSA-174178	APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1	7	1.09E-09
		GO:0050821	Protein stabilization	6	0.00014	HSA-174184	Cdc20:Phospho-APC/C mediated degradation of Cyclin A	7	1.09E-09
	WM29 (Salmon)	GO:0051603	Proteolysis involved in cellular protein catabolic process	14	2.85E-08	HSA-392499	Metabolism of proteins	24	7.61E-10
		GO:1901565	Organonitrogen compound catabolic process	16	0.00000135	HSA-597592	Post-translational protein modification	20	3.06E-09
		GO:0044265	Cellular macromolecule catabolic process	15	0.00000184	HSA-2467813	Separation of sister chromatids	9	0.00000084
		GO:0006508	proteolysis	15	0.0000141	HSA-69275	G2/M transition	9	9.14E-08
		GO:0036503	ERAD pathway	6	0.0000141	HSA-8852276	The role of G1S1 in G2/M progression after G2 checkpoint	7	9.14E-08
		GO:1901575	Organic substance catabolic process	17	0.0000141	HSA-68886	M Phase	10	0.00000471
		GO:0044248	Cellular catabolic process	17	0.0000165	HSA-69278	Cell cycle, mitotic	11	0.00000838
		GO:0019941	Modification-dependent protein catabolic process	10	0.0000175	HSA-162909	Host interactions of HIV factors	7	0.00000927
GO:0043687		Post-translational protein modification	9	0.0000212	HSA-69206	G1/S transition	7	0.00000123	
GO:0044267		Cellular protein metabolic process	24	0.0000409	HSA-195253	Degradation of beta-catenin by the destruction complex	6	0.00000187	

Figure 6. Top pathways enriched for the protein core networks obtained for AIS, MIA, and LPA concerning Biological Process (GO) and Reactome pathways.

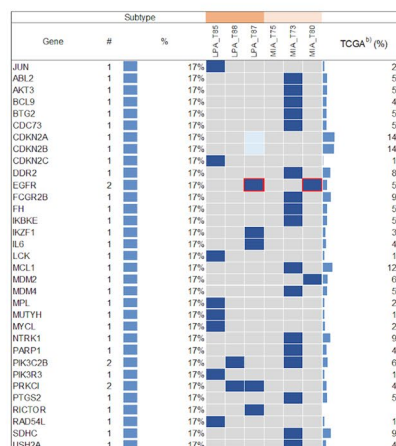
involved in mesenchyme migration, cell cycle, mitotic/recruitment of NuMA to mitotic centrosomes and regulation of Polo-like kinase 1 (PLK1) activity at G2/M Transition. The subnetwork (2) including the hub protein small nuclear ribonucleoprotein Sm D2 (SM-D2) encoded by *SNRPD2* belongs to the spliceosomal 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- κ B) 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 translation³². 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)³³; its overexpression has been identified in numerous cancers, including colorectal carcinoma³⁴, pancreatic cancer³⁵, cholangiocellular carcinoma³⁶ and ovarian cancer³⁷, and it is possibly associated with poor survival outcomes in patients with cancer³⁸. *ST13* encodes ST13 HSP70s-interacting protein (HIP)/putative tumour suppressor ST13/suppression of tumorigenicity 13 protein, which is also involved in the regulation of heat shock factor 1 (HSF1)-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.³⁹ reported that signal recognition particle receptor subunit beta

A. Somatic mutation



B. Copy number variation (CNV)



^{a)}Variant identified in the COSMIC (Catalogue of Somatic Mutations in Cancer).

^{b)}TCGA lung adenocarcinoma (M stage: M0, Mx, NA; N stage: N0, NA; sample number: $n=471$) taken from TCGA, Nat Genet 2016.

C. Copy number loss

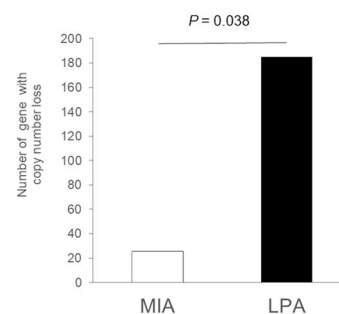


Figure 7. Genomic alterations obtained for MIA and LPA. **(A)** Somatic mutations, and **(B)** copy number variations (CNVs) together with the TCGA lung adenocarcinoma datasets (sample size: $n = 471$). **(C)** Analysis of copy number loss between MIA and LPA. There are significant more genes with copy number loss ($CN < 2$) found in the LPA group ($p = 0.038$).

(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- κ B 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 (ADC) 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 LPA1. 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 studies⁴¹. Aside from the tumour which exhibited the APOBEC-like mutation signature, in our cohort, a high percentage of MIAs and LPAs 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 control⁴³. 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 subpopulation 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)⁴⁴ 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^{11,12}.

The mutational landscape shown in the larger study conducted by Qian et al.⁴² 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.⁴¹ 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 study⁴².

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^{41,42}. However, together with our protein network-based profiles and the previous OPLS discriminant analysis of mutant proteomes¹¹, 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 subtypes⁴⁵ 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)⁴⁶. 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- μ m-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- μ L low-binding tube. Proteins were extracted and digested with trypsin using LIQUID TISSUE MS Protein Prep kits (Onco-

Plex Diagnostics Inc.) according to the manufacturer's protocol⁴⁷. The procedures have been described in detail elsewhere^{12,13}.

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 μ L 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 analysis has been described in detail previously¹³. The expressions of the identified proteins were assessed by spectral count-based protein quantification⁴⁸.

Tumour sequencing and analysis of genetic alterations. The NGS was performed with the ACTONCO panel. The extracted genomic DNA was amplified by a polymerase chain reaction to enrich the targeting coding exons of the analysed genes. For variant analysis, raw reads were mapped to the hg19 reference genome using the ION TORRENT SUITE version 5.2 software program (Thermo Fisher Scientific). Coverage depth was calculated using TORRENT COVERAGE ANALYSIS PLUG-IN. Single-nucleotide variants and short insertion/deletions were identified using the TORRENT VARIANT CALLER PLUG-IN (version 5.2; Thermo Fisher Scientific). The coverage was down-sampled to 4,000. The VARIANT EFFECT PREDICTOR version 88 software program (European Bioinformatics Institute, Cambridge, UK) was used to annotate variants using datasets from the Catalogue Of Somatic Mutations In Cancer (COSMIC) version 86 resource and the GENOME AGGREGATION database (version 2.0.2; MacArthur Lab, Boston, MA, USA). Variants with coverage ≥ 25 and an allele frequency of $\geq 10\%$ were retained for further analysis. Variants reported in the GENOME AGGREGATION database with a minor allele frequency of $> 1\%$ were considered as polymorphisms. The in-house peripheral blood mononuclear cell database of ACT Genomics was used to determine technical errors.

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-package¹⁷.

Protein-protein interaction network construction. To construct a protein interaction network for a protein module, we used the STRING database (version 11.0)²², which accumulates information on protein-protein interactions from various other databases such as IntAct, Reactome, DIP, BioGRID, MINT, 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 (<https://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

Published online: 12 August 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. & Saji, H. Worldwide trend of increasing primary adenocarcinoma of the lung. *Surg. Today.* **44**, 1004–1012 (2014).

5. Koike, T. *et al.* Japanese Joint Committee for Lung Cancer Registration. Improvements in surgical results for lung cancer from 1989 to 1999 in Japan. *J. Thorac. Oncol.* **4**, 1364–1369 (2009).
6. Travis, W. D. *et al.* International association for the study of lung cancer/American thoracic society/European respiratory society international multidisciplinary classification of lung adenocarcinoma. *J. Thoracic. Oncol.* **6**, 244–285 (2011).
7. Travis, W. D., Brambilla, E. & Riely, G. J. New pathologic classification of lung cancer: relevance for clinical practice and clinical trials. *J. Clin. Oncol.* **31**, 992–1001 (2013).
8. Yoshizawa, A. *et al.* Impact of proposed IASLC/ATS/ERS classification of lung adenocarcinoma: prognostic subgroups and implications for further revision of staging based on analysis of 514 stage I cases. *Mod. Pathol.* **24**, 653–664 (2011).
9. Yoshizawa, A. *et al.* Validation of the IASLC/ATS/ERS lung adenocarcinoma classification for prognosis and association with EGFR and KRAS gene mutations: analysis of 440 Japanese patients. *J. Thoracic Oncol.* **8**, 52–61 (2013).
10. Gu, J. *et al.* Prognostic significance of the IASLC/ATS/ERS classification in Chinese patients—a single-institution retrospective study of 292 lung adenocarcinoma. *J. Surg. Oncol.* **107**, 474–480 (2013).
11. Nishimura, T. *et al.* Current status of clinical proteogenomics in lung cancer. *Expert Rev. Proteom.* **16**, 761–772 (2019).
12. Kato, Y. *et al.* A proteomic profiling of laser-microdissected lung adenocarcinoma cells of early lepidic-types. *Clin. Transl. Med.* **4**, 64. <https://doi.org/10.1186/s40169-015-0064-3> (2015).
13. Fujii, K. *et al.* Differential proteomic analysis between small cell lung carcinoma (SCLC) and pulmonary carcinoid tumors reveals molecular signatures for malignancy in lung cancer. *Proteom. Clin. Appl.* **12**, e1800015. <https://doi.org/10.1002/prca.201800015> (2018).
14. Nakamura, H. *et al.* Identification of key modules and hub genes for small-cell lung carcinoma and large-cell neuroendocrine lung carcinoma by weighted gene co-expression network analysis of clinical tissue-proteomes. *PLoS ONE* **14**, e0217105 (2019).
15. Wu, K., Huang, R. S., House, L. & Cho, W. C. Next-generation sequencing for lung cancer. *Future Oncol.* **9**, 1323–1336 (2013).
16. Devarakonda, S., Masood, A. & Govindan, R. Next-generation sequencing of lung cancers: lessons learned and future directions. *Hematol. Oncol. Clin. N. Am.* **31**, 1–12 (2017).
17. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559. <https://doi.org/10.1186/1471-2105-9-559> (2008).
18. Liu, X., Hu, A. X., Zhao, J. L. & Chen, F. L. Identification of key gene modules in human osteosarcoma by co-expression analysis weighted gene co-expression network analysis (WGCNA). *J. Cell. Biochem.* **118**, 3953–3959 (2017).
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. *J. Cell. Biochem.* **119**, 2102–2109 (2018).
20. Mi, H. *et al.* PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
21. Sokal, R. R. & Rohlf, F. J. *Biometry: The Principles and Practice of Statistics in Biological Research* 729–731 (W.H. Freeman, New York, 1995).
22. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
23. Chin, C. H. *et al.* cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **8**(Suppl 4), S11. <https://doi.org/10.1186/1752-0509-8-S4-S11> (2014).
24. Douglass, S., Goyal, A. & Iozzo, R. V. The role of perlecan and endorepellin in the control of tumor angiogenesis and endothelial cell autophagy. *Connect. Tissue Res.* **56**, 381–391 (2015).
25. Mollinedo, F. Neutrophil degranulation, plasticity, and cancer metastasis. *Trends Immunol.* **40**, 228–242 (2019).
26. Liu, H. *et al.* PTP1B promotes cell proliferation and metastasis through activating src and ERK1/2 in non-small cell lung cancer. *Cancer Lett.* **359**, 218–225 (2015).
27. Kostrzewa, T. *et al.* Inhibitors of protein tyrosine phosphatase PTP1B with anticancer potential. *Anticancer Res.* **39**, 3379–3384 (2019).
28. Shi, S. *et al.* HPIP silencing inhibits TGF- β 1-induced EMT in lung cancer cells. *Int. J. Mol. Med.* **39**, 479–483 (2017).
29. Yang, S. *et al.* Diacetyl/L-xylulose reductase mediates chemical redox cycling in lung epithelial cells. *Chem. Res. Toxicol.* **30**, 1406–1418 (2017).
30. Verdugo-Sivianes, E. M. *et al.* Coordinated downregulation of Spinophilin and the catalytic subunits of PP1, PPP1CA/B/C, contributes to a worse prognosis in lung cancer. *Oncotarget.* **8**, 105196–105210 (2017).
31. Aghajian, Y., Levy, M. L., Malicki, D. M. & Crawford, J. R. Novel fusion protein in a high-grade glioma of infancy. *BMJ Case Rep.* <https://doi.org/10.1136/bcr-2016-217189> (2016).
32. Vera, M. *et al.* The translation elongation factor eEF1A1 couples transcription to translation during heat shock response. *Elife* **3**, e03164. <https://doi.org/10.7554/eLife.03164> (2014).
33. Xu, Y. W. *et al.* Serum autoantibodies against stip1 as a potential biomarker in the diagnosis of esophageal squamous cell carcinoma. *Dis. Mark.* **2017**, 5384091. <https://doi.org/10.1155/2017/5384091> (2017).
34. Kubota, H. *et al.* Increased expression of co-chaperone HOP with HSP90s and HSC70 and complex formation in human colonic carcinoma. *Cell Stress Chaperones.* **15**, 1003–1011 (2010).
35. Walsh, N. *et al.* Identification of pancreatic cancer invasion-related proteins by proteomic analysis. *Proteome Sci.* **7**, 3. <https://doi.org/10.1186/1477-5956-7-3> (2009).
36. Padden, J. *et al.* Identification of novel biomarker candidates for the immunohistochemical diagnosis of cholangiocellular carcinoma. *Mol. Cell. Proteom.* **13**, 2661–2672 (2014).
37. Cho, H. *et al.* Expression of stress-induced phosphoprotein1 (STIP1) is associated with tumour progression and poor prognosis in epithelial ovarian cancer. *Genes Chromosom. Cancer.* **53**, 277–288 (2014).
38. Chao, A. *et al.* Tumour stress-induced phosphoprotein1 (STIP1) as a prognostic biomarker in ovarian cancer. *PLoS ONE* **8**, e57084. <https://doi.org/10.1371/journal.pone.0057084> (2013).
39. Ma, Q. *et al.* SERP1 is a novel marker of poor prognosis in pancreatic ductal adenocarcinoma patients via anti-apoptosis and regulating SRPRB/NF- κ B axis. *Int. J. Oncol.* **51**, 1104–1114 (2017).
40. 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).
41. 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).
42. 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).
43. Willis, R. E. 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).
44. Bylesjö, M. *et al.* OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J. Chemom.* **20**, 341–351 (2006).
45. 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).
46. 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. Zybailov, 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 and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020