SUPPLEMENTARY MATERIALS

Integrative analysis of imaging and transcriptomic data for immune landscape associated with tumor metabolism in lung adenocarcinoma: Clinical and prognostic implications

Supplementary Methods

The preprocessing step for transcriptome data

For identifying tumor metabolism-associated gene coexpression network module, we made the training set with two microarray datasets from Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/) [1]. A microarray dataset with ^{18}F-Fluorodeoxyglucose (FDG) positron emission tomography (PET) image data (accession number GSE28827 [2, 3]) was included in the training set. Because GSE28827 includes few lung adenocarcinoma (LUAD) samples for conducting gene coexpression network analysis, we merged additional microarray dataset (accession number GSE31210 [4, 5]). The normalized gene expression data of GSE28827 was downloaded using 'GEOquery' R package [6]. The raw gene expression data of GSE31210 was downloaded from the Gene Expression Omnibus data repository and called and normalized using the robust multichip average method using the 'affy' R package [7]. Since two datasets included multiple histologic types of non-small cell lung cancer, only LUAD samples were extracted for further preprocessing step. On a study-by-study basis, we removed invalid and duplicated probe sets by 'featureFilter' function in 'genefilter' R package [8], and mapped array probe sets for the respective gene symbols. As we combined microarray data from different studies, we performed additional normalization using Combat algorithm in order to eliminate potential batch effect [9]. Lastly, to remove poor quality probes, we filtered out probe sets with low expression level (signal intensity < $\log_2(100)$ in at least 25% of samples within at least one study) and low variability (interquartile range < 0.75). As a result, the training set contained 4010 genes from 246 LUAD samples including 20 samples with available FDG PET image.

For validation of the tumor metabolism-associated gene coexpression network modules, we used mRNA transcriptome data of LUAD from The Cancer Genome Atlas projects

(TCGA) [10]. Using 'TCGABiolinks' R package [11], we downloaded the level three RNA sequence data of LUAD from TCGA data portal (https://portal.gdc.cancer.gov/), which consisted of 21022 genes from 515 samples obtained with Illumina HiSeq RNASeqV2 (Illumina, San Diego, CA, USA). Clinical information, including vital status, follow-up time, and time of death was also collected in the same manner. We searched for possible outlier samples from the raw expression data by calculating array-array intensity correlation based on the Pearson's correlation coefficient for all samples; consequently, twenty-five outliers were removed from the raw expression data. We then normalized mRNA transcripts using 'TCGAAnalyze_Normalization' function and the expression data of 18323 genes from 490 samples were included for the validation test.

FDG PET/CT Data and Image Processing

In this study, we used FDG-PET/CT data of both training and validation sets provided by The Cancer Imaging Archive [2, 12, 13]. We identified 20 and 17 patients having both transcriptome and FDG PET data available from the training and validation set, respectively. For the training set (GSE28827), FDG was injected with a dose between 370 and 629 MBq depending on patients' weight. Scans were acquired by using a Discovery STE or LS PET/CT scanner (GE Healthcare) (section thicknesses, between 3 and 5 mm) with an iterative algorithm (ordered subset expectation maximization, OSEM). For the validation set (TCGA data), patients were administered mean 579.5 MBq (range: 518-724 MBq) FDG and images were acquired 60 minutes after administration. PET data were reconstructed by an iterative algorithm (OSEM). The acquisition and reconstruction parameters such as matrix size were different according to the imaging protocol of institute.

To characterize tumor metabolism, the maximum standardized uptake value was calculated.

A manually drawn spherical volume-of-interest around the tumor lesion was used for measuring maximum standardized uptake value. Image parameters were obtained by Metavol package [14].

Gene ontology enrichment analysis

The enrichment of the gene ontology terms in tumor metabolism-associated module was evaluated based on the hypergeometric test using ‘clusterProfiler’ R package [15]. The gene ontology biological process terms at false discovery rate under < 0.05 in each tumor metabolism-associated module were regarded as significantly enriched terms.

LUAD molecular subtypes classification

The LUAD centroid subtypes (bronchioid, magnoid and squamoid) were assigned to all samples of TCGA [16]. Previously published classifier employed the nearest centroid classification based on 506 genes, which included several missing gene expression data in TCGA samples. Thus, for subtype classification, common genes of the classifier and TCGA samples were selected and the Pearson correlation was used as the similarity metric. A subtype with the maximum correlation coefficient was assigned to each sample as the previous TCGA study [10].

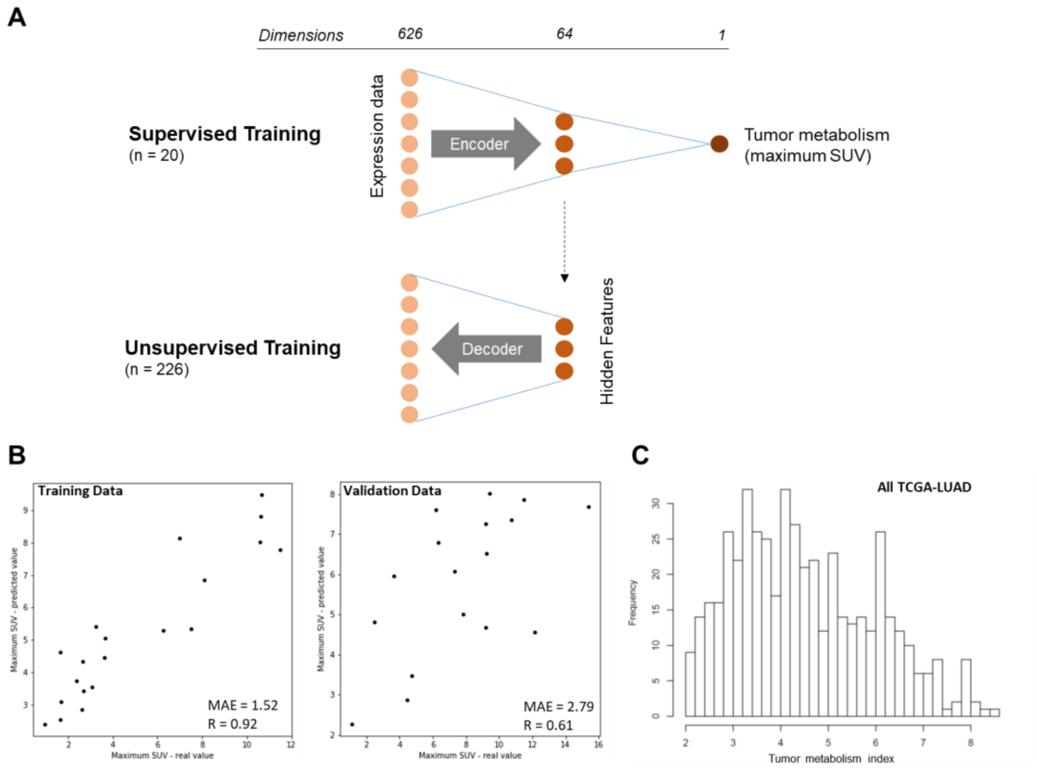
Glucose metabolism signatures

Tumor metabolism index (TMI) was compared with gene signatures representing glucose metabolism. Glucose metabolism signatures were obtained by two different methods. Firstly, mean expression value of manually selected genes associated with glycolysis and gluconeogenesis was used as a metabolic signature [17]. Secondly, we used Reactome to select genes of glycolysis pathway [18]. To obtain enrichment score, we used single sample gene set enrichment analysis (ssGSEA) which provide pathway activity for each sample [19]. The output of ssGSEA was normalized by z-score across samples and compared with TMI. The Spearman's method was used for the correlation analysis.

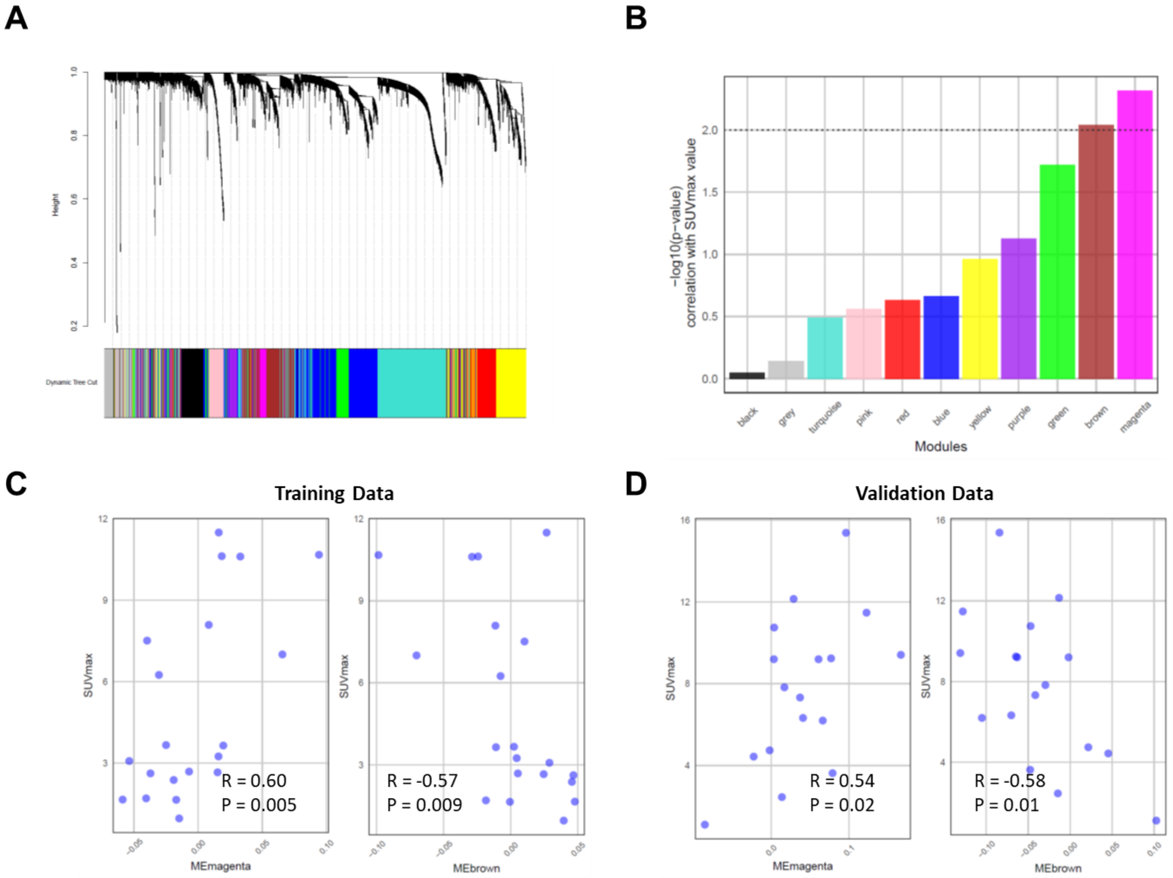
Validation in an independent cohort

To verify the association between immune landscape, tumor metabolism, and prognosis, the analyses were additionally performed in an independent lung adenocarcinoma cohort [20] (GSE41271). The normalized gene expression data of GSE41271 was downloaded and TMI and cell type enrichment scores of all lung adenocarcinoma samples were calculated by the trained model, and xCell [21], respectively. To define the clusters based on TCGA data, we obtained centers of each cluster from cell types enrichment scores of TCGA data. We calculated Euclidean distance between cell types enrichment score of each sample of the independent data and center of each cluster, and then assigned the cluster with the lowest distance to each sample. TMI and ImmuneScore of clusters were compared by one-way ANOVA followed by *post hoc* Tukey's test. The association of overall survival and variables including TMI, ImmuneScore, and clusters was analyzed by the Cox regression analysis. The survival rate of the groups was depicted with the Kaplan-Meier's method and compared with the log-rank test. To define risk groups, TMI and ImmuneScore were dichotomized using the median value of each variable in the validation set.

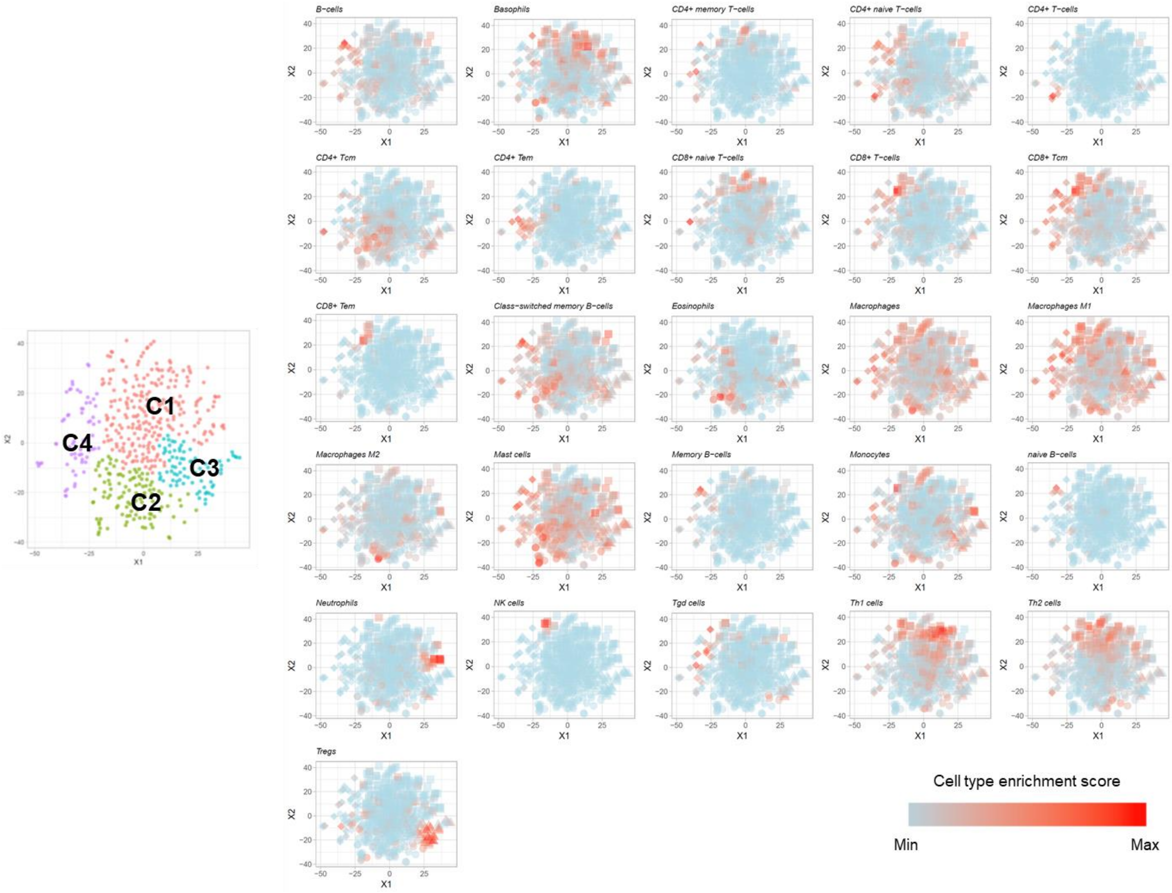
Supplementary Figures



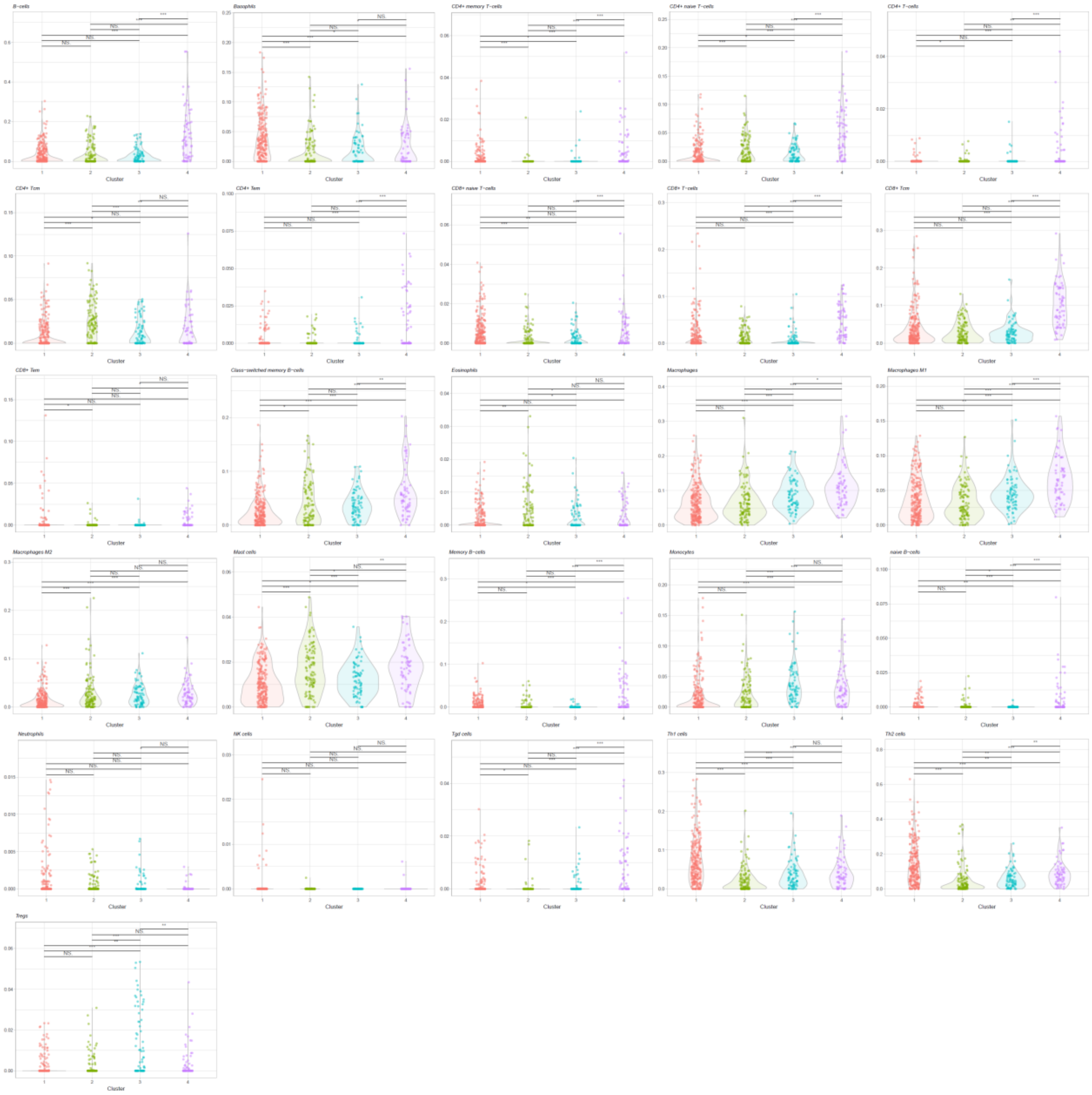
Supplementary Figure 1. Tumor metabolism estimation model. (A) The neural network model predicted tumor metabolism estimated by FDG PET. The input of the neural network was gene expression data of two tumor metabolism-associated modules. As the training data consist of gene expression data with or without matched PET data, the parameters of neural network were updated by unsupervised and supervised training. The supervised training was aimed at minimizing the error between tumor metabolism predicted by the model and measured by FDG PET. The gene expression data without PET data were used for training the robust feature layer with unsupervised learning, denoising autoencoder. (B) After the training, the model was applied to TCGA data, an independent data with large samples, for validation. The performance of tumor metabolism estimation model in both the training set and TCGA data was presented (MAE: mean absolute error). (C) The histogram shows the distribution of tumor metabolism index of all samples of TCGA projects estimated by our model.



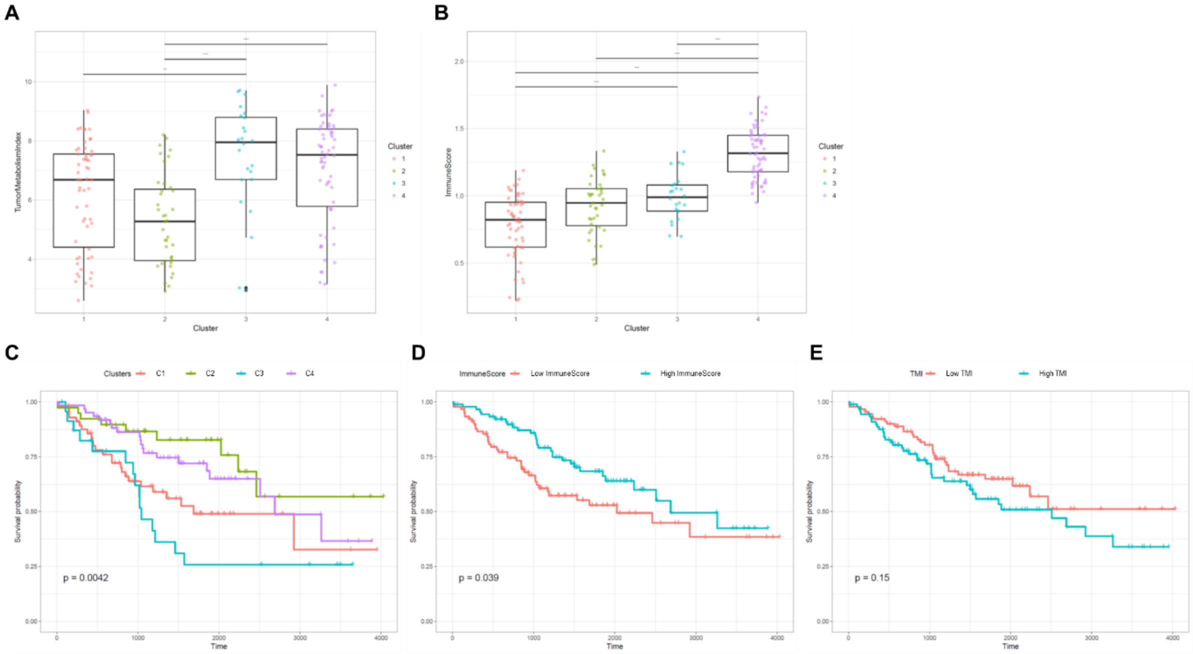
Supplementary Figure 2. Identification of tumor metabolism-associated gene coexpression network modules. (A) Gene coexpression network modules identification from the training set using weighted gene coexpression network analysis. Total 10 gene network modules were identified, except the gray color representing genes not assigned to any module. (B) The p-value of the correlation test with training set was shown in the bar plot. The dotted line represents statistical significance threshold (false discovery rate-corrected p-value = 0.05); magenta and brown modules were significantly correlated with maximum SUV. (C, D) The scatterplot shows the correlation between module eigenegene and maximum SUV in the training set (C) and TCGA data (D).



Supplementary Figure 3. Two dimensional tumor microenvironment landscape map with individual immune cell enrichment scores. The 2D projection of tumor microenvironment cellular landscape was visualized with each immune cell type enrichment score. The left panel showed the tumor microenvironment cell type-based clusters.



Supplementary Figure 4. Distribution of individual immune cell enrichment scores between cell type-based clusters. Scatter plots were drawn for each immune cell type enrichment score. The comparison between two paired clusters was performed by the nonparametric Dunn test (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).



Supplementary Figure 5. Independent validation of the association of TMI, ImmuneScore and survival analysis. The validation of the analyses was performed by an independent cohort. TMI (A) and ImmuneScore (B) were different between clusters and the pattern of difference was consistent with the results of TCGA data (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). (C) As TCGA data, C2 and C3 were associated with favorable prognosis. (D) The subjects with high ImmuneScore showed significantly better prognosis. (E) A trend of poor prognosis in high TMI tumors was found as results of TCGA data.

Supplementary Tables

Supplementary Table 1. Demographic and baseline clinical characteristics of TCGA LUAD data

Variables		TCGA LUAD data (n = 490)	
			Available data
Sex	Female:Male	265:225 (54.1%:45.9%)	490
Age (years)		65.9 ± 10.0 (38.5 – 88.8)	459
Race	American Indian	1 (0.2%)	427
	Asian	8 (1.7%)	
	Black	49 (10.3%)	
	White	369 (87.8%)	
Stage	1	262 (54.4%)	482
	2	116 (24.1%)	
	3	81 (16.8%)	
	4	23 (4.7%)	
Status	Death:Alive	311 : 179 (63.5% : 36.5%)	490
Survival time (months)		21.8 (0.1 – 241.6)	481

TCGA = The Cancer Genome Atlas; LUAD = Lung Adenocarcinoma

Supplementary Table 2. Top 10 GO biological process terms of the two tumor metabolism-associated modules

	ID	Biological process	Count	q-value
Magenta module	GO:0030198	Extracellular matrix organization	55	1.1×10^{-25}
	GO:0043062	Extracellular structure organization	55	1.1×10^{-25}
	GO:0030574	Collagen catabolic process	24	1.5×10^{-18}
	GO:0044243	Multicellular organism catabolic process	24	1.7×10^{-17}
	GO:0032963	Collagen metabolic process	26	1.5×10^{-15}
	GO:0044259	Multicellular organismal macromolecule metabolic process	26	4.3×10^{-15}
	GO:0044236	Multicellular organism metabolic process	26	1.2×10^{-13}
	GO:0030199	Collagen fibril organization	15	6.7×10^{-12}
	GO:0048514	Blood vessel morphogenesis	39	1.3×10^{-7} 7
	GO:0001525	Angiogenesis	35	2.0×10^{-7}
Brown module	GO:0060485	Mesenchyme development	15	1.2×10^{-6}
	GO:0030198	Extracellular matrix organization	15	2.7×10^{-5}
	GO:0043062	Extracellular structure organization	15	2.7×10^{-5}
	GO:0050673	Epithelial cell proliferation	15	2.7×10^{-5}
	GO:0001503	Ossification	15	5.6×10^{-5}
	GO:0001501	Skeletal system development	17	7.0×10^{-5}
	GO:0050678	Regulation of epithelial cell proliferation	13	8.5×10^{-5}
	GO:0048762	Mesenchymal cell differentiation	10	4.2×10^{-4}
	GO:0001837	Epithelial to mesenchymal transition	8	5.0×10^{-4}
	GO:0006024	Glycosaminoglycan biosynthetic process	8	5.0×10^{-4}

GO = Gene Ontology

Supplementary Table 3. Tumor metabolism index of each cluster and result of *post hoc* analysis

	C1 (n = 240)	C2 (n = 109)	C3 (n = 77)	C4 (n = 64)
Tumor Metabolism Index (Mean +- SD)	4.71±1.37	3.51±1.24	5.25±1.44	4.50±1.34
p-value for statistical comparison (Tukey's *post hoc* test)				
C1		<1E-8	0.01	0.68
C2			<1E-8	<0.0001
C3				0.005
C4				

Supplementary Table 4. ImmuneScore of each cluster and result of *post hoc* analysis

	C1 (n = 240)	C2 (n = 109)	C3 (n = 77)	C4 (n = 64)
ImmuneScore (Mean +- SD)	0.12±0.10	0.14±0.09	0.15±0.08	0.28±0.13
p-value for statistical comparison (Tukey's *post hoc* test)				
C1		0.35	0.02	<1E-8
C2			0.57	<1E-8
C3				<1E-8
C4				

Supplementary References

1. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41:D991-5.
2. Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, et al. Non–small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results. Radiology. 2012;264:387-96.
3. Nair VS, Gevaert O, Davidzon G, Napel S, Graves EE, Hoang CD, et al. Prognostic PET 18F-FDG uptake imaging features are associated with major oncogenomic alterations in patients with resected non-small cell lung cancer. Cancer Res. 2012;72:3725-34.
4. Yamauchi M, Yamaguchi R, Nakata A, Kohno T, Nagasaki M, Shimamura T, et al. Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma. PloS one. 2012;7:e43923.
5. Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, Iwakawa R, et al. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res. 2012;72:100-11.
6. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23:1846-7.
7. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20:307-15.
8. Gentleman R, Carey V, Huber W, Hahne F. Genefilter: Methods for filtering genes from microarray experiments. R package version. 2011;1.
9. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118-27.
10. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543-50.
11. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44:e71.
12. Albertina B, Watson, M., Holback, C. et al. Radiology Data from The Cancer Genome Atlas Lung Adenocarcinoma [TCGA-LUAD] collection. The Cancer Imaging Archive. 2016.
13. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045-57.
14. Hirata K, Kobayashi K, Wong K-P, Manabe O, Surmak A, Tamaki N, et al. A semi-automated technique determining the liver standardized uptake value reference for tumor delineation in FDG PET-CT. PloS one. 2014;9:e105682.
15. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284-7.
16. Wilkerson MD, Yin X, Walter V, Zhao N, Cabanski CR, Hayward MC, et al. Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. PloS one. 2012;7:e36530.
17. Gaude E, Frezza C. Tissue-specific and convergent metabolic transformation of cancer correlates with metastatic potential and patient survival. Nat Commun. 2016;7:13041.
18. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33:D428-32.
19. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA

interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462:108-12.

20. Sato M, Larsen JE, Lee W, Sun H, Shames DS, Dalvi MP, et al. Human lung epithelial cells progressed to malignancy through specific oncogenic manipulations. Mol Cancer Res. 2013;11:638-50.

21. Aran D, Hu Z, Butte AJ. xCell: Digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18:220.