nature medicine

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41591-018-0177-5

In the format provided by the authors and unedited.

Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning

Nicolas Coudray[1,2,9], Paolo Santiago Ocampo[3,9], Theodore Sakellaropoulos[4], Navneet Narula[3], Matija Snuderl[3], David Fenyö[5,6], Andre L. Moreira[3,7], Narges Razavian[8*] and Aristotelis Tsirigos[1,3*]

[1]Applied Bioinformatics Laboratories, New York University School of Medicine, New York, NY, USA. [2]Skirball Institute, Department of Cell Biology, New York University School of Medicine, New York, NY, USA. [3]Department of Pathology, New York University School of Medicine, New York, NY, USA. [4]School of Mechanical Engineering, National Technical University of Athens, Zografou, Greece. [5]Institute for Systems Genetics, New York University School of Medicine, New York, NY, USA. [6]Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, NY, USA. [7]Center for Biospecimen Research and Development, New York University, New York, NY, USA. [8]Department of Population Health and the Center for Healthcare Innovation and Delivery Science, New York University School of Medicine, New York, NY, USA. [9]These authors contributed equally to this work: Nicolas Coudray, Paolo Santiago Ocampo. *e-mail: narges.razavian@nyumc.org; aristotelis.tsirigos@nyumc.org

Supplementary Figure Legends

Supplementary Figure 1. Workflow of the computational analysis of H&E scans of lung tissues. The images are first run through a first classifier to determine the lung cancer types and identify the regions where LUAD cancer is present. Then, the mutation prediction network is run on those regions.

Supplementary Figure 2. Accurate classification of lung cancer histopathology images. (a) Per-slide Receiver Operating Characteristic (ROC) curves after classification of normal versus tumor images (using 20x magnified tiles) resulted in an almost error-free classification. **(b)** The ROC curves obtained after transfer learning for LUAD vs LUSC images classification show inferior performance compared to those obtained from **(c)** fully trained network. Aggregation was either done by averaging the probability scores (purple ROC curves) or by counting the percentage of properly classified tiles (green ROC curves). In all panels, the crosses correspond to the manual classification of LUAD vs LUSC slides by pathologists. **(d)** Multi-class ROC of the Normal vs LUAD vs LUSC classification yields the best result for overall classification of cancer types. Dotted lines are negative control trained and tested after random label assignments. In **(e)** and **(f)**, training and testing in (c) and (d) were replicated using tiles at 5x magnification instead of 20x. The ROC curves show that performance is similar for both magnifications. n=244 slides for **b,c,e** and n=170 slides for **a,d,f,** all from 137 patients**. (g)** Comparison of AUCs obtained with different techniques for classification of normal and **(h)** of cancer type slides (For Terry et a.[1]l, IHC stands for Immunoschemistry. For Khosravi et al.[2],data from inter-images tests on the TCGA and Stanford Tissue Microarray databases are displayed). **(i)** Proportion of LUAD and LUSC slides misclassified by the pathologists as a function of the true positive probability assigned in (c). The number of slides are indicated on the bars.

Supplementary Figure 3. Impact of tumor selection on model performance. (a) Tumor content distributions and AUCs across datasets after manual tumor selection, no selection and automatic selection using a deep learning model (for Frozen, FFPE and Biopsies respectively, n=98, 140 and 102 biologically independent slides; whiskers represent the minima and maxima. The middle line within the box represents the median; the AUC values are shown with the error bars representing the 95% CIs), **(b)** Difference in AUC compared to manual tumor selection (20x magnification).

Supplementary Figure 4. Relationship between the number of tiles in biopsy slides versus the accuracy of the three-way classifier. R-squared is shown for the linear fit obtained by linear regression (black line) of the LUAD (red) and LUSC (blue) data points (n=102 biologically independent slides). Also, the dataset was split in 3 equal sets (same number of tiles) and AUCs were computed for slides with a low, medium or high number of tiles. AUCs are shown above the graphs.

Supplementary Figure 5. Gene mutation prediction from histopathology slides give promising results for at least 6 genes: (a) Mutation probability distribution for slides where each mutation is present and absent after tile aggregation done by counting the percentage of tiles properly classified. n=62 slides from 59 patients. p-values estimated with two-tailed Mann-Whitney U-test are shown as ns ($p>0.05$), * ($p\leq0.05$), ** ($p\leq0.01$) or *** ($p\leq0.001$). Whiskers represent the minima and maxima. The middle line within the box represents the median. **(b)** ROC curves associated with (a).

Supplementary Figure 6. Illustration of gene mutations learned by deep-learning projected to 2 dimensions for visualization via the t-SNE algorithm using values of the last fully connected layer. (a) Scatterplots where each point represents a tile where the color is proportional to the mutation probability generated by the deep learning network. **(b)** Tile-embedded t-SNE representation with zooms on clusters having specific mutation predictions. n=24,144 tiles of 62 slides from 59 patients.

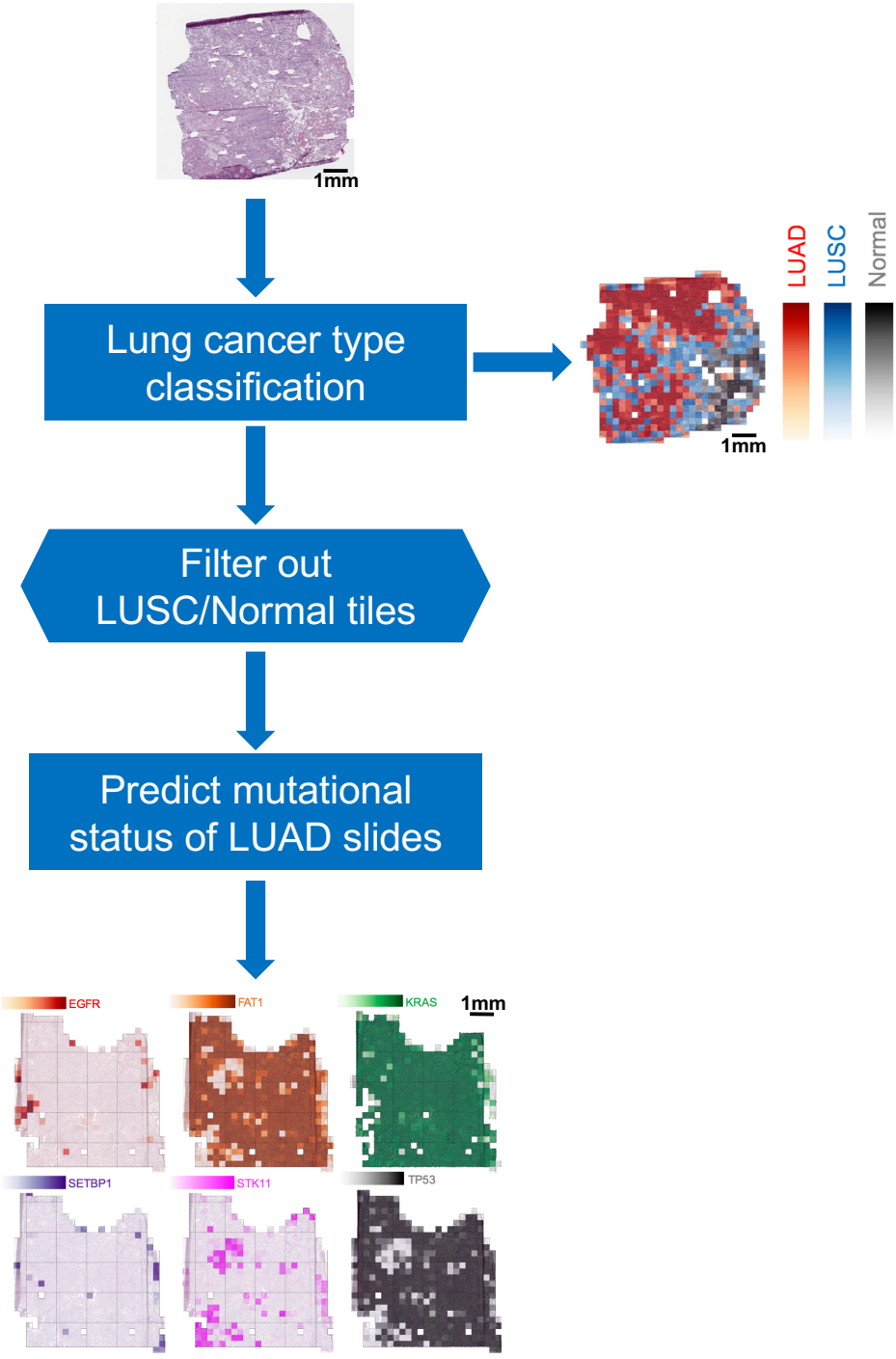
Supplementary Figure 7. Example of cases classified correctly by the algorithm but misclassified by at least one pathologist. For each case, we show the original image with a 250x250um zoom from the center of the image, the heatmap generated by the LUAD/LUSC binary classifier and the heatmaps from the Norma/LUAD/LUSC classifier. Training was done once.

Supplementary Figure 8. Evolution of train loss and validation accuracy during the training of the modified architecture inception v3 for the prediction of mutations. n~212,000 tiles from 320 slides for the training set and n~24,400 tiles from 48 slides for the validation set.

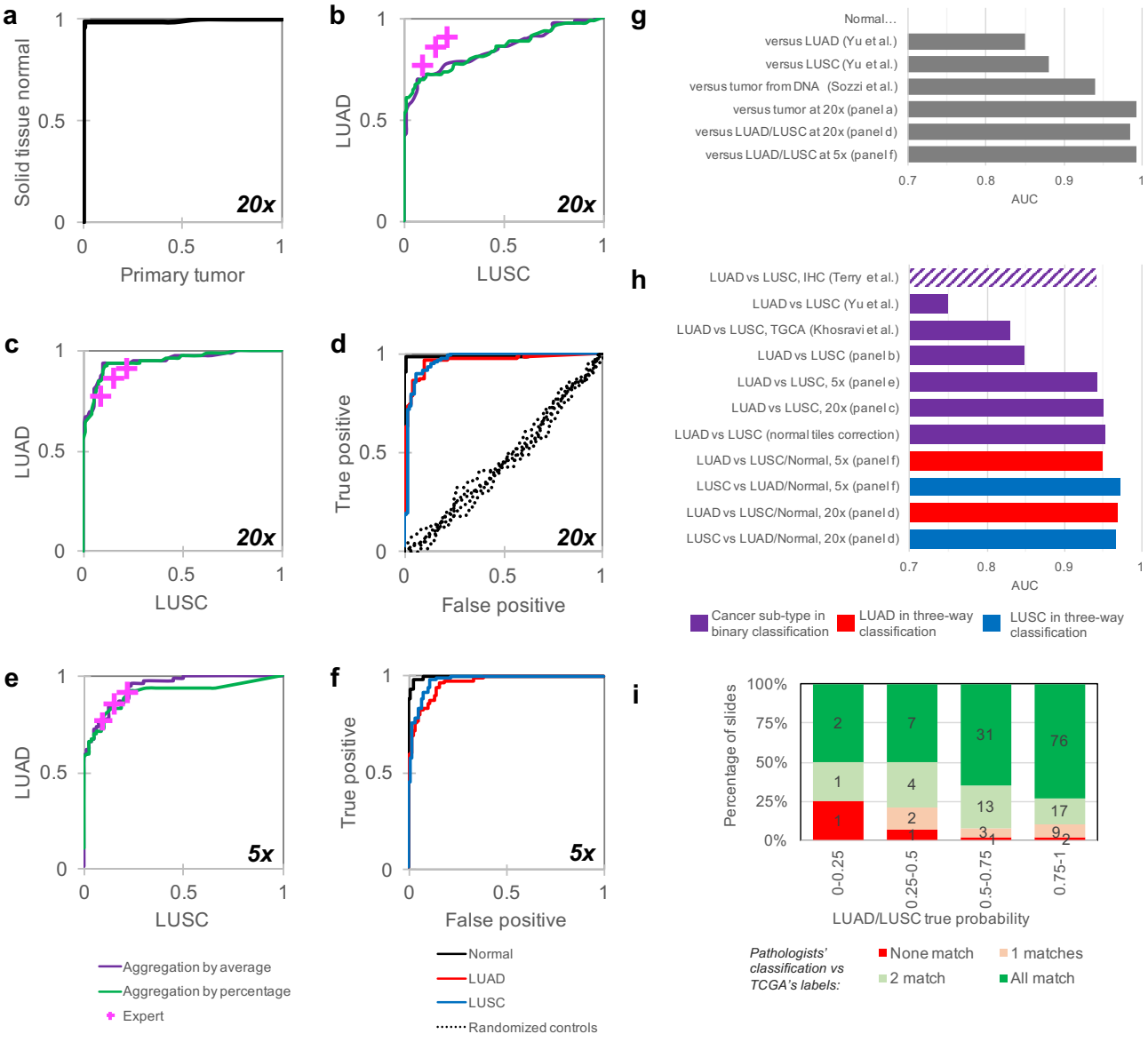
Supplementary Figure 9. Heatmaps for classification of Normal vs LUAD vs LUSC: (a) and **(b)** show typical examples of LUAD and LUSC whole-slide images. **(c)** and **(d)** show the corresponding heatmaps with probabilities of the winning class assigned to each tile such as: red for tiles classified as LUAD, blue for LUSC and grey for Normal. Training was done once

Supplementary Figure 10. Illustration of three-way classifier learned by deep-learning projected to 2 dimensions for visualization via the t-SNE algorithm using values of the last fully connected layer. (a) Scatterplots where each point represents a tile where the color is proportional to the probability generated by the deep learning network for each class. **(b)** Tile-embedded t-SNE representation with insets showing a random selection of tiles for different regions. n=149,790 tiles of 244 slides from 137 patients.

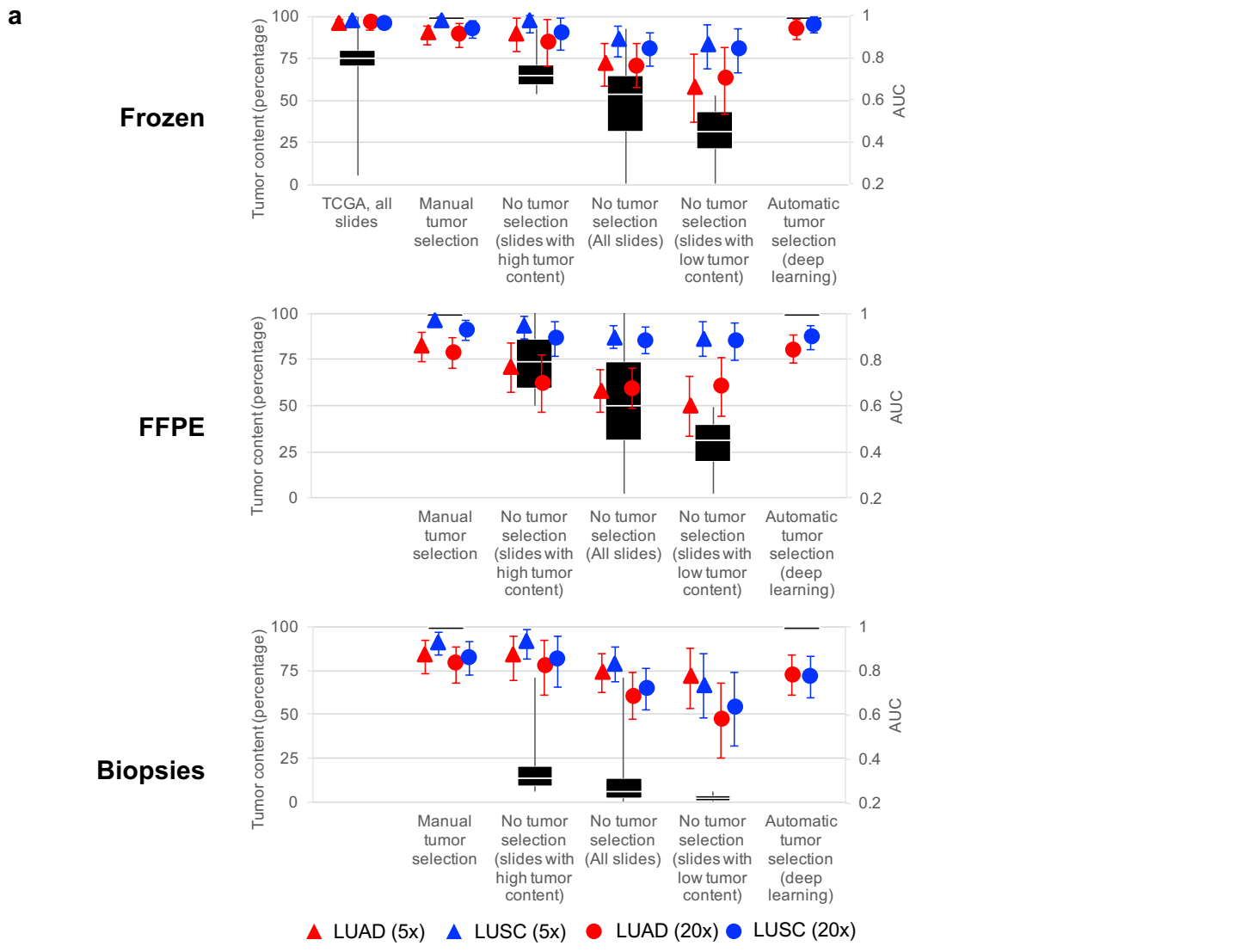
Supplementary Figure 1



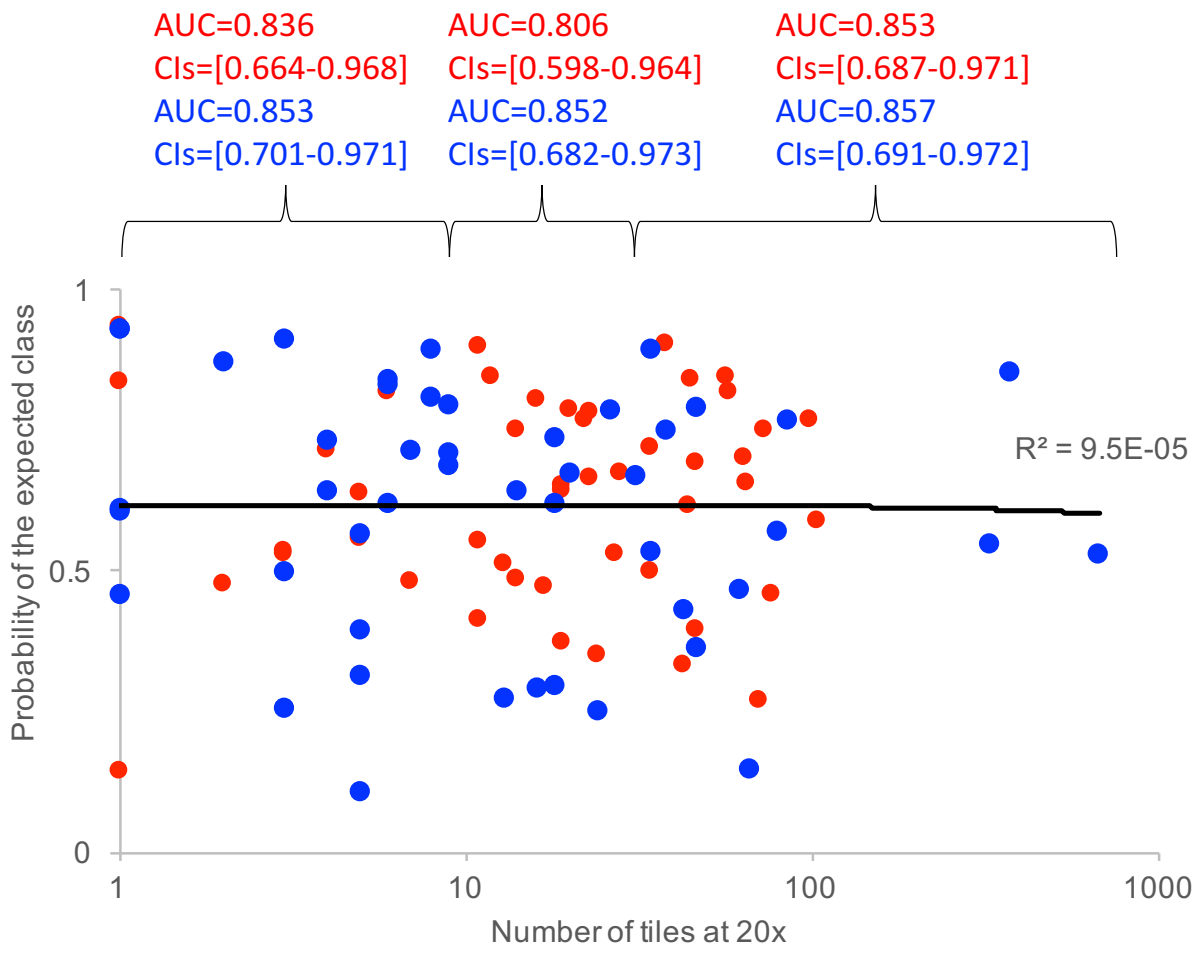
Supplementary Figure 2



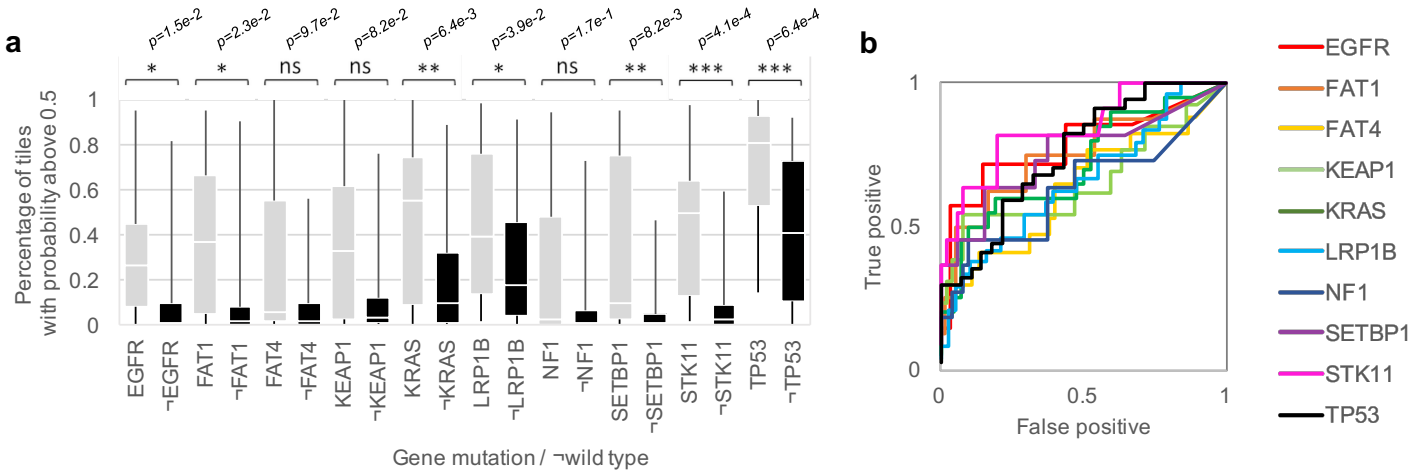
Supplementary Figure 3



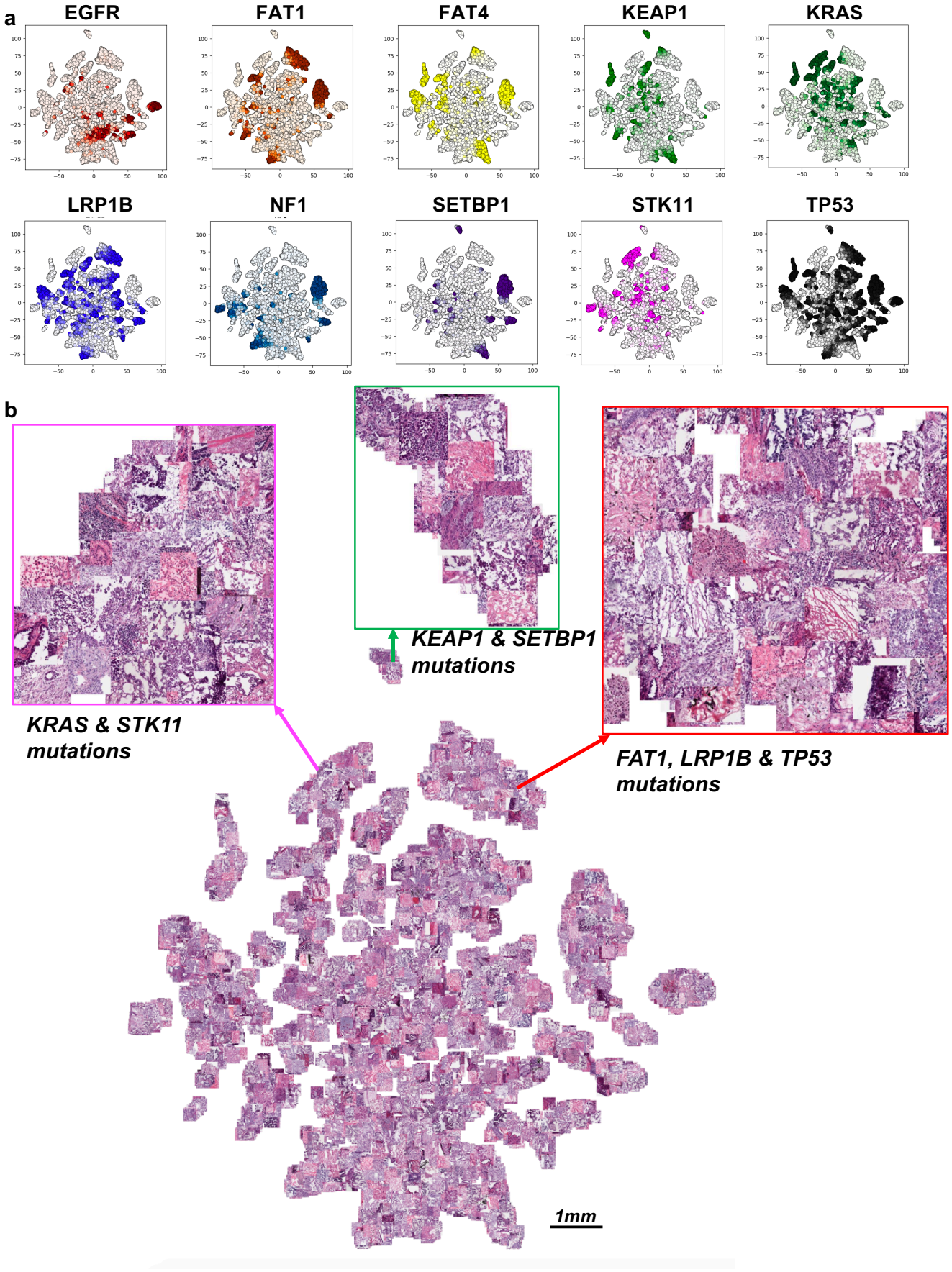
Supplementary Figure 4



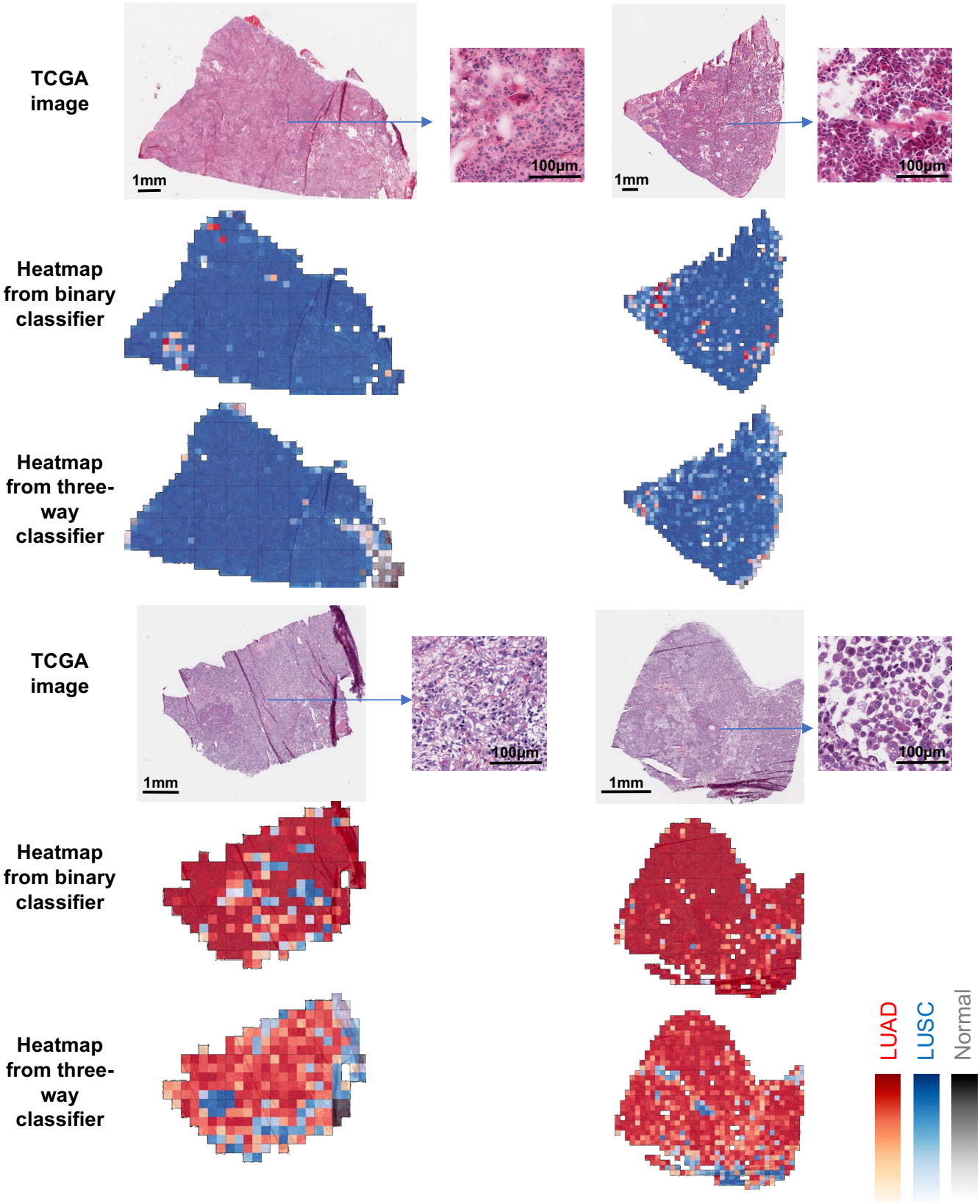
Supplementary Figure 5



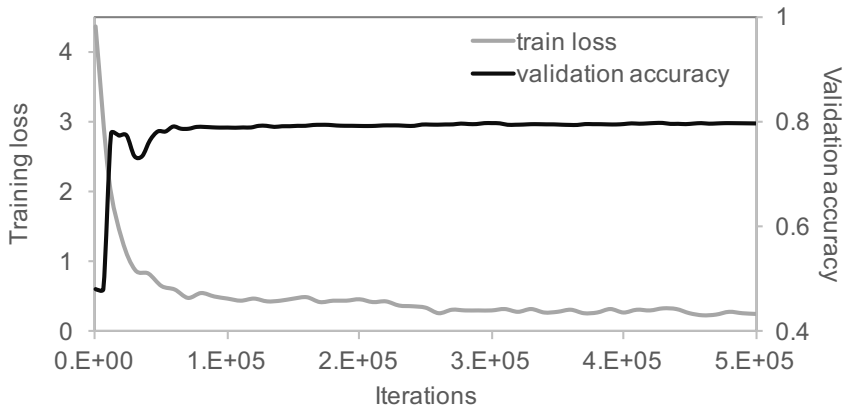
Supplementary Figure 6



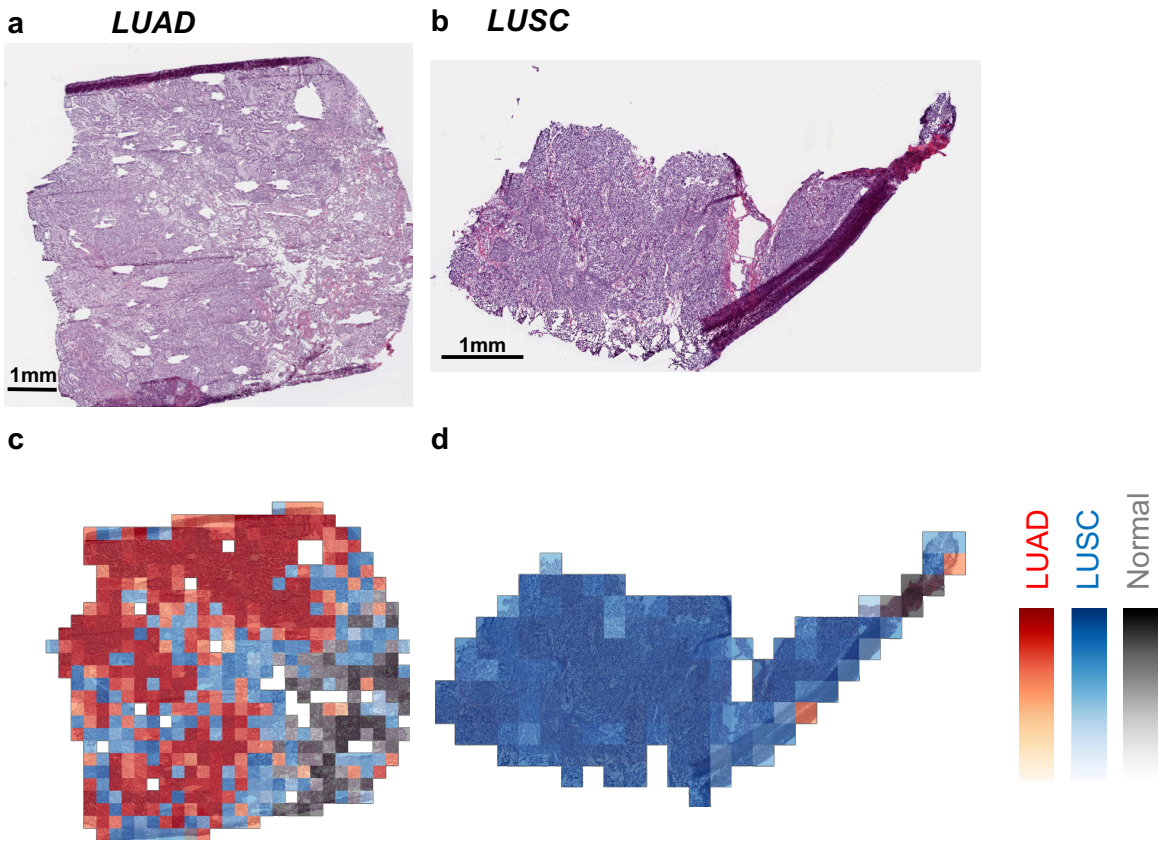
Supplementary Figure 7



Supplementary Figure 8



Supplementary Figure 9



Supplementary Figure 10

Supplementary Tables

Supplementary Table 1. Area Under the Curve (AUC) achieved by the different classifiers (with 95% CIs)

Classification	Information	AUC after aggregation by…	
		… average predicted probability	… percentage of positively classified tiles
Normal vs Tumor (20x tiles)	a) Inception v3, fully-trained	0.993 [0.974-1.000]	0.990 [0.969-1.000]
LUAD vs LUSC (20x tiles)	b) Inception v3, transfer learning	0.847 [0.782-0.906]	0.844 [0.777-0.904]
	c) Inception v3, fully-trained	0.950 [0.913-0.980]	0.947 [0.911-0.978]
	d) Same as (c) but aggregation done solely on tiles classified as “tumor” by A	0.952 [0.915-0.981]	0.949 [0.912-0.980]
LUAD vs LUSC (5x tiles)	Inception v3, fully-trained	**0.942 [0.907-0.971]**	0.906 [0.851-0.951]
3 classes. Normal vs LUAD vs LUSC at 20x	Normal	0.984 [0.947-1.000]	0.985 [0.953-1.000]
	LUAD	**0.969 [0.933-0.994]**	0.970 [0.937-0.993]
	LUSC	0.966 [0.935-0.990]	0.964 [0.932-0.989]
	Micro-average	0.970 [0.950-0.986]	0.969 [0.949-0.985]
	Macro-average	0.976 [0.949-0.993]	0.976 [0.950-0.993]
3 classes. Normal vs LUAD vs LUSC at 5x	Normal	**0.997 [0.993-0.998]**	0.988 [0.962-1.000]
	LUAD	0.965 [0.942-0.983]	0.938 [0.896-0.971]
	LUSC	**0.977 [0.960-0.991]**	0.964 [0.937-0.986]
	Micro-average	**0.980 [0.972-0.987]**	0.966 [0.948-0.980]
	Macro-average	**0.981 [0.968-0.991]**	0.964 [0.939-0.980]

n=244 slides for LUAD vs LUSC classifiers and n=170 slides for the others, all from 137 patients.

Supplementary Table 2. Diagnostic performance based on molecular profiling data.

Author	Method	Normal vs NSCLC	LUAD vs LUSC	Cohort size	Test on independent cohorts
Girard et al.[3]	62-gene microarray panel	Accuracy=86% Sensitivity=83% Specificity=100%	Accuracy=93% Sensitivity=95% Specificity=89%	1337 lung cancer; 191 healthy controls	yes
Charkiewicz et al.[4]	53-gene microarray panel		Accuracy=92.7% Sensitivity=100% Specificity=88%	152 LUSC and LUAD tissue	yes
Hou et al.[5]	5-gene microarray panel	Accuracy = 97%	Accuracy=84%	91 NSCLC; 65 adjacent normal lung tissue	yes
Wilkerson et al.[6]	57-gene microarray panel		Accuracy=78% (additional categories)	442 lung cancer with adjacent normal lung tissue	yes
Bhattacharjee et al.[7]	52-gene microarray panel	Accuracy = 85% (81% - 89%)	Accuracy=85% (additional categories)	186 lung cancer, 17 normal lung tissue	no
Cuezva et al.[8]	protein expression of a 3-gene panel from tissue samples	Accuracy=91.4% Sensitivity=97.3% (LUAD vs normal)		90 LUAD, 10 normal lung tissue	no
Amachika et al.[9]	RT-qPCR for NOK mRNA in peripheral blood	Sensitivity=80.5% Specificity=92.3%		41 lung cancer; 13 healthy controls	no
Du et al.[10]	RT-qPCR for STC1 mRNA levels in peripheral blood	AUC=0.969		65 lung cancer; 52 healthy controls	no
Cheng et al.[11]	RT-qPCR for LunX mRNA levels in peripheral blood	Sensitivity=92.9% Specificity=75.0%		44 lung lung cancer; 15 healthy controls	no
Sheu et al.[12]	RT-qPCR for 3-gene mRNA panel in peripheral blood	AUC=0.887		69 lung cancer; 100 healthy controls	no

Supplementary Table 3. Inter-pathologists and binary deep-learning method variability estimated with the Cohen's Kappa statistic.

	*Pathologist 1**	*Pathologist 2***	*Pathologist 3**	*Consensus between pathologists*	*Deep-learning*
TCGA	0.67 CIs=[0.56-8.78]	0.70 CIs=[0.60-0.81]	0.70 CIs=[0.59-0.81]	0.78 CIs=[0.69-0.88]	0.82 CIs=[0.74-0.91]
Pathologist 1		0.52 CIs=[0.39-0.65]	0.55 CIs=[0.42-0.67]	0.56 CIs=[0.44-0.69]	0.64 CIs=[0.52-0.75]
Pathologist 2			0.78 CIs=[0.69-0.88]	0.65 CIs=[0.54-0.77]	0.63 CIs=[0.52-0.75]
Pathologist 3				0.75 CIs=[65-0.86]	0.60 CIs=[0.48-0.72]
Consensus between 3 pathologists					0.77 CIs=[0.68-0.87]

n=170 slides from 137 patients
** thoracic pathologists; ** anatomic pathologist*

Supplementary Table 4. Comparison of LUAD/LUSC classifications (number and percentage of cases shown for each confusion matrix; for the LUAD/LUSC classifier, optimal threshold of 0.4/0.6 was selected).

TCGA dataset		Pathologist 1		Pathologist 2		Pathologist 3		LUAD/LUSC deep-learning classifier*	
		LUAD	LUSC	LUAD	LUSC	LUAD	LUSC	LUAD	LUSC
TCGA database	LUAD	72 (42%)	7 (4%)	67 (39%)	12 (7%)	62 (36%)	17 (10%)	**73 (43%)**	**6 (4%)**
	LUSC	21 (13%)	70 (41%)	13 (8%)	78 (46%)	8 (5%)	83 (49%)	**9 (5%)**	**82 (48%)**
Pathologist 1	LUAD			66 (39%)	27 (16%)	62 (36%)	31 (18%)	72 (42%)	21 (12%)
	LUSC			14 (8%)	63 (37%)	8 (5%)	69 (41%)	10 (6%)	67 (40%)
Pathologist 2	LUAD					66 (39%)	13 (8%)	65 (38%)	14 (8%)
	LUSC					5 (3%)	86 (50%)	17 (10%)	74 (44%)
Pathologist 3	LUAD							59 (35%)	11 (6%)
	LUSC							23 (14%)	77 (45%)

Supplementary Table 5. Dataset information for normal vs tumor classification (number of tiles / slides in each category).

	Training	Validation	Testing
Normal	132,185 / 332	28,403 / 53	28,741 / 74
Primary tumor	556,449 / 825	121,094 / 181	121,059 / 170

Supplementary Table 6. Dataset information for LUAD vs LUSC classification (number of tiles / slides in each category).

	Training	Validation	Testing
LUAD	255,975 / 403	55,721 / 85	55,210 / 79
LUSC	300,474 / 422	65,373 / 96	65,849 / 91

Supplementary Table 7. Gene included in the multi-output classification and the percentage of patients with LUAD in the database where the genes are mutated.

Gene mutated	TP53	LRP1B	KRAS	KEAP1	FAT4	STK11	EGFR	FAT1	NF1	SETBP1
%Patients	50	34	28	18	16	15	12	11	11	11

Supplementary References

1 Terry, J. *et al.* Optimal immunohistochemical markers for distinguishing lung adenocarcinomas from squamous cell carcinomas in small tumor samples. *The American journal of surgical pathology* **34**, 1805-1811 (2010).

2 Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O. & Hajirasouliha, I. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. *EBioMedicine*, doi:10.1016/j.ebiom.2017.12.026. (2017).

3 Girard, L. *et al.* An expression signature as an aid to the histologic classification of non–small cell lung cancer. *Clinical Cancer Research* **22**, 4880-4889 (2016).

4 Charkiewicz, R. *et al.* Gene Expression Signature Differentiates Histology But Not Progression Status of Early-Stage NSCLC. *Translational oncology* **10**, 450-458 (2017).

5 Hou, J. *et al.* Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PloS one* **5**, e10312 (2010).

6 Wilkerson, M. D. *et al.* Prediction of lung cancer histological types by RT-qPCR gene expression in FFPE specimens. *The Journal of Molecular Diagnostics* **15**, 485-497 (2013).

7 Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences* **98**, 13790-13795 (2001).

8 Cuezva, J. M. *et al.* The bioenergetic signature of lung adenocarcinomas is a molecular marker of cancer diagnosis and prognosis. *Carcinogenesis* **25**, 1157-1163 (2004).

9 Amachika, T., Kobayashi, D., Moriai, R., Tsuji, N. & Watanabe, N. Diagnostic relevance of overexpressed mRNA of novel oncogene with kinase-domain (NOK) in lung cancers. *Lung Cancer* **56**, 337-340 (2007).

10 Du, Y. Z., GU, X. H., Li, L. & Gao, F. The diagnostic value of circulating stanniocalcin-1 mRNA in non-small cell lung cancer. *Journal of surgical oncology* **104**, 836-840 (2011).

11 Cheng, M., Chen, Y., Yu, X., Tian, Z. & Wei, H. Diagnostic utility of LunX mRNA in peripheral blood and pleural fluid in patients with primary non-small cell lung cancer. *BMC cancer* **8**, 156 (2008).

12 Sheu, C.-C. *et al.* Combined detection of CEA, CK-19 and c-met mRNAs in peripheral blood: a highly sensitive panel for potential molecular diagnosis of non-small cell lung cancer. *Oncology* **70**, 203-211 (2006).