A metastasis prediction model in non-small cell lung cancer using GLCM_contrast and epithelial mesenchymal transition related genes

Byung-Chul Kim
Department of Nuclear Medicine, Korea Institute of Radiological and Medical Sciences, Seoul
Ilhan Lim
Department of Nuclear Medicine, Korea Institute of Radiological and Medical Sciences, Seoul
Byung Hyun Byun
Department of Nuclear Medicine, Korea Institute of Radiological and Medical Sciences, Seoul
Jingyu Kim
Radiological & Medico-Oncological Sciences, University of Science & Technology, Daejeon
Sang-Keun Woo (skwoo@kirams.re.kr)
Department of Nuclear Medicine, Korea Institute of Radiological and Medical Sciences, Seoul

Research Article

Keywords: Non-small cell lung cancer, Metastasis, RNA-seq, 18F-FDG PET, GLCM_contrast, Prediction model, radiogenomics

Posted Date: March 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-268933/v1

Abstract

Purpose: The aim of this study was to estimate a metastasis prediction model in non-small cell lung cancer by correlation next generation sequence gene expression level and fluorine-18-2-fluoro-2-deoxy-D-glucose positron emission tomography image features from non-small cell lung cancer patients.

Methods: RNA-sequencing data and ^{18}F-FDG PET images of 63 patients with NSCLC (29 metastasis and 34 non-metastasis) from The Cancer Imaging Archive and The Cancer Genome Atlas Program databases were used in a combined analysis. Weighted correlation network analysis was performed to identify gene groups were related metastasis. Module was selected with high module significance. Genes selection was performed by gene function related metastasis and high AUC (AUC > 0.6). A total of 47 image features were extracted from PET images as radiomics. The relationship of Gene expression and image features were calculated by using a hypergeometric distribution test with the Pearson correlation method. Metastasis prediction model was validated by random forest algorithm using image texture features related gene expression.

Results: 36 modules were identified by gene expression pattern with WGCNA assay. The modules had highest module significance was selected assay. 7 genes from selected module were identified to involve in the epithelial mesenchymal transition pathway that have important role in the cancer metastasis and had high AUC. Also, expression of these genes was related to quantitative of image feature (GLCM_contrast, -log10 P-value: 2.45~3.89). The AUC value (accuracy: 0.856 ± 0.06, AUC: 0.868 ± 0.05) was shown from the EMT-related gene and GLCM_contrast model and AUC value (accuracy: 0.842 ± 0.06, AUC: 0.838 ± 0.09) was shown from GLCM_contrast image texture model.

Conclusion: GLCM_contrast image texture feature shows relationship with EMT related gene expression. We developed a model for predicting metastasis of non-small cell lung cancer using ^{18}F-FDG PET image feature and evaluated its accuracy.

Introduction

Non-small cell lung cancer (NSCLC) has a high incidence among cancers that can occur in modern people with large molecular heterogeneity in tissues[1,2]. Its molecular heterogeneity was shown to be different between patients and intratumor and intertumor regions[3]. Intratumor heterogeneity is known to be linked to the development of primary tumors and metastases[4]. It is possible to diagnose cancer by analyzing intracellular gene expression events and finding a suitable treatment method for each cancer[5]. Many studies have been conducted to search for methods to diagnose cancers having different genotypes and to find a treatment for each cancer: image features that analyze phenotypes based on genotype, next generation sequencing (NGS) for large-scale gene analysis, and radiogenomics that uses fluorine-18-2-fluoro-2-deoxy-D-glucose positron emission tomography (^{18}F-FDG PET) image features and NGS in combination.

NGS is a high-throughput sequencing analysis method that is capable of accurately quantifying large amounts of gene information compared to conventional gene analysis methods[6]. In the past, gene expression was characterized one by one with electrophoresis after PCR, a time-consuming, expensive procedure and limitation of sample amounts. Recently, advances in NGS technology have made it possible to analyze total RNA in single cells. Studies of genes involved in NSCLC metastasis have also been conducted using NGS, and genes that play important roles in metastasis, such as EFGR, had been identified[7]. However, this method has some disadvantages: time-consuming sequencing, painful invasive biopsies, and identification the genes from the sampled tissue but not necessarily from the entire tissue[8]. The classical image technique uses radiation to image the affected area without causing pain to the patient, grasping the overall characteristics of the affected area, and has the advantage of quick analysis[9] but only showing the cancer phenotype. Radiogenomics is a study that combines image feature technology for analyzing images and NGS technology for mass analysis of genes, revealing the relationship between expression of specific genes related to cancer and image features present. By combining the two analysis methods, diagnosis and prediction of cancer without any invasive method is possible[10].

^{18}F-FDG PET/CT has the advantage of evaluating metabolic processes in cancer. It is an advanced technique compared to CT, a traditional imaging technique: ^{18}F-FDG is absorbed during glucose metabolism, and it is possible to estimate glucose metabolism by imaging FDG remaining in the cell. Depending on the degree of cancer progression, glucose uptake and FDG concentration remaining in the cells are different. Because the residual FDG concentration in the initial cancer is low and increases as the cancer progresses, the degree of cancer progression can be evaluated through FDG imaging. This method is also suitable for evaluating cancer metastasis[11]. ^{18}F-FDG PET/CT imaging was used to predict the chemotherapy response after treatment with an anticancer drug in NSCLC[12]. In other studies, ^{18}F-FDG PET/CT imaging can be used for prognosticating survival in NSCLC by analyzing image features[13].

The following is a case study of NSCLC that recently utilized radiogenomics. Research on gene expression specific to NSCLC has already been conducted, and it is well known that the *EGFR* gene plays an important role in metastasis when mutation occurs[14]. A recent study has shown that ^{18}F-FDG PET/CT image features are correlated with *EGFR* mutation status in NSCLC[15]. In this study, patient DNA was collected to distinguish patients with *EGFR* mutations, and image features of CT images were analyzed to determine whether the features (SUVmax, SUVmean, and SUVpeak) were related to the *EGFR* mutation. A metastasis prediction model was estimated with these results. In another study, mRNA extracted from NSCLC tissues was analyzed by NGS to find metagenes, and image features from CT images were used for analysis by searching for correlations between NGS and CT image features[16]. The relationship and action of the expressed metagenes and image features for cancer cell proliferation were studied. Epithelial mesenchymal transition (EMT) plays a most important role in cancer metastasis. In NSCLC cells, activation of EMT induces cell migration, proliferation, and invasion[17].

In this study, we estimated correlation between the expression of genes in metastasis of NSCLC and the quantitative ^{18}F-FDG PET image texture features. The NSCLC metastasis prediction model was developed by image texture features have relation with gene expression.

Results

In this study, ^{18}F-FDG PET data and RNA-sequencing data from 63 patients with NSCLC were used for analysis. The average age of the patients was 67.5 years, and the ratio of men and women was approximately 8:2. (Table 1). The process of development of the relationship between the RNA-sequencing data and ^{18}F-FDG PET image features are schematically described in Fig. 1.

Table 1
List of clinical data for NSCLC patients. The patient's age, gender, type of cancer, smoking status, EGFR mutation, KRAS mutation and cancer progression are shown.

Characteristic	Result	Rate
Average age	67.5	
Sex		
male	54	79%
female	14	21%
Histology		
Adenocarcinoma	48	71%
NSCLC NOS (not otherwise specified)	3	4%
Squamous cell carcinoma	17	25%
Smoking status		
Current	10	15%
Former	45	66%
Non-smoker	13	19%
EGFR mutation		
wild	48	71%
mutation	9	13%
unknown	11	16%
KRAS mutation		
wild	43	63%
mutation	14	21%
unknown	11	16%
Pathological M stage		
M0	63	93%
M1a	1	1%
M1b	4	6%
Pathological N stage		
N0	51	75%

Characteristic	Result	Rate
N1	7	10%
N2	10	15%

Gene modulation and hub gene assay

To search for hub genes, have important role in the metastasis, WGCNA was used first to construct a gene module with a similar expression pattern, and a network analysis was performed to search for hub genes. A total of 36 gene modules were obtained (Fig. 2). The module with the highest significance in the metastasis group was selected. To confirm the function of the gene module, GO term analysis was performed. A total of 7 genes were selected as EMT-related genes with high GS scores (GS > 0.8) and high AUC value (AUC > 0.6).

Hub gene and image feature associations

The analysis was performed using 47 radiomics and 7 EMT-related genes. Results regarding the relationship between expression levels of the factors were obtained. Among the relationships between image features and gene expression levels, the top 50 genes were selected to show the total relationship in the highest order and visualized as a heatmap (Fig. 3). The results show one image feature (GLCM_contrast) that was expressed deeply in relation (P-value < 0.05) to the expression of seven genes (*NME1.NME2, LST1, KAT7, BMX, CLIC1, KANSL2*, and *UFL1*) (Table. 2).

Table 2
List of the seven genes that are related to image features in NSCLC metastasis. P-value was calculated using the hypergeometric distribution method. Genes were selected with the smaller P-values and related metastasis and EMT function (value was normalized by -log10).

GENE	SUVmax	SUVpeak	TLG	Entropy_log10	GLCM_contrast
BMX	0.96	0.72	0.08	0.49	2.81
NME1.NME2	1.07	0.69	0.16	0.32	3.89
LST1	0.92	0.48	0.25	0.05	3.17
KAT7	0.84	0.44	0.26	0.12	2.81
CLIC1	0.82	0.5	0.19	0.29	2.61
TAP2	0.67	0.34	0.25	0.04	2.52
PSMB9	0.57	0.29	0.24	0.07	2.45

SUVmax: Maximum Standard Uptake Value, SUVpeak: Peak Standardized Uptake Value, TLG: Total lesion glycolysis, GLCM_contrast: Gray-Level Co-occurrence Matrix contrast.

Estimation of the prediction model

Genetic expression levels and features extracted from PET/CT images were used to create a model for predicting metastasis of NSCLC. The EMT-related gene (7) model precision, recall, AUC, and accuracy score were 0.860 ± 0.16, 0.642 ± 0.2, 0.766 ± 0.09, and 0.799 ± 0.06, respectively. The histogram first order (15) model precision, recall, AUC, and accuracy were 0.77 ± 0.14, 0.713 ± 0.04, 0.713 ± 0.04, and 0.794 ± 0.07, respectively. The texture (32) mod el precision, recall, AUC, and accuracy score were 0.80 ± 0.13, 0.642 ± 0.18, 0.766 ± 0.08, and 0.805 ± 0.07, respectively. The EMT gene (7) and radiomics (47) model precision, recall, AUC, and accuracy score were 0.840 ± 0.10, 0.814 ± 0.13, 0.856 ± 0.06, and 0.868 ± 0.05, respectively. Finally, the GLCM_contrast model precision, recall, AUC, and accuracy score were 0.759 ± 0.04, 0.828 ± 0.21 0.838 ± 0.09, and 0.842 ± 0.06, respectively (Table 3).

Table 3
Precision, recall, AUC, and accuracy values of predictive models created using meta-related genes and image extraction factors expressed using the random forest algorithm.

Random Forest (N = 63) Test	EMT-related gene (7)	Histogram_first order (15)	Texture (32)	EMT gene (7) & Radiomics (47)	GLCM_contrast
Precision	0.860 ± 0.16	0.77 ± 0.14	0.80 ± 0.13	0.840 ± 0.10	0.759 ± 0.04
Recall	0.642 ± 0.2	0.713 ± 0.04	0.642 ± 0.18	0.814 ± 0.13	0.828 ± 0.21
AUC	0.766 ± 0.09	0.713 ± 0.04	0.766 ± 0.08	0.856 ± 0.06	0.838 ± 0.09
Accuracy	0.799 ± 0.06	0.794 ± 0.07	0.805 ± 0.07	0.868 ± 0.05	0.842 ± 0.06

Discussion

In this study, RNA-sequencing and ^{18}F-FDG PET/CT images of non-small cell lung cancer patients were used to search for gene groups expression related to non-small cell lung cancer metastasis and imaging features related to the expression of gene groups. The gene group involved in metastasis has an EMT function known to be induce metastasis. It was observed that one of the imaging features, GLCM_contrast, was expressed in relation to the expression of EMT function. This is a clue that can predict the metastasis of non-small cell lung cancer through the analysis of imaging features.

Recently, a combination of two analysis methods, NGS and PET CT imaging, has been studied to overcome the limitations of each. The prediction and diagnosis of lung cancer metastasis is related to serious problems for patients because lung cancer shows no symptoms or pain until the late stages and

has spread to other organs, with a high probability of being at a late stage when diagnosed[18]. Development of a composite diagnosis method for genes and images has the advantage of being noninvasive[19] and fast compared to existing diagnostic methods, and is also capable of diagnosing overall cancer. In terms of genetic analysis, two methods were used to reduce the number of genes used for analysis. The first was to select genes with significant differences between the two groups using a t-test[20] and the second was to use the hub gene assay to select genes with the desired functions. A t-test was performed for more efficient analysis to remove genes with low P-values using mathematical calculations[20]. Genes were divided into modules according to the gene expression pattern through WGCNA analysis, and each module was assigned a significant value according to its contribution to the module. One module selected had the highest gene significance. A total of 7 genes were identified as EMT-related genes from the selected module (GS > 0.8 and AUC > 0.6). The hypergeometric distribution method[21] was used to identify which EMT-related genes are associated with image features extracted from the genetics. The relevance of image features and genes was calculated by P-value and was listed from low values. P-values greater than 0.05 were excluded. Gene expressed levels were compared in patients with and without metastasis of each gene to identify differences in both conditions. A total of seven genes were identified as having a high relationship with one radiomics: GLCM_contrast. The seven identified genes, *NME1.NME2, LST1, KAT7, BMX, CLIC1, TAP2* and *PSMB9* are known to be involved in EMT. Bone marrow X-linked kinase (*BMX*) has been reported to be involved in EMT, such as cell growth, transformation, migration, survival, apoptosis, and tumorigenicity[22–25]. Nucleoside diphosphate kinase A (*NME1*) and nucleoside diphosphate kinase B (*NME2*) form *the complex unit NM23 (NME1.NME2)* and have the nucleoside diphosphate kinase activity, which catalyzes the phosphorylation of nucleoside diphosphates to the corresponding nucleoside triphosphates. *NME1.NME2* is the first metastasis suppressor in lung cancer. A decrease in *NME1.NME2* increases cancer metastasis[26]. The function or mechanism of leukocyte-specific transcript 1 protein (*LST1*) has not been well studied, but high expression of *LST1* in metastasized lung cancer has been reported[27]. Chloride intracellular channel 1 (*CLIC1*) has the ability of the antiangiogenic peptide *CLT1* on proliferating endothelial cells [28]. *CLIC1* is mainly overexpressed in the tumor vasculature, and overexpression has been observed in breast, lung, and liver cancer patients[29,30]. *CLIC1* has been shown to promote regular invasion and proliferation of tumor and endothelial cells, but the underlying mechanism is unclear[31]. Transporter associated with antigen processing 1 (*TPA1*) regulates *WISP2*, which can affect TGF-b signaling. TGB-b signaling is one of the most important roles of EMT in breast cancer[32]. Proteasome subunit beta type-9 (*PSMB9)* is co-expressed with *RARRES3* and is a well-known metastasis suppressor in breast cancer cells[33].

EMT is an evolutionarily conserved process in which cells undergo the conversion from epithelial cells to mesenchymal cells. EMT was found in a study on the development of embryo stem cells. EMT is a major activity during embryo stem cell development, gastrulation, neural nests, and development of the heart and other tissues and organs[34]. Recent studies have shown that EMT is also implicated in cancer progression and metastasis. Studies on breast cancer metastasis suggest that EMT is also involved in the acquisition of characteristics of cancer stem-like cells (CSCs)[35]. CSCs are cancer cells that have the

characteristics of embryonic stem cells of self-renewal, regeneration, and differentiation to diverse types of cancer cells. CSCs are thought to be crucial for the initiation and maintenance of tumors as well as their metastasis[36]. Many studies using NGS for NSCLC have been performed because of the ability to determine the molecular characteristics of the cancer state for diagnosis or treatment[37]. NGS is a technology that can analyze gene expression levels at a fast and large scale compared to conventional gene analysis methods. However, a limitation is biopsies are need for sampling, which is not available all cancer cases because of cancer location[38]. Another limitation is representativeness[39]. Cancer tissues have a high heterogeneity; biopsy samples cannot represent all cancer regions. To overcome this limitation, image features had to be introduced into the analysis.

PET/CT images have become a popular research topic for the diagnosis of NSCLC in recent days. Features extracted from the images were used for analysis. Each feature is represented by a call status such as cell shape, cell surface texture, and cell density. These features were digitized for cancer analysis using a mathematical method[40]. Many studies have been published on the possibility of tumor classification by analysis of PET/CT texture features with ^{18}F-FDG PET/CT. The development of ^{18}FDG PET/CT imaging technology and techniques for analyzing digitized features from images have information on cell activity [41]. A limitation of the PET/CT imaging method is the lack of information from image analysis. Imaging factors of cells or tissues can only provide information on cell morphology and the texture of the cell surface. Some cancers with a unique phenotype can be diagnosed, but accurate diagnosis is not possible for most cancers using a phenotype because it cannot represent the genotype[42].

GLCM_contrast is a feature from image feature analysis. It is considered a texture feature from the LIFEx image analysis tool. In general, features such as SUVmax, SUVpeak, TLG, and ENTROPY were used for radiogenomics analysis for cancer prediction or cancer metastasis prediction[43]. However, in this study, the correlation (P-value) of SUVmax, SUVpeak, TLG, and ENTROPY was lower than that of GLCM_contrast. This result shows that new factors such as GLCM_contrast can be used to develop a model for predicting metastasis of NSCLC using radiogenomics. One of the limitations of our study that although we provide the evidence that EMT related gene has relation to GLCM_contrast in NSCLC but do not provide mechanistic studies. While this was not the goal of this study, future investigations could be directed toward to uncover the mechanisms of operation of genes that play an important role in NSCLC metastasis, and to elucidate the correlation of expression of imaging features. Large scale of follow-up studies with molecular mechanism of metastasis in NSCLC could strengthen the study and further confirm and extend our findings. In addition, it was possible to search for radiomics related to EMT genes in this study and it will be possible to search for imaging biomarkers for diagnosis and prognosis by analyzing genetic functions related to other cancers or diseases.

Conclusion

In this study, we confirmed through RNA-sequencing analysis that the group genes involved in the NSCLC metastasis were related to EMT function. The expression of these group genes was related to the image texture feature like GLCM_Contrast. It was confirmed that the accuracy of the prediction model developed using two factor that was consist of the EMT-related group genes and GLCM_Contrast and GLCM_Contrast only by the Random Forest algorithm was high. These results reveal the possibility of a prediction model using image text features related to gene expression in NSCLC metastasis.

Material And Methods

NSCLC NGS data processing

All experiments were performed according to institutional guidelines and approved by the Korea institute of Radiological & Medical Science institutional (IRB, e-IRB number: kirams 2021-03-001).

informed consent was waived by IRB.
RNA-sequencing data, clinical data of patients, and ^{18}F-FDG PET images were acquired from the TCIA/TCGA database (NGS data accession number: GSE103584, PET image data: http://doi.org/10.7937/K9/TCIA.2017.7hs46erv - DOI). Patient data were classified in a binary manner between metastasis (n = 29) and non-metastasis (n = 34) groups based on clinical data and ^{18}F-FDG PET images. The classification in the metastasis and non-metastasis models was performed with reference to clinical data from TCGA. Patients in the N1 and N2 stages were placed in the metastasis group, and those in the N0 stage were placed in non-metastasis group. Patient information is summarized in Table 1. Acquired data were normalized by FRKM. The genes with zero FRKM values from all the samples were trimmed for fast analysis[44]. For differentially expressed gene (DEG) analysis, the Deseq2 tool of the R packaged was used[45]. Input data groups followed the metastasis and non-metastasis groups. DEG analysis results were visualized in volcano plots by ggplot in R [46].

Weighted gene co-expression networks and modules associated with clinical traits

To analyze the correlation between expressed genes and features extracted from images, gene selection was conducted at first. A total of 22,125 genes were analyzed by DEG and the selected only those genes with significant differences[47]. To obtain the gene module with the greatest influence on determining metastasis, WGCNA analysis was performed[48]. The genes were separated into several modules using the WGCNA tool in the R package. A soft threshold for network construction was selected for gene clustering. In the soft threshold, the adjacency matrix forms a continuous range of values between 0 and 1. The constructed network conforms to the power-law distribution and is closer to a real biological network state. A scale-free network was constructed using the blockwise module function, followed by module partition analysis to identify gene co-expression modules, which grouped genes with similar expression patterns. The modules were defined by cutting the clustering tree into branches using a dynamic tree cutting algorithm and assigned to different colors for visualization[49]. The module eigengene (ME) of each module was calculated. ME represents the expression level for each module. The correlation between ME and clinical traits in each module was calculated. Finally, the gene significance (GS) that

represented the correlation between genes and samples was further calculated. Genes from selected modules with a GS value of 0.8 or more and a P-value of 0.05 or less were selected[50]. Each gene's AUC value was calculated, and genes have high AUC values (AUC > 0.6) were selected for correlation assays.

Functional and pathway enrichment analyses of selected modules

Genes from selected modules were used for functional analysis. DAVID 6.8[51] software was used for the GO term, biological process (BP), molecular function (MF), and cellular component (CC)[52] in each module. A P-value < 0.05 was selected as the threshold for the identification of significant GO terms and pathways. Go terms were visualized using the revigo web tool[53].

^{18}F-FDG PET imaging

Tumor volumes were segmented and radiomics features in the defined tumors were subsequently extracted using the Local Image Features Extraction (LIFEx) version 4.0 software package[54]. The tumor region was drawn using a semi-automated segmentation method with a threshold SUV of 2.0 based on our previous report[55] in three-dimensional (3D) images. In segmented tumors, SUVmax, SUVmean, SUVpeak, metabolic tumor volume (MTV), total lesion glycolysis (TLG), and features from shape and histogram were calculated as the first order features. For texture feature calculation, the number of intensity levels was resampled using 64 discrete values between zero and 20 SUVs, corresponding to a sampling bin width of 0.3125 SUV[41,56]. Spatial resampling was 4.1 mm (X-direction), 4.1 mm (Y-direction), and 2.5 mm (Z-direction) in Cartesian coordinates [14]. Texture features were assessed using four texture matrices: co-occurrence matrix (CM), gray-level run length matrix (GRLM), gray-level zone length matrix (GZLM), and neighborhood gray-level different matrix (NGLDM). The CM was calculated in 13 directions with one voxel distance relationship between neighboring voxels, and each texture feature calculated from this matrix was the average of the features over the 13 directions in space (X, Y, Z). The GRLM was also calculated for 13 directions via a similar method, whereas the GZLM was computed directly in 3D. The NGLDM was computed from the difference in gray levels between one voxel and its 26 neighbors in 3D, and each texture feature was calculated from this matrix[57]. A total of 47 features were extracted from the PET image data.

Hub gene and image feature correlation

A total of 47 image features and 7genes were used to estimate the relationship between all table factors, which was calculated using a hypergeometric distribution test with the Pearson correlation method. The hypergeometric P-value was calculated using the equation p = (kCx) ((n − k)C(n − x))/NCn, where N is the number of total genes in the genome, k is the number of expression values identified in gene expression, n is the expression value of features identified in the images, x is the number of overlapping genes, and kCx is the number of possible genes and features from image combinations[58]. The image features and genes for estimation of the metastasis prediction model were selected by the P-value of correlation (P-value < 0.05). The selected image features were compared with image values that are generally used for validation of radiogenomics.

Evaluation of the metastasis prediction model

To predict the patient's outcome in terms of metastasis, we used a machine learning approach[59] called random forest (RF)[60]. The machine learning prediction model was used to evaluate the accuracy, precision, and recall score using test data. Prediction was performed 10 times to obtain an average value[61]. A radiomics (47) only prediction model, an EMT-related gene (7) model, a histogram first order (15) model, a texture (32) model, an EMT-related gene (7) and radiomics (47) model, and a GLCM_contrast model was used for estimation of the machine learning method using the random forest algorithm.

Declarations

Authors' contributions

BC design of this experiment, acquire and analysis of patient data with RNA-sequencing analysis tools and write this article. IH and BH advised about the Classification of clinical data for metastasis and non-metastasis and trimming of data. JK analysis of image features and machine learning analysis. SK supervise the total process as a corresponding author. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. 2020M2D9A1094070, No. 2019M2D2A1A02057204)

References

1. Chen, Z., Fillmore, C. M., Hammerman, P. S., Kim, C. F. & Wong, K. K. J. N. R. C. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat Rev Cancer.* **14**, 535–546 (2014).
2. Cai, M. *et al.* Adam17, a target of Mir-326, promotes Emt-induced cells invasion in lung adenocarcinoma. *Cell Physiol Biochem.* **36**, 1175–1185 (2015).
3. Marino, F. Z. *et al.* Molecular heterogeneity in lung cancer: from mechanisms of origin to clinical implications. *Int J Med Sci.* **16**, 981 (2019).
4. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. J. N. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature.* **501**, 338–345 (2013).
5. Kamel, H. F. M. & Al-Amodi, H. S. A. B. Exploitation of gene expression and cancer biomarkers in paving the path to era of personalized medicine. *Genomics, proteomics & bioinformatics.* **15**, 220–235 (2017).
6. Ambardar, S., Gupta, R., Trakroo, D., Lal, R. & Vakhlu, J. High throughput sequencing: an overview of sequencing chemistry. *Indian journal of microbiology.* **56**, 394–404 (2016).

7. Cardnell, R. *et al.* Proteomic Markers of DNA Repair and PI3K Pathway Activation Predict Response to the PARP Inhibitor BMN 673 in Small Cell Lung Cancer. *Clin Cancer Res.* **9** (22), 6322–6328 (2013).
8. Ari, Å. & Arikan, M. in Plant omics: Trends and applications 109–135(Springer, 2016).
9. O'Brien, B. & van der Putten, W. Quantification of risk-benefit in interventional radiology. *Radiation protection dosimetry.* **129**, 59–62 (2008).
10. Sala, E. *et al.* Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clinical radiology.* **72**, 3–10 (2017).
11. Németh, Z., Boér, K., Borbély, K. J. P. & Research, O. Advantages of 18 F FDG-PET/CT over Conventional Staging for Sarcoma Patients. *Pathol Oncol Res.* **25**, 131–136 (2019).
12. Lee, D. H. *et al.* Early prediction of response to first-line therapy using integrated 18F-FDG PET/CT for patients with advanced/metastatic non-small cell lung cancer. *J Thorac Oncol.* **4**, 816–821 (2009).
13. Cung, C., Wong, W., Martin, R. & Shon, I. H. J. J. o. N. M. Radiomic analysis of pre-treatment FDG PET prognosticates survival in non-small cell lung cancer. *Journal of Nuclear Medicine.* **61**, 277–277 (2020).
14. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science.* **304**, 1497–1500 (2004).
15. Xu, W. J. F. i. o. Predictive power of a radiomic signature based on 18F-FDG PET/CT images for EGFR mutational status in NSCLC. *Front Oncol.* **9**, 1062 (2019).
16. Zhou, M. *et al.* Non–small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. *Radiology.* **286**, 307–315 (2018).
17. Liao, T. T. & Yang, M. H. J. C. Hybrid Epithelial/Mesenchymal State in Cancer Metastasis: Clinical Significance and Regulatory Mechanisms. *Cells.* **9**, 623 (2020).
18. Mulvenna, P. *et al.* Dexamethasone and supportive care with or without whole brain radiotherapy in treating patients with non-small cell lung cancer with brain metastases unsuitable for resection or stereotactic radiotherapy (QUARTZ): results from a phase 3, non-inferiority, randomised trial. *Lancet.* **388**, 2004–2014 (2016).
19. Jansen, R. W. *et al.* Non-invasive tumor genotyping using radiogenomic biomarkers, a systematic review and oncology-wide pathway analysis. *Oncotarget.* **9**, 20134 (2018).
20. Lai, Y. Differential expression analysis of Digital Gene Expression data: RNA-tag filtering, comparison of t-type tests and their genome-wide co-expression based adjustments. *International journal of bioinformatics research and applications.* **6**, 353–365 (2010).
21. Yamamoto, S. *et al.* Radiogenomic analysis demonstrates associations between 18F-fluoro-2-deoxyglucose PET, prognosis, and epithelial-mesenchymal transition in non–small cell lung cancer. *Radiology.* **280**, 261–270 (2016).
22. Fujisawa, Y. *et al.* Ligand-independent activation of the arylhydrocarbon receptor by ETK (Bmx) tyrosine kinase helps MCF10AT1 breast cancer cells to survive in an apoptosis-inducing

environment. *Biological chemistry.* **392**, 897–908 (2011).

23. Guo, S. *et al.* Tyrosine kinase ETK/BMX is up-regulated in bladder cancer and predicts poor prognosis in patients with cystectomy. *PLoS One.* **6**, e17778 (2011).
24. Holopainen, T. *et al.* Deletion of the endothelial Bmx tyrosine kinase decreases tumor angiogenesis and growth. *Cancer research.* **72**, 3512–3521 (2012).
25. Fox, J. L. & Storey, A. BMX negatively regulates BAK function, thereby increasing apoptotic resistance to chemotherapeutic drugs. *Cancer research.* **75**, 1345–1355 (2015).
26. Zhao, R. *et al.* nm23-H1 is a negative regulator of TGF-β1-dependent induction of epithelial–mesenchymal transition. *Exp. Cell Res.* **319**, 740–749 (2013).
27. Liu, X. *et al.* Synaptotagmin 7 in twist-related protein 1-mediated epithelial–Mesenchymal transition of non-small cell lung cancer. *EBioMedicine.* **46**, 42–53 (2019).
28. Knowles, L. M., Malik, G., Hood, B. L., Conrads, T. P. & Pilch, J. CLT1 targets angiogenic endothelium through CLIC1 and fibronectin. *Angiogenesis.* **15**, 115–129 (2012).
29. Hill, J. J. *et al.* Identification of vascular breast tumor markers by laser capture microdissection and label-free lc-ms. *Journal of proteome research.* **10**, 2479–2493 (2011).
30. Li, R. K. *et al.* Chloride intracellular channel 1 is an important factor in the lymphatic metastasis of hepatocarcinoma. *Biomed. Pharmacother.* **66**, 167–172 (2012).
31. Tung, J. J. & Kitajewski, J. Chloride intracellular channel 1 functions in endothelial cell growth and migration. *Journal of angiogenesis research.* **2**, 23 (2010).
32. Akalay, I. *et al.* Targeting WNT1-inducible signaling pathway protein 2 alters human breast cancer cell susceptibility to specific lysis through regulation of KLF-4 and miR-7 expression. *Oncogene.* **34**, 2261–2271 (2015).
33. Anderson, A. M. *et al.* The metastasis suppressor RARRES3 as an endogenous inhibitor of the immunoproteasome expression in breast cancer cells. *Sci. Rep.* **7**, 1–13 (2017).
34. Thiery, J. P., Acloque, H., Huang, R. Y. & Nieto, M. A. J. c. Epithelial-mesenchymal transitions in development and disease. *Cell.* **139**, 871–890 (2009).
35. Morel, A. P. *et al.* Generation of breast cancer stem cells through epithelial-mesenchymal transition. *PLoS One.* **3**, e2888 (2008).
36. Charafe-Jauffret, E. *et al.* Breast cancer cell lines contain functional cancer stem cells with metastatic capacity and a distinct molecular signature. *Cancer Res.* **69**, 1302–1313 (2009).
37. Inamura, K. J. C. Diagnostic and therapeutic potential of microRNAs in lung cancer. cancers 9, 49(2017).
38. Puech, P. *et al.* Prostate cancer diagnosis: multiparametric MR-targeted biopsy with cognitive and transrectal US–MR fusion guidance versus systematic biopsy—prospective multicenter study. *Radiology.* **268**, 461–469 (2013).
39. Pirrelli, M. *et al.* Are biopsy specimens predictive of HER2 status in gastric cancer patients? *Digestive Diseases and Sciences.* **58**, 397–404 (2013).

40. Thawani, R. *et al.* Radiomics and radiogenomics in lung cancer: a review for the clinician. *Lung Cancer.* **115**, 34–41 (2018).
41. Orlhac, F., Soussan, M., Chouahnia, K., Martinod, E. & Buvat, I. J. P. O. 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. *PLoS One.* **10**, e0145063 (2015).
42. Rothlauf, F. in Representations for Genetic and Evolutionary Algorithms 9–32(Springer, 2006).
43. Kramer, G. M. *et al.* Repeatability of quantitative whole-body 18F-FDG PET/CT uptake measures as function of uptake interval and lesion selection in non–small cell lung cancer patients. *J Nucl Med.* **57**, 1343–1349 (2016).
44. Williams, C. R., Baccarella, A., Parrish, J. Z. & Kim, C. C. J. B. b. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics volume.* **17**, 103 (2016).
45. Love, M. I., Huber, W. & Anders, S. J. G. b. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
46. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols.* **7**, 562–578 (2012).
47. Chu, F. & Wang, L. J. I. j. o. n. s. Applications of support vector machines to cancer classification with microarray data. *International Journal of Neural Systems.* **15**, 475–484 (2005).
48. Pei, G., Chen, L. & Zhang, W. in Methods in enzymology Vol. 585 135–158(Elsevier, 2017).
49. Langfelder, P., Zhang, B. & Horvath, S. J. B. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics.* **24**, 719–720 (2008).
50. Langfelder, P., Mischel, P. S. & Horvath, S. J. P. o. When is hub gene selection better than standard meta-analysis? *PLoS One.* **8**, e61505 (2013).
51. Sherman, B. T. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8**, R183 (2007).
52. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genetics.* **25**, 25–29 (2000).
53. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. J. P. o. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One.* **6**, e21800 (2011).
54. Nioche, C. *et al.* LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Res.* **78**, 4786–4789 (2018).
55. Byun, B. H. *et al.* Initial metabolic tumor volume measured by 18F-FDG PET/CT can predict the outcome of osteosarcoma of the extremities. *J Nucl Med.* **54**, 1725–1732 (2013).
56. Orlhac, F. *et al.* Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *J Nucl Med.* **55**, 414–422 (2014).

57. Sheen, H. *et al.* Metastasis risk prediction model in osteosarcoma using metabolic imaging phenotypes: A multivariable radiomics model. *PLoS One.* **14**, e0225242 (2019).
58. Luesse, D. R., Wilson, M. E. & Haswell, E. S. RNA sequencing analysis of the msl2msl3, crl, and ggps1 mutants indicates that diverse sources of plastid dysfunction do not alter leaf morphology through a common signaling pathway. *Frontiers in plant science.* **6**, 1148 (2015).
59. Way, G. P. *et al.* A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma. *BMC Genomics.* **18**, 1–11 (2017).
60. Zhang, Y., Deng, Q., Liang, W. & Zou, X. J. B. r. i. An efficient feature selection strategy based on multiple support vector machine technology with gene expression data. BioMed Research International 2018 (2018).
61. Wu, J., Roy, J. & Stewart, W. F. J. M. c. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Med Care.* **48**, S106–S113 (2010).

Figures

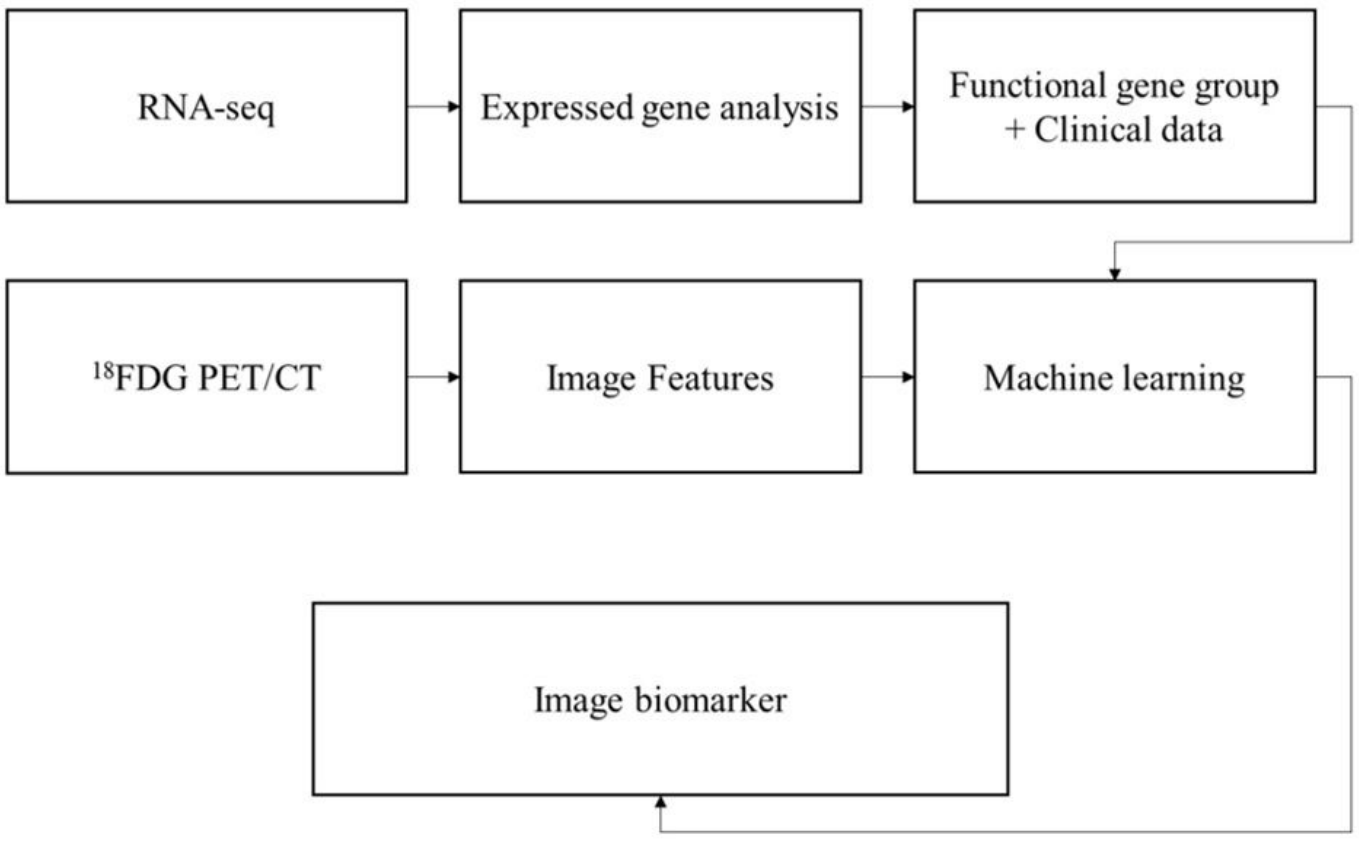

Figure 1

Schematic outline of prediction model flow revealing significant associations between 47 image features from PET/CT and 147 EMT-related genes having strong relationship with semantic features from images

in NSCLC. The degree of relevance was evaluated by P-value.

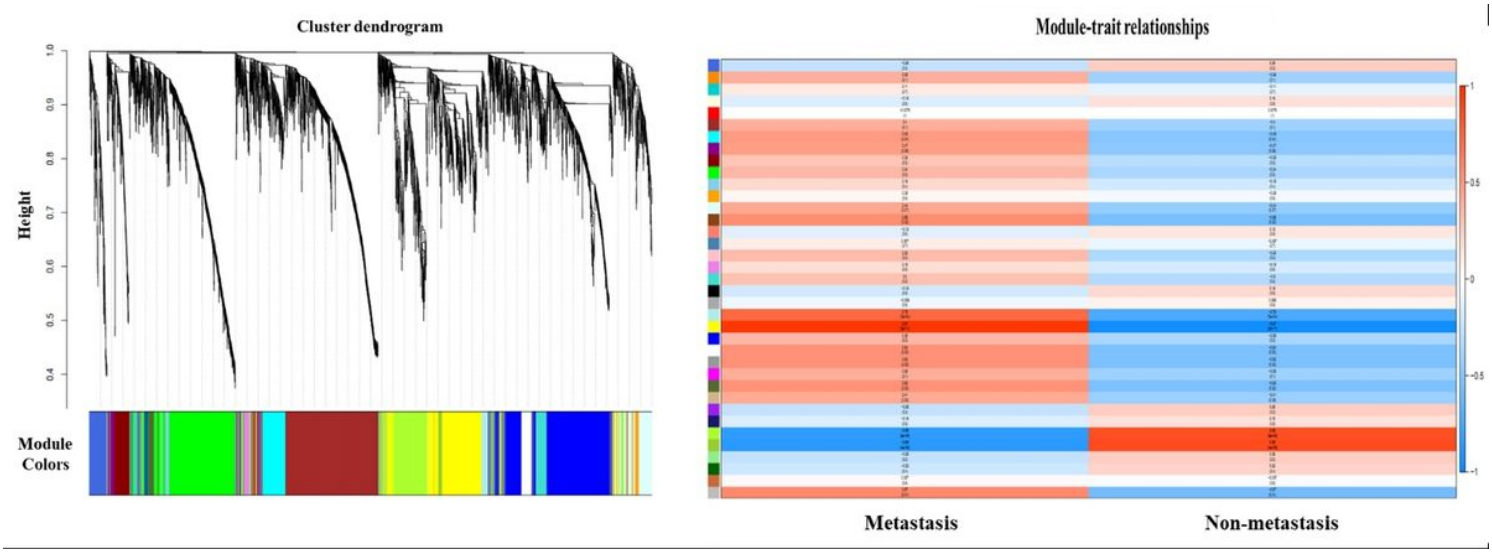

Figure 2

Gene regulation was completed through clustering. A total of 36 gene modules were generated, each module consisting of genes with similar expression patterns (left), and the relationship between each module and metastasis and non-transient functions is shown as a heat map (right). The module most relevant to the transition is the yellow module (module meaning = 0.97, P-value = -3e-11), and the module most relevant to the non-transient model is dark green module (module meaning = 0.88, P-value) = 4e -06).

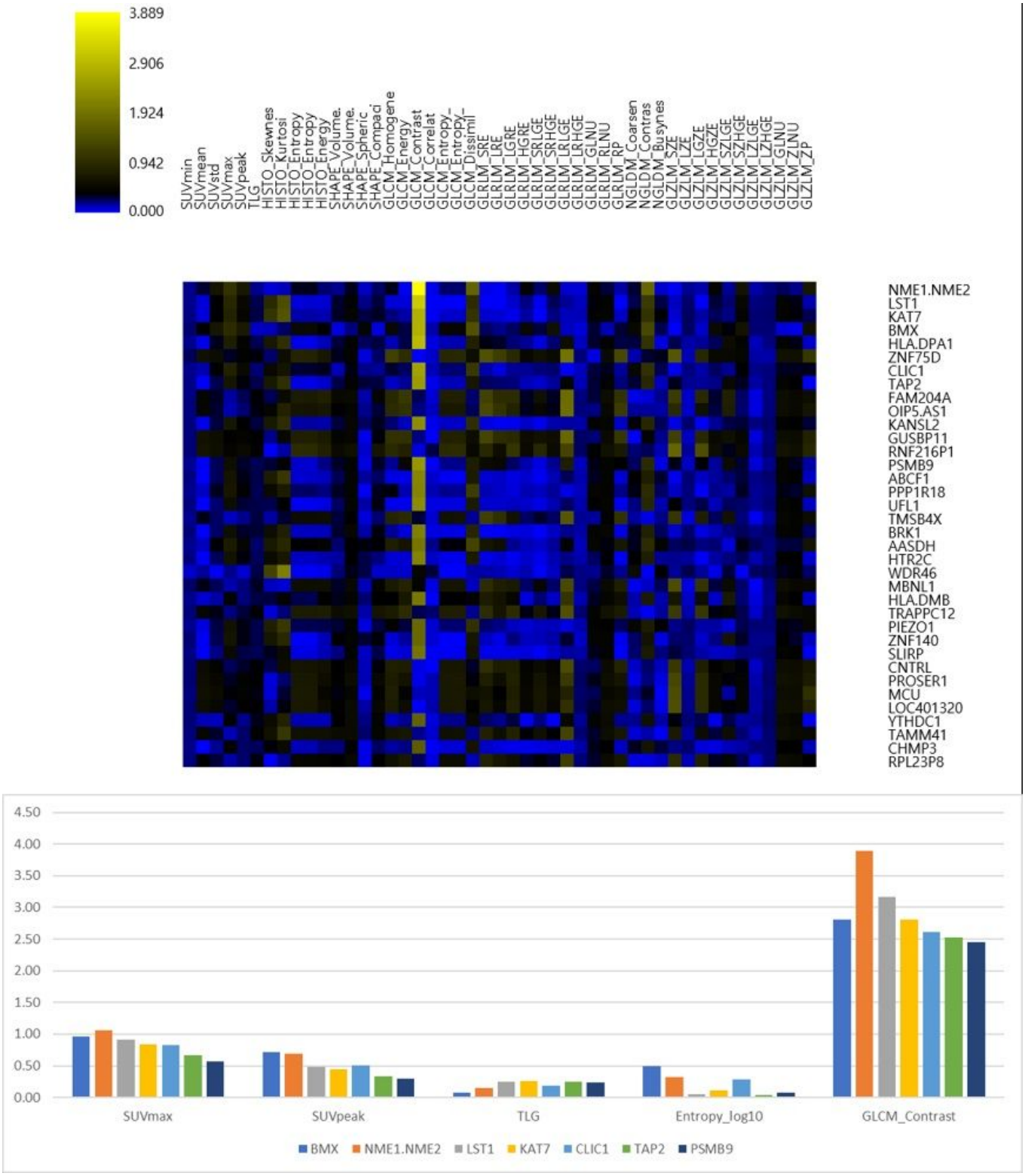

Figure 3

Radiogenomics map revealing significant associations between 47 semantic features from PET/CT and top 50 genes having strong relationship with semantic features from image in NSCLC. P-value was used for display correlation (upper panel). Correlation of SUVmax, SUVpeak, TLG, Entropy log10 and GLCM_contrast with EMT-related genes (lower panel). All P-value was normalized by -log10.