Universität Rostock Traditio et Innovatio

Universitätsmedizin Rostock

Institute for Biostatistics and Informatics in Medicine and Ageing Research

Head: Prof. Georg Fuellen

A Radiogenomic Approach in Lung Adenocarcinoma to Unravel Molecular Mechanisms Underlying CT Imaging Phenotypes

Dissertation

for

the Academic Degree

Doctor of Medicine

of the

University Medical Center Rostock

From

Nicolas Luca Spath | born 15.03.1996 in Regensburg

from Rostock

https://doi.org/10.18453/rosdok_id00005084

YEAR OF SUBMISSION: 2024

YEAR OF DEFENSE: 2024

First Reviewer: Prof. Georg Fuellen

Institute: Institute for Biostatistics and Informatics in Medicine and Ageing Research, University Rostock

Second Reviewer: Prof. Dr. med. Marc-André Weber

Institute: Institute for Diagnostic and Interventional Radiology, University Rostock

Third Reviewer: Prof. Jan Hövener

Institute: Molecular Imaging North Competence Center, University Kiel

ACKNOWLEDGMENT

I thank Dr. Mohamed Hamed, Sarah Fischer and Prof. Georg Fuellen for their outstanding support and assistance with this dissertation.

Thanks to my family and Charlotte for encouraging support, critical corrections, and comments on the dissertation.

ABSTRACT

The field of radiogenomics encompasses the connection between radiologic imaging (radiomics) and molecular data (e.g., transcriptomics and methylomics). However, it is not easy to establish meaningful connections between them because of massive data sets and different acquisition techniques. The herein-developed bioinformatics workflow aims to improve the interpretability of results from a multi-omics integration by extracting and enriching significant correlations between image traits, transcriptomics and methylomics of the TCGA and TCIA lung adenocarcinoma cohort.

Extraction of 86 image features from the CT image series describing tumor characteristics (morphology, shape, texture, size) is followed by differential expression/methylation analysis. The integrative analysis comprises 22 matching patients (median age 67.5 years, from 42 to 80 years). To find molecular signatures behind tumor phenotypes, radiogenomic association maps (RAMs) were constructed. These link image traits to biological correlates (GO terms).

Particularly, *regulation of signaling* and *cellular response to organic substance* were reflected in CT image phenotypes. Hypomethylation (*HOXA3, HOXB3 or HOXB4*) showed a distinct shift in tumor texture.

In addition, a gene regulatory network comprised of these biological processes involves similar transcription factors to acute leukemia (*TAL1, EZH2, TGFBR2*). These may reflect tumor texture and morphology in CT images.

The visualization of the results also aims to identify potential imaging and molecular biomarker candidates to classify the heterogeneity of tumors. The described approach is generalizable to other cancer entities or non-malignant diseases to broaden the interpretability of future radiogenomics studies.

TABLE OF CONTENTS

FIGURES

TABLES

ABBREVIATIONS

BP	Biological Process
CNV	Copy Number Variation
CT	Computer Tomography
DEG	Differentially Expressed Gene
DEM	Differentially Expressed micro RNA
DICOM	Digital Imaging and Communications in Medicine
DMR	Differentially Methylated Regions
DNA	Deoxyribonucleic Acid
DRF	Differentially Represented (Imaging) Features
EU	European Union
FDR	False Discovery Rate
GLCM	Gray-Level Co-occurrence Matrix
GLRLM	Gray-Level Run-Length Matrix
GLSZM	Gray-Level Size Zone Matrix
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
IARC	International Agency for Research on Cancer
IASLC	International Association for the Study of Lung Cancer
LUAD	Lung Adenocarcinoma
miRNA	micro Ribonucleic acid
MRI	Magnetic Resonance Imaging
mRNA	messenger Ribonucleic acid
NGS	Next Generation Sequencing
NGTDM	Neighborhood Gray-Tone Difference Matrix
NIH	National Institute of Health
NSCLC	Non-Small Cell Lung Cancer
ORA	Over Representation Analysis
PCR	Polymerase Chain Reaction
PET	Positron Emission Tomography
RAM	Radiogenomic Association Map
ROI	Region of Interest
SNV	Single Nucleotide Variants
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive

TF	Transcription Factor
TNM	T (Tumor) N (Node) M (Metastasis) Classification
UICC	Union internationale contre le cancer
VOI	Volume of Interest
WGS	Whole Genome Sequencing
WHO	World Health Organization

1 INTRODUCTION

As described in the global cancer statistics 2020, lung cancer is the leading cause of death, considering the mortality of all cancer types. [1] With its estimated high incidence (2,206,771 with 11.4% of total new cancer cases worldwide) and high mortality (1,796,144 with 18% of total deaths from cancer worldwide), lung cancer is a priority problem in the struggle for the betterment of global health.

Lung cancer mortality leads the cause of cancer morbidity and mortality in men. In women, lung cancer ranks third for mortality after breast and colorectal cancer. The incidence and mortality rates of lung cancer are 3 to 4 times lower in transitioning countries than in transitioned countries. Given that 80% of smokers aged over 15 years reside in low- and middle-income countries, this may change as the tobacco epidemic evolves. While the highest incidence rates occur in Northern and Western Europe, Northern America, Micronesia/Polynesia and Australia/New Zealand, rates are also high in Eastern Asia.

The high prevalence of lung cancer in Chinese women largely contributes to the high prevalence in Eastern Asia. This is attributed to the ambient pollution and inhalation of carcinogenic agents from the extensive use of charcoal and solid fuels for heating and cooking. [2]

Lung cancer is an outstanding example of carcinogenesis by environmental carcinogens, like those contained in tobacco. [3] According to the IARC Monograph on the Evaluation of Carcinogenic Risks to Humans vol. 83 from the WHO [4], smoking cigarettes increases the risk of all histological types of lung cancer. In addition, tobacco smoke is the cause of other cancer entities such as the esophagus, larynx, stomach, liver, ovary cavities, and pancreas.

Typical clinical symptoms vary but often include coughing, dyspnea, weight loss, hemoptysis, chest pain and fever.[5] The S3 clinical guideline on lung cancer published by the *“Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften”* in Germany outlines possible diagnostic pathways.[6] The diagnosis considers a combination of the clinical symptoms just listed, followed by computer tomography. A variety of PET and CT can be applied if diagnostic precision needs to be added. Typically, a biopsy and (immuno)histology of a tissue sample follow to classify the suspected malignant node. In addition, lymphatic and organ metastasis can be found by applying imaging technologies such as sonography and whole body (PET)-CT. [7] Today's classification determines the tumor grade and stage according to the 8th edition of the UICC/AJCC classification for lung cancer utilizing the TNM

system. [8], [9] The main histologic type is small cell carcinoma, which develops from neuroendocrine tissue and sets it apart from non-small cell lung cancer, again subdivided into squamous cell carcinoma and adenocarcinoma of the lung. Further histologic subclassification can be applied according to the 2015 WHO classification of lung tumors.[10] Genetic variance of cancer plays a crucial role in finding targeted therapies. In non-small cell lung cancer (NSCLC), treatment possibilities are affected by:

(i) relevant expression of *KRAS* [11], *HER2* [12] and *PD-L1* [13],

(ii) *EML4-ALK* [14] or *ROS1* translocation [15] or

(iii) mutations in *EGFR, EPHA3, KDR* and *NTRK* genes [16]

Improvements in survival by combining platin-adjacent chemotherapy with Pembrolizumab to counter immune evasion have been shown. [13] Lately, double immune checkpoint inhibition with Durvalumab and Tremelimumab show promising results in improving overall survival. [17] Crizotinib, a tyrosine kinase inhibitor for *ALK*-positive lung cancer [18], *HER2* inhibitors like Trastuzumab [19] or *EGFR* wildtype receptor inhibitors like Gefitinib [20] add new facets of therapy possibilities in NSCLC. Even given these vast advancements in therapeutic options, significant impact on overall survival is limited. [21] Numerous protocols have been tested in high-risk populations to detect lung cancer in a surgically operable and thus curable early stage [22], [23], [24]. Often, these include a (low dose) CT to discriminate lung nodules. Usually, there is a follow-up imaging of suspected nodules to measure volumetric growth and a biopsy to rule out malignant tissue.

2 BACKGROUND

2.1 DATATYPES AND LABORATORY TECHNOLOGIES USED IN THIS STUDY

Various omics data can be generated using different laboratory techniques to understand cancer biology. Transcriptome profiling (mRNA and miRNA sequencing) as whole genome or whole exome sequencing offers deep insights into the tumor biology. Over the past decade, this field has had numerous advancements, from microarray over next-generation sequencing (NGS) to third-generation sequencing, which is still in development. Next-generation sequencing allows for high throughput experiments due to its fast speed, low cost and high accuracy. It furthermore allows the detection of previously unknown sequences in comparison to microarrays. The third-generation sequencing technologies allow the generation of full-length cDNA from single RNA molecules in vitro, reducing the false positives of splicing sites and enabling the analysis of isoforms. Further development led to single-cell RNA-sequencing to profile the transcriptome of one single cell. In addition, current methods arose using DNA barcode beads to add spatial information of the cell's transcriptomes in histologic slides. Another interesting method is nanopore sequencing, which uses the DNA's electronic charge, which differs depending on each base. This method works in real-time, allowing for the selection of a specific subset of genes to read while doing the analysis, accelerating the speed and reducing cost at scale.[25]

Numerous advances have been made in the past years to find methylated cytosine bases. Even though the protocols are getting more refined, they still use one common step, the polymorphism of unmethylated cytosine to uracil under treatment with bisulfite. After bisulfite treatment, the DNA sequencing will reveal unmethylated sites due to the C/U base exchange. As sequencing is expensive and thus not always affordable by all laboratories, microarrays remain to play a significant role in whole genome assessments of methylation. The typical microarray workflow is being applied here, although protocols of the various available methods differ. After bisulfite treatment, whole genome amplification followed by enzyme fragmentation and hybridization on a chip is carried out. In addition, new approaches like the Illumina 450k methylation array using two chips with different chemical assays promise to evaluate more CpG sites while running in parallel. After the raw read of the intensity signal, normalization and reduction of batch effects are applied to the dataset. When calculating the ratio between the two allele signals, the resulting value is called the beta value. It can range from 0 (homozygous unmethylated) to 1 (homozygous methylated). [26], [27]

Mutation analysis in research is often carried out with a whole exome sequencing (WES) approach, followed by alignment to a reference genome. To detect SNVs, mutation calling is applied. There are various bioinformatics tools designed for this purpose, for example,

MuTect2 developed by the Broad Institute [28], VarScan2 [29], Strelka2 [30] or tools like VaDir [31], which detects SNVs in RNA. First, the laboratory method to prepare and sequence samples influence the quality of the results. Adjustment of the calling strategy is needed when utilizing fresh frozen DNA compared to formalin fixed paraffin embedded samples, often used in clinical environments when performing the histologic evaluation. [32] Furthermore, the tools differ in performance considering the sensitivity and false positive rate. For the tools, MuTect2 and Strelka2, the sensitivity and false positive rate depend on the mutation frequency and sequencing depth. This directly impacts clinical research as those tools are widely used across research fields [33].

2.2 BIG DATA APPROACHES

The laboratory techniques mentioned above produce vast amounts of raw and heterogeneous data. The analysis and integration of the various data types from different patients participate in the world of big data to upcoming healthcare information systems.[34] Herein lies the promise of omics data in healthcare: To create models that can significantly advance predictive, personalized and preventive medicine. The big challenge is integrating high-quality multi-omics data and electronic medical patient records, which might be structured or unstructured, discrete or continuous.[35] Another prominent issue in the last decade is data privacy and security following the utilization of large data repositories by companies like Google and Facebook. Governments have passed several laws like the EU in the form of the general data privacy act or the California Consumer Privacy Act in the USA to give back some of the decision-making to the individuals the data belongs to. Medical data is highly sensitive, and the distribution and analysis require high standards to ensure compliance with general medical ethics and laws. In addition, the data cannot be fully anonymized to link the gained insights back to the patient and use the data in the environment of medical facilities like hospitals. The method in question is pseudo-anonymization which is typically complex and only ensures a certain amount of anonymization that might lack function when being cross analyzed with data from, for example, social media. [34]

2.3 IMPORTANCE OF INTRATUMORAL HETEROGENEITY

Intratumoral heterogeneity is the critical driver for the evolution of more malignant and drug-resistant tumors. The focus lies within the expansion of cell clones that carry the potential to escape growth and proliferation inhibition and start to self-renew without control of cell-cell adherence or external stimulants. This is a risk in multicellular organisms. Cancer is not evolving linearly, but often simultaneously, multiple subclones co-exist, showcasing a fast

parallel evolution. This is a significant problem since it would need adapted treatment options for each clone. As the mutational burden rises, the co-existent clones are under heavy selective pressure in this microenvironment which might lead to a starburst architecture of clone development. At the same time, each cell tries to reach its fitness peak. So, viewing cancer development from an evolutionary standpoint gives insight into the potential lack or rebounds of cancer therapies leading to recurrence or explains minimal residual disease. [36] The TRACERx study in NSCLC tried to follow the evolutionary path of tumor development in a clinical setting. [37] One significant finding was a common and almost universal branched evolution and intratumoral heterogeneity across cohorts. This provides important implications for treatment and recurrence control. Clonal early mutations happened exclusively in EGFR, MET and BRAF, explaining the uniform responses across different tumor sites when targeting these alterations. Considering the subclonal driver mutations that are not targeted may lead over time to further evolution and are specific to only a subset of the whole tumor.

To derive knowledge of this subclonal variability, it is necessary to survey the tumor as a whole. At the same time, that might not be possible when total surgical resection of the tumor is often not feasible due to tumor spread and infiltration of neighboring structures. Therefore, radiomics proposes an answer. A current hypothesis in research is to utilize modern imaging modalities like CT and MRI to measure the signal of the tumor in 3-dimensional space and all its different infiltrated nodes by image features that refer to grey distribution or extract spatial information. Information from the extracted features must then be associated with the underlying biological data to allow clinical applications. This would enable the tracking of intratumoral heterogeneity non-invasively and adjust the targeted treatment or allow conclusions on the overall prognosis. Thus, radiomics is a new and rapidly developing field and is probably indispensable in individualized medicine in the future. [38], [39]

2.4 ANALYSIS OF RADIOLOGIC IMAGES

As computer image analysis is a wide field of ongoing research, numerous algorithms and extraction methods exist. There is no common classification of image features into specific groups yet. Several descriptors allow for a rough grouping and comparison between studies. In the field of medical image analysis, several studies are defining the landscape of feature extraction. Recently there has been a study to standardize quantitative radiomics through cross-evaluation between research teams with varying acquisition methods to find robust image features. [40] These features originate from Aerts and Grossmann et al. They assigned their radiomic features into 5 groups: (i) tumor intensity, (ii) shape, (iii) texture, (iv) wavelet, and (v) Laplace of Gaussian features. [41], [42] The assignment in this study roughly uses the same categories. Texture features are calculated from three dimensional matrices. Tumor

intensity is renamed more descriptively to morphology and refers to first-order statistics like *mean intensity* or *variance* but contains the same features. The shape category contains geometric features, while the size category evolves into its own feature set and is a new category introduced.

Table 1 Comparison of the groups of image features in different studies.
This work tries to group image features under descriptive terms to improve readability and interpretability. For detailed references, check Table 10. 1-First-order statistics (e.g., variance, mean, kurtosis) of voxel intensity values in the given VOI; 2-Geometric features (e.g., compactness, eccentricity, geometric moments); 3-Texture features based on matrix-raster calculations that are gray-level co-occurrence matrix (GLCM), Gray-Level Size Zone Matrix (GLSZM), the gray-level run-length matrix (GLRLM), neighborhood gray-tone difference Matrix (NGTDM); 4-Assessment of features 1-3 after wavelet decomposition of the image; 5-Application of Laplace of Gaussian transformation to the image and recalculation of first-order statistics; 6-Size and volume of the tumor; VOI – the volume of interest

Aerts et al. 2014 [41]	**Grossmann et al. 2017** [42]	**Spath & Fischer et al. 2023** [43]
1) tumor intensity	1) tumor intensity	1) morphology
2) shape	2) shape	2) shape
3) texture (no NGTDM)	3) texture	3) texture
4) wavelet	4) wavelet	6) size
	5) Laplace of Gaussian (LoG) features	

2.5 BIOLOGICAL REPOSITORIES AND BIOINFORMATICS TOOLS USED IN THIS STUDY

2.5.1 THE CANCER GENOME ATLAS

The Cancer Genome Atlas (TCGA) project started officially in December 2005, receiving a $100 Mio. pilot grant from the National Institute of Health (USA). In the consecutive five years the Cancer Genome Characterization Centers and Data Coordinating Center as components of TCGA were announced. The first cancers to be mapped are lung, brain and ovarian cancers, followed by multiple other cancer types, including breast cancer, various types of squamous cell cancer and organ-specific cancers like hepatocellular cancer, as well as endocrine cancer types like pheochromocytomas. In 2010 the data was first made publicly available as “level 3 data access”. Each data set released was accompanied by a marker paper in high-ranking journals as a first-of-its-kind data mining effort, integrating various big data sets to deepen the understanding of the underlying genetics of the selected cancer tissues. With such great sources available, many studies used these datasets to push our boundaries of

knowledge further. Since 2018 TCGA data generation ended, resulting in a giant public data repository that data scientists worldwide continue to use.[44]

2.5.2 THE CANCER IMAGING ARCHIVE

The Cancer Imaging Archive (TCIA) launched in 2011 as several other predecessors (mostly the national biomedical imaging archive) showcased the need to provide a versatile image repository developed to support the needs of the cancer research community. It is an extension of the National Biomedical Imaging Archive (NBIA). In 2010 the Washington University's (Saint Louis, Missouri) Electronic Radiology Laboratory received a grant from the National Cancer Institute. It thus started developing TCIA as a support platform for the various projects that needed managed anonymized imaging data. Until now, they encourage and support researchers to curate and publish their datasets and analysis results, like image segmentations, to strengthen publicly available medical images.[45]

2.5.3 BIOCONDUCTOR

Bioconductor was initiated at the Dana Farber Cancer Institute in 2001 and was initially funded by the High-Tech Industry Multidisciplinary Research Fund. Grants from various NIH and the Chan-Zuckerberg foundation projects mainly secure today's funding. It provides the research community with a platform to write, test, publish, distribute and document bioinformatic tools for high throughput data analysis mainly written in R. In 2022, the number of released packages was up to 2183. [46], [47], [48]

2.5.4 DESEQ2

DESeq2 implements a mathematical model using the negative binomial distribution to calculate differential expression in high-throughput RNA-seq data. It achieves high accuracy in the estimation of variance-mean dependence and dispersion. This study uses an implementation of this method as a Bioconductor package to receive the differential expression of genes and miRNAs. [49]

2.5.5 TCGABIOLINKS

This is yet another Bioconductor package to provide a convenient way to obtain TCGA data through an R programming interface. Several methods are implemented to derive differentially expressed genes and miRNAs or differentially methylated CpG sites, as well as methods to

integrate mutation data sets with expression data. Furthermore, it coerces data into the SummarizedExperiment [50] package and thus provides a well-documented R data object for large-scale genomic experiments. This study uses the “TCGAanalyze_DMR” function to compute differentially methylated regions.

2.5.6 COMPLEXHEATMAPS

The ComplexHeatmaps package can be used to plot heatmaps easily, enriching them with additional information and combining them on the go. It is included here as this study heavily relies on the flexibility and ability to add and visualize data of analysis results in a quick manner.[51]

2.5.7 PIANO

This Bioconductor package implements methods to integrate various gene-level statistics, gene set collections and other wide-range omics data sets on the level of a directionality analysis using a gene set enrichment analysis (GSEA). This enables easy integration of very different data sets with precise results. This study applies it to determine a direction of association between image features and gene ontology biological processes. [52]

2.5.8 G:PROFILER

This web-based implementation of an enrichment tool finds biological categories in gene lists. It can be used for overrepresentation analysis (ORA) in a transcriptomic data set. It uses the quarterly releases of the Ensemble database to provide up-to-date information. A variation of the Benjamini-Hochberg method for correction in areas of multiple hypothesis testing, called gSCS, is commonly used to limit false discoveries. [53]

3 METHODS

This study implements a data-driven pipeline to mine big genomic data sets. It mainly consists of methods of already published work to filter large omics data sets until a relevant small portion of the data is left, followed by integration and visualization. Aggregation is achieved through construction of the radiogenomic association maps and points of interest are further investigated and discussed.

3.1 A GENERAL NOTICE ABOUT MULTI OMICS INTEGRATION

To integrate data in omics studies, a common underlying sample or patient from where the data originates is often required. There are several approaches to overcome limitations in studies where data availability is problematic. A typical question is whether to reduce datasets to common samples early or late in the pipeline. As seen in drug response prediction with deep learning, a late integration of multi-omics data led to the improved overall performance of the model compared to similar studies.[54] This study aims to reduce the number of patient samples as late as possible. After the intersection with all datatypes used, this approach results in a relatively small batch size of 22 samples. These are used for the following integration steps. Although this is a very small number of patients to analyze, it must be mentioned that each of the different data types used is from one tissue sample of one patient. This gives a unique insight into the radiogenomics of lung adenocarcinoma as the results align in time and space.

3.2 ORIGIN OF THE DATA

Clinical data, gene and miRNA expression profiles for lung adenocarcinoma were downloaded from the TCGA-LUAD project of The Cancer Genome Atlas portal (GDC data portal). Genomic datasets were received at level three. The matching CT studies were obtained from The Cancer Imaging Archive (TCIA). [55], [56] All available patients with primary tumors and normal samples were used in the preliminary pipeline using a late integration approach. Through the later reduction of the samples, which have all data types available, the separate analysis of the genomic data prevents error accumulation through a small sample set. This is due to the availability of large sample sets for genomic data (>500 samples/data type) versus the relatively small overlap with imaging data (69 image series).

3.3 DATA PROCESSING

3.3.1 CLINICAL DATA ANALYSIS

As described in the introduction, environmental toxins play a significant role as lung cancer risk factors. The selection of clinical data follows the outline of the major risk factors, including pack-years smoked and patients' age. Those were sufficiently available and thus compiled into readable formats from TCGA as the data source. However, not for every patient of the TCGA LUAD project data points for mRNA, miRNA sequencing, Methylation or Imaging were available.

3.3.2 MUTATION DATA ANALYSIS

After obtaining mutation data, filtering for moderate and severe mutations in known lung adenocarcinoma driver genes, ALK, EGFR, KRAS and TP53, according to the annotation, was carried out. Furthermore, mutation frequencies within those genes were measured.

3.3.3 MRNA AND MIRNA EXPRESSION ANALYSIS

Gene and miRNA expression profiles were preprocessed by normalizing the read counts. If multiple tissues of the same tumor node of the same patient ID were available, the mean of gene expression was compiled into one patient expression profile. The DESeq2 v. 1.12.4 R package [57] was used to identify genes and miRNAs that exhibited at least 2-fold change and a p-value cutoff of 0.05 as differentially expressed genes (DEGs) and miRNAs (DEMs) between normal and tumor samples. P-values were adjusted using the Benjamini-Hochberg [58] procedure to limit the false discovery rate to 5%.

3.3.4 METHYLATION DATA ANALYSIS

Only primary tumors were considered. To eliminate bias from sex, CpG sites of X- and Y-chromosomes were removed from the originally annotated data set. The "TCGAanalyze_DMR" function calculated the differentially methylated regions (DMRs) for epigenetic profiling in the TCGAbiolinks R package v. 1.21.0 [59] between normal and tumor samples. CpG sites with a Benjamini-Hochberg [60] corrected p-value of 0.05, and a cutoff of a difference in methylation of 0.25 were classified as differentially methylated. The two kilobase pairs range up- and downstream of the transcription start site (TSS) was used to define the promotor region and find the CpG sites associated gene.

3.3.5 IMAGING DATA ANALYSIS

The CT DICOM images were loaded as image sequences utilizing the ImageJ2 software [61] and segmented using the segmentation manager plugin of Fiji version 8 [62] to create the region of interests (3D ROIs), delineating the tumor in each CT slide. The resulting ROIs were saved in TIFF format. The 86 defined image features (Table 10) of the 3D tumor were extracted using the Fiji 3D-ROI Manager plugin [63] and the texture toolbox [64] by loading the TIFF ROIs (TIFF-stack library) into MATLAB ver. R2020b. Organizing the image traits into four groups as a best semantic fit approach was made to achieve a better overview of the results. These are Size, Shape, Morphology and Texture. [65] For a detailed explanation of how the different features are calculated, please see at the texture toolbox of Vallières et al. on GitHub. [66] MATLAB's uniform quantization was used to calculate the texture features. The CT images were scaled to their houndsfields units to generate the gray level matrices utilizing transformation information in the DICOM files. Voxels were normalized to 1 mm^3. Various authors describe the used gray-level matrices: There is the gray-level co-occurrence matrix (GLCM) from Haralick [67], Gray-Level Size Zone Matrix (GLSZM) published by Thibault et al. [68], the gray-level run-length matrix (GLRLM) described by Galloway [69] and the neighborhood gray-tone difference Matrix (NGTDM) outlined by Amadasun and King [70]. Essentially these transform image information into smaller-sized matrices to extract spatial information by a predefined mathematical function. All the named methods to generate texture features were used in this study.

3.3.6 ENRICHMENT ANALYSIS OF DIFFERENTIALLY EXPRESSED GENES, MIRNAS AND METHYLATION SITES

Utilizing the found differentially expressed genes (DEGs) and miRNAs (DEMs) the GOSt tool gProfiler (Ensembl 102, Ensembl Genomes 49) was used to find functionally enriched gene ontology biological processes terms (GO BP) with the R command line interface. [71] The p-value was adjusted using the custom gSCS method of gProfiler, a variation of the Benjamini-Hochberg method, due to the limitations of classical methods of multiple hypothesis testing in the environment of gene overrepresentation analysis (ORA). The significant ($p < 0.05$) gene ontology biological processes terms were considered for in-depth analysis. In the following, the patient cohort was reduced to 22 patients, with data available in the transcriptomic data sets and suitable CT image series. Utilizing the Gene Set Enrichment Analysis (GSEA) implemented in the Piano v. 2.8.0 R package, functional enrichments (GO BP) can be assigned distinct directions (positive or negative) in comparison to image features. [72] Therefore, the spearman correlation coefficient as the distance between image features and genes or miRNA

expression in each term was determined. Through 10.000 random permutations of gene or miRNA sets, an FDR-corrected p-value of lower than 0.05 in Piano's distance calculation was deemed significant.

A benefit of this method is the directionality of the association between the functional term and the image feature in the up or down direction, revealing positive and negative associations between the transcriptomic expression profiles and the imaging feature.

For each GO term, the results are plotted into categorical heatmaps using the ComplexHeatmaps Bioconductor package v. 2.2.0, visualizing the associations of image features with biological processes and further adding information on image feature groups.[51]

3.3.7 CONSTRUCTION OF RADIOGENOMIC ASSOCIATION MAPS

Radiogenomic association maps were constructed for all DEGs and DEMs (Transcriptome). The methylome is analyzed separately due to restrictions in available integration methods. The intersecting terms of the ORA in the transcriptomic sets were selected. The gene or miRNA collection of the selected functional term was chosen and patients were hierarchically clustered (Euclidean distance and complete method) by the corresponding expression values. The top 50 significant genes/miRNAs are visualized as a heatmap adding further information about clinical data points, like pack years, and the incidence of frequent mutations in lung adenocarcinoma: *ALK, EGFR, KRAS* and *TP53*. In addition, the T and N status of the patient's tumor was added. The resulting complex heatmap reveals gene and miRNA expression associations for all enriched terms. To uncover small changes between the two clusters, the expression values were displayed in descending order of significance, enabling to examine of genes that might serve as a candidate for subclassification of LUAD.

3.3.8 IDENTIFICATION OF DIFFERENTIALLY REPRESENTATIVE FEATURES

To find differentially represented image features (DRFs) that might serve as future biomarkers in lung adenocarcinoma, the fold change (FC) of each image trait was calculated between the two groups using the unpaired statistical t-test considering only image features with a significance level of $p < 0.05$ (Benjamini-Hochberg corrected). [58]

3.3.9 GENE REGULATORY NETWORK CONSTRUCTION

To construct the gene regulatory network (GRN), the TFmiR2 web server [73] was utilized. Incorporated were genes and miRNAs with significant associations with the previously examined GO terms with a p-value smaller than 0.01. The disease attribute “non-small cell lung carcinoma” was selected to limit the discovery to lung cancer. Only molecular interactions with experimental evidence were considered. The output was visualized by Cytoscape V.3.7.1 [74], highlighting edges/interactions specific to lung cancer and tissue.

4 RESULTS

4.1 OVERVIEW OF THE RADIOGENOMIC APPROACH

The developed bioinformatics workflow integrates gene (mRNA) expression, miRNA expression, methylation, clinical and imaging data (Figure 1). [43] Figure 2 summarizes the results from different pipeline steps. Patients with primary tumors were included. The intersection of data sets leads to 22 common patients used for the integrative analysis with a mean age of 67.5 (Min-Max, 42-80) (Table 2). All data sets are preprocessed individually. Image preparation is done by manually segmenting each CT slide's ROI (region of interest). Automated image processing tools achieve the extraction and storage of established image phenotypes and refer to tumor size, texture, morphology and shape. Global analysis of gene expression, miRNA expression, and methylation sites results in differentially expressed genes (DEGs), miRNAs (DEMs) and methylated regions (DMRs). Those were further investigated to identify over-represented Gene Ontology (GO) biological processes with gProfiler. [71] Gene Set Enrichment Analysis (GSEA) via Piano associated the extracted image features with the functional biological processes. [72] The intersection of the terms from the enrichment analysis shows overlapping biological processes of different data origins. Patients of these overlapping terms were clustered using their gene or miRNA signatures, resulting in diverse subgroups visualized as radiogenomic association maps (RAMs). Differentially represented image features (DRFs) between these subgroups are calculated to determine CT image correlates. A separate analysis of the methylome is carried out. To explain phenotypic differences/common ground between patient subgroups, a gene regulatory network (GRN) was constructed via TFmiR2. [73]

Table 2 Overview of demographic and clinical characteristics of the cohort

The “No data” row refers to the non-availability of the data. 22 Patients are in common when intersecting the data sets for integration, as noted in the “common” column. Student's t-test shows no differences between the patient cohorts regarding their clinical characteristics. Taken from Spath and Fischer et al. [43]

		mRNA Seq LUAD (*n* = 515)	**mRNA Seq Normal (*n* = 59)**	**mRNA LUAD-vs. Normal**	**miRNA Seq LUAD (*n* = 513)**	**miRNA Seq Normal (*n* = 46)**	**miRNA LUAD vs. Normal**	**Common (*n* = 22)**	**Common vs. mRNA LUAD**	**Common vs. miRNA LUAD**
Cohort size	Clinical data	515	59		513	46		22		
	No data	0	0		0	0		0		
Age (years)	Median	66	66	−0.558 $p = 0.578$	66	67	−0.171 $p = 0.865$	67.5	0.294 $p = 0.722$	0.229 $p = 0.821$
	Min–Max	33–88	42–86		38–88	47–85		42–80		
	No data	19	0		19	0		0		
Gender	Female	276	34	/	274	26	/	14	/	/
	Male	239	25		239	20		8		
Pack-years smoked	Median	40	48	0.181 $p = 0.857$	40	40	0.290 $p = 0.773$	25	−0.574 $p = 0.581$	−0.591 $p = 0.57$
	Min–Max	0.15–154	5–94		0.15–154	2–124		10–120		
	No data	163	26		163	10		13		
Vital status at last follow-up	Alive	389	37	/	388	41	/	14	/	/
	Dead	126	22		125	5		8		
Last Follow-up	Median days	157	306	−1.73 $p = 0.091$	157.5	182	−0.345 $p = 0.732$	242	−1.18 $p = 0.251$	−1.24 $p = 0.229$
	No data	134	19		133	4		8		
KRAS mutation	Tested and mutation found	23	3		23	2		3		
	Tested and no mutation found	36	3		36	2		8		
EGFR mutation	Tested and mutation found	23	1		23	-		2		
	Tested and no mutation found	57	4		57	3		10		

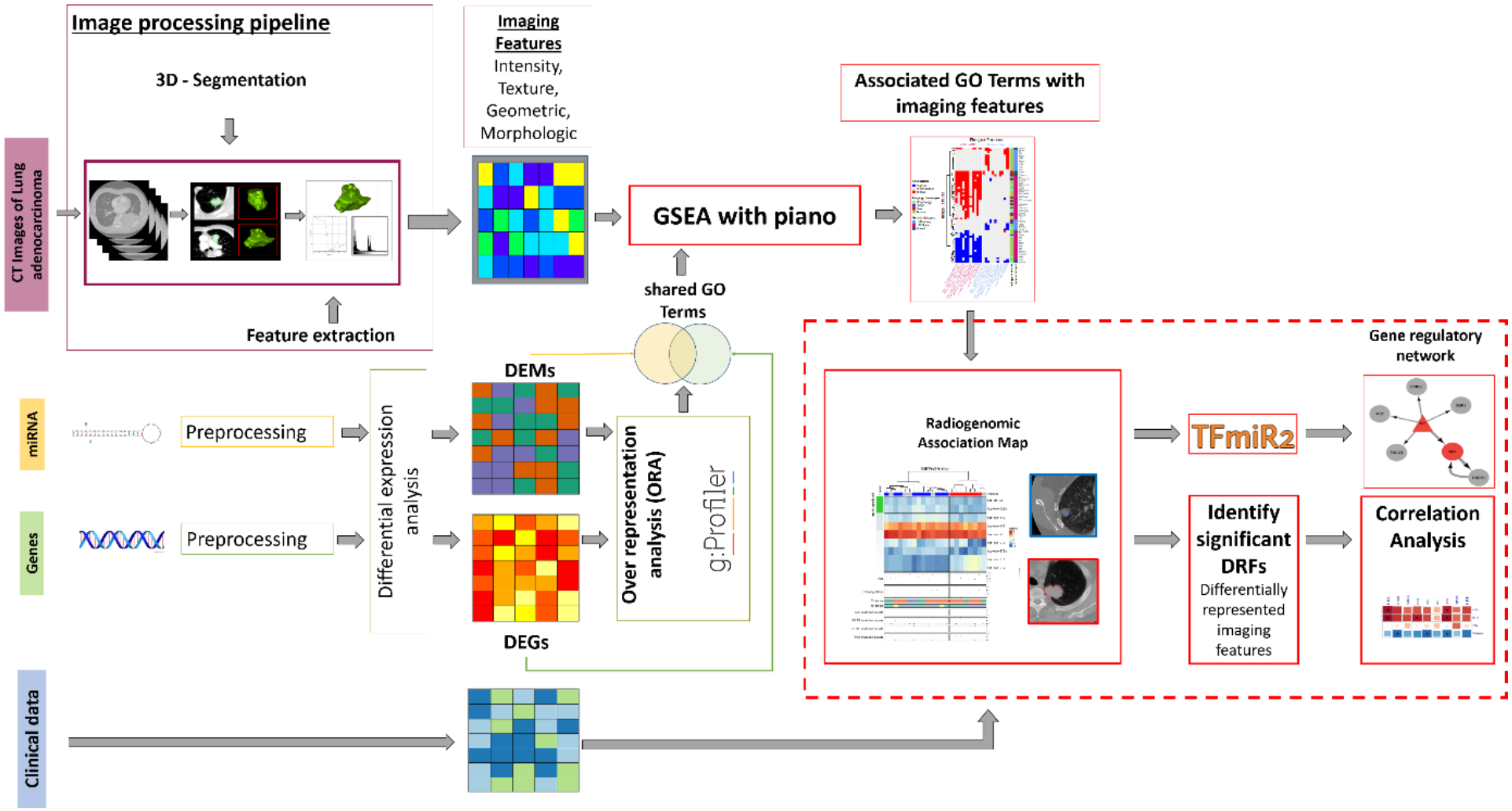


Figure 1 Schematic diagram of the workflow

It shows the image and transcriptomic data (pre)processing and the integration of the four different datasets into radiogenomic association maps. In the preprocessing step, sample sets (515 mRNA samples, 513 miRNA samples and 69 CT image series) are reduced to the same patients by intersection, resulting in 22 patients for the integration steps. Then, the evaluation occurs with a gene regulatory network construction and correlation analysis. Taken from Spath and Fischer et al. [43]

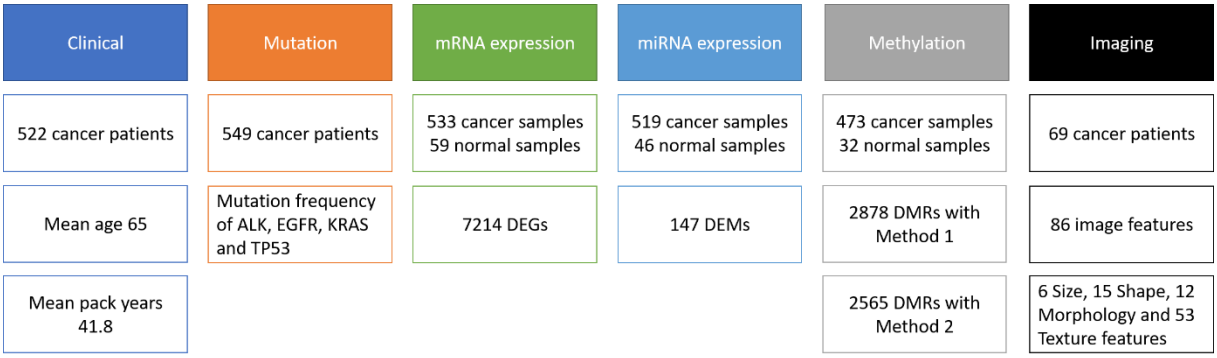


Figure 2 Overview of this study's datasets and cohorts
DEGs (differentially expressed genes/mRNAs), DEMs (differentially expressed miRNAs), DMRs (differentially methylated regions)

4.2 RESULTS FROM DATA PROCESSING

4.2.1 METHYLATION

Removing common bias results in 358.710 CpG sites for further analysis. The mean methylation of the whole genome only marginally differs between control and cancer samples ($p < 0.05$, Figure 3).

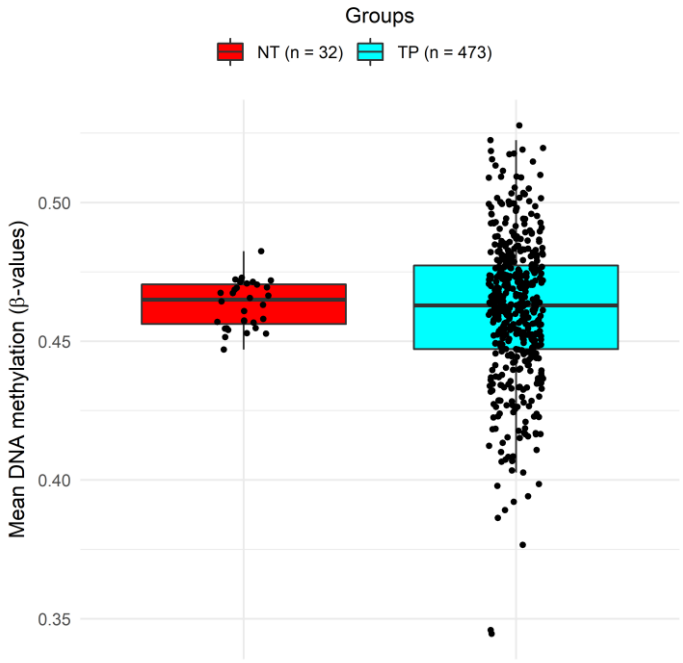


Figure 3 Mean methylation compared between normal tissue (NT) and primary tumor tissue (TP)
Each dot represents a sample. The mean methylation of the whole genome does not vary much between healthy and malignant tissue, although the methylation variance between the tumor samples is vast.

4.2.2 IMAGING

Quantitative automated processing of ROIs leads to 86 image traits. There are 6 size, 15 shape, 12 morphology and 53 texture features (Figure 2, Table 10).

4.2.3 DIFFERENTIALLY EXPRESSED GENES AND MIRNAS

After normalization of read counts, differential expression analysis revealed 7214 differentially expressed genes (DEGs) and 147 differentially expressed miRNAs (DEMs).

4.2.4 DIFFERENTIALLY METHYLATED REGIONS

Using the TCGAbiolinks R package, hypo- and hypermethylated CpG sites were determined. Methylation islands that reside within the promotor of a gene have the most impact on its expression. [75] There were 2878 differentially methylated regions (DMRs, Figure 4) left after filtering CpG sites within a promotor region where each site has one or more associated genes. This resulted in 3606 genes due to region overlapping, so one CpG site could remain in promotor regions of different genes.

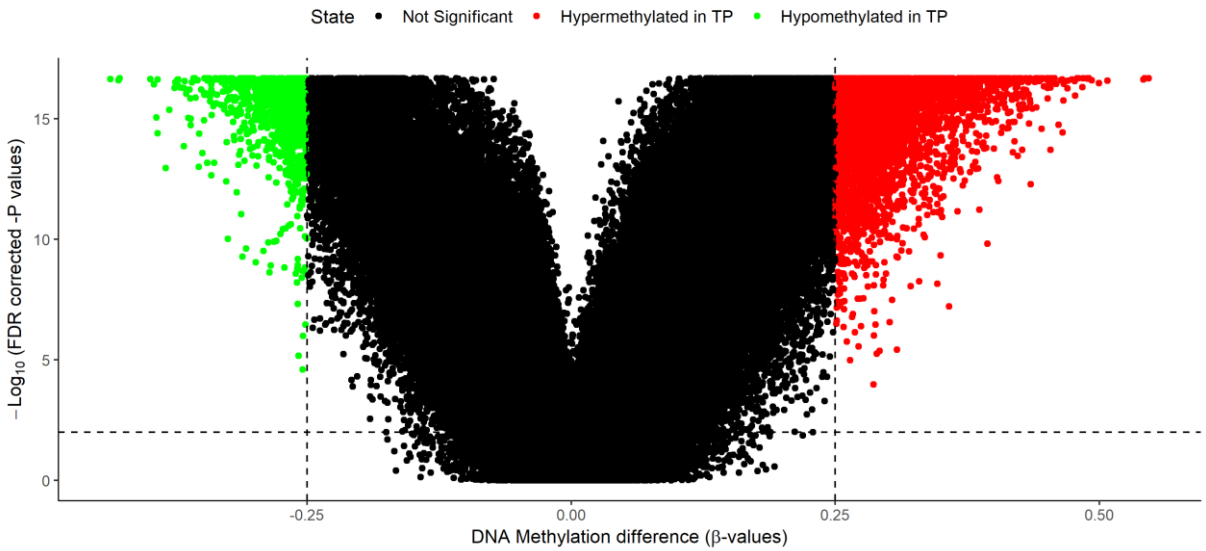


Figure 4 Visualization of the differentially methylated CpG sites

Hypo- or hypermethylation cutoffs are presented as a dashed line with a difference in methylation beta value of 0.25 compared to normal tissue samples and a significance level of 0.01 (dashed lines). Significance levels approach an asymptote of $p <= 10^{-17}$. Most of the methylation sites are not differentially methylated.

4.2.5 OVERREPRESENTATION ANALYSIS

The postulated functional roles of these dysregulated DEGs, DEMs and DMRs were revealed, resulting in a list of biological processes (GO BP). DEGs revealed 317 significant terms (see Table 8, Page 67 for term names), while the miRNAs showed 538 enriched terms (see Table 9, Page 76 for term names). With data from DMRs, gProfiler returns 353 enriched terms (see Table 6, Page 54 for term names).

4.3 RESULTS FROM THE INTEGRATIVE ANALYSIS

4.3.1 INTEGRATION OF IMAGE FEATURES AND EPIGENETICS

Each gene in the gene set of the GO term was replaced with the corresponding CpG site in the promotor region. Only GO terms with more than one CpG site left were considered, leaving 353 gene sets. As distance for the GSEA with Piano, the Spearman correlation between the beta values and the image features was computed. In total, 4552 significant associations between image features and gene sets ($p < 0.05$) were calculated, in which 337 gene sets of the 353 (96%) participated (Figure 17, Table 6).
After ordering the count of associations of any image feature in any direction, the top 5 were evaluated further (Figure 5, Table 3). These biological processes are highly associated with radiomic phenotypes.

Table 3 Biological processes involving methylation

The top five biological processes of the piano analysis from the methylation data. They are shown with the count of their significantly associated image features. For the whole table, see Table 6.

Biological processes	Count of significant radiogenomic associations
rhombomere development	44
peripheral nervous system neuron development	43
peripheral nervous system neuron differentiation	43
thyroid gland development	41
glossopharyngeal nerve morphogenesis	40

Term names generally concern the initial development and tissue forming during the embryonal and fetal life cycle. Differentiated cells should not express gene sets that are common in undifferentiated cells yet to develop a specific function. Of 93 CpG sites in the 5

terms, 23 sites are common (25%). All these 23 sites are within the promotor region of HOXA3, HOXB3 or HOXB4. HOX genes are significantly regulated by methylation in human cancer tissue, mediating chemotherapy resistance and tumor proliferation. They are organized in 4 clusters (A-D) on chromosomes 7, 17, 12 and 2. A different subset is expressed during each embryonal and fetal development phase. [76], [77] HOXB3 and HOXB4, along with other HOX genes, are proposed as epigenetic candidate biomarkers in lung adenocarcinoma. [78]

55 of all 86 image features showed an association with these 5 GO terms (64%). Further analysis of the 55 image features with associations in these 5 GO terms showed a complete intersection of 30 (Table 4). These specific image features even provide the same association direction in every of the 5 GO terms. Association direction describes the mean correlation coefficient between beta values of CpG sites within a biological process and an image feature. It is therefore either negative or positive. The CpG sites within a biological process may either be hyper- or hypomethylated or both. This graph is not designed to display single correlations between image features and CpG sites. This further underlines a common methylation-image-phenotype link with similar biological processes determining quantitative, measurable image features.

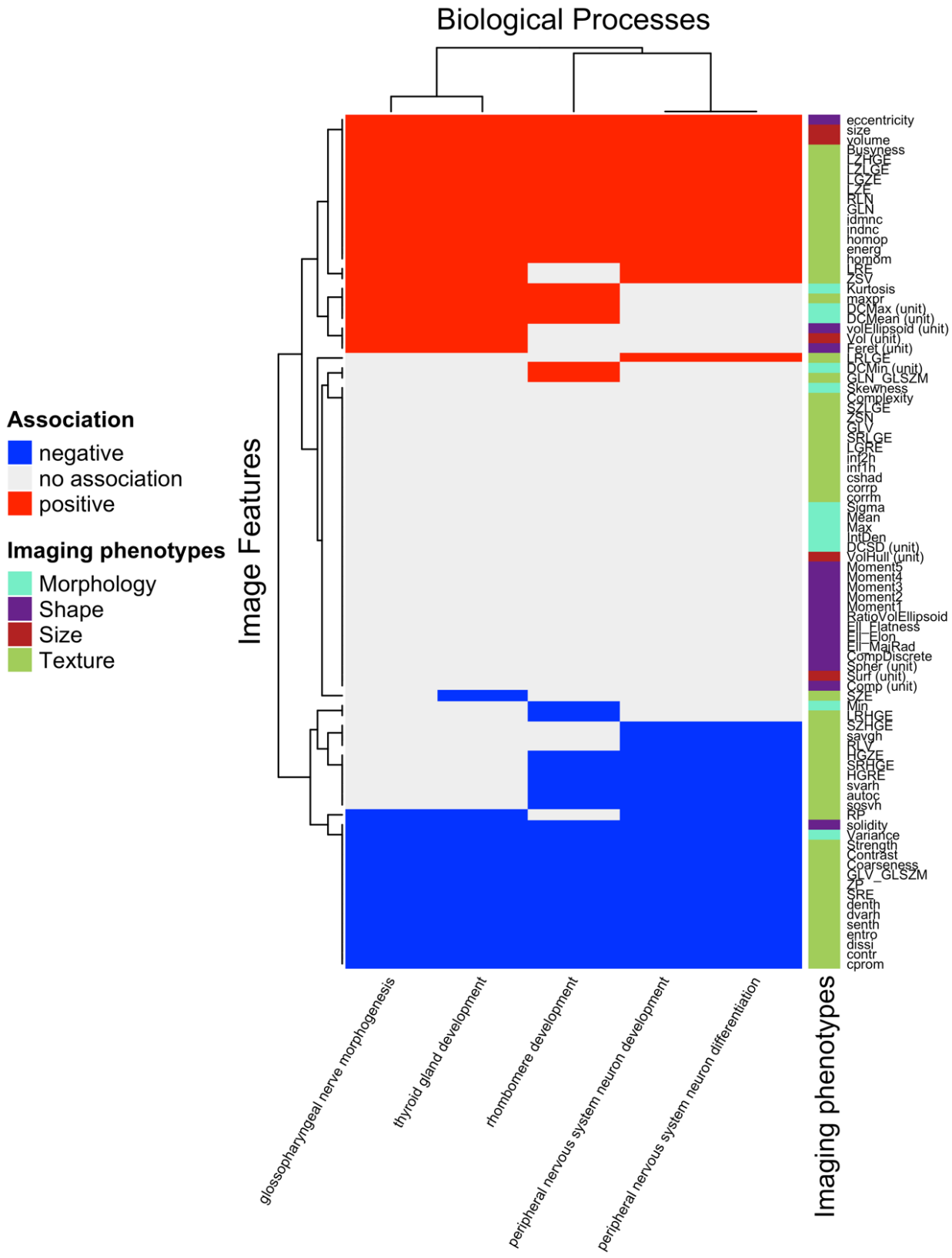


Figure 5 Visualization of the piano results from methylation processes and image features

Displayed are the five biological processes with the highest association count. The right panel includes the hierarchically clustered image features with their associated categories. The directions from the piano analysis show a clear pattern, dividing the features into two groups.

Table 4 Image features associated with methylation processes

Image features with their assigned group and ordered by association direction. All these image features are common in the five GO BP terms from DMR data with the highest association count.

Image Feature	Direction	Group
volume	+	Size
size	+	Size
RLN	+	Texture
LZLGE	+	Texture
LZHGE	+	Texture
LZE	+	Texture
LGZE	+	Texture
indnc	+	Texture
idmnc	+	Texture
homop	+	Texture
homom	+	Texture
GLN	+	Texture
energ	+	Texture
eccentricity	+	Shape
Busyness	+	Texture
ZP	-	Texture
Variance	-	Morphology
Strength	-	Texture
SRE	-	Texture
solidity	-	Shape
senth	-	Texture
GLV_GLSZM	-	Texture
entro	-	Texture
dvarh	-	Texture
dissi	-	Texture
denth	-	Texture
cprom	-	Texture
Contrast	-	Texture
contr	-	Texture
Coarseness	-	Texture

4.3.2 INTEGRATION OF TRANSCRIPTOMICS AND RADIOMICS

The GSEA results from the DEGs show 7634 significantly associated biological processes with image features ($p < 0.05$, Table 8). 83 image features (98%) and 311 from 317 enriched biological processes participated (98%) with at least one association. Looking at the GSEA analysis utilizing the DEMs, 1156 significant associations were returned ($p < 0.05$, Table 9). 280 from 538 biological processes (52%) and 43 image features (50%) have at least one significant association. Where almost all biological processes calculated from overrepresented gene expression have at least one significantly associated image feature, half of the overrepresented biological processes from overrepresented miRNA expression are associated with half of the image features. For each data set, the GSEA was visualized for all terms that have significant associations (Figure 15, Figure 16). Following the outlined pipeline, the number of associations was counted, sorting the terms in descending order to the number of image feature associations. Looking at the top 4 terms from the DEG data set (*nuclear division, meiotic cell cycle, meiotic cell cycle process, meiotic nuclear division*), we found that all of them have connections to processes regarding cell growth and proliferation which is of particular interest in cancer in general (Table 8). In addition, the GO terms from the DEM data set see many associations in the cytokine inflammation axis. Interleukin-6-involved biological processes have the most associations taking the top 5 spots (*cellular response to interleukin-6, interleukin-6-mediated signaling pathway, negative regulation of interleukin-6-mediated signaling pathway, regulation of interleukin-6-mediated signaling pathway and response to interleukin-6*). Furthermore, the top 9 terms refer to the cytokine response of the tumor cells in general, leaving a trace in CT images according to their associations count (Table 9).

The intersection of the enriched biological processes from gene and miRNA expression revealed 11 common terms where mRNA expression and miRNA expression define a radiomic phenotype together. Most of these 11 biological processes encompass proliferation and differentiation-specific biological processes, like *cell population proliferation* or *positive regulation of developmental processes*.

Close consideration of the connection between the four studied tumor phenotypes (*morphology, shape, texture, size*) and their association with dysregulated genes (DEGs) and miRNAs (DEMs) shows how the underlying biological processes influence different image features. DEGs present associations with mostly *tumor size* and *morphology* (Figure 6). Moreover, DEGs correlate negatively to several image features calculated based on the *Neighborhood Gray-Tone Difference Matrix.* In contrast, DEMs correlate positively with the texture features calculated from the *Gray-Level Run-Length Matrix*.

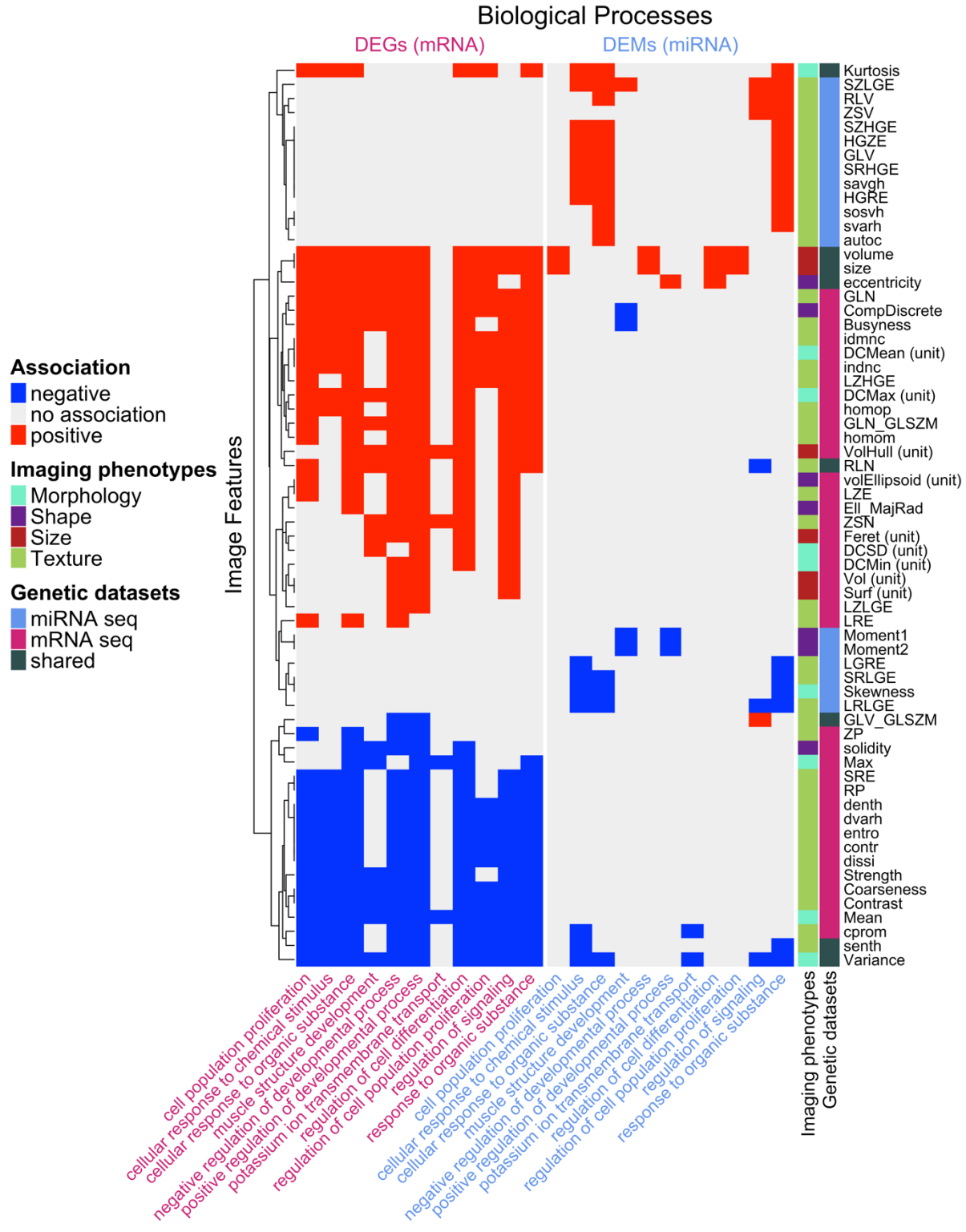


Figure 6 Visualization of the piano analysis of the transcriptome and associated image traits

The graph's content is the categorical association (positive, negative and no association) from the Piano GSEA results. Each colored column (DEGs, DEMs) showcases the associations of image features and biological processes within one dataset. Imaging phenotypes refer to the grouping of image features. The Genetic datasets legend shows if an image feature has associations with biological processes solely within one dataset or if associations in multiple datasets appear (shared). Taken from Spath and Fischer et al. [43]

To find underlying commonalities and differences in the patient cohort, hierarchical clustering using the gene or miRNA expression for each biological process results in two patient clusters. The top 50 significant genes/miRNAs are visualized as a heatmap adding further information about clinical data points, like pack years, and the incidence of frequent mutations in lung adenocarcinoma: *ALK, EGFR, KRAS* and *TP53*. In addition, the T and N status of the patient's tumor was added. To uncover small changes between the two clusters, the expression values were displayed in descending order of significance, enabling to examine of genes that might serve as a candidate for subclassification of lung adenocarcinoma. Calculating the fold change (FC) between the two groups for each image feature allows searching for differentially represented image features (DRFs). To validate the findings of these features, they are correlated to the gene sets of the clustered term, narrowing the results to a few gene- or miRNA-image feature correlations and compare to experimental results in the literature.

The similarity between patient clusters was measured by the intersection of common patients to narrow the analysis to points of interest. Points of interest to highlight are (Table 5):

1.) The biological process with the highest similarity is *regulation of signaling* with 17 of 22 possible similar patients (77%).
2.) The biological process with the highest number of associations between radiomic phenotypes and transcriptomics is *cellular response to organic substance* (Figure 7).

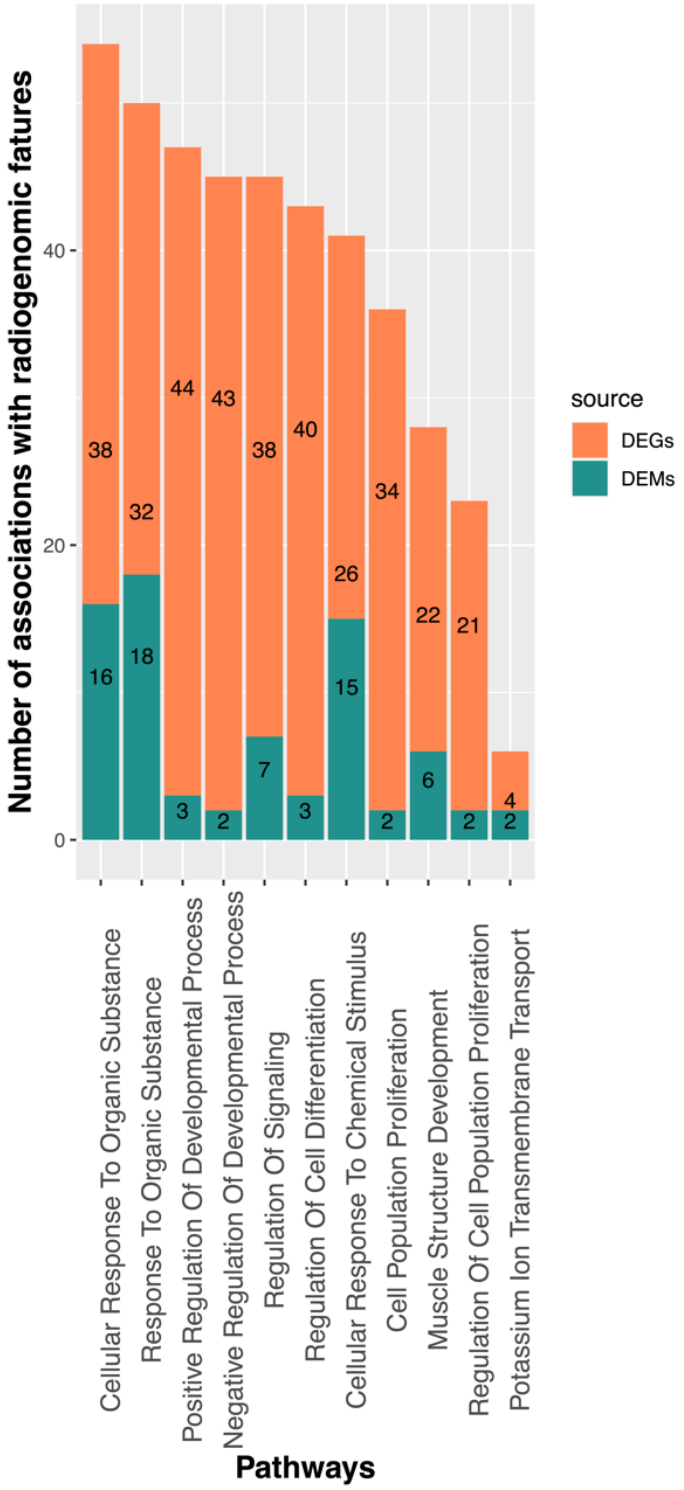


Figure 7 Visualization of the impact of the transcriptome-derived biological processes
Illustrated is a bar plot where the count of significant associations between image features and GO terms is visualized. They are split into differentially expressed genes (DEGs) and differentially expressed miRNAs (DEMs). The term with the most associations is “cellular response to organic substances”. Taken from Spath and Fischer et al. [43]

For all 11 common biological processes, radiogenomic association maps were constructed (Figure 18 to Figure 35), while *regulation of signaling* and *cellular response to organic substance* were analyzed in greater detail.

Table 5 Biological processes retrieved from the GSEA of mRNA and miRNA data sets

After construction of two patient clusters using hierarchical clustering, the number of similar patients in each group of the 11 previously selected BP terms was determined by intersection. Similar patients were counted. The percentage overlap of patients was calculated by summing similar patients, dividing by their total number (n=22). For each biological process, DRFs (differentially represented features) were calculated between patient clusters utilizing the fold change ($p < 0.05$). Taken from Spath and Fischer et al. [43]

GO Terms	Patient cluster 1	Patient cluster 2	overlap	Percent overlap	DRFs associated with DEGs	DRFs associated with DEM
cell population proliferation	11	1	12/22	55 %	Variance	Variance
regulation of cell population proliferation	11	1	12/22	55 %	Variance	RLN, GLN_GLSZM, ZSN
positive regulation of developmental process	7	1	8/22	36%	RLN, GLN_GLSZM, ZSN	RLV, volume, size, solidity, eccentricity
potassium ion transmembrane transport	9	5	14/22	64%	Ell_Flatness, Moment5, cprom	Moment4
cellular response to chemical stimulus	4	3	7/22	32%	Variance	LRLGE, SZLGE, Complexity
response to organic substance	5	2	7/22	32%	Variance	No DRF
muscle structure development	11	1	12/22	55 %	RLN, GLN_GLSZM, ZSN	Moment4
negative regulation of developmental process	10	7	17/22	77%	Variance	Variance
regulation of cell differentiation	10	7	17/22	77%	Variance	Variance
regulation of signaling	10	7	17/22	77%	Variance	Variance
cellular response to organic substance	4	2	6/22	27%	Variance	contr, dissi, homom, homop, dvarh, indnc, idmnc, SRE, LRE, RP, SZLGE, Coarseness, Complexity, Strength

4.3.3 REGULATION OF SIGNALING

Varying radiomic tumor phenotypes are significantly associated with the *regulation of signaling*. While clustering patients based on their expression signatures of signaling genes/miRNAs (Figure 8, Figure 9), significant differences in tumor morphology (*Variance*) were found. Many signaling genes and miRNAs that correlate to tumor variance appear to have an inflammatory function (52%), have already shown to correlate to overall survival (0.8%), or can be used as drug targets (0.6%). The two patient groups had no apparent differences in clinical data, tumor stage or mutation status.

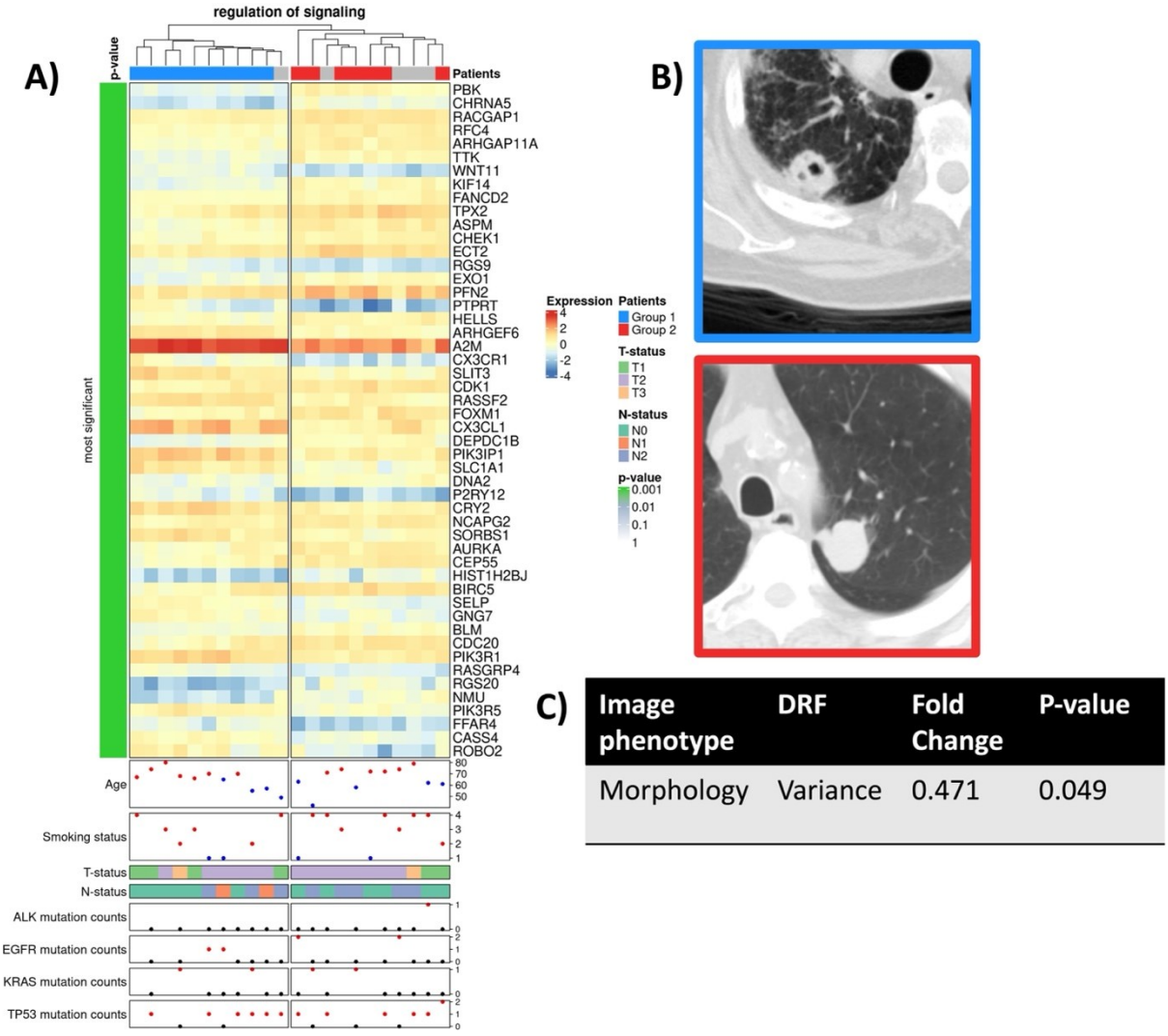


Figure 8 Radiogenomic association map of “regulation of signaling” of the gene expression

Radiogenomic association map (RAM) of gene expression of the DEG data set ordered by p-value A) for the top 50 genes in the term “regulation of signaling”. This results from clustering 1001 genes, of which 322 have a $p < 0.05$ (32,2%). The greyed-out patient's annotation marks different clusters from the DEM data set. Smoking status is presented as: 1-Lifelong Non-Smoker, 2-Current Smoker, 3-Current Reformed Smoker for more than 15 years, 4-Current Reformed Smoker for less than 15 years. According to T and N status or mutation status, the pathological tumor stage does not seem to have a visible pattern. B) CT images to showcase the radiologic imaging of the two clustered groups. The CT window width is at 1500, as the level is -600 for all images presented. “Variance” cannot be evaluated by the human eye. C) The differentially represented image feature is “Variance” with a $p < 0.05$. Adapted from Spath and Fischer et al. [43]

Outstanding are the miRNAs *hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3*, which have previously been proposed as biomarkers in lung adenocarcinoma. [79] Their upregulation correlates with poor overall survival. [80] Another study shows findings of enhanced sensitivity of tumor nodes to ionizing radiation with higher expression levels of mir-9 mediated through the downregulation of *NFκB1*. [81]

There is a significant correlation of *Variance* with *hsa-mir-182* ($p < 0.05$) in patient cluster 2 (Figure 10A). *hsa-mir-182* downregulates *RGS17* expression, which results in inhibition of cell proliferation, invasion and migration. [82]

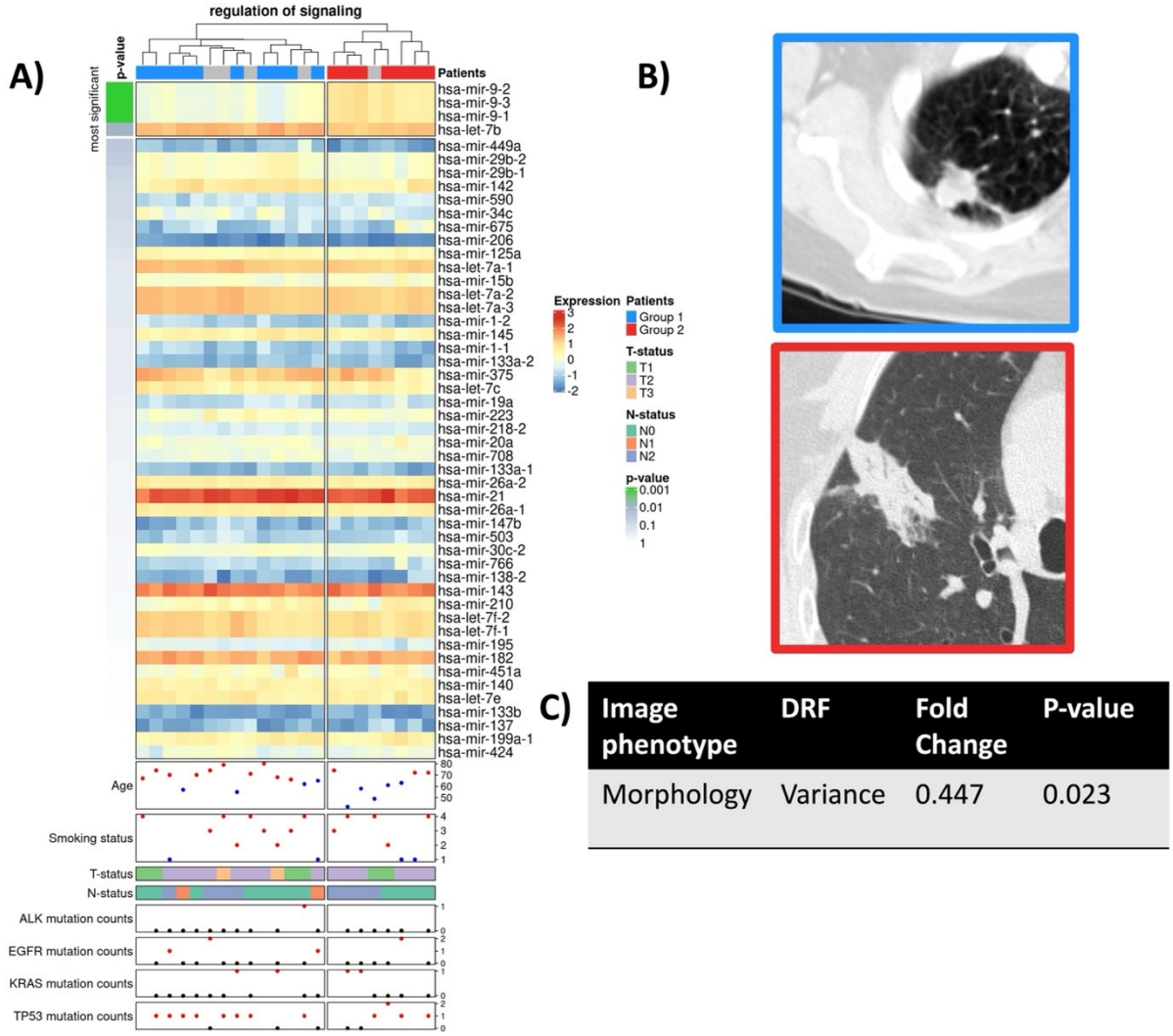


Image phenotype	DRF	Fold Change	P-value
Morphology	Variance	0.447	0.023

Figure 9 Radiogenomic association map (RAM) of “regulation of signaling” of the miRNA expression

Radiogenomic association map (RAM) of miRNA expression of the DEM data set ordered by p-value A) for the top 50 miRNAs in the term “regulation of signaling”. This RAM showcases four miRNAs with a $p < 0.05$ from 53 miRNAs used for clustering (7,5%). The greyed-out patient's annotation marks different clusters from the DEG data set. Smoking status is presented as: 1-Lifelong Non-Smoker, 2-Current Smoker, 3-Current Reformed Smoker for more than 15 years, 4-Current Reformed Smoker for less than 15 years. The 4 miRNAs on top (hsa-mir-9-1, hsa-mir-9-2, hsa-mir-9-3, hsa-let-7b) influence the clustering. Group 1 (blue) has a higher mean age than group 2 (red). Mutation status does not seem to have a visible pattern. B) CT images to showcase the radiologic imaging of the two clustered groups. The CT window width is at 1500, as the level is -600 for all images presented. “Variance” cannot be evaluated by the human eye. C) The differentially represented image feature is Variance with a $p < 0.05$. Adapted from Spath and Fischer et al. [43]

The gene *DEPTOR* correlates negatively with variance in patient cluster one. Experimental evidence supports the *DEPTOR*-mediated inhibition of tumorigenesis in lung cancer by inactivating *EGFR-mTOR* signals.[83]

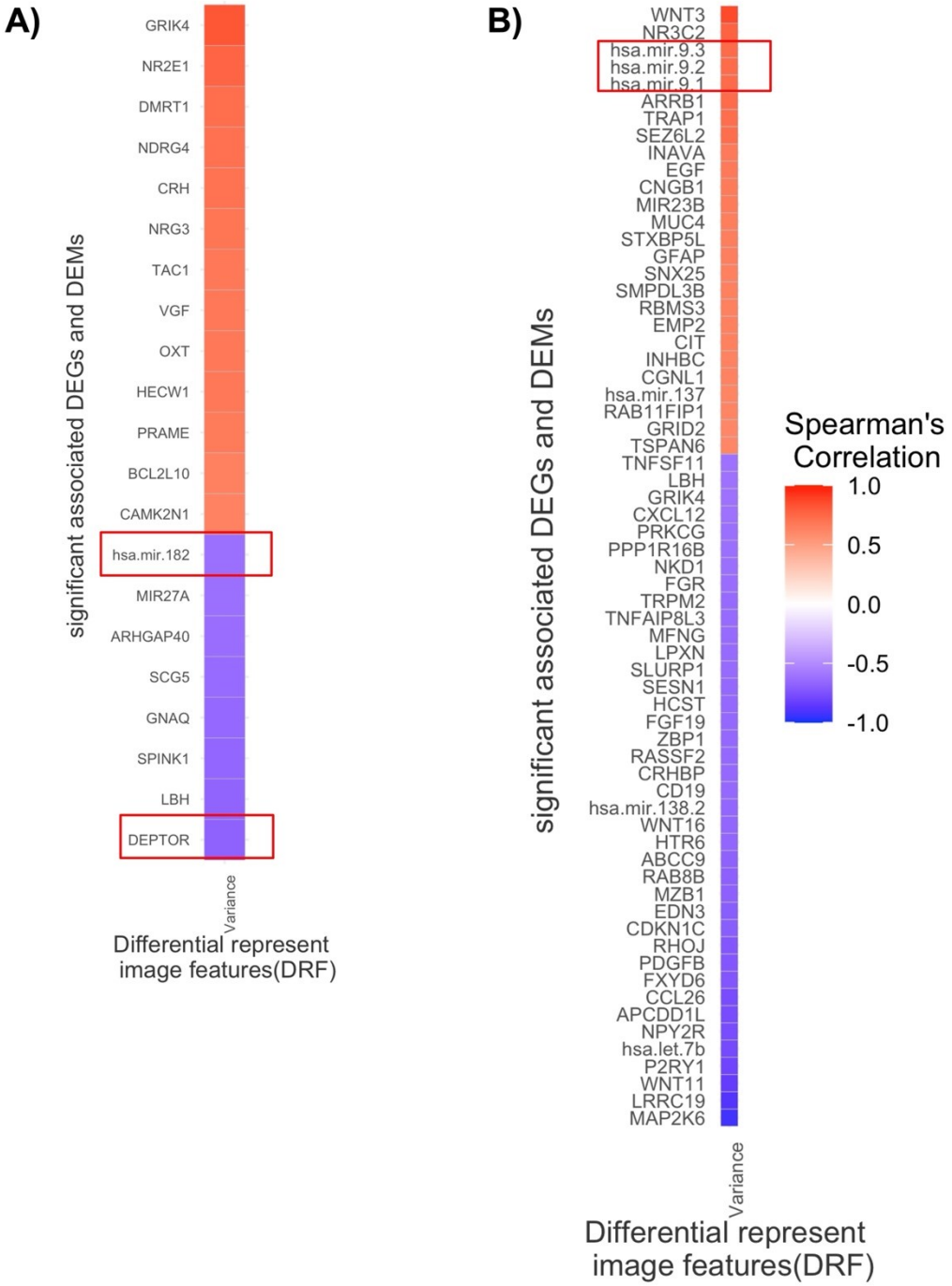


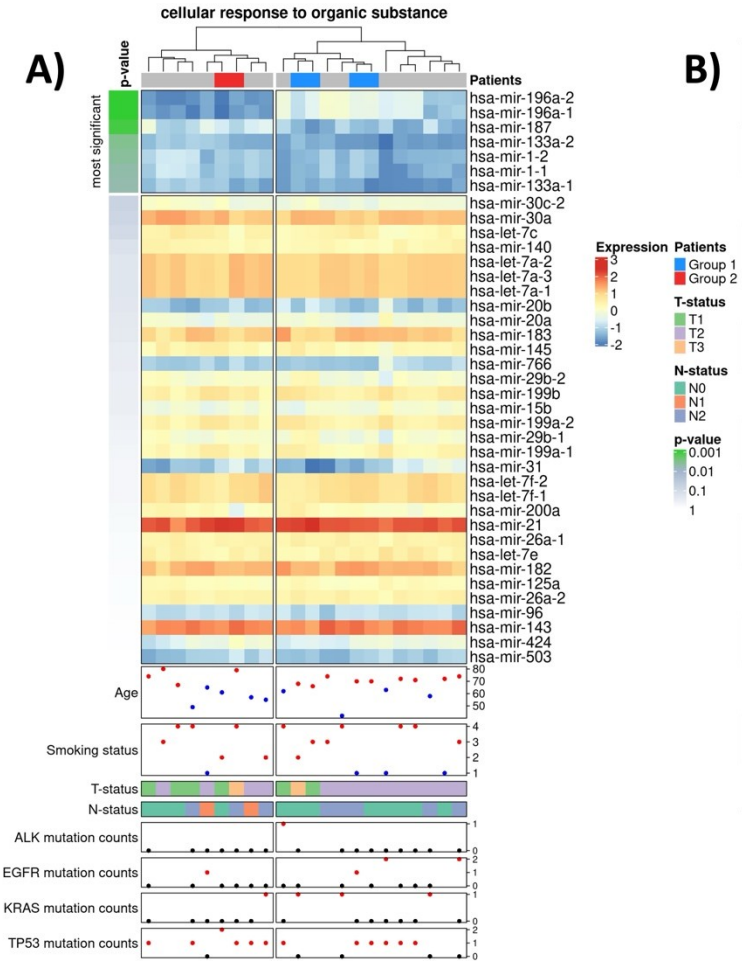
Figure 10 Spearman correlation coefficient between Variance and gene/miRNA expression of “regulation of signaling”

A) Spearman correlation coefficient ($p<.05$) between the DRF Variance and gene/miRNA expression within the term “regulation of signaling” of patient cluster 1. B) Spearman correlation coefficient ($p < 0.05$) between the DRF Variance and gene/miRNA expression within the term “regulation of signaling” of patient cluster 2. Red boxes highlight confirmation from other studies. Adapted from Spath and Fischer et al. [43]

4.3.4 CELLULAR RESPONSE TO ORGANIC SUBSTANCE

The GSEA yielded the most associations of image features and transcriptomics with biological processes encompassing response to organic substances. This is in concordance with the leading risk factor of lung cancer, tobacco smoking. It contains carcinogenic substances such as organic cyclic compounds. [4] Texture features correlated with the gene signature are mainly negative, while significant correlations with the miRNA signature are primarily positive (Figure 6). Clustering patients based on the miRNA expression revealed differences in texture features such as *homogeneity, contrast and coarseness* (Figure 11B). Patients exposed to organic substances seem to have dysregulated miRNAs critical to tumor heterogeneity in CT images (Figure 11A).

A)



B)

DRF	Fold Change	P-value
SZLGE	0.497	0.008
Complexity	0.605	0.014
Strength	0.369	0.016
Coarseness	0.318	0.018
RP	0.895	0.022
LRE	1.664	0.025
dvarh	0.419	0.029
contr	0.419	0.029
homom	1.222	0.031
SRE	0.939	0.032
homop	1.309	0.033
dissi	0.63	0.033
idmnc	1.01	0.034
indnc	1.029	0.035

Figure 11 Radiogenomic association map (RAM) of “cellular response to organic substance” of the gene expression

Radiogenomic association map (RAM) of miRNA expression of the DEM data set ordered by p-value A) for the top 50 miRNAs in the term “cellular response to organic substance”. This RAM showcases seven miRNAs with a $p < 0.05$ out of 39 miRNAs used for clustering (18%). The greyed-out patient's annotation marks different clusters from the DEM data set. Smoking status is presented as: 1-Lifelong Non-Smoker, 2-Current Smoker, 3-Current Reformed Smoker for more than 15 years, 4-Current Reformed Smoker for less than 15 years. Mutation status does not seem to have a visible pattern. B) Differentially represented image features all belong to the texture category. Texture features cannot be evaluated by the human eye. Adapted from Spath and Fischer et al. [43]

Patients' gene expression signatures of the two clustered groups show differences in tumor morphology (Figure 12C). Notably, the overall positive correlation between inflammatory activity and previous exposure to organic cyclic compounds concord with tumor heterogeneity. As with the biological process "regulation of signaling" clinical data, tumor stage or mutation status between the two patient groups does not adhere to a visible pattern.

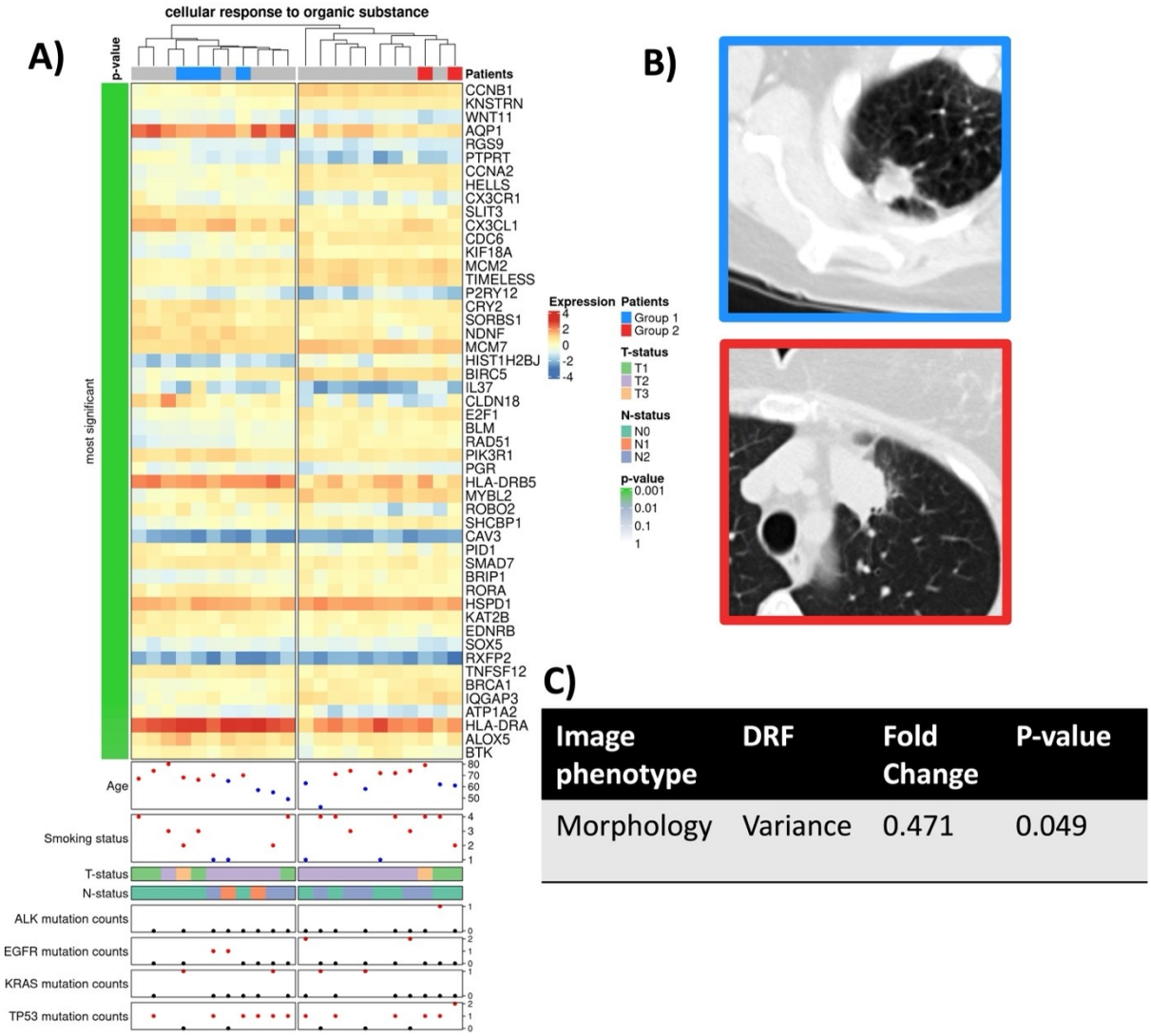


Image phenotype	DRF	Fold Change	P-value
Morphology	Variance	0.471	0.049

Figure 12 Radiogenomic association map (RAM) of "cellular response to organic substance" of the miRNA expression

Radiogenomic association map (RAM) of mRNA expression of the DEG data set ordered by p-value A) for the top 50 mRNAs in the term "cellular response to organic substance". It results from hierarchical clustering of 794 genes, from which 246 have a $p < .05$ (31%). The greyed-out patient's annotation marks different clusters from the DEM data set. Smoking status is presented as: 1-Lifelong Non-Smoker, 2-Current Smoker, 3-Current Reformed Smoker for more than 15 years, 4-Current Reformed Smoker for less than 15 years. Mutation status does not seem to have a visible pattern. B) CT images to showcase the radiologic imaging of the two clustered groups. The CT window width is at 1500, as the level is -600 for all images presented. "Variance" cannot be evaluated by the human eye. C) The differentially represented image feature is Variance with a $p < 0.05$. Adapted from Spath and Fischer et al. [43]

The correlation analysis supports previous publications (Figure 13). In both patient clusters, *ONECUT2* shows different correlation patterns with texture image features. *ONECUT2* is a

lineage-specific and context-dependent oncogene in lung adenocarcinoma. [84] Moreover, it is a potential therapeutic target for these tumors. In patient cluster two, *FOXH1* positively correlates with *Strength* and *Coarseness*. It promotes lung cancer progression or tumor growth and metastasis via regulation of the *WNT/beta catenin* signaling pathway. [85]

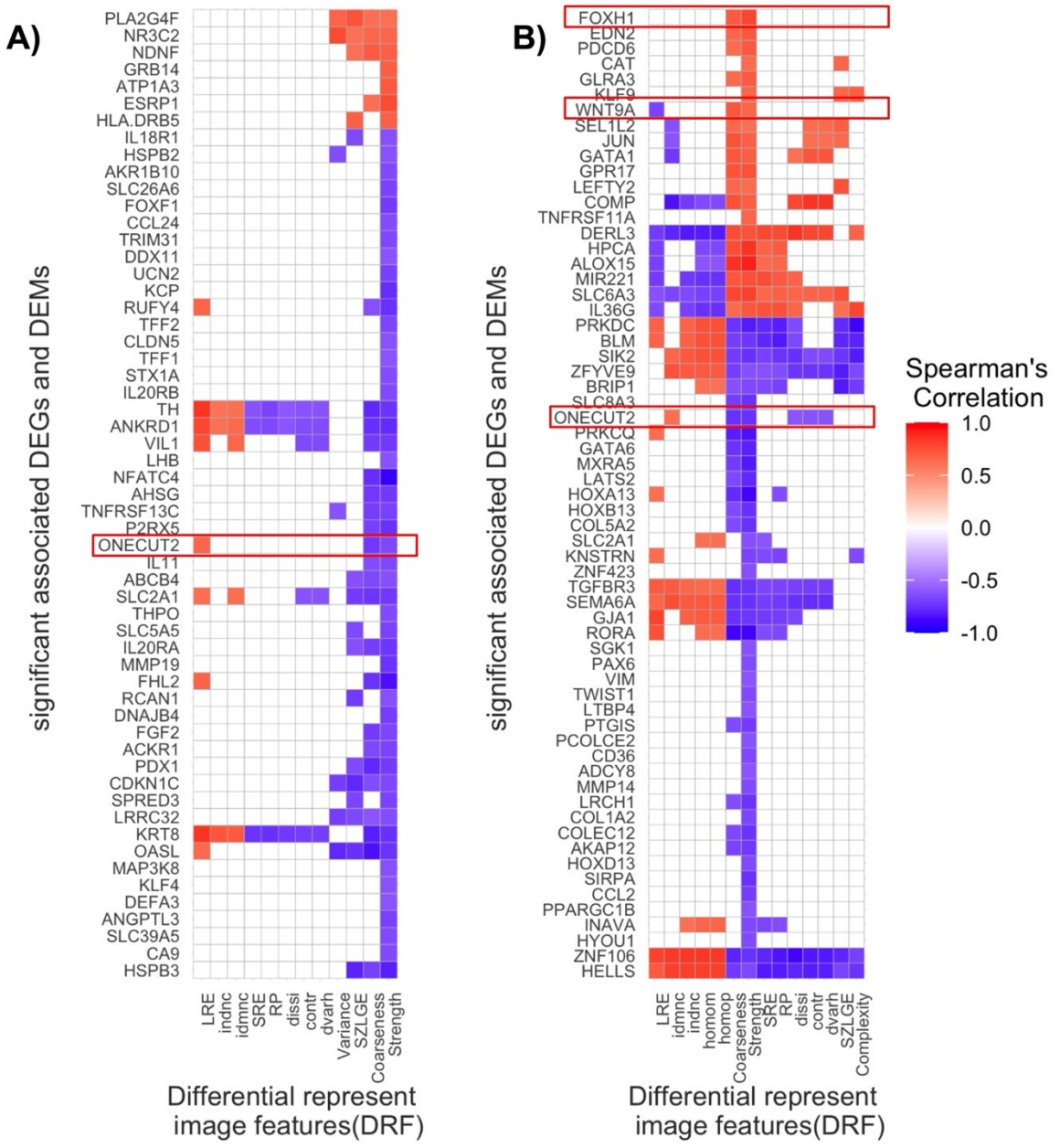


Figure 13 Spearman correlation coefficient between the DRFs and gene/miRNA of “cellular response to organic substance”

A) Spearman correlation coefficient (p<.05) between the DRFs and gene/miRNA expression within the term “cellular response to organic substance” of patient cluster one. B) Spearman correlation coefficient ($p < 0.05$) between the DRFs and gene/miRNA expression within the term “cellular response to organic substance” of patient cluster two. Red boxes highlight confirmation from other studies. Adapted from Spath and Fischer et al. [43]

4.3.5 GENE REGULATORY NETWORKS

To explain phenotypic differences within the two examined biological processes construction of a transcription factor-miRNA regulatory network was carried out. It combines transcriptional and post-transcriptional interactions between DEGs and DEMs (Figure 14).

The molecular interactions can be defined in three distinct types: 1.) transcription factor -> target gene, 2.) transcription factor -> target miRNA, and 3.) miRNA -> target gene. The main hub genes contribute the most to the regulation of the network. "*Regulation of signaling"* presented two main hub genes: *TAL1* and *TGFBR2* (Figure 14A).

Additionally, the downregulated transcript *TGFBR2* was identified as the main hub gene in the "*cellular response to organic substances*" term (Figure 14B). *TGFBR2* is annotated as a tumor suppressor gene. [86] The regulatory network shows a direct regulation of *TGFBR2* through *TAL1,* which is associated with lung carcinoma, and in this case, lung-specific. Previous work describes *TAL1* as a regulator of normal myeloid differentiation and is a potential target in treating T-cell acute lymphoblastic leukemia. [87]

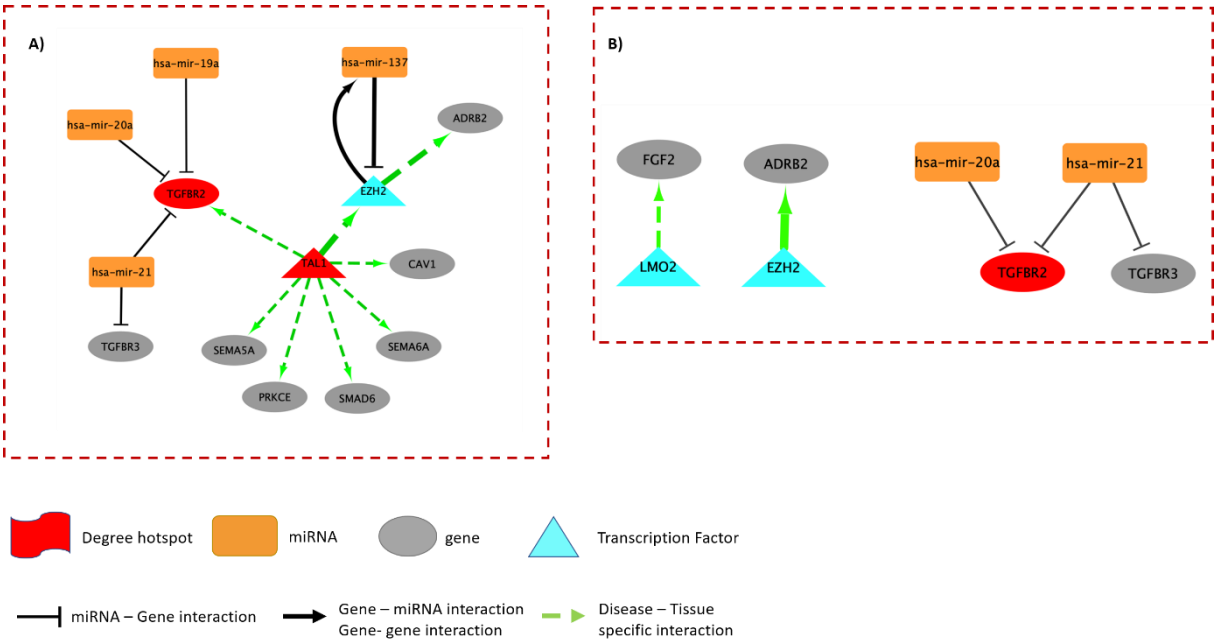


Figure 14 Gene regulatory network (Tfmir2) of the biological processes "regulation of signaling" and "cellular response to organic substance

A) "regulation of signaling" and B) "cellular response to organic substance". TAL1 and TGFBR2 are regulatory hotspots based on the degree of centrality of the edges. Green arrows show tissue specificity for non-small cell lung carcinoma. Directly influenced target genes are grey, whereas orange presents miRNAs that act as inhibitors. TGF-beta receptor types 2 and 3 are both inhibited by hsa-mir-21. Taken from Spath and Fischer et al. [43]

5 DISCUSSION

5.1 CONSTRUCTING A RADIOGENOMIC ASSOCIATION MAP (RAM)

Previous studies showed associations of radiomic signatures with underlying genomic processes in lung adenocarcinoma. [42] This is especially interesting as image features could possibly serve as non-invasive biomarkers providing enough diagnostic and therapeutic value in case biopsies are not feasible or when questions arise about untypical (nonmalignant) nodules of the lung. A diagnosis without biopsy and histopathology requires detailed knowledge of how molecular interactions define image phenotypes to deduce the underlying genomic information. The proposed data-driven pipeline to mine large genomic and imaging datasets can reveal associations of image traits with underlying biological processes regulated by gene expression, miRNA expression and methylation patterns in so-called Radiogenomic Association Maps (RAMs). Various studies have showcased the correlation of mutation status of *EGFR, KRAS* or *ALK/ROS1* translocation with radiomic features, showing prognostic, diagnostic or therapeutic value and thus having a potential clinical application. [88], [89], [90] The lack of data on the larger genomic background narrows these analyses to only a few image-phenotype-mutation correlations. Clinical imaging techniques (CT/MRI/PET) provide a thorough 3-dimensional characterization of tumor nodules and the phenotype of the surrounding area. This representation can offer more information on the whole tumor and its environment than localized biopsy-based tumor slices. [41], [42] Associations of signatures of imaging phenotypes with biological processes on the molecular level further add knowledge to underlying gene and miRNA expression, focusing mainly on the influence on image phenotypes. [91] As an extension of existing radiologic tools, this could have a direct clinical impact when the analysis would be available in real-time, adding to existing approaches to individualized medicine, like mutation analysis from biopsies. The proposed custom approach to radiogenomic association maps aims to improve the interpretability of the results of large-scale radiogenomic studies and highlight interrelationships between molecular data and image data by further visualization. Systematic correlation between genomic and imaging data following this pipeline grants an overview of the image feature association with genomics large picture of lung adenocarcinoma. Furthermore, the method is generalized and reusable for routine applications and studies of other tumor entities or diseases. When the understanding of imaging and molecular interactions deepens, it could provide clinicians with a more detailed diagnosis and treatment options for patients in the field of individualized medicine.

However, the most significant limitations of this study are the available resources. Complete open datasets are rare; thus, this study's cohort size is relatively small. In addition,

standardization problems arise from the imaging data, originating in the acquisition from different institutes using different CT scanners. This results in CT scans with different parameters, like reconstruction algorithms and slice thickness, although intravenous contrast enhancement was always present in the dataset. Another issue are manually segmented ROIs. This leaves further bias and requires robust automatic segmentation to eliminate and make this study reproducible. Therefore, validating the results using a prospective study design to minimize the impact of systematic errors is warranted. Finally, it must be added that the findings of this study do not derive causal associations of radiomic features with biological processes on a molecular level.

5.2 EPIGENETIC MODIFICATIONS OF HOX GENES MAY BE REPRESENTED IN CT IMAGING

Differential methylation of *HOX* genes is widespread in a wide variety of cancer types. They are organized in 4 clusters *A-D* and their encoded proteins may act as proto-oncogenes or tumor suppressors. Their hyper- or hypomethylation leads to tissue-specific effects, including apoptosis, angiogenesis, and proliferation. [92] This work finds a similar molecular signature to previously described *HOX* genes in lung adenocarcinoma. *HOX* genes have been part of molecular signatures describing tumor evolution [93] and future biomarkers in LUAD. [94] *HOXA3* is involved in molecular mechanisms underlying cancer progression. [95] Since methylation modifications are reversible by DNA-methyltransferases, they make an attractive pharmacological target with drugs like Hydralazin, which are already in use for other human diseases. [92] This work adds to these already-known DNA methylation modifications of *HOX* genes and new insights into their representation in CT imaging. A set of image traits has been described that is associated with biological processes mainly containing housekeeping genes. *HOXA3, HOXB3* and *HOXB4* aberrant methylation in lung adenocarcinoma are a part of the developmental biological processes described. There seem to be quantitative image features that are either positively or negatively correlated to the enriched GO terms. Texture features mainly seem to represent the imaging phenotype although size, shape and morphologic traits are also represented. As the texture features are calculated from different matrices, it is noticeable that overrepresented associated quantitative textures have the *Gray-Level Co-Occurence Matrix* by Haralick et al. [67] as a basis for calculation. These are known to be robust to bias from different methods and data quality and have already been featured in various publications in radiologic image analysis. [67], [89], [96], [97] On the other hand, texture features that originate from all four different matrix calculations (*Gray-Level Co-Occurrence Matrix, Gray-Level Run-Length Matrix, Gray-Level Size Zone Matrix, Neighborhood Gray-Tone Difference Matrix*) are present with positive and negative associations too. In conclusion, the set of image traits has a wide variety of features which adds to its robustness, though this

needs to be further evaluated with a larger independent cohort. With this newly described image trait-methylation link, it might be possible to acquire methylation states of tumor nodes in LUAD by calculating quantitative image features. Combined with (semi)automatic segmentation of suspected cancerous nodes in the lung, these image features could contribute significantly to predicting patients' individual survival or treatment options without extensive molecular profiling.

5.3 BIOLOGICAL PROCESSES ARE CORRELATED WITH SPECIFIC IMAGE PHENOTYPES

The results of this study support the hypothesis that common biological processes tend to associate with image features of specific categories. The gene sets originating from the differentially expressed genes link with all four categories described but mainly reside with the texture group. As the expressed genes largely contribute to the phenotype of a tumor, it is expected that such a broad representation in imaging phenotypes exists. For future studies, a further split into genes involved in pathways with mutation driver genes would define a more specific association of image categories and (dys)regulated pathways.
As for the miRNAs, there is an apparent adherence to the texture category. Regulatory miRNAs are crucial in defining tumor heterogeneity in CT imaging represented by the texture category. Measuring texture features as representatives of heterogeneity may allow the classification of a tumor’s miRNA expression to find vital regulatory pathways through radiologic imaging non-invasively.

With these newly described mRNA or miRNA expression associations with image categories, this study provides a further step into a possible generalization and application in the clinical context. Previously not described adherences of underlying biological processes to image trait categories add new perspectives on how tumor heterogeneity and, thus, intratumoral somatic evolution can be detected using well-known quantitative features. As every biopsy only contains a small part of the whole tumor, it also only includes a sample of the genes expressed, leaving tumor heterogeneity as a clear barrier for individualized treatment. This could be overcome by three-dimensional calculation of image biomarkers, including every part of the tumor, but only if clinically validated radiomic signatures allow discrete evaluation of the underlying tumor genomics in the future.

5.4 CELL CYCLE PROCESSES MAY HAVE A SIGNIFICANT IMPACT ON IMAGE PHENOTYPES

The predominance of image traits correlated with genes of cell cycle processes may help predict cancer proliferation activity, evaluate drug efficacy, and determine treatment failure. For example, drugs like Abemaciclib or Erlotinib (cyclin-dependent kinase inhibitor and EGFR inhibitor) may show future treatment options in the late stages of lung cancer. [98] Even though a recent phase III clinical study (JUNIPER) comparing Abemaciclib and Erlotinib in NSCLC did not meet its primary endpoint of improved overall survival, it met secondary endpoints in overall response rate and progression-free survival. [99] Thus, these drugs might enable clinicians to offer more variable and safe treatment options in the late stages of lung cancer, lowering the dose of classical platin adjacent chemotherapies to reduce side effects. With the association of image features to the biological processes involving the cell cycle, it might be possible to use CT imaging to predict and monitor the therapy efficacy of CDK inhibitors without invasive biopsies and molecular profiling.

5.5 THE ROLE OF SYSTEMIC AND LOCAL INFLAMMATION IN LUNG CANCER

Numeral studies measure systemic inflammation as a predictor of patient outcome. A common score is the advanced lung cancer inflammation index (ALI). The original study evaluated this score which contains body mass index, serum albumin and the neutrophil-lymphocyte ratio at the hand of metastatic NSCLC. [100] Later studies showed prognostic value in stage IA, typically treated with surgical resection, as a predictor of recurrence. [101] In other tumor entities like colorectal cancer, it has been proposed as a prognostic indicator of clinical outcomes. [102] With this in mind, systemic inflammation clearly correlates with cancer patients' outcomes.

The role of intratumoral inflammation activity remains mainly unknown in lung adenocarcinoma. One study follows lung adenocarcinoma evolution by determining co-mutation events in KRAS-mutant lung cancer. Moreover, they describe immune heterogeneity as a distinct driver for intratumoral development events, either aborting tumor proliferation or contributing to its progress. [103] In addition, the TNF alpha axis seems to be involved in MHC-II expression in alveolar type II cells from which inflammation-driven lung adenocarcinoma may originate. [104] In this micro-inflammatory environment, MHC-II overexpression triggers tumor expansion and thus TNF-α contributes as a factor for pro-cancer activity in this specific case. Another study tried to find radiomic features which measure intratumoral heterogeneity and relate these features to histologic images. [105] Interestingly, these iodine contrast-enhanced

CT images showed less significant correlations with the pathologic heterogeneity index than non-contrast-enhanced images.

5.6 INFLAMMATORY CYTOKINE-DEPENDENT PROCESSES MAY IMPACT CT IMAGE TEXTURE

As presented in the results, biological processes that involve cytokine-mediated signaling often correlate with texture features. In these processes, miRNAs like *mir-9* play a key role. Thus, it is imaginable to measure intratumoral inflammatory activity with CT imaging. Furthermore, the known mechanisms of immune evasion of PD-L1 positive lung cancer and their treatment with PD-L1 inhibitors like Pembrolizumab or Durvalumab could be measured using the described texture traits (e.g., *RLV, ZSV, autocorrelation, LGRE*). However, another independent study is warranted to provide more information on this hypothesis. This would be very useful to monitor treatment and could contribute as low-dose CT to prospective studies evaluating the treatment failure of drugs like these non-invasively.

5.7 SUBCLASSIFICATION OF THE PATIENT COHORT REVEALS THE IMPORTANCE OF TUMOR MORPHOLOGY

As mentioned in previous paragraphs, the classification of drug-sensitizing mutations and prediction of overall survival or tumor stage are common applications of radiogenomics. The goal of the stratification of the patient cohort in this study is to determine differentially represented image features (DRFs), which can be measured to classify patients into two groups based on a common biological process. This would reveal slight differences in the regulation of one process and deepen the understanding of why certain patients, even though they have the same tumor grade and mutation status, still have different overall outcomes. If this could be determined in CT imaging, it would be crucial in treatment adjustments to tackle more aggressive cancer. However, this hypothesis would need to be confirmed by a survival analysis with Kaplan Meier plots of these patients. The relevance of the results for this approach is consolidated by the significant correlations of the found DRFs and gene or miRNA expression of the biological processes "*regulation of signaling*" and "*cellular response to organic substance*".

5.7.1 REGULATION OF SIGNALING

Interestingly, patient classification based on gene and miRNA expression involved in the "*regulation of signaling*" exhibits the same image feature. "*Variance*" as DRF reveals correlations with known genes and miRNAs, which play a role in tumor development and as

drug targets. This would hint at how image traits can be used to find regulating mechanisms within a specific subset of biological processes. The correlations of "*Variance"* with genes like *DEPTOR* or miRNAs like *hsa-mir-182 and hsa-mir-9* have already been described in lung cancer, some with a possible impact as drug targets, have the potential to serve as patient classifiers. [80], [81], [83] Currently, *Variance* as a proposed imaging surrogate for gene expression is preliminary and unconfirmed by an independent data set. Due to the small cohort, there has not been any validation. A future vision might be to take medical images of patients, cluster them based on the known DRFs and then predict their gene or miRNA expression. Generally, studies involving high-throughput omics experiments are expensive. While CT imaging is often available in the clinical context aside from studies, high-throughput omics measurements are not performed. Narrowing down the needed genes to sequence to answer a specific question in radiogenomics can make them much more feasible. Thus, low-cost experiments can validate the findings in this work when only determining the described DEGs or DEMs.

5.7.2 CELLULAR RESPONSE TO ORGANIC SUBSTANCE

As tobacco smoke contains organic substances and leads to lung cancer [4] it is unsurprising to see enrichment for genes and miRNAs in the tumor tissue of smoking patients. The results show that CT image phenotypes reflect this, too: Most dysregulated genes in this biological process correlate negatively to image texture features, while miRNAs correlate positively. Consequently, it should be expected that clustering patients based on these genes and miRNAs would yield two groups of patients, one who smokes heavily and the other one who does not or never has smoked. However, this was not the case. An explanation could be that almost all patients in this cohort have at least smoked tobacco continuously for over 5-10 years. Interestingly, clustering patients based on the gene set of "*cellular response to organic substance*" shows differences in "*Variance"* too. It may seem that tumor morphology defines the imaging phenotype and changes according to the expression of various genes in different biological processes. In this case, the biological processes of *regulation of signaling, cellular response to organic substance, regulation of cell differentiation, negative regulation of developmental process, response to organic substance, cellular response to chemical stimulus, regulation of cell population proliferation, cell population proliferation* contribute to tumor morphology. It further supports "*Variance"* as the proposed image surrogate of differential gene expression with the restrictions mentioned earlier due to lack of validation.

It is difficult to measure the inflammatory microenvironment of the tumor using laboratory techniques. The results demonstrate evidence that organic cyclic compounds (smoking)

correlate with inflammatory activity and tumor texture. Moreover, the miRNAs: *hsa-mir-196a, hsa-mir-187, hsa-mir-133a* and *hsa-mir-1* could be regulators for tumor heterogeneity and differentiate between patient groups. An application could be the measurement of inflammation in the tumor microenvironment, enabling a prediction of treatment response before drug application.

5.8 TAL1 AND TGFBR2 ARE REGULATORY HOTSPOTS

The gene regulatory network shows common regulatory mechanisms in the biological processes "*regulation of signaling"* and "*cellular response to organic substance"*. The signaling axis involving the inhibition of *TGFBR2* by *EZH2* has already been confirmed experimentally. [86] This dysregulation results in tumor growth mediated by a cluster of miRNAs (*miR-25, 93 and 106b*). Although this GRN cannot confirm these miRNAs, it unveils another set of miRNAs (*hsa-mir-19a*, *20a* and *21*) which might act in addition to the previously described cluster. Moreover, the network suggests stimulation of *TGFBR2* by *TAL1* in lung cancer. There is no experimental evidence as of now that can support this interaction. This computational result could be further investigated in an experimental laboratory approach to confirm or discard its relevance.

The enhanced transcription of *ADRB2* by the *TAL1-EZH2* axis is visualized in the network graph. It is a target gene that encodes for beta-adrenoreceptors. The controversial discussion of its importance for angiogenesis, proliferation, TKI resistance and distant metastasis has not yet been resolved. [106] The visualized gene regulatory network cannot differentiate if *ADRB2* is a bystander target not involved as a tumor driver or plays a crucial role in lung cancer progression.

5.9 CONCLUSION

In conclusion, genetic data sets are represented in CT imaging and link to specific biological processes leveraging a multi-omics integration GSEA. Depending on the underlying data, this has a possible meaning for clinical and scientific purposes. These associations have not been previously described and broaden the horizon of knowledge of radiogenomics in lung adenocarcinoma, giving new insights into the association of CT imaging underlying genetics, providing answers, raising further questions and paving the way to move procedures like this pipeline into direct clinical use.

Furthermore, the proposed data-driven pipeline can be as such utilized in a wide variety of diseases, given the data availability. It can help to investigate and discover radiogenomics in other research fields. This generalization provides another step for cross-validation and enhances the interpretability of results across multiple studies.

Some results concur with previous studies, while others require confirmation with a large independent cohort. However, the interpretation of complex high-dimensionality data sets remains an issue as often overgeneralization and overfitting of the data due to unjustified methods occur. Moreover, applying complex pipelines like this to a small patient cohort may result in poor performance when applied to another cohort. This highlights the need to revalidate the herein proposed radiogenomic association maps again.

6 THESES

In this study, the following phrases and questions are posed, investigated and discussed:

1. With proper visualization, it may be possible to integrate different omics data sets on a descriptive level and develop a bioinformatics pipeline to build radiogenomic association maps for large-scale omics experiments.
2. Lung cancer exhibits enriched gene sets that correlate to certain CT image traits. This allows the description of the underlying transcriptome/methylome utilizing non-invasive CT imaging.
3. Classifying lung cancer patients according to their transcriptome and mutations yields different subgroups which CT image traits can differentiate.
4. The construction of gene regulatory networks from gene ontology terms describes the relationship between genes and miRNAs in lung cancer.

In addition, new questions arise when those findings prove to be different or conflict with past studies or curated knowledge.

7 REFERENCES

[1] H. Sung *et al.*, “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/CAAC.21660.

[2] M. C. Turner *et al.*, “Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations,” *CA Cancer J Clin*, vol. 70, no. 6, pp. 460–479, Nov. 2020, doi: 10.3322/CAAC.21632.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA Cancer J Clin*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[4] IARC Working Group on the Evaluation of Carcinogenic Risks to Humans., World Health Organization., and International Agency for Research on Cancer., *Tobacco Smoke and Involuntary Smoking*. IARC Press, 2004.

[5] S. G. Spiro, M. K. Gould, and G. L. Colice, “Initial evaluation of the patient with lung cancer: Symptoms, signs, laboratory tests, and paraneoplastic syndromes. ACCP evidenced-based clinical practice guidelines (2nd edition),” *Chest*, vol. 132, no. 3 SUPPL., pp. 149S-160S, 2007, doi: 10.1378/chest.07-1358.

[6] D. K. Leitlinienprogramm Onkologie (Deutsche Krebsgesellschaft and AWMF):, “Prävention, Diagnostik, Therapie und Nachsorge des Lungenkarzinoms, Langversion 1.0,” *AWMF-Registernummer: 020/007OL, http://leitlinienprogramm-onkologie.de/Lungenkarzinom.98.0.html*, 201AD.

[7] K. M. Latimer and T. F. Mott, “Lung cancer: Diagnosis, treatment principles, and screening,” *Am Fam Physician*, 2015.

[8] P. Goldstraw *et al.*, “The IASLC lung cancer staging project: Proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM Classification for lung cancer,” *Journal of Thoracic Oncology*, vol. 11, no. 1, pp. 39–51, 2016, doi: 10.1016/j.jtho.2015.09.009.

[9] K. Chansky *et al.*, “The IASLC Lung Cancer Staging Project: External Validation of the Revision of the TNM Stage Groupings in the Eighth Edition of the TNM Classification of Lung Cancer,” *Journal of Thoracic Oncology*, vol. 12, no. 7, pp. 1109–1121, Jul. 2017, doi: 10.1016/j.jtho.2017.04.011.

[10] W. D. Travis *et al.*, “The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances since the 2004 Classification,” *Journal of Thoracic Oncology*, vol. 10, no. 9. Elsevier Inc, pp. 1243–1260, Sep. 26, 2015. doi: 10.1097/JTO.0000000000000630.

[11] T. C. G. A. R. Network, “Comprehensive molecular profiling of lung adenocarcinoma,” *Nature*, vol. 511, no. 7511, pp. 543–550, Jul. 2014, doi: 10.1038/nature13385.

[12] J. Mazières *et al.*, “Lung cancer that harbors an HER2 Mutation: Epidemiologic characteristics and therapeutic perspectives,” *Journal of Clinical Oncology*, vol. 31, no. 16, pp. 1997–2003, Jun. 2013, doi: 10.1200/JCO.2012.45.6095.

[13] M. Reck *et al.*, “Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer,” *N Engl J Med*, vol. 375, no. 19, pp. 1823–1833, 2016, doi: 10.1056/NEJMoa1606774.

[14] M. Soda *et al.*, “Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer,” *Nature*, vol. 448, no. 7153, pp. 561–566, Aug. 2007, doi: 10.1038/nature05945.

[15] J. J. Lin and A. T. Shaw, “Recent Advances in Targeting ROS1 in Lung Cancer,” *Journal of Thoracic Oncology*, vol. 12, no. 11. Elsevier Inc, pp. 1611–1625, Nov. 01, 2017. doi: 10.1016/j.jtho.2017.08.002.

[16] E. A. Collisson *et al.*, “Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network,” *Nature*, vol. 511, no. 7511, pp. 543–550, Jul. 2014, doi: 10.1038/nature13385.

[17] M. L. Johnson *et al.*, “Durvalumab With or Without Tremelimumab in Combination With Chemotherapy as First-Line Therapy for Metastatic Non-Small-Cell Lung Cancer: The Phase III POSEIDON Study,” *J Clin Oncol*, vol. 41, no. 6, pp. 1213–1227, Feb. 2023, doi: 10.1200/JCO.22.00975.

[18] A. T. Shaw *et al.*, “Crizotinib versus chemotherapy in advanced ALK-positive lung cancer,” *N Engl J Med*, vol. 368, no. 25, pp. 2385–2394, 2013, doi: 10.1056/NEJMoa1214886.

[19] J. Mazéres *et al.*, “Lung cancer patients with HER2 mutations treated with chemotherapy and HER2-targeted drugs: Results from the European EUHER2 cohort,” *Annals of Oncology*, vol. 27, no. 2, pp. 281–286, Feb. 2016, doi: 10.1093/annonc/mdv573.

[20] M. Maemondo *et al.*, “Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR,” *N Engl J Med*, vol. 362, no. 25, pp. 2380–2388, 2010, doi: 10.1056/NEJMoa0909530.

[21] S. Bagcchi, “Lung cancer survival only increases by a small amount despite recent treatment advances,” *The Lancet. Respiratory medicine*, vol. 5, no. 3. Lancet Respir Med, p. 169, Mar. 01, 2017. doi: 10.1016/S2213-2600(17)30041-3.

[22] E. Martínez Pérez *et al.*, “Lung Cancer Screening: Use of Low-Dose Computed Tomography,” *Archivos de Bronconeumología (English Edition)*, vol. 55, no. 10, pp. 526–531, 2019, doi: 10.1016/j.arbr.2019.08.001.

[23] H. J. de Koning *et al.*, “Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial,” *N Engl J Med*, vol. 382, no. 6, pp. 503–513, Feb. 2020, doi: 10.1056/NEJMOA1911793.

[24] National Lung Screening Trial Research Team, “Lung Cancer Incidence and Mortality with Extended Follow-up in the National Lung Screening Trial,” *J Thorac Oncol*, vol. 14, no. 10, pp. 1732–1742, Oct. 2019, doi: 10.1016/J.JTHO.2019.05.044.

[25] M. Hong *et al.*, “RNA sequencing: new technologies and applications in cancer research,” *Journal of Hematology and Oncology*, vol. 13, no. 1. BioMed Central Ltd, Dec. 01, 2020. doi: 10.1186/s13045-020-01005-x.

[26] S. Dedeurwaerder, M. Defrance, M. Bizet, E. Calonne, G. Bontempi, and F. Fuks, “A comprehensive overview of Infinium Human Methylation450 data processing,” *Brief Bioinform*, vol. 15, no. 6, pp. 929–941, Aug. 2013, doi: 10.1093/bib/bbt054.

[27] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks, “Evaluation of the Infinium Methylation 450K technology,” *Epigenomics*, vol. 3, no. 6, pp. 771–784, Dec. 2011, doi: 10.2217/epi.11.105.

[28] A. McKenna *et al.*, “The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data,” *Genome Res*, vol. 20, no. 9, pp. 1297–1303, Sep. 2010, doi: 10.1101/gr.107524.110.

[29] D. C. Koboldt *et al.*, “VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing,” *Genome Res*, vol. 22, no. 3, pp. 568–576, Mar. 2012, doi: 10.1101/gr.129684.111.

[30] S. Kim *et al.*, “Strelka2: fast and accurate calling of germline and somatic variants,” *Nat Methods*, vol. 15, no. 8, pp. 591–594, Aug. 2018, doi: 10.1038/s41592-018-0051-x.

[31] L. Neums *et al.*, “VaDiR: An integrated approach to Variant Detection in RNA,” *Gigascience*, vol. 7, no. 2, pp. 1–13, Feb. 2018, doi: 10.1093/gigascience/gix122.

[32] L. De Schaetzen Van Brienen *et al.*, “Comparative analysis of somatic variant calling on matched FF and FFPE WGS samples,” *BMC Med Genomics*, vol. 13, no. 1, Jul. 2020, doi: 10.1186/s12920-020-00746-5.

[33] Z. Chen *et al.*, “Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency,” *Sci Rep*, vol. 10, no. 1, pp. 1–9, Dec. 2020, doi: 10.1038/s41598-020-60559-5.

[34] M. Viceconti, P. Hunter, and R. Hose, “Big Data, Big Knowledge: Big Data for Personalized Healthcare,” *IEEE J Biomed Health Inform*, vol. 19, no. 4, pp. 1209–1215, Jul. 2015, doi: 10.1109/JBHI.2015.2406883.

[35] C. Yang, C. Li, Q. Wang, D. Chung, and H. Zhao, “Implications of pleiotropy: Challenges and opportunities for mining Big Data in biomedicine,” *Frontiers in Genetics*, vol. 6, no. JUN. Frontiers Media S.A., 2015. doi: 10.3389/fgene.2015.00229.

[36] M. Greaves, “Evolutionary determinants of cancer,” *Cancer Discovery*, vol. 5, no. 8. American Association for Cancer Research Inc., pp. 806–821, Aug. 01, 2015. doi: 10.1158/2159-8290.CD-15-0439.

[37] M. Jamal-Hanjani *et al.*, “Tracking the Evolution of Non–Small-Cell Lung Cancer,” *New England Journal of Medicine*, vol. 376, no. 22, pp. 2109–2121, Jun. 2017, doi: 10.1056/nejmoa1616288.

[38] R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: Images are more than pictures, they are data,” *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016, doi: 10.1148/radiol.2015151169.

[39] P. Lambin *et al.*, “Radiomics: The bridge between medical imaging and personalized medicine,” *Nature Reviews Clinical Oncology*, vol. 14, no. 12. Nature Publishing Group, pp. 749–762, Dec. 01, 2017. doi: 10.1038/nrclinonc.2017.141.

[40] A. Zwanenburg *et al.*, “The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping,” *Radiology*, vol. 295, no. 2, pp. 328–338, May 2020, doi: 10.1148/RADIOL.2020191145/ASSET/IMAGES/LARGE/RADIOL.2020191145.FIG5.JPEG.

[41] H. J. W. L. Aerts *et al.*, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nat Commun*, vol. 5, no. 1, pp. 1–9, 2014, doi: 10.1038/ncomms5006.

[42] P. Grossmann *et al.*, “Defining the biological basis of radiomic phenotypes in lung cancer,” *Elife*, vol. 6, 2017, doi: 10.7554/eLife.23421.

[43] S. Fischer, N. Spath, and M. Hamed, “Data-Driven Radiogenomic Approach for Deciphering Molecular Mechanisms Underlying Imaging Phenotypes in Lung Adenocarcinoma: A Pilot Study,” *International Journal of Molecular Sciences 2023, Vol. 24, Page 4947*, vol. 24, no. 5, p. 4947, Mar. 2023, doi: 10.3390/IJMS24054947.

[44] “The Cancer Genome Atlas - Timeline and Milestones - National Cancer Institute.” Accessed: Mar. 31, 2021. [Online]. Available: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/timeline

[45] K. Clark *et al.*, “The cancer imaging archive (TCIA): Maintaining and operating a public information repository,” *J Digit Imaging*, 2013, doi: 10.1007/s10278-013-9622-7.

[46] R. Gentleman, “2002 Annual Report for the Bioconductor Project, DFCI,” 2008.

[47] W. Huber *et al.*, “Orchestrating high-throughput genomic analysis with Bioconductor,” *Nat Methods*, vol. 12, no. 2, pp. 115–121, Jan. 2015, doi: 10.1038/nmeth.3252.

[48] Lori Kern and Vince Carey, “Bioconductor Annual Report 2022,” 2023.

[49] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biol*, 2014, doi: 10.1186/s13059-014-0550-8.

[50] P. H. Morgan M, Obenchain V, Hester J, “SummarizedExperiment: SummarizedExperiment container. R package version 1.20.0,” 2020.

[51] Z. Gu, R. Eils, and M. Schlesner, “Complex heatmaps reveal patterns and correlations in multidimensional genomic data,” *Bioinformatics*, vol. 32, no. 18, pp. 2847–2849, Sep. 2016, doi: 10.1093/bioinformatics/btw313.

[52] L. Väremo, J. Nielsen, and I. Nookaew, “Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods,” *Nucleic Acids Res*, vol. 41, no. 8, pp. 4378–4391, Apr. 2013, doi: 10.1093/nar/gkt111.

[53] U. Raudvere *et al.*, “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update),” *Nucleic Acids Res*, vol. 47, no. W1, pp. W191–W198, Jul. 2019, doi: 10.1093/nar/gkz369.

[54] H. Sharifi-Noghabi, O. Zolotareva, C. C. Collins, and M. Ester, “MOLI: Multi-omics late integration with deep neural networks for drug response prediction,” in *Bioinformatics*, Oxford University Press, Jul. 2019, pp. i501–i509. doi: 10.1093/bioinformatics/btz318.

[55] K. Clark *et al.*, “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository,” *J Digit Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013, doi: 10.1007/s10278-013-9622-7.

[56] B. Albertina *et al.*, “Radiology Data from The Cancer Genome Atlas Lung Adenocarcinoma [TCGA-LUAD] collection.” The Cancer Imaging Archive. doi: 10.7937/K9/TCIA.2016.JGNIHEP5.

[57] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biol*, 2014, doi: 10.1186/s13059-014-0550-8.

[58] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[59] A. Colaprico *et al.*, “TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data,” *Nucleic Acids Res*, vol. 44, no. 8, p. e71, 2016, doi: 10.1093/nar/gkv1507.

[60] Y. Benjamini and Y. Hochberg, “Benjamini-1995.pdf,” *Journal of the Royal Statistical Society B*, vol. 57, no. 1. pp. 289–300, 1995. doi: 10.2307/2346101.

[61] C. T. Rueden *et al.*, “ImageJ2: ImageJ for the next generation of scientific image data,” *BMC Bioinformatics*, vol. 18, no. 1, p. 529, 2017, doi: 10.1186/s12859-017-1934-z.

[62] J. Schindelin *et al.*, “Fiji: an open-source platform for biological-image analysis,” *Nat Methods*, vol. 9, no. 7, pp. 676–682, 2012, doi: 10.1038/nmeth.2019.

[63] J. Ollion, J. Cochennec, F. Loll, C. Escudé, and T. Boudier, “TANGO: a generic tool for high-throughput 3D image analysis for studying nuclear organization,” *Bioinformatics*, vol. 29, no. 14, pp. 1840–1841, Jul. 2013, doi: 10.1093/bioinformatics/btt276.

[64] H. Zhou *et al.*, “MRI features predict survival and molecular markers in diffuse lower-grade gliomas,” *Neuro Oncol*, vol. 19, no. 6, pp. 862–870, 2017, doi: 10.1093/neuonc/now256.

[65] A. C. Yeh *et al.*, “Radiogenomics of breast cancer using dynamic contrast enhanced MRI and gene expression profiling,” *Cancer Imaging*, vol. 19, no. 1, p. 48, Jul. 2019, doi: 10.1186/s40644-019-0233-5.

[66] “radiomics/TextureToolbox at master · mvallieres/radiomics · GitHub.” Accessed: Mar. 14, 2021. [Online]. Available: https://github.com/mvallieres/radiomics/tree/master/TextureToolbox

[67] Robert M. Haralick, K. Sam Shanmugam, and Its’hak Dinstein, “Textural Features for Image Classification,” *IEEE Trans. Systems, Man, and Cybernetics*, vol. 3, pp. 610–621, 1973.

[68] G. Thibault *et al.*, “Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification,” in *Pattern Recognition and Information Processing*, 2009. doi: Artn 1357002\rDoi 10.1142/S0218001413570024.

[69] M. M. Galloway, “Texture analysis using gray level run lengths,” *Computer Graphics and Image Processing*, vol. 4, no. 2, pp. 172–179, 1975, doi: 10.1016/S0146-664X(75)80008-6.

[70] M. Amadasun and R. King, “Textural features corresponding to textural properties,” *IEEE Trans Syst Man Cybern*, vol. 19, no. 5, pp. 1264–1274, 1989.

[71] U. Raudvere *et al.*, “G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update),” *Nucleic Acids Res*, vol. 47, no. W1, pp. W191–W198, Jul. 2019, doi: 10.1093/nar/gkz369.

[72] L. Väremo, J. Nielsen, and I. Nookaew, “Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods,” *Nucleic Acids Res*, vol. 41, no. 8, pp. 4378–4391, Apr. 2013, doi: 10.1093/nar/gkt111.

[73] M. Nazarieh, M. Hamed, C. Spaniol, T. Will, and V. Helms, “TFmiR2: constructing and analyzing disease-, tissue- and process-specific transcription factor and microRNA co-

regulatory networks," *Bioinformatics*, vol. 36, no. 7, pp. 2300–2302, Apr. 2020, doi: 10.1093/bioinformatics/btz871.

[74] P. Shannon *et al.*, "Cytoscape: A software Environment for integrated models of biomolecular interaction networks," *Genome Res*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.

[75] W. M *et al.*, "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome," *Nat Genet*, vol. 39, no. 4, pp. 457–466, Apr. 2007, doi: 10.1038/NG1990.

[76] A. Paço and R. Freitas, "HOX genes as transcriptional and epigenetic regulators during tumorigenesis and their value as therapeutic targets," *Epigenomics*, vol. 11, no. 13. Future Medicine Ltd., pp. 1539–1552, Oct. 01, 2019. doi: 10.2217/epi-2019-0090.

[77] A. Paço, S. A. de Bessa Garcia, and R. Freitas, "Methylation in HOX Clusters and Its Applications in Cancer Therapy," *Cells*, vol. 9, no. 7. NLM (Medline), Jul. 03, 2020. doi: 10.3390/cells9071613.

[78] I. Daugaard *et al.*, "Identification and validation of candidate epigenetic biomarkers in lung adenocarcinoma," *Sci Rep*, vol. 6, Oct. 2016, doi: 10.1038/srep35807.

[79] Z. P. Ren *et al.*, "Identification of nine microRNAs as potential biomarkers for lung adenocarcinoma," *FEBS Open Bio*, vol. 9, no. 2, pp. 315–327, Feb. 2019, doi: 10.1002/2211-5463.12572.

[80] T. Xu, X. Liu, L. Han, H. Shen, L. Liu, and Y. Shu, "Up-regulation of miR-9 expression as a poor prognostic biomarker in patients with non-small cell lung cancer," *Clinical and Translational Oncology*, vol. 16, no. 5, pp. 469–475, Sep. 2014, doi: 10.1007/s12094-013-1106-1.

[81] H. Arora, R. Qureshi, S. Jin, A. K. Park, and W. Y. Park, "MiR-9 and let-7g enhance the sensitivity to ionizing radiation by suppression of NFκB1," *Exp Mol Med*, vol. 43, no. 5, pp. 298–304, Apr. 2011, doi: 10.3858/emm.2011.43.5.031.

[82] Y. Sun *et al.*, "Hsa-mir-182 suppresses lung tumorigenesis through down regulation of RGS17 expression in vitro," *Biochem Biophys Res Commun*, vol. 396, no. 2, pp. 501–507, May 2010, doi: 10.1016/j.bbrc.2010.04.127.

[83] L. Gong *et al.*, "DEPTOR inhibits lung tumorigenesis by inactivating the EGFR-mTOR signals," *Cancer Lett*, vol. 519, pp. 263–276, Oct. 2021, doi: 10.1016/J.CANLET.2021.07.031.

[84] Q. Ma *et al.*, "ONECUT2 overexpression promotes RAS-driven lung adenocarcinoma progression," *Scientific Reports 2019 9:1*, vol. 9, no. 1, pp. 1–12, Dec. 2019, doi: 10.1038/s41598-019-56277-2.

[85] J. Zhang, X. Zhang, S. Yang, Y. Bao, D. Xu, and L. Liu, “FOXH1 promotes lung cancer progression by activating the Wnt/β-catenin signaling pathway,” *Cancer Cell Int*, vol. 21, no. 1, pp. 1–13, Dec. 2021, doi: 10.1186/S12935-021-01995-9/FIGURES/6.

[86] F. lo Sardo *et al.*, “YAP/TAZ and EZH2 synergize to impair tumor suppressor activity of TGFBR2 in non-small cell lung cancer,” *Cancer Lett*, vol. 500, pp. 51–63, Mar. 2021, doi: 10.1016/J.CANLET.2020.11.037.

[87] E. R. Vagapova, P. v. Spirin, T. D. Lebedev, and V. S. Prassolov, “The Role of TAL1 in Hematopoiesis and Leukemogenesis,” *Acta Naturae*, vol. 10, no. 1, p. 15, 2018, doi: 10.32607/20758251-2018-10-1-15-23.

[88] J. Lv *et al.*, “Comparison of CT radiogenomic and clinical characteristics between EGFR and KRAS mutations in lung adenocarcinomas,” *Clin Radiol*, vol. 73, no. 6, pp. 590.e1-590.e8, Jun. 2018, doi: 10.1016/j.crad.2018.01.009.

[89] E. Ozkan *et al.*, “CT gray-level texture analysis as a quantitative imaging biomarker of epidermal growth factor receptor mutation status in adenocarcinoma of the lung,” *American Journal of Roentgenology*, vol. 205, no. 5, pp. 1016–1025, Nov. 2015, doi: 10.2214/AJR.14.14147.

[90] J. Y. Zhou *et al.*, “Comparative analysis of clinicoradiologic characteristics of lung adenocarcinomas with ALK rearrangements or EGFR mutations,” *Eur Radiol*, vol. 25, no. 5, pp. 1257–1266, May 2015, doi: 10.1007/s00330-014-3516-z.

[91] S. Yamamoto, D. D. Maki, R. L. Korn, and M. D. Kuo, “Radiogenomic analysis of breast cancer using MRI: A preliminary study to define the landscape,” *American Journal of Roentgenology*, 2012, doi: 10.2214/AJR.11.7824.

[92] A. Paço, S. A. de Bessa Garcia, and R. Freitas, “Methylation in HOX Clusters and Its Applications in Cancer Therapy,” *Cells*, vol. 9, no. 7. NLM (Medline), Jul. 03, 2020. doi: 10.3390/cells9071613.

[93] S. A. Selamat *et al.*, “DNA Methylation changes in atypical adenomatous hyperplasia, adenocarcinoma in situ, and lung adenocarcinoma,” *PLoS One*, vol. 6, no. 6, p. e21443, 2011, doi: 10.1371/journal.pone.0021443.

[94] I. Daugaard *et al.*, “Identification and validation of candidate epigenetic biomarkers in lung adenocarcinoma,” *Sci Rep*, vol. 6, Oct. 2016, doi: 10.1038/srep35807.

[95] B. L. Gan, R. Q. He, Y. Zhang, D. M. Wei, X. H. Hu, and G. Chen, “Downregulation of HOXA3 in lung adenocarcinoma and its relevant molecular mechanism analysed by RT-qPCR, TCGA and in silico analysis,” *Int J Oncol*, vol. 53, no. 4, pp. 1557–1579, Oct. 2018, doi: 10.3892/ijo.2018.4508.

[96] M. Vallières *et al.*, “Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer,” *Sci Rep*, vol. 7, no. 1, p. 10117, 2017, doi: 10.1038/s41598-017-10371-5.

[97] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa, “A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities,” *Phys Med Biol*, vol. 60, no. 14, pp. 5471–5496, Jul. 2015, doi: 10.1088/0031-9155/60/14/5471.

[98] A. Patnaik *et al.*, “Efficacy and safety of Abemaciclib, an inhibitor of CDK4 and CDK6, for patients with breast cancer, non–small cell lung cancer, and other solid tumors,” *Cancer Discov*, vol. 6, no. 7, pp. 740–753, Jul. 2016, doi: 10.1158/2159-8290.CD-16-0095.

[99] J. W. Goldman *et al.*, “A Randomized Phase III Study of Abemaciclib Versus Erlotinib in Patients with Stage IV Non-small Cell Lung Cancer With a Detectable KRAS Mutation Who Failed Prior Platinum-Based Therapy: JUNIPER,” *Front Oncol*, vol. 10, Oct. 2020, doi: 10.3389/fonc.2020.578756.

[100] S. H. Jafri, R. Shi, and G. Mills, “Advance lung cancer inflammation index (ALI) at diagnosis is a prognostic marker in patients with metastatic non-small cell lung cancer (NSCLC): A retrospective review,” *BMC Cancer*, vol. 13, Mar. 2013, doi: 10.1186/1471-2407-13-158.

[101] S. Kobayashi *et al.*, “Advanced Lung Cancer Inflammation Index Predicts Outcomes of Patients with Pathological Stage IA Lung Adenocarcinoma Following Surgical Resection,” *Ann Thorac Cardiovasc Surg*, vol. 25, no. 2, pp. 87–94, Apr. 2019, doi: 10.5761/atcs.oa.18-00158.

[102] K. Kusunoki *et al.*, “Advanced lung cancer inflammation index predicts outcomes of patients with colorectal cancer after surgical resection,” *Dis Colon Rectum*, vol. 63, no. 9, pp. 1242–1250, 2020, doi: 10.1097/DCR.0000000000001658.

[103] S. A. Best *et al.*, “Distinct initiating events underpin the immune and metabolic heterogeneity of KRAS-mutant lung adenocarcinoma,” *Nat Commun*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/s41467-019-12164-y.

[104] N. Guo *et al.*, “Lung adenocarcinoma-related TNF-α-dependent inflammation upregulates MHC-II on alveolar type II cells through CXCR-2 to contribute to Treg expansion,” *FASEB Journal*, vol. 34, no. 9, pp. 12197–12213, Sep. 2020, doi: 10.1096/fj.202000166RR.

[105] E. R. Choi *et al.*, “Quantitative image variables reflect the intratumoral pathologic heterogeneity of lung adenocarcinoma,” *Oncotarget*, vol. 7, no. 41, pp. 67302–67313, 2016, doi: 10.18632/oncotarget.11693.

[106] Q. Huang *et al.*, “The role of adrenergic receptors in lung cancer,” *Am J Cancer Res*, vol. 8, no. 11, p. 2227, 2018, Accessed: Apr. 02, 2022. [Online]. Available: /pmc/articles/PMC6291649/

[107] L.-K. Soh and C. Tsatsoulis, “Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices,” *IEEE Transactions on geoscience and remote sensing*, vol. 37, no. 2, pp. 780–795, 1999.

[108] D. A. Clausi, “An analysis of co-occurrence texture statistics as a function of grey level quantization,” *Canadian Journal of Remote Sensing*, vol. 28, no. 1, pp. 45–62, 2002, doi: 10.5589/m02-004.

[109] A. Chu, C. M. Sehgal, and J. F. Greenleaf, “Use of gray value distribution of run lengths for texture analysis,” *Pattern Recognit Lett*, vol. 11, no. 6, pp. 415–419, 1990.

[110] B. V Dasarathy and E. B. Holder, “Image characterizations based on joint gray level—run length distributions,” *Pattern Recognit Lett*, vol. 12, no. 8, pp. 497–502, 1991.

[111] Q. Li and J. G. Griffiths, “Least squares ellipsoid specific fitting,” in *Geometric modeling and processing, 2004. proceedings*, 2004, pp. 335–340.

[112] P. Andrey *et al.*, “Statistical analysis of 3D images detects regular spatial distributions of centromeres and chromocenters in animal and plant nuclei,” *PLoS Comput Biol*, vol. 6, no. 7, 2010.

[113] Sheets KG, Bokkyoo J, Zhou Y, Winkler J, Zhu M, Petasis N, Gordon WC, Bazan NG, “Topical Neuroprotectin D1 Attenuates Experimental CNV and Induces Activated Microglia Redistribution: ARVO Meeting Abstracts April 22, 2011,” 2011.

8 SUPPLEMENTARY MATERIAL

Table 6 Significant radiogenomic associations with methylome

The number of significant radiogenomic associations and GO BP terms from integrating methylation data and image features.

Pathway	Number of significant radiomic associations
rhombomere development	44
peripheral nervous system neuron development	43
peripheral nervous system neuron differentiation	43
thyroid gland development	41
glossopharyngeal nerve morphogenesis	40
learning	37
cochlea development	35
hindlimb morphogenesis	35
learning or memory	35
cognition	34
embryonic skeletal system development	33
embryonic skeletal system morphogenesis	33
camera-type eye morphogenesis	32
endocardial cushion morphogenesis	32
mesoderm formation	31
cranial nerve morphogenesis	30
embryonic organ morphogenesis	30
endocardial cushion development	30
forelimb morphogenesis	30
mesenchyme morphogenesis	30
mesoderm development	30
proximal/distal pattern formation	30
skeletal system morphogenesis	30
chordate embryonic development	29
eye morphogenesis	29
glial cell fate commitment	29
skeletal system development	29

gland development	28
stem cell differentiation	28
embryo development ending in birth or egg hatching	27
formation of primary germ layer	27
spinal cord motor neuron cell fate specification	27
appendage morphogenesis	26
embryonic appendage morphogenesis	26
embryonic limb morphogenesis	26
limb morphogenesis	26
neuropeptide signaling pathway	26
outflow tract morphogenesis	26
telencephalon development	26
camera-type eye development	25
embryonic forelimb morphogenesis	25
endocrine system development	25
appendage development	24
cartilage development	24
dorsal spinal cord development	24
embryonic organ development	24
eye development	24
limb development	24
neural precursor cell proliferation	24
neuron migration	24
sensory system development	24
visual system development	24
embryonic morphogenesis	23
pallium development	23
pattern specification process	23
animal organ morphogenesis	22
embryo development	22
gastrulation	22
limbic system development	22
mesenchymal cell differentiation	22
negative regulation of transcription by RNA polymerase II	22

peripheral nervous system development	22
positive regulation of neuron differentiation	22
positive regulation of transcription by RNA polymerase II	22
inner ear development	21
locomotory behavior	21
negative regulation of biosynthetic process	21
negative regulation of cellular biosynthetic process	21
negative regulation of cellular macromolecule biosynthetic process	21
negative regulation of macromolecule biosynthetic process	21
negative regulation of nucleobase-containing compound metabolic process	21
negative regulation of RNA metabolic process	21
regulation of neurogenesis	21
skeletal muscle tissue development	21
branching morphogenesis of an epithelial tube	20
cranial nerve development	20
glutamate receptor signaling pathway	20
negative regulation of nucleic acid-templated transcription	20
negative regulation of RNA biosynthetic process	20
negative regulation of transcription, DNA-templated	20
ureter morphogenesis	20
anterior/posterior pattern specification	19
dorsal/ventral pattern formation	19
blood vessel morphogenesis	18
connective tissue development	18
nerve development	18
regionalization	18
skeletal muscle organ development	18
commitment of neuronal cell to specific neuron type in forebrain	17
forebrain generation of neurons	17
morphogenesis of a branching epithelium	17
negative regulation of cell differentiation	17

neural crest cell differentiation	17
pancreatic A cell differentiation	17
specification of animal organ identity	17
anatomical structure morphogenesis	16
cardiac ventricle development	16
cell fate specification	16
inner ear morphogenesis	16
mesenchymal cell proliferation	16
morphogenesis of a branching structure	16
positive regulation of nucleic acid-templated transcription	16
positive regulation of RNA biosynthetic process	16
positive regulation of RNA metabolic process	16
positive regulation of transcription, DNA-templated	16
roof of mouth development	16
sensory organ morphogenesis	16
branching involved in ureteric bud morphogenesis	15
digestive tract development	15
positive regulation of nucleobase-containing compound metabolic process	15
skeletal muscle cell differentiation	15
tube development	15
ventral spinal cord development	15
associative learning	14
cell fate commitment	14
cellular response to growth factor stimulus	14
G protein-coupled receptor signaling pathway	14
positive regulation of biosynthetic process	14
positive regulation of cellular biosynthetic process	14
positive regulation of macromolecule biosynthetic process	14
regulation of cell differentiation	14
regulation of developmental process	14
response to growth factor	14
sensory organ development	14
ureter development	14

anatomical structure formation involved in morphogenesis	13
cell differentiation in spinal cord	13
cell fate commitment involved in formation of primary germ layer	13
digestive system development	13
ear morphogenesis	13
forebrain neuron differentiation	13
negative regulation of neuron differentiation	13
neural retina development	13
positive regulation of cell population proliferation	13
positive regulation of macromolecule metabolic process	13
regulation of cell development	13
regulation of macromolecule metabolic process	13
regulation of nervous system development	13
renal system development	13
tube morphogenesis	13
circulatory system development	12
muscle tissue development	12
nucleic acid-templated transcription	12
pancreas development	12
positive regulation of biological process	12
regulation of multicellular organismal development	12
regulation of RNA metabolic process	12
regulation of transcription by RNA polymerase II	12
RNA metabolic process	12
spinal cord development	12
striated muscle tissue development	12
transcription by RNA polymerase II	12
transcription, DNA-templated	12
adenylate cyclase-inhibiting G protein-coupled receptor signaling pathway	11
biological regulation	11
brain development	11
cardiac muscle tissue development	11

cell population proliferation	11
cellular macromolecule biosynthetic process	11
cellular response to BMP stimulus	11
forebrain development	11
head development	11
mesonephric epithelium development	11
mesonephric tubule development	11
metanephric nephron development	11
pituitary gland development	11
positive regulation of animal organ morphogenesis	11
positive regulation of nitrogen compound metabolic process	11
regulation of cellular macromolecule biosynthetic process	11
regulation of macromolecule biosynthetic process	11
regulation of nucleic acid-templated transcription	11
regulation of RNA biosynthetic process	11
regulation of transcription, DNA-templated	11
response to BMP	11
RNA biosynthetic process	11
ureteric bud development	11
urogenital system development	11
adenylate cyclase-modulating G protein-coupled receptor signaling pathway	10
axon guidance	10
cell differentiation	10
cell differentiation involved in kidney development	10
cell morphogenesis involved in neuron differentiation	10
cellular nitrogen compound biosynthetic process	10
developmental induction	10
developmental process	10
gene expression	10
heterocycle biosynthetic process	10
hindbrain development	10
macromolecule biosynthetic process	10
neurogenesis	10

neurotransmitter secretion	10
positive regulation of cellular metabolic process	10
regulation of biosynthetic process	10
regulation of cell population proliferation	10
regulation of cellular biosynthetic process	10
regulation of cellular metabolic process	10
regulation of gene expression	10
regulation of ion transmembrane transport	10
regulation of metabolic process	10
regulation of neuron differentiation	10
regulation of nucleobase-containing compound metabolic process	10
regulation of transmembrane transport	10
signal release from synapse	10
tissue development	10
animal organ development	9
animal organ formation	9
aromatic compound biosynthetic process	9
cardiac chamber morphogenesis	9
cell development	9
cellular developmental process	9
central nervous system development	9
central nervous system neuron differentiation	9
embryonic hindlimb morphogenesis	9
generation of neurons	9
heart morphogenesis	9
kidney epithelium development	9
mesenchyme development	9
mesonephros development	9
metanephros development	9
multicellular organism development	9
nephron epithelium development	9
neuron differentiation	9
noradrenergic neuron differentiation	9

nucleobase-containing compound biosynthetic process	9
organic cyclic compound biosynthetic process	9
positive regulation of metabolic process	9
regulation of membrane potential	9
regulation of nitrogen compound metabolic process	9
regulation of primary metabolic process	9
system development	9
adult behavior	8
anatomical structure development	8
cation transmembrane transport	8
cell migration	8
cell morphogenesis involved in differentiation	8
cell motility	8
developmental growth	8
diencephalon development	8
heart development	8
kidney development	8
localization of cell	8
multicellular organismal process	8
nephron development	8
nervous system development	8
positive regulation of cell differentiation	8
positive regulation of cellular process	8
positive regulation of developmental process	8
regulation of biological process	8
regulation of multicellular organismal process	8
sympathetic nervous system development	8
synapse assembly	8
synaptic signaling	8
anterograde trans-synaptic signaling	7
autonomic nervous system development	7
axon development	7
axonogenesis	7
cardiac chamber development	7

cardiac ventricle morphogenesis	7
central nervous system neuron development	7
chemical synaptic transmission	7
ear development	7
epithelial tube morphogenesis	7
kidney morphogenesis	7
mesonephric tubule morphogenesis	7
modulation of chemical synaptic transmission	7
muscle organ development	7
nephron epithelium morphogenesis	7
nephron morphogenesis	7
nephron tubule development	7
nephron tubule morphogenesis	7
neuron development	7
neuron projection guidance	7
regulation of trans-synaptic signaling	7
renal tubule development	7
renal tubule morphogenesis	7
ureteric bud morphogenesis	7
amine transport	6
cardiac right ventricle morphogenesis	6
dopaminergic neuron differentiation	6
locomotion	6
negative regulation of developmental process	6
regulation of postsynaptic membrane potential	6
regulation of synaptic plasticity	6
tissue morphogenesis	6
behavior	5
catecholamine secretion	5
cell part morphogenesis	5
cell projection morphogenesis	5
inorganic cation transmembrane transport	5
inorganic ion transmembrane transport	5
neuron fate determination	5

neuron projection morphogenesis	5
plasma membrane bounded cell projection morphogenesis	5
positive regulation of nervous system development	5
potassium ion transport	5
regulation of amine transport	5
regulation of catecholamine secretion	5
regulation of cellular process	5
regulation of neurotransmitter levels	5
regulation of secretion	5
regulation of synapse assembly	5
spinal cord motor neuron differentiation	5
trans-synaptic signaling	5
cell morphogenesis	4
cellular component morphogenesis	4
cellular response to endogenous stimulus	4
dopamine transport	4
morphogenesis of an epithelium	4
movement of cell or subcellular component	4
neuron projection development	4
positive regulation of multicellular organismal process	4
regulation of transport	4
system process	4
catecholamine transport	3
cation transport	3
chemotaxis	3
homophilic cell adhesion via plasma membrane adhesion molecules	3
ion transmembrane transport	3
monoamine transport	3
neuron fate commitment	3
neurotransmitter transport	3
potassium ion transmembrane transport	3
signal release	3
taxis	3

cell fate determination	2
cell-cell adhesion via plasma-membrane adhesion molecules	2
epithelium development	2
nervous system process	2
neuron fate specification	2
parasympathetic nervous system development	2
response to endogenous stimulus	2
trigeminal nerve development	2

Table 7 Assignments of image features to their represented category.

There are 12 features residing in morphology, 6 assigned to size, shape accounts for 15 and the category texture has 53 assignments.

Image Feature	**Category**
DCMin (unit)	Morphology
DCMax (unit)	Morphology
DCMean (unit)	Morphology
DCSD (unit)	Morphology
IntDen	Morphology
Min	Morphology
Max	Morphology
Mean	Morphology
Sigma	Morphology
Variance	Morphology
Skewness	Morphology
Kurtosis	Morphology
Comp (unit)	Shape
Spher (unit)	Shape
CompDiscrete	Shape
Ell_MajRad	Shape
Ell_Elon	Shape
Ell_Flatness	Shape
volEllipsoid (unit)	Shape
RatioVolEllipsoid	Shape

Moment1	Shape
Moment2	Shape
Moment3	Shape
Moment4	Shape
Moment5	Shape
Solidity	Shape
Eccentricity	Shape
Vol (unit)	Size
Surf (unit)	Size
Feret (unit)	Size
VolHull (unit)	Size
volume	Size
Size	Size
autoc	Texture
contr	Texture
corrm	Texture
corrp	Texture
cprom	Texture
cshad	Texture
dissi	Texture
energ	Texture
entro	Texture
homom	Texture
homop	Texture
maxpr	Texture
sosvh	Texture
savgh	Texture
svarh	Texture
senth	Texture
dvarh	Texture
denth	Texture
inf1h	Texture
inf2h	Texture
indnc	Texture

idmnc	Texture
SRE	Texture
LRE	Texture
GLN	Texture
RLN	Texture
RP	Texture
LGRE	Texture
HGRE	Texture
SRLGE	Texture
SRHGE	Texture
LRLGE	Texture
LRHGE	Texture
GLV	Texture
RLV	Texture
SZE	Texture
LZE	Texture
GLN_GLSZM	Texture
ZSN	Texture
ZP	Texture
LGZE	Texture
HGZE	Texture
SZLGE	Texture
SZHGE	Texture
LZLGE	Texture
LZHGE	Texture
GLV_GLSZM	Texture
ZSV	Texture
Coarseness	Texture
Contrast	Texture
Busyness	Texture
Complexity	Texture
Strength	Texture

Table 8 Significant radiogenomic associations with gene expression

The number of significant radiogenomic associations and GO BP terms from integrating mRNA expression data and image features.

Pathway	**Number of significant radiomic associations**
nuclear division	63
meiotic cell cycle	59
meiotic cell cycle process	57
meiotic nuclear division	55
external encapsulating structure organization	51
extracellular structure organization	51
embryonic skeletal system morphogenesis	50
extracellular matrix organization	50
skeletal system development	50
skeletal system morphogenesis	50
animal organ morphogenesis	49
central nervous system neuron differentiation	49
biological adhesion	48
cardiac chamber development	48
cardiac ventricle development	48
cell adhesion	48
collagen metabolic process	48
embryo development	48
urogenital system development	48
anatomical structure formation involved in morphogenesis	47
appendage development	47
circulatory system development	47
head development	47
limb development	47
mesenchyme development	47
regulation of anatomical structure morphogenesis	47
anatomical structure morphogenesis	46
brain development	46
kidney development	46

regulation of developmental process	46
regulation of multicellular organismal process	46
tube development	46
animal organ development	45
developmental process	45
embryonic morphogenesis	45
embryonic skeletal system development	45
heart development	45
regulation of nervous system development	45
system development	45
tube morphogenesis	45
ureteric bud development	45
anatomical structure development	44
blood vessel development	44
central nervous system development	44
homophilic cell adhesion via plasma membrane adhesion molecules	44
morphogenesis of a branching epithelium	44
multicellular organism development	44
multicellular organismal process	44
neuron migration	44
positive regulation of developmental process	44
regulation of multicellular organismal development	44
renal system development	44
tissue morphogenesis	44
vasculature development	44
blood vessel morphogenesis	43
branching morphogenesis of an epithelial tube	43
cardiac chamber morphogenesis	43
cell differentiation	43
cell junction organization	43
cell morphogenesis	43
cellular developmental process	43
digestive tract development	43

morphogenesis of a branching structure	43
negative regulation of developmental process	43
pattern specification process	43
tissue development	43
cell part morphogenesis	42
cell projection morphogenesis	42
embryonic organ development	42
embryonic organ morphogenesis	42
mesonephros development	42
nervous system development	42
plasma membrane bounded cell projection morphogenesis	42
regionalization	42
regulation of localization	42
angiogenesis	41
axon development	41
axonogenesis	41
cell-cell adhesion via plasma-membrane adhesion molecules	41
forebrain development	41
regulation of cell adhesion	41
synapse organization	41
cell communication	40
cell migration	40
neurogenesis	40
regulation of cell differentiation	40
signaling	40
cell development	39
cellular component morphogenesis	39
neuron projection morphogenesis	39
cell fate specification	38
cellular response to organic substance	38
extracellular matrix disassembly	38
negative regulation of multicellular organismal process	38
positive regulation of multicellular organismal process	38
regulation of cell development	38

regulation of cellular component movement	38
regulation of signaling	38
cardiac ventricle morphogenesis	37
cell junction assembly	37
cellular response to stimulus	37
kidney epithelium development	37
neural precursor cell proliferation	37
generation of neurons	36
ossification	36
response to stimulus	36
cell surface receptor signaling pathway	35
epithelium development	35
regulation of locomotion	35
telencephalon development	35
cell population proliferation	34
cell-cell adhesion	34
cell morphogenesis involved in differentiation	33
cell motility	33
localization of cell	33
locomotion	33
cell fate commitment	32
neuron differentiation	32
positive regulation of cell differentiation	32
positive regulation of nervous system development	32
regulation of synapse assembly	32
response to endogenous stimulus	32
response to organic substance	32
carboxylic acid transport	31
cell morphogenesis involved in neuron differentiation	31
neuron fate commitment	31
actin filament-based process	30
neuron projection development	30
response to wounding	30
cytoskeleton organization	29

epithelial cell differentiation	29
neuron development	29
regulation of biological quality	29
developmental process involved in reproduction	28
neuron projection guidance	28
regulation of wound healing	28
cell differentiation in spinal cord	27
diencephalon development	27
positive regulation of cell population proliferation	27
regulation of ERK1 and ERK2 cascade	27
response to chemical	27
wound healing	27
cell-cell signaling	26
cellular response to chemical stimulus	26
chemotaxis	26
neuropeptide signaling pathway	26
sensory organ development	26
sensory organ morphogenesis	26
sensory perception of mechanical stimulus	26
taxis	26
circulatory system process	24
ERK1 and ERK2 cascade	24
negative regulation of wound healing	24
positive regulation of secretion by cell	24
reproduction	24
reproductive process	24
cornification	23
localization	23
movement of cell or subcellular component	23
negative regulation of response to wounding	23
positive regulation of ERK1 and ERK2 cascade	23
axon guidance	22
muscle structure development	22
signal release	22

blood circulation	21
regulation of cell population proliferation	21
regulation of ossification	21
muscle tissue development	20
negative regulation of blood coagulation	20
positive regulation of secretion	20
system process	20
homeostatic process	19
regulation of secretion	18
regulation of transport	18
vascular process in circulatory system	18
calcium ion homeostasis	17
negative regulation of hemostasis	17
organic anion transport	17
positive regulation of MAPK cascade	17
regulation of secretion by cell	17
response to external stimulus	17
secretion	17
transmission of nerve impulse	17
action potential	16
biomineral tissue development	16
lipid transport	16
neutrophil chemotaxis	16
organic hydroxy compound transport	16
regulation of ion transport	16
striated muscle tissue development	16
cellular calcium ion homeostasis	15
export from cell	15
monocarboxylic acid transport	15
multicellular organismal signaling	15
organic acid transport	15
regulation of body fluid levels	15
regulation of system process	15
response to lipid	15

response to oxygen-containing compound	15
chemical homeostasis	14
positive regulation of calcium ion transport	14
response to corticosteroid	14
retinoid metabolic process	14
secretion by cell	14
cell-cell junction organization	13
divalent inorganic cation homeostasis	13
excretion	13
heart process	13
regulation of blood circulation	13
regulation of blood pressure	12
sensory perception of pain	12
cellular divalent inorganic cation homeostasis	11
sodium ion transmembrane transport	11
sodium ion transport	11
adenylate cyclase-activating G protein-coupled receptor signaling pathway	10
antimicrobial humoral response	10
bone remodeling	10
calcium ion transport	10
digestion	10
drug metabolic process	10
endocrine process	10
multicellular organismal reproductive process	10
regulation of calcium ion transport	10
regulation of cation transmembrane transport	10
regulation of heart contraction	10
regulation of transporter activity	10
response to hormone	10
digestive system process	9
multicellular organism reproduction	9
phospholipase C-activating G protein-coupled receptor signaling pathway	9

regulation of metal ion transport	9
terpenoid metabolic process	9
cellular metal ion homeostasis	8
heart contraction	8
muscle organ development	8
nervous system process	8
regulation of renal system process	8
response to xenobiotic stimulus	8
sexual reproduction	8
amine transport	7
behavior	7
cation homeostasis	7
cellular cation homeostasis	7
cellular chemical homeostasis	7
cognition	7
icosanoid metabolic process	7
metal ion homeostasis	7
metal ion transport	7
multicellular organismal homeostasis	7
positive regulation of cytosolic calcium ion concentration	7
positive regulation of transport	7
regulation of ion transmembrane transport	7
regulation of transmembrane transport	7
tissue homeostasis	7
adenylate cyclase-modulating G protein-coupled receptor signaling pathway	6
learning or memory	6
multi-organism process	6
multi-organism reproductive process	6
positive regulation of icosanoid secretion	6
regulation of amine transport	6
regulation of cytosolic calcium ion concentration	6
regulation of membrane potential	6
regulation of systemic arterial blood pressure	6

cellular ion homeostasis	5
icosanoid transport	5
inorganic ion homeostasis	5
ion homeostasis	5
modulation of chemical synaptic transmission	5
muscle contraction	5
muscle system process	5
regulation of hormone levels	5
response to ketone	5
anterograde trans-synaptic signaling	4
cation transmembrane transport	4
cation transport	4
cellular homeostasis	4
cellular response to xenobiotic stimulus	4
chemical synaptic transmission	4
inorganic cation import across plasma membrane	4
inorganic ion import across plasma membrane	4
negative regulation of endopeptidase activity	4
negative regulation of peptidase activity	4
positive regulation of prostaglandin secretion	4
potassium ion transmembrane transport	4
potassium ion transport	4
regulation of bone remodeling	4
regulation of trans-synaptic signaling	4
response to mechanical stimulus	4
response to organic cyclic compound	4
response to organophosphorus	4
trans-synaptic signaling	4
transmembrane transport	4
anatomical structure maturation	3
developmental maturation	3
neuron fate specification	3
positive regulation of protein tyrosine kinase activity	3
response to cAMP	3

synaptic signaling	3
actin-mediated cell contraction	2
catecholamine secretion	2
feeding behavior	2
inorganic cation transmembrane transport	2
inorganic ion transmembrane transport	2
ion transmembrane transport	2
lipid export from cell	2
regulation of catecholamine secretion	2

Table 9 Significant radiogenomic associations with miRNA expression

The number of significant radiogenomic associations and GO BP terms from integrating miRNA expression data and image features.

Pathway	**Number of significant radiomic associations**
cellular response to interleukin-6	31
interleukin-6-mediated signaling pathway	31
negative regulation of interleukin-6-mediated signaling pathway	31
regulation of interleukin-6-mediated signaling pathway	31
response to interleukin-6	31
negative regulation of cytokine-mediated signaling pathway	20
regulation of cytokine-mediated signaling pathway	20
negative regulation of response to cytokine stimulus	19
regulation of response to cytokine stimulus	19
response to organic substance	18
cellular response to organic substance	16
cellular response to chemical stimulus	15
cellular response to transforming growth factor beta stimulus	13
positive regulation of cell migration involved in sprouting angiogenesis	13
response to transforming growth factor beta	13
transforming growth factor beta receptor signaling pathway	13

transmembrane receptor protein serine/threonine kinase signaling pathway	10
negative regulation of cell cycle G1/S phase transition	9
negative regulation of cell cycle process	9
negative regulation of peptide secretion	9
negative regulation of protein secretion	9
negative regulation of secretion by cell	9
positive regulation of apoptotic process	9
positive regulation of cell death	9
positive regulation of programmed cell death	9
negative regulation of protein kinase B signaling	8
negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	8
regulation of MAPK cascade	8
cell cycle G1/S phase transition	7
negative regulation of cell cycle	7
negative regulation of cell cycle phase transition	7
negative regulation of G1/S transition of mitotic cell cycle	7
negative regulation of muscle cell apoptotic process	7
regulation of cell communication	7
regulation of cell cycle G1/S phase transition	7
regulation of signaling	7
regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	7
amide biosynthetic process	6
miRNA mediated inhibition of translation	6
muscle structure development	6
negative regulation of mitotic cell cycle phase transition	6
negative regulation of secretion	6
negative regulation of translation	6
negative regulation of translation, ncRNA-mediated	6
peptide biosynthetic process	6
regulation of cellular macromolecule biosynthetic process	6
regulation of macromolecule biosynthetic process	6

regulation of translation	6
regulation of translation, ncRNA-mediated	6
translation	6
cell cycle phase transition	5
cellular amide metabolic process	5
G1/S transition of mitotic cell cycle	5
matrix metallopeptidase secretion	5
muscle cell differentiation	5
muscle cell proliferation	5
negative regulation of cellular amide metabolic process	5
negative regulation of matrix metallopeptidase secretion	5
negative regulation of transporter activity	5
peptide metabolic process	5
positive regulation of cellular process	5
positive regulation of endothelial cell apoptotic process	5
regulation of cell cycle phase transition	5
regulation of cell cycle process	5
regulation of cellular amide metabolic process	5
regulation of G1/S transition of mitotic cell cycle	5
regulation of matrix metallopeptidase secretion	5
regulation of response to stress	5
regulation of signal transduction	5
cardiac muscle cell membrane repolarization	4
epithelial cell apoptotic process	4
intracellular signal transduction	4
membrane repolarization	4
membrane repolarization during action potential	4
membrane repolarization during cardiac muscle cell action potential	4
negative regulation of calcium ion transmembrane transport	4
negative regulation of cardiac muscle cell apoptotic process	4
negative regulation of cellular macromolecule biosynthetic process	4
negative regulation of cellular protein metabolic process	4
negative regulation of DNA-binding transcription factor activity	4

negative regulation of macromolecule biosynthetic process	4
negative regulation of MAPK cascade	4
negative regulation of membrane repolarization during cardiac muscle cell action potential	4
negative regulation of response to stimulus	4
negative regulation of smooth muscle cell proliferation	4
negative regulation of striated muscle cell apoptotic process	4
positive regulation of biological process	4
positive regulation of cell fate commitment	4
positive regulation of epithelial cell apoptotic process	4
positive regulation of muscle cell differentiation	4
regulation of cardiac muscle cell membrane repolarization	4
regulation of cell fate commitment	4
regulation of DNA-binding transcription factor activity	4
regulation of epithelial cell apoptotic process	4
regulation of membrane repolarization	4
regulation of membrane repolarization during action potential	4
regulation of membrane repolarization during cardiac muscle cell action potential	4
regulation of muscle cell differentiation	4
regulation of response to stimulus	4
regulation of smooth muscle cell proliferation	4
smooth muscle cell proliferation	4
apoptotic process	3
cardiac muscle cell apoptotic process	3
cardiocyte differentiation	3
cell death	3
cell death in response to oxidative stress	3
cellular response to vascular endothelial growth factor stimulus	3
cellular response to virus	3
defense response	3
enzyme linked receptor protein signaling pathway	3
mitotic cell cycle phase transition	3
negative regulation of cardiac muscle hypertrophy	3

negative regulation of catalytic activity	3
negative regulation of cation channel activity	3
negative regulation of cell communication	3
negative regulation of cell differentiation	3
negative regulation of cell migration involved in sprouting angiogenesis	3
negative regulation of cell population proliferation	3
negative regulation of ion transmembrane transporter activity	3
negative regulation of mitotic cell cycle	3
negative regulation of muscle cell differentiation	3
negative regulation of muscle hypertrophy	3
negative regulation of oxidative stress-induced cell death	3
negative regulation of oxidoreductase activity	3
negative regulation of potassium ion transmembrane transport	3
negative regulation of potassium ion transmembrane transporter activity	3
negative regulation of potassium ion transport	3
negative regulation of signal transduction	3
negative regulation of signaling	3
negative regulation of vascular associated smooth muscle cell proliferation	3
negative regulation of voltage-gated potassium channel activity	3
neuroinflammatory response	3
organonitrogen compound biosynthetic process	3
osteoblast differentiation	3
positive regulation of cell migration	3
positive regulation of cell motility	3
positive regulation of cellular component movement	3
positive regulation of developmental process	3
positive regulation of locomotion	3
positive regulation of myotube differentiation	3
programmed cell death	3
regulation of apoptotic process	3
regulation of cardiac muscle cell apoptotic process	3

regulation of cell death	3
regulation of cell differentiation	3
regulation of cellular protein metabolic process	3
regulation of mitotic cell cycle phase transition	3
regulation of myotube differentiation	3
regulation of neuroinflammatory response	3
regulation of oxidative stress-induced cell death	3
regulation of oxidoreductase activity	3
regulation of potassium ion transmembrane transporter activity	3
regulation of programmed cell death	3
regulation of striated muscle cell apoptotic process	3
regulation of tumor necrosis factor production	3
regulation of tumor necrosis factor superfamily cytokine production	3
striated muscle cell apoptotic process	3
tumor necrosis factor production	3
tumor necrosis factor superfamily cytokine production	3
cardiac muscle tissue regeneration	2
cell population proliferation	2
interleukin-17 production	2
interleukin-6 production	2
negative regulation of cation transmembrane transport	2
negative regulation of cell adhesion molecule production	2
negative regulation of cell death	2
negative regulation of cell migration	2
negative regulation of cell motility	2
negative regulation of cellular component movement	2
negative regulation of developmental process	2
negative regulation of ERK1 and ERK2 cascade	2
negative regulation of interleukin-17 production	2
negative regulation of interleukin-6 production	2
negative regulation of intracellular signal transduction	2
negative regulation of ion transmembrane transport	2
negative regulation of lipid metabolic process	2

negative regulation of locomotion	2
negative regulation of molecular function	2
negative regulation of NIK/NF-kappaB signaling	2
negative regulation of nitrogen compound metabolic process	2
negative regulation of osteoblast differentiation	2
negative regulation of osteoblast proliferation	2
negative regulation of protein metabolic process	2
negative regulation of response to external stimulus	2
negative regulation of transmembrane transport	2
negative regulation of tumor necrosis factor production	2
negative regulation of tumor necrosis factor superfamily cytokine production	2
osteoblast proliferation	2
positive regulation of blood vessel endothelial cell migration	2
positive regulation of cardiac muscle tissue regeneration	2
potassium ion transmembrane transport	2
protein kinase B signaling	2
regulation of cardiac muscle tissue regeneration	2
regulation of cell population proliferation	2
regulation of cellular response to growth factor stimulus	2
regulation of cellular response to vascular endothelial growth factor stimulus	2
regulation of defense response	2
regulation of inflammatory response	2
regulation of interleukin-17 production	2
regulation of interleukin-6 production	2
regulation of intracellular signal transduction	2
regulation of mitotic cell cycle	2
regulation of molecular function	2
regulation of nitrogen compound metabolic process	2
regulation of osteoblast differentiation	2
regulation of osteoblast proliferation	2
regulation of potassium ion transmembrane transport	2
regulation of potassium ion transport	2

regulation of protein kinase B signaling	2
regulation of response to external stimulus	2
regulation of vascular associated smooth muscle cell differentiation	2
regulation of vascular associated smooth muscle cell proliferation	2
regulation of vascular endothelial growth factor signaling pathway	2
tissue regeneration	2
vascular associated smooth muscle cell differentiation	2
vascular associated smooth muscle cell proliferation	2

Table 10 Detailed references of the various image features utilized in this study.

Texture Features	GLCM (Gray-Level Co-Occurence Matrix): - Autocorrelation (autoc)[67], [69] - Contrast (contr) [67], [107] - Correlation Matlab (corrm) [108] - Correlation (corrp) [67], [107] - Cluster Prominence (cprom) [107] - Cluster Shade (cshad) [107] - Dissimilarity (dissi) [107] - Energy matlab (energ) [67], [107] - Entropy (entro) [107] - Homogeneity from matlab (homom) [108] - Homogeneity (homop) [107] - Maximum probability (maxpr) [107] - Sum of squares: Variance (sosvh) [67] - Sum average (savgh) [67] - Sum variance (svarh) [67] - Sum entropy (senth) [67] - Difference variance (dvarh) [67] - Difference entropy (denth) [67] - Information measure of correlation1 (inf1h) [67] - Information measure of correlation2 (inf2h) [67] - Inverse difference (INV) [108] - Inverse difference normalized (INN) (indnc) [108] - Inverse difference moment normalized (idmnc) [108]

Texture Features	GLSZM (Gray-Level Size Zone Matrix): - Small Zone Emphasis (SZE) [68], [69] - Large Zone Emphasis (LZE) [68], [69] - Gray-Level Nonuniformity (GLN), adapted from Galloway and Thibault et. al [68], [69] - Zone-Size Nonuniformity (ZSN), adapted from Galloway and Thibault et. al [68], [69] - Zone Percentage (ZP), adapted from Galloway [68], [69] - Low Gray-Level Zone Emphasis (LGZE) [68], [109] - High Gray-Level Zone Emphasis (HGZE) [68], [109] - Small Zone Low Gray-Level Emphasis (SZLGE) [68], [110] - Small Zone High Gray-Level Emphasis (SZHGE) [68], [110] - Large Zone Low Gray-Level Emphasis (LZLGE) [68], [110] - Large Zone High Gray-Level Emphasis (LZHGE) [68], [110] - Gray-Level Variance (GLV), adapted from Thibault et. al [68] - Zone-Size Variance (ZSV), adapted from Thibault et. al [68]
Texture Features	GLRLM (Gray-Level Run-Length Matrix): - Short Run Emphasis (SRE) [69] - Long Run Emphasis (LRE) [69] - Gray-Level Nonuniformity (GLN), adapted from Galloway [69] - Run-Length Nonuniformity (RLN), adapted from Galloway [69] - Run Percentage (RP), adapted from Galloway [69] - Low Gray-Level Run Emphasis (LGRE) [109] - High Gray-Level Run Emphasis (HGRE) [109] - Short Run Low Gray-Level Emphasis (SRLGE) [110] - Short Run High Gray-Level Emphasis (SRHGE) [110] - Long Run Low Gray-Level Emphasis (LRLGE) [110] - Long Run High Gray-Level Emphasis (LRHGE) [110] - Gray-Level Variance (GLV), adapted from Thibault [68] - Run-Length Variance (RLV), adapted from Thibault [68]
Texture Features	NGTDM (Neighborhood Gray-Tone Difference Matrix): - Coarseness [70] - Contrast [70] - Busyness [70] - Complexity [70]

	- Strength [70]
Shape	Geometric measurements: - 3D Moments: [63] o Moment1, Moment2, Moment3, Moment4, Moment5 - Compactness (Comp (unit)) [63] - Compactness Discrete (CompDiscrete) [63] - Eccentricity (eccentricity) from Matlab - Ellongation Major Radius (Ell_MajRad) [63] - Ellongation (Ell_Elon) [63] - Ellongation Flatness (Ell_Flatness) [63] - Fit Ellipse 3-dimensional fitting of an ellipse within the tumor volume (Spher (unit)) [111] - Fitted volume for an ellipsoid (volEllipsoid (unit)) [63] - RatioVolEllipsoid [63]
Morphology	- Density: [112] o DCMin (unit), DCMax (unit), DCMean (unit), DCSD (unit) - Integrated Density (IntDen) [112] - Kurtosis from Matlab - Min Grey Value (Min) [112] - Max Grey Value (Max) [112] - Mean Gray Value (Mean) [112] - Skewness from Matlab - Standard deviation Grey Value (Sigma) [112] - Variance from Matlab
Size	- Volume (Vol (unit) from ImageJ) [63] - volume from Matlab - Feret (Feret (unit)) [63] - Convex Hull (VolHull (unit)) [113] - size from Matlab, computes the longest diameter - Surface (Surf (unit)) [111]

Table 11 Software and R packages with versions utilized

Software	Version	Method
MATLAB	R2018b	Execute the Texture Toolbox
Fiji	8	Segmentation
3D-ROI Manager plugin for Fiji	3.94	3D intensity-, quantitative- and distance measurements
Texture Toolbox for MATLAB https://github.com/mvallieres/radiomics/tree/master/TextureToolbox	Commit id: 9a8c5cb	GLCM, GLRLM, GLSZM, NGTDM, Global
R statistical programming language	3.6.3	See R sessionInfo()
g:Profiler	Ensembl 102, Ensembl Genomes 49	Over-representation analysis
TFmiR2	The webserver was accessed on 15.06.2021	Gene regulatory network construction
Cytoscape	3.7.1	Visualization of TFmiR2 output
piano	2.20	Gene set analysis
stringr	1.4.0	Wrapper for processing tools for character text
psych	2.0.9	Functions for statistical analysis
ggplot2	3.3.5	Creating graphics
DESeq2	1.12.4	Identification of differentially expressed genes and miRNAs
reshape2	1.4.4	aggregation and restructuring of data
ComplexHeatmap	2.2.0	Visualization of Heatmaps

dplyr	1.0.7	Working with data frames
tidyr	1.1.4	Creation of tidy data for working with it
ggcorrplot	0.1.3	Visualizing correlation matrices
XML	3.99-0.3	Parsing XML files
org.Hs.eg.db	3.10.0	Annotation of Human genes
readr	1.4.0	reading csv files
vsn	3.54.0	Normalization of data
TCGAbiolinks	2.12.6	Facilitate access to TCGA data
data.table	1.13.0	Fast working with large data structures
SummarizedExperiment	1.14.1	Container structure

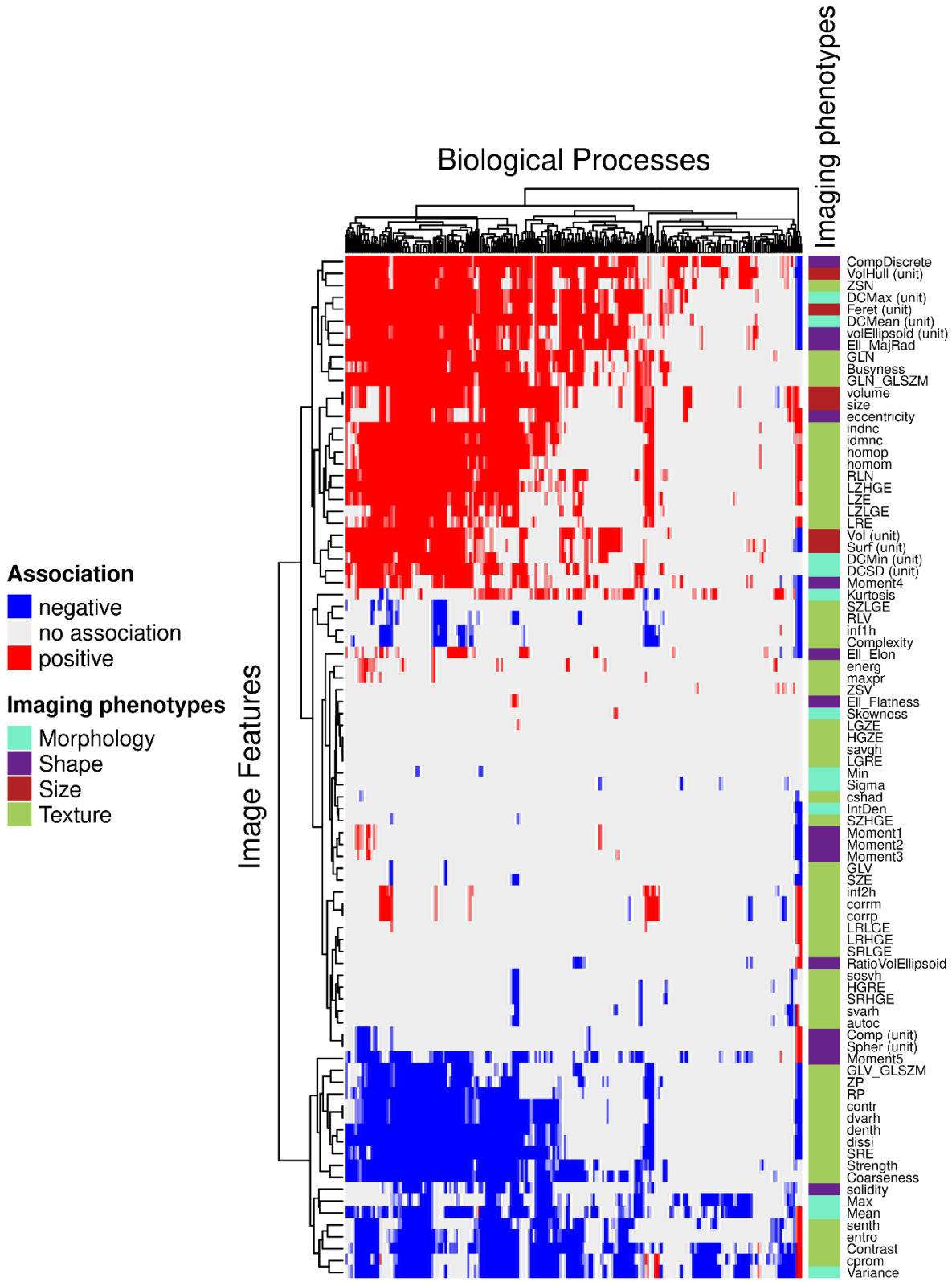


Figure 15 Piano results of the DEG data set for biological processes

The GO BP Terms show a clustering of (from left to right) highly positive and highly negative, and low positive with low negative association count. Allmost all image features have at least one associated image feature. The categories of image features do not show a specific pattern. Taken from Spath and Fischer et al. [43]

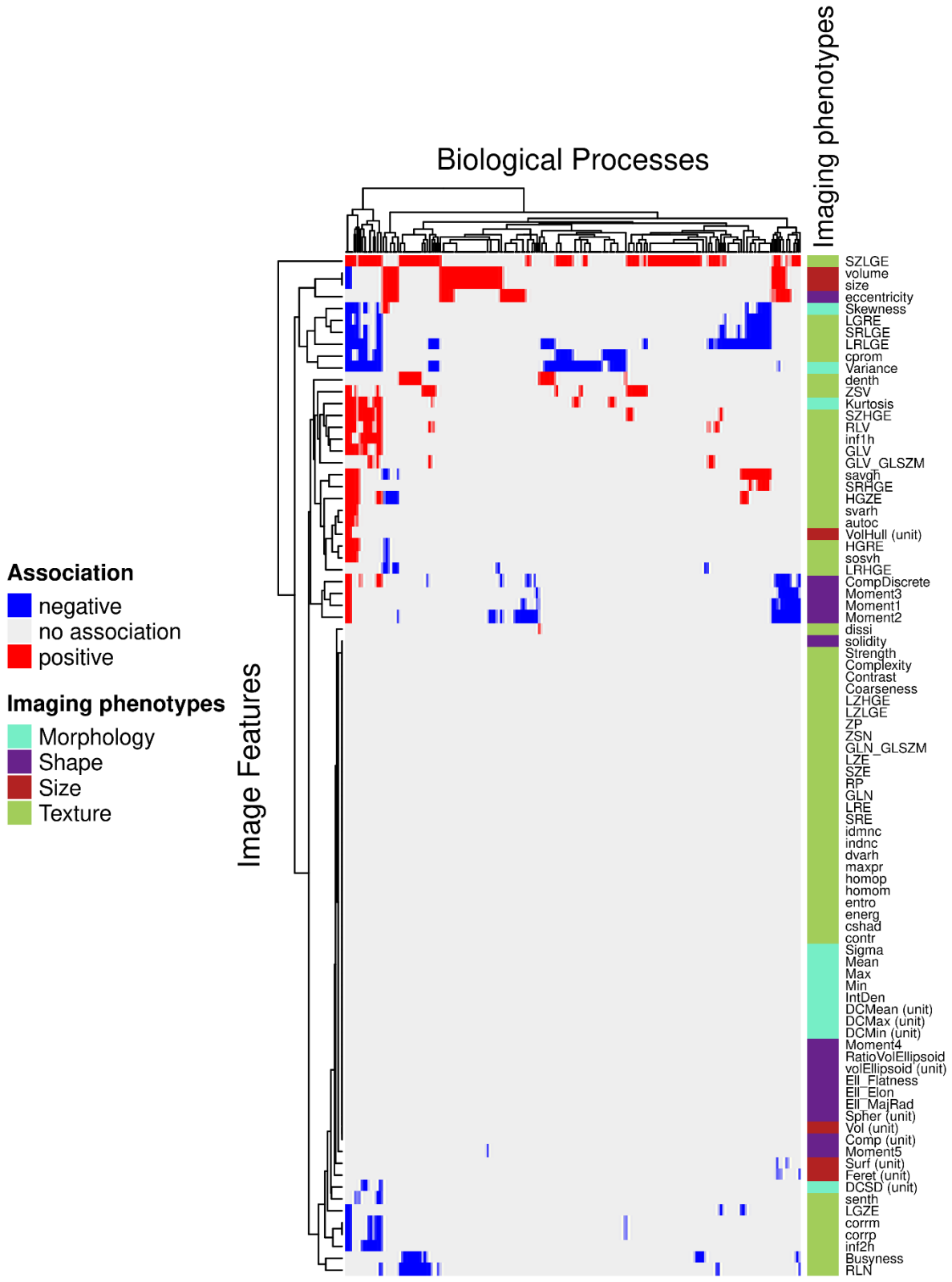


Figure 16 Piano results of the DEM data set for biological processes

The GO BP Terms show a clustering of (from left to right) highly positive and highly negative, low negative and highly positive and low positive with highly negative association count. Interestingly a group of image features (CompDiscrete, Moment1, Moment2, Moment3, volume, VollHull (unit), size, eccentricity, skewness, HGRE, sosvh, savgh, HGZE) show negative and positive associations depending on the pathways. The category morphology is relatively underrepresented. Taken from Spath and Fischer et al. [43]

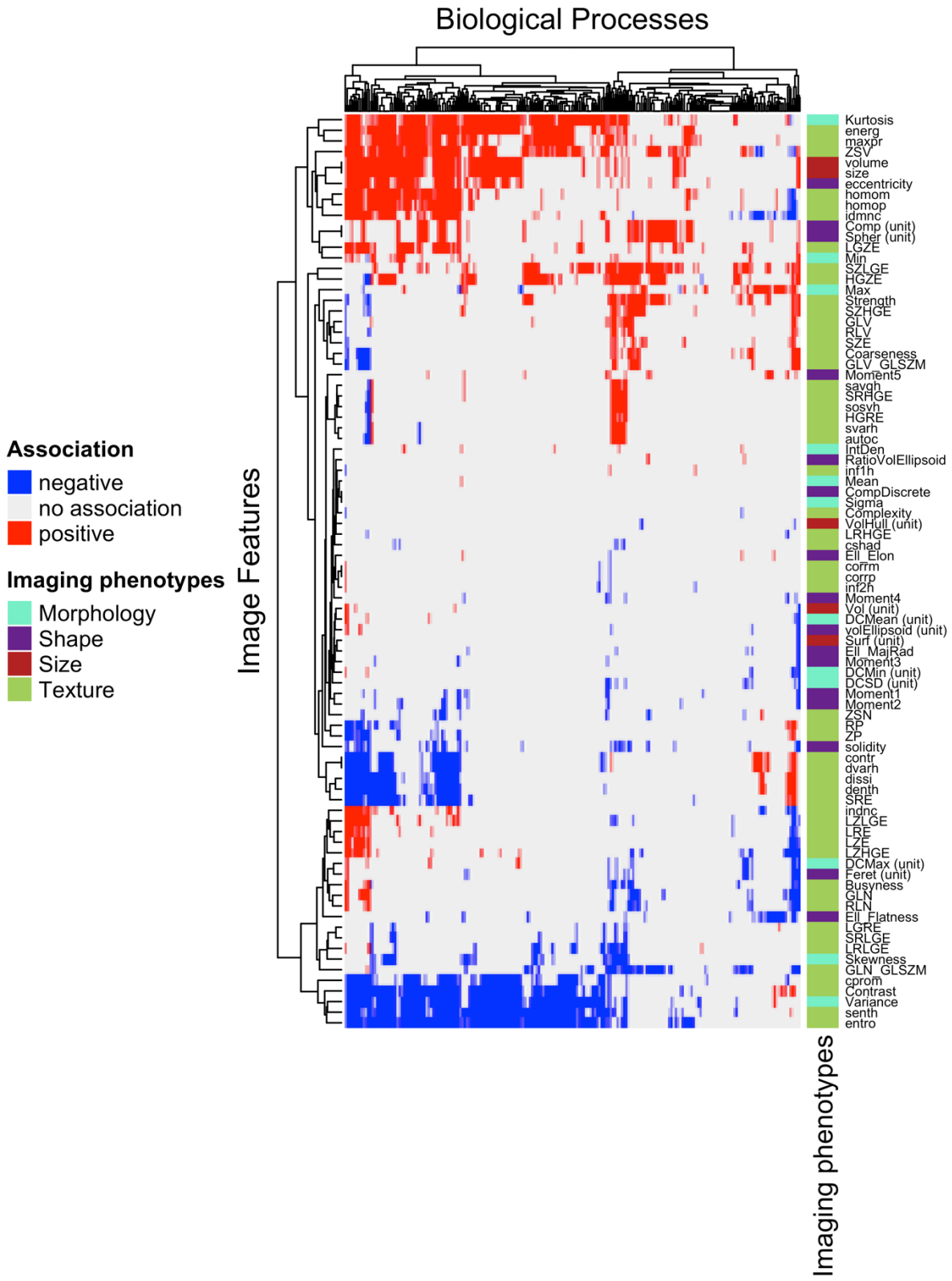


Figure 17 Piano results of the methylation data set for biological processes

The GO BP Terms show a clustering of (from left to right) highly positive to low positive association count. The negative associations are more evenly distributed over the terms. Most image features are very strict in their association direction, regardless of the pathway. Exemptions are Moment4, Moment5, ZSV maxpr, homop, homom, energ, entro and Ell_Elon. Taken from Spath and Fischer et al. [43]

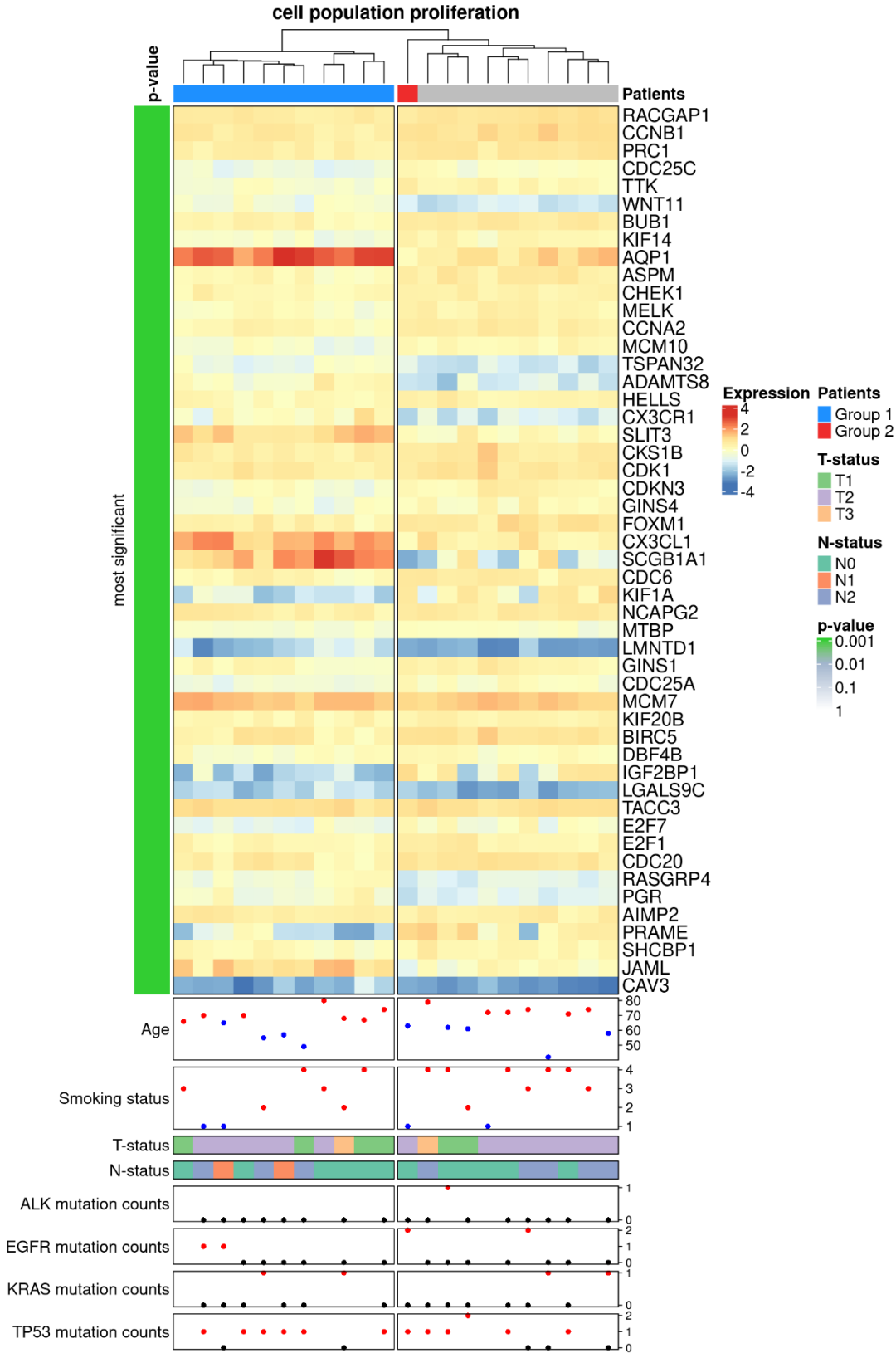


Figure 18 RAM for cell population proliferation based on mRNA expression clustering
Taken from Spath and Fischer et al. [43]

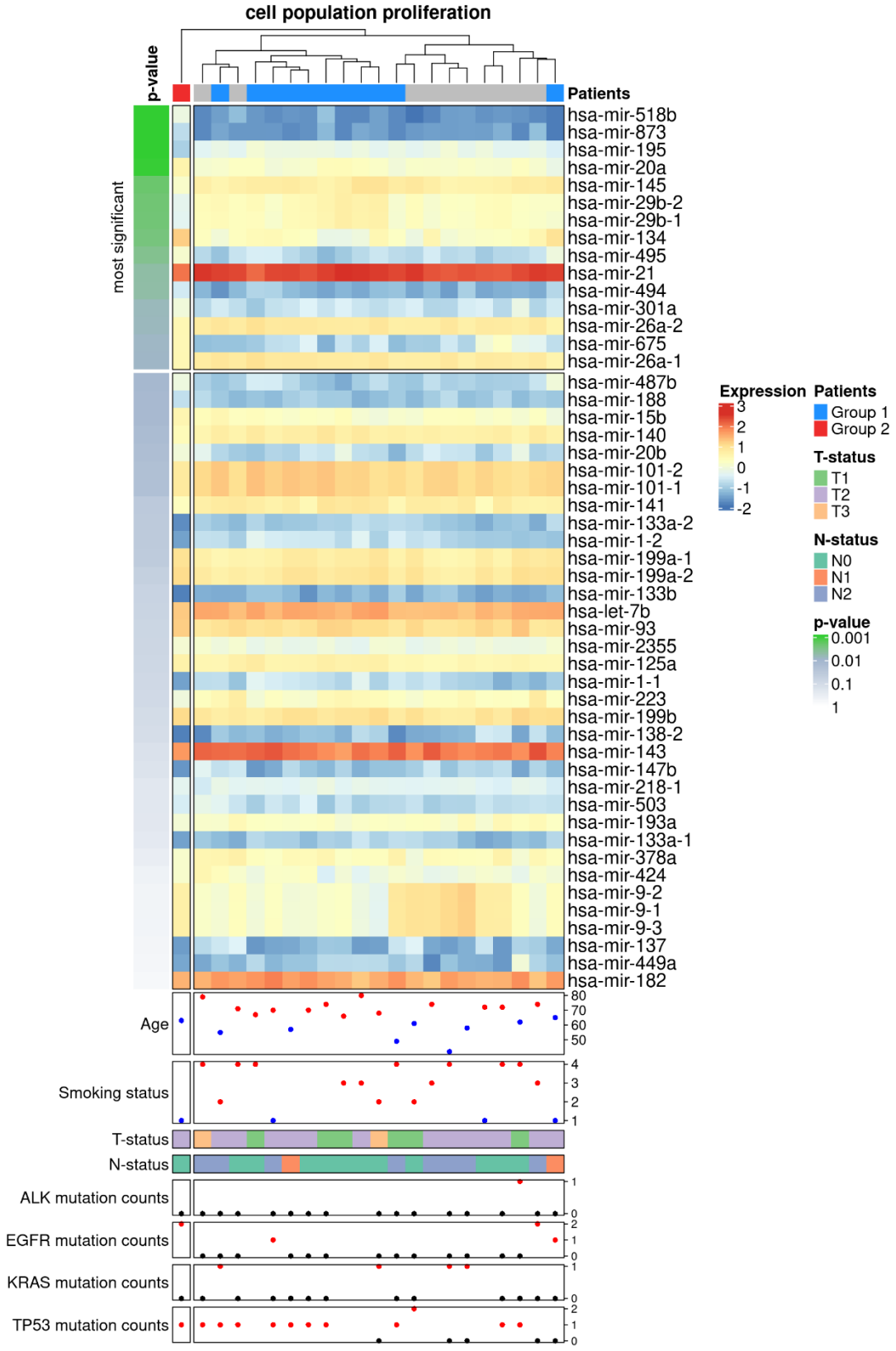


Figure 19 RAM for cell population proliferation based on miRNA expression clustering
Taken from Spath and Fischer et al. [43]

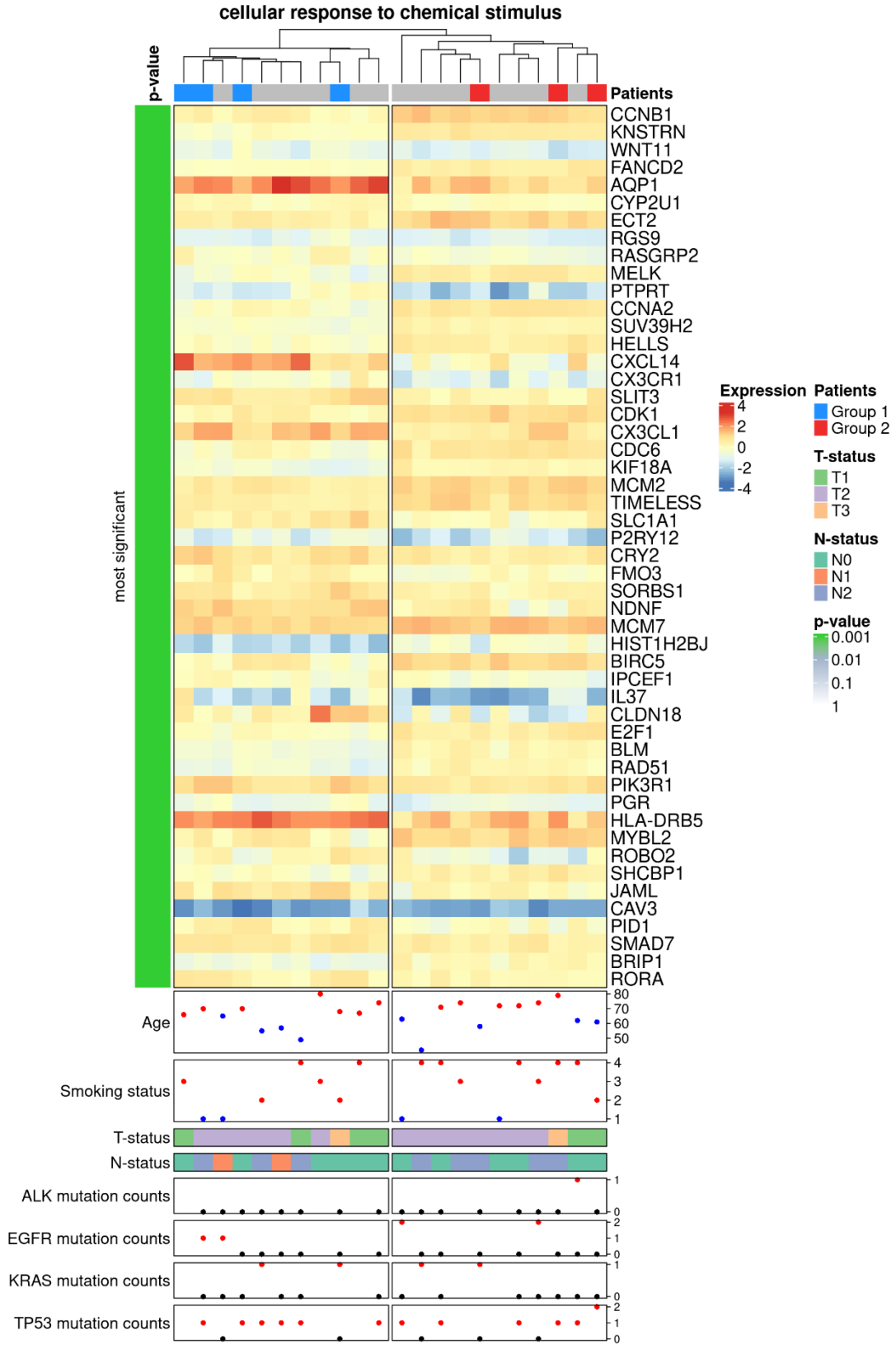


Figure 20 RAM for cellular response to chemical stimulus based on mRNA expression clustering
Taken from Spath and Fischer et al. [43]

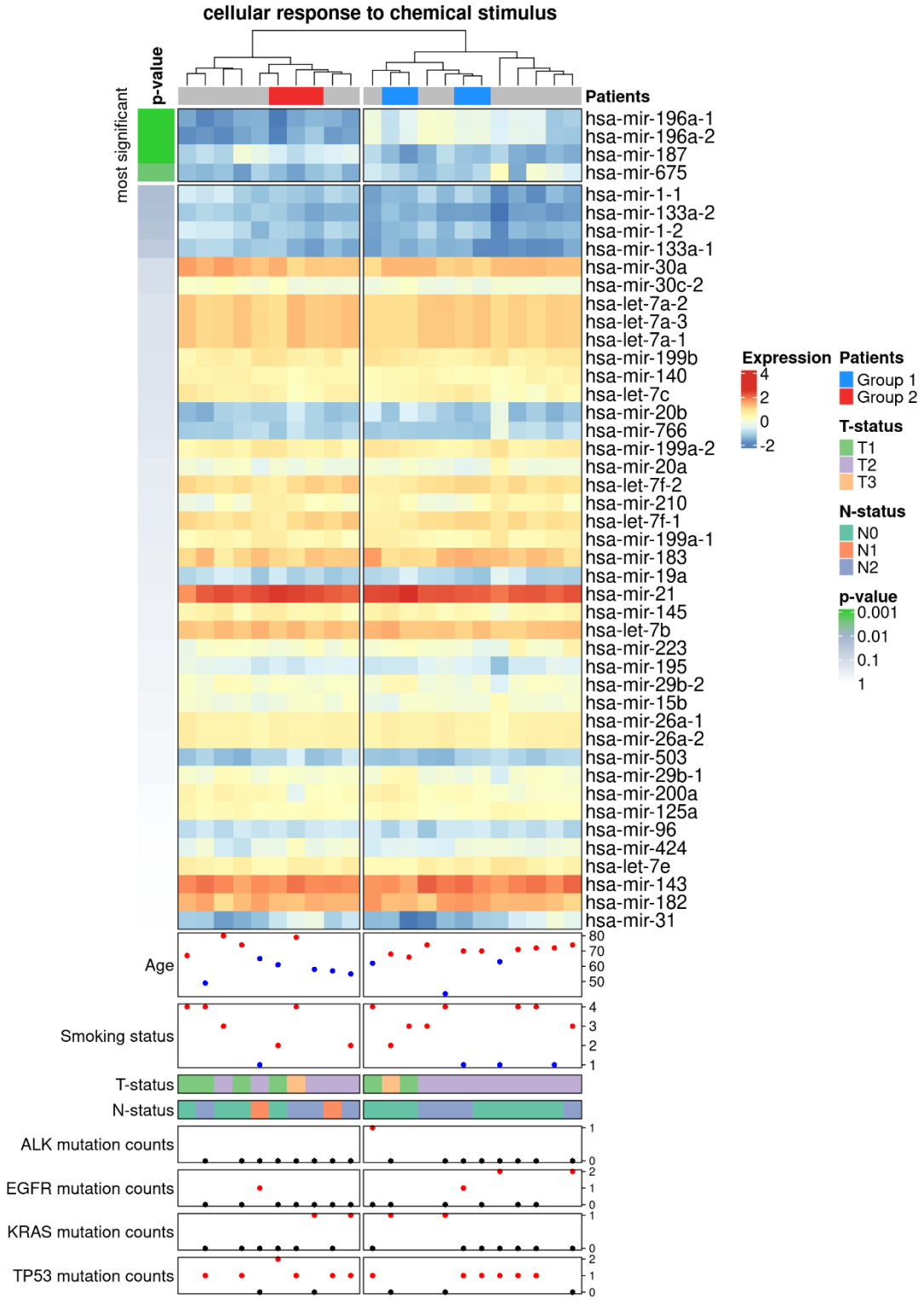


Figure 21 RAM for cellular response to chemical stimulus based on miRNA expression clustering
Taken from Spath and Fischer et al. [43]

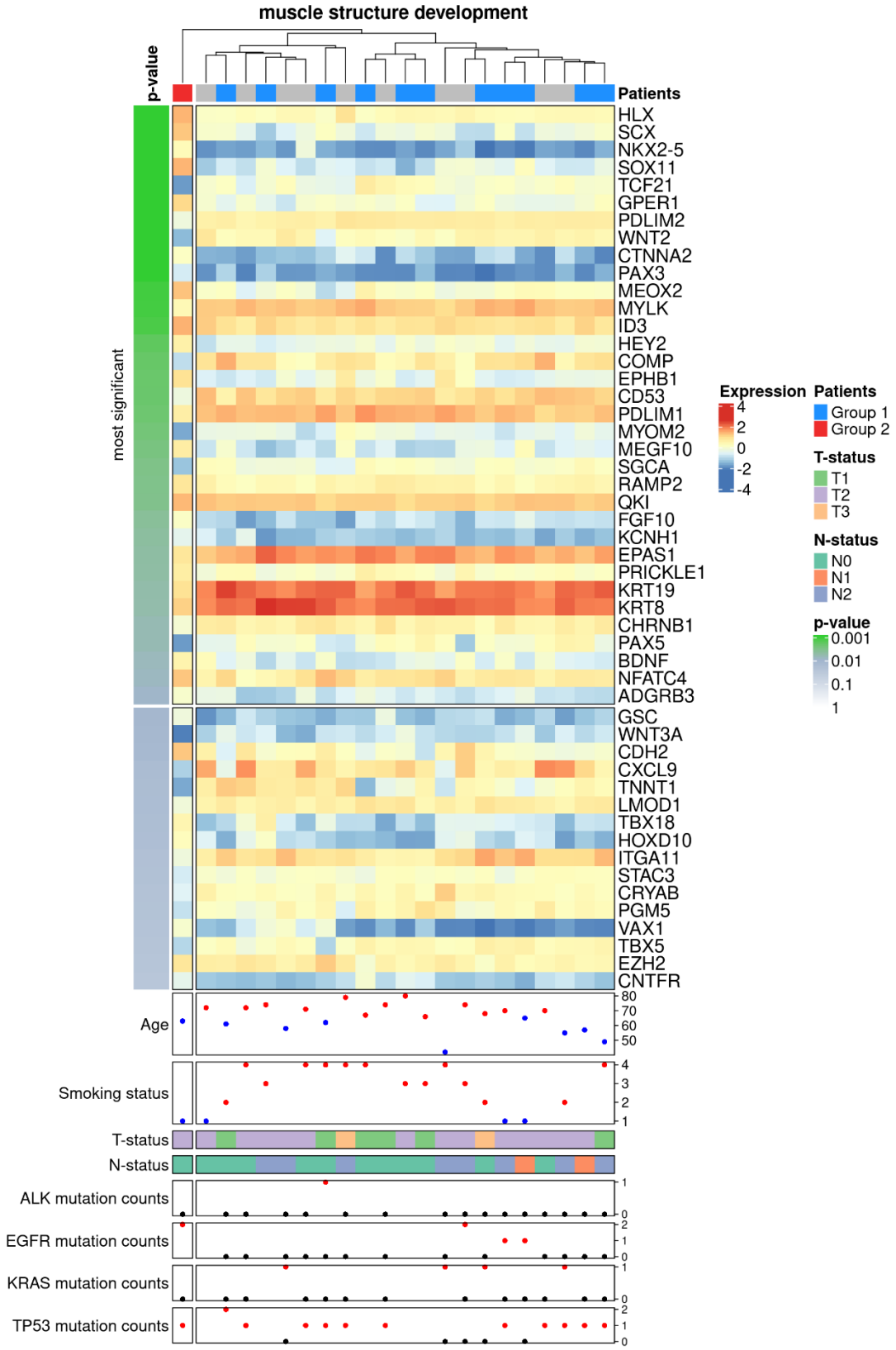


Figure 22 RAM for muscle structure development based on mRNA expression clustering
Taken from Spath and Fischer et al. [43]

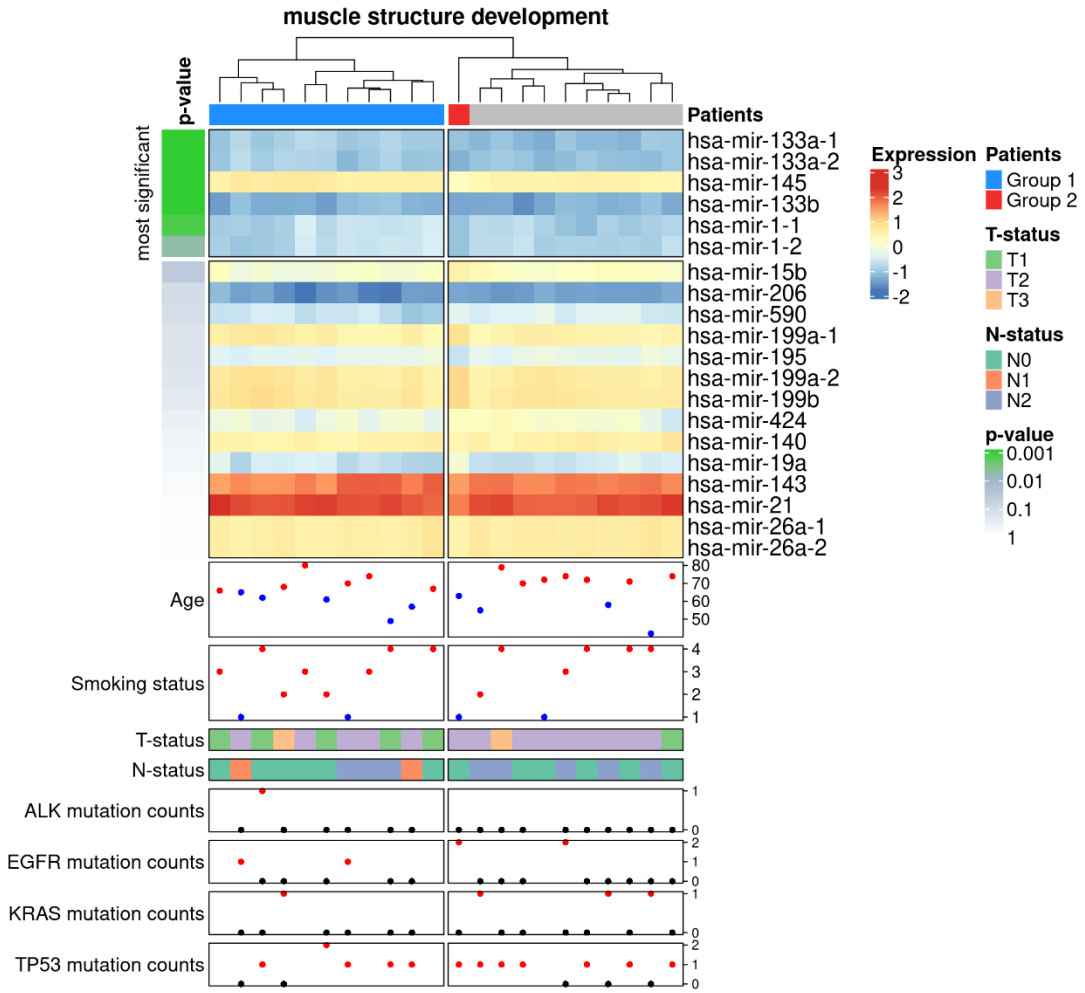


Figure 23 RAM for muscle structure development based on miRNA expression clustering
Taken from Spath and Fischer et al. [43]

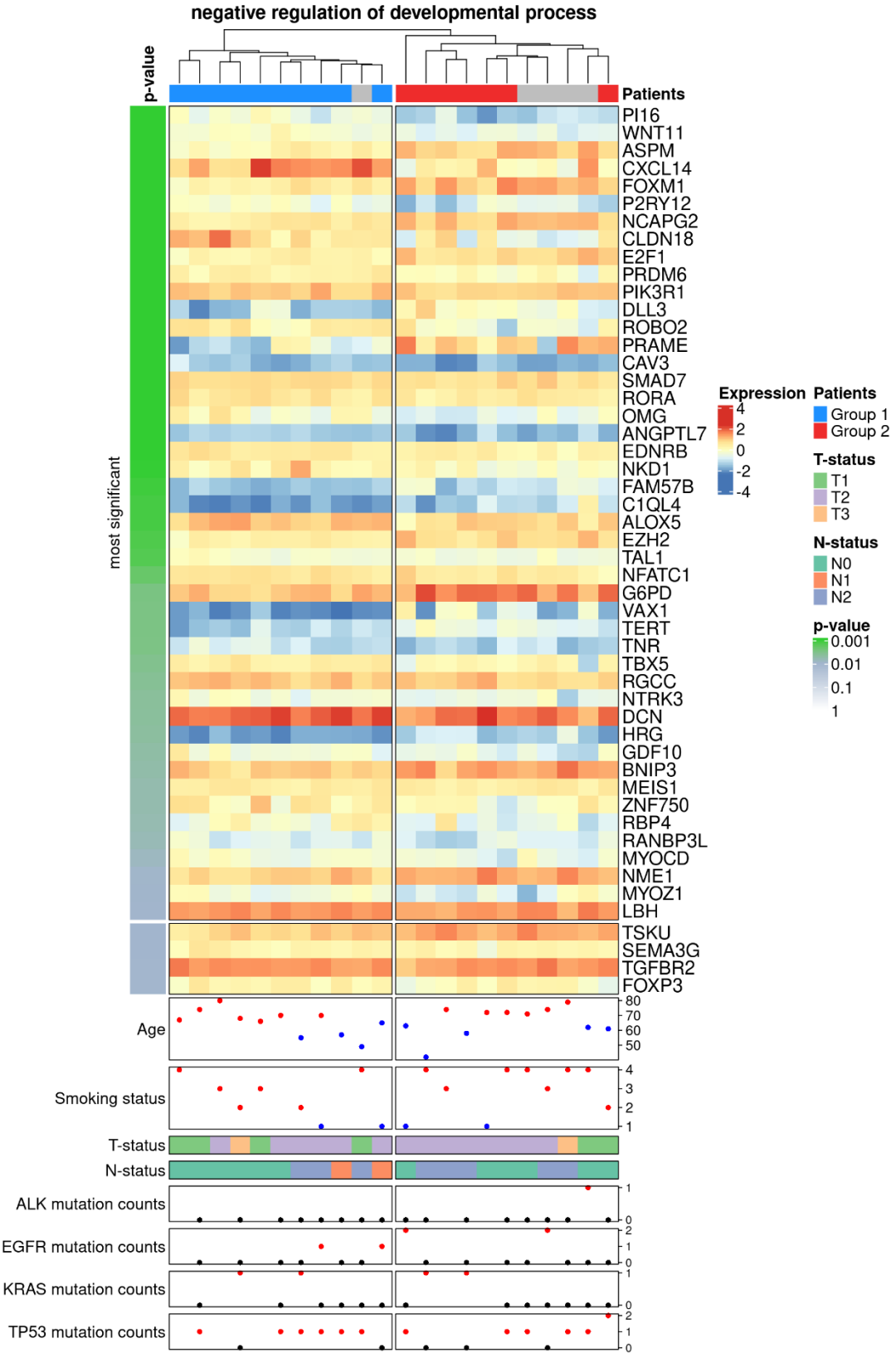


Figure 24 RAM for negative regulation of developmental process based on mRNA expression clustering Taken from Spath and Fischer et al. [43]

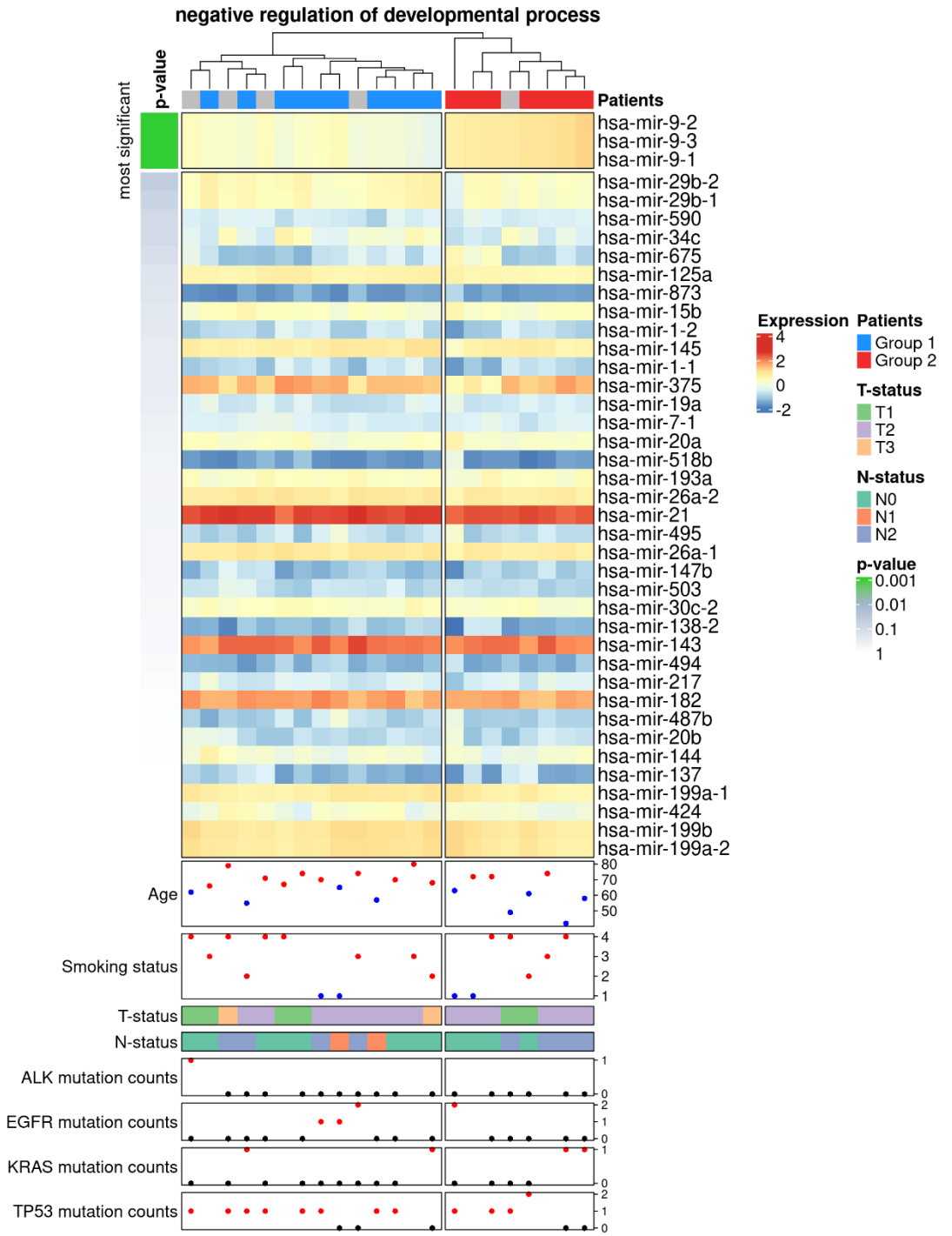


Figure 25 RAM for negative regulation of developmental process based on mRNA expression clustering
Taken from Spath and Fischer et al. [43]

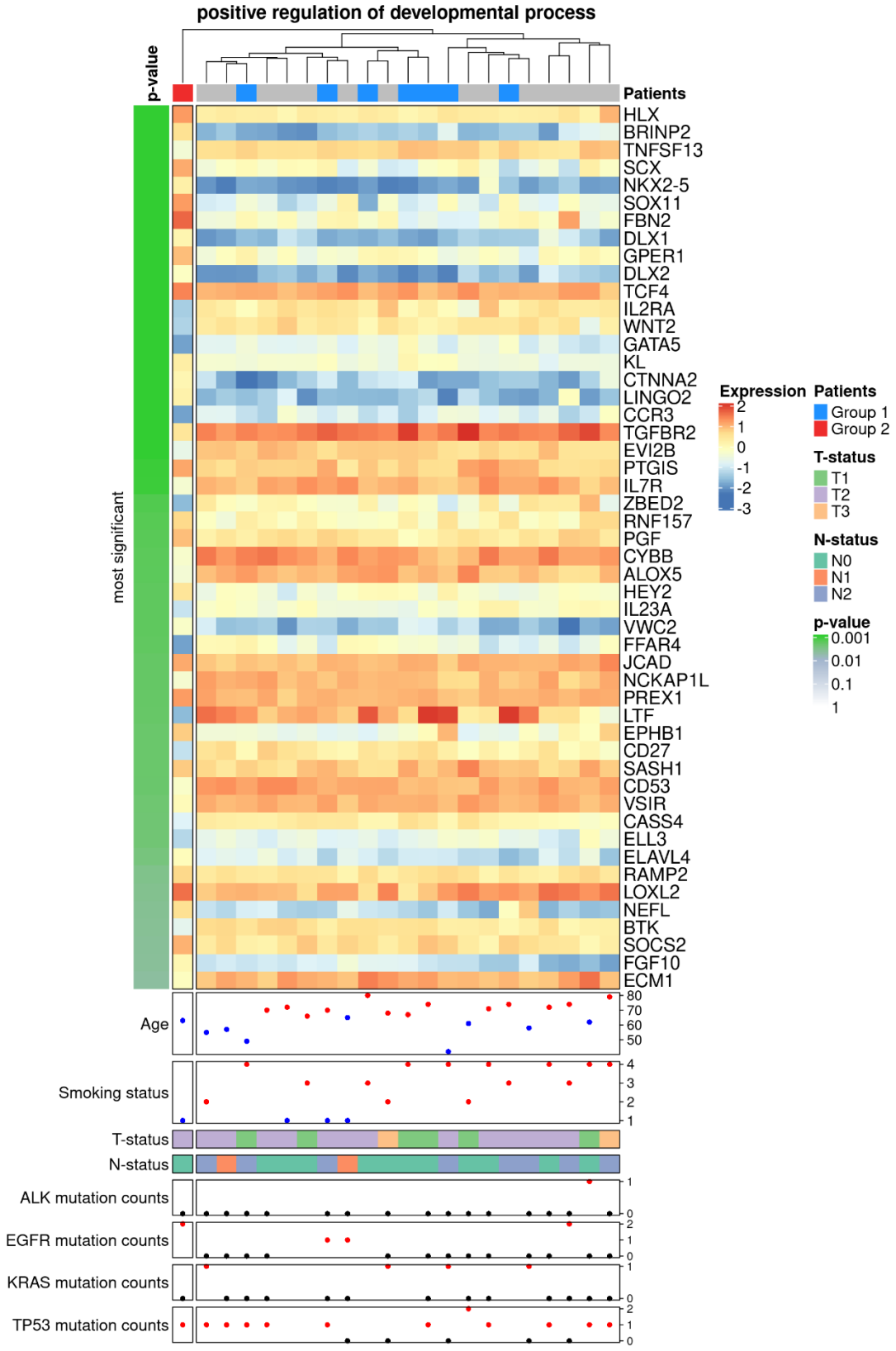


Figure 26 RAM for positive regulation of developmental process based on mRNA expression clustering Taken from Spath and Fischer et al. [43]

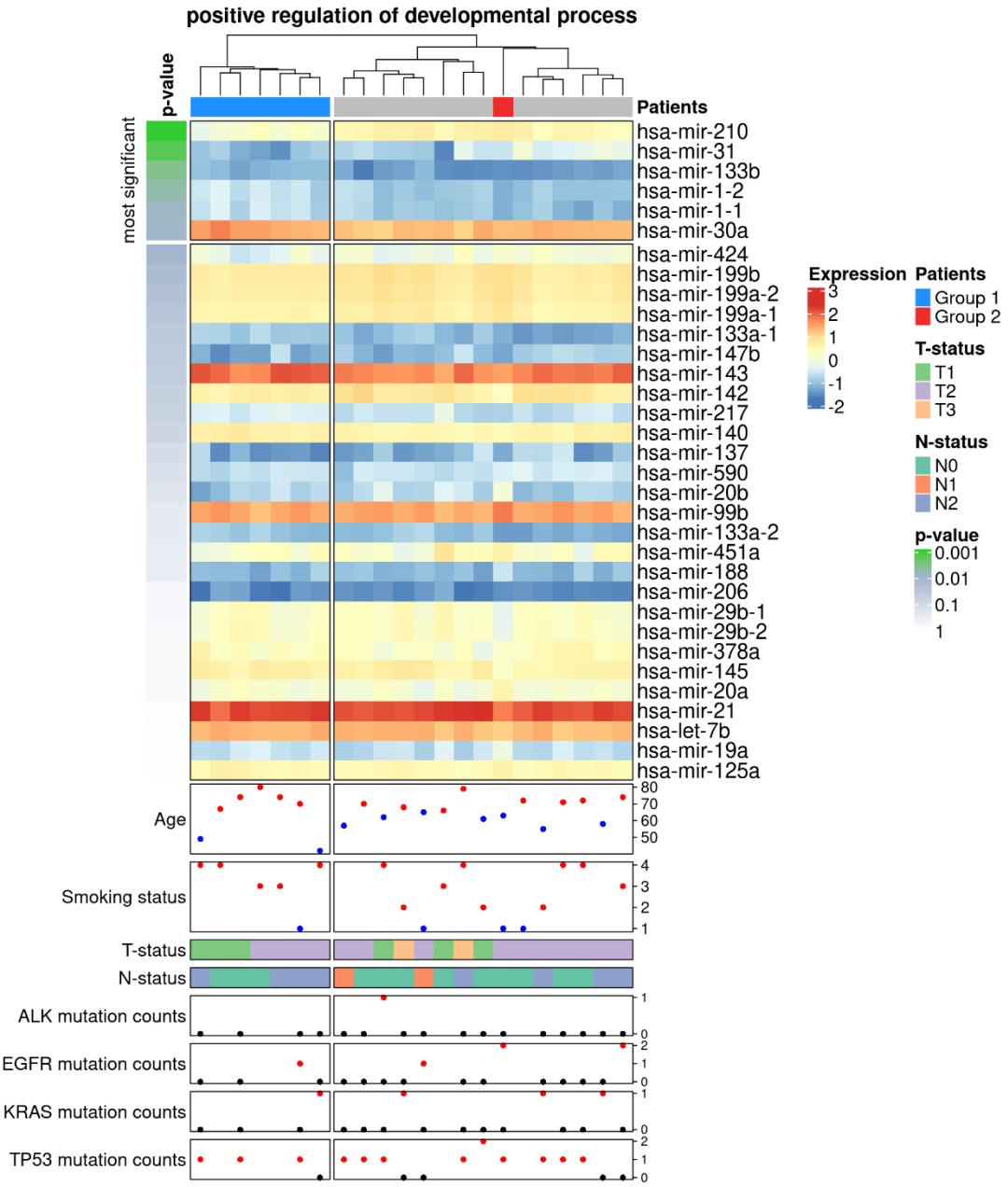


Figure 27 RAM for positive regulation of developmental process based on miRNA expression clustering
Taken from Spath and Fischer et al. [43]

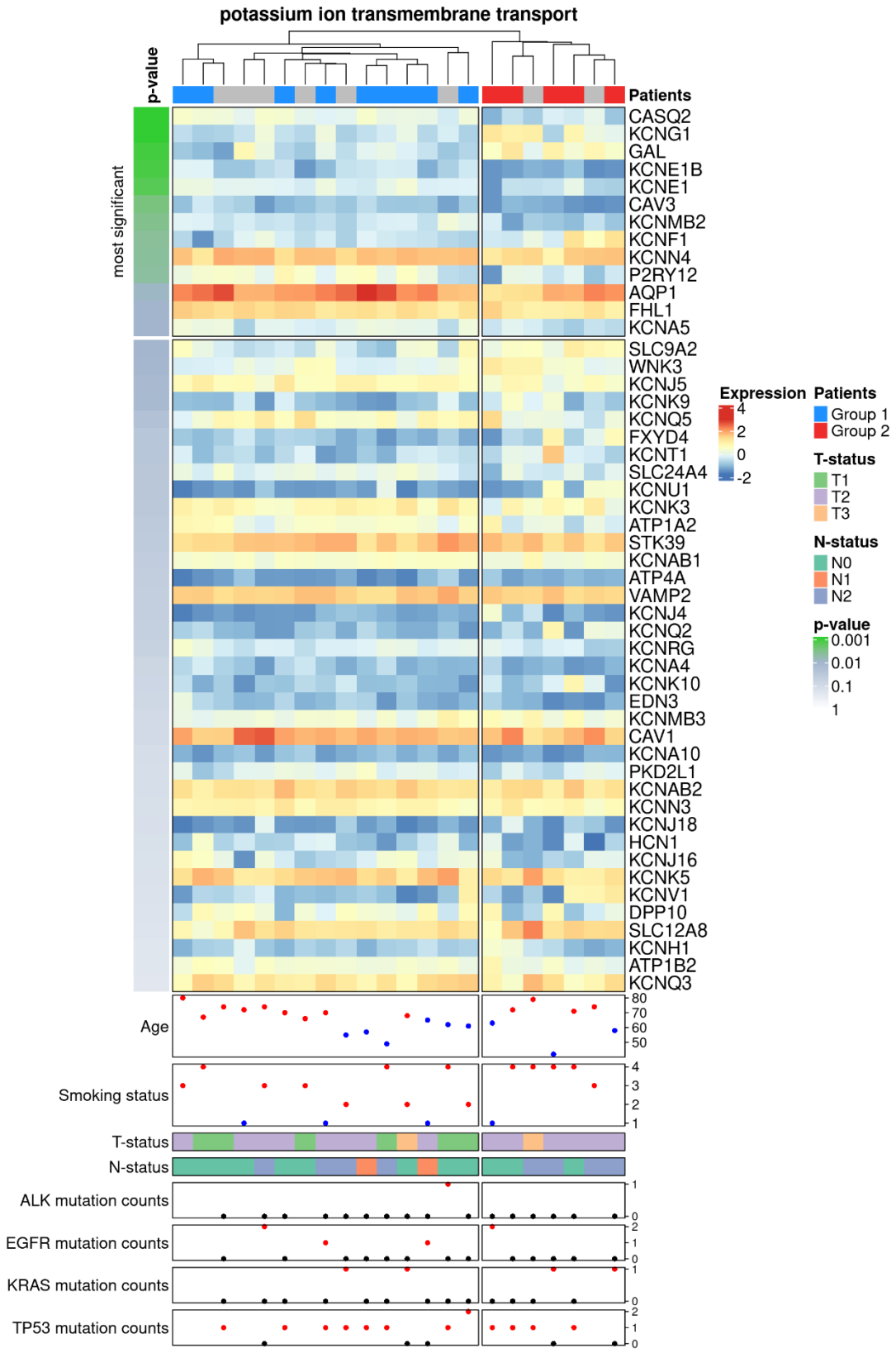


Figure 28 RAM for potassium ion transmembrane transport based on mRNA expression clustering
Taken from Spath and Fischer et al. [43]

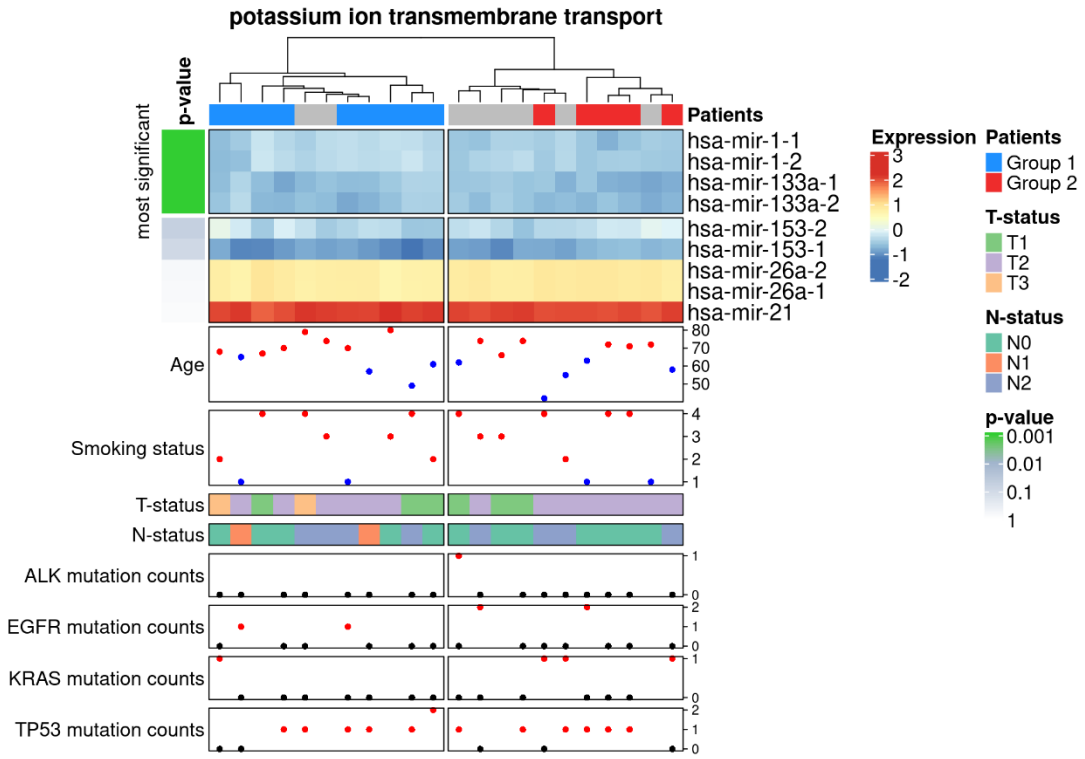


Figure 29 RAM for potassium ion transmembrane transport based on miRNA expression clustering
Taken from Spath and Fischer et al. [43]

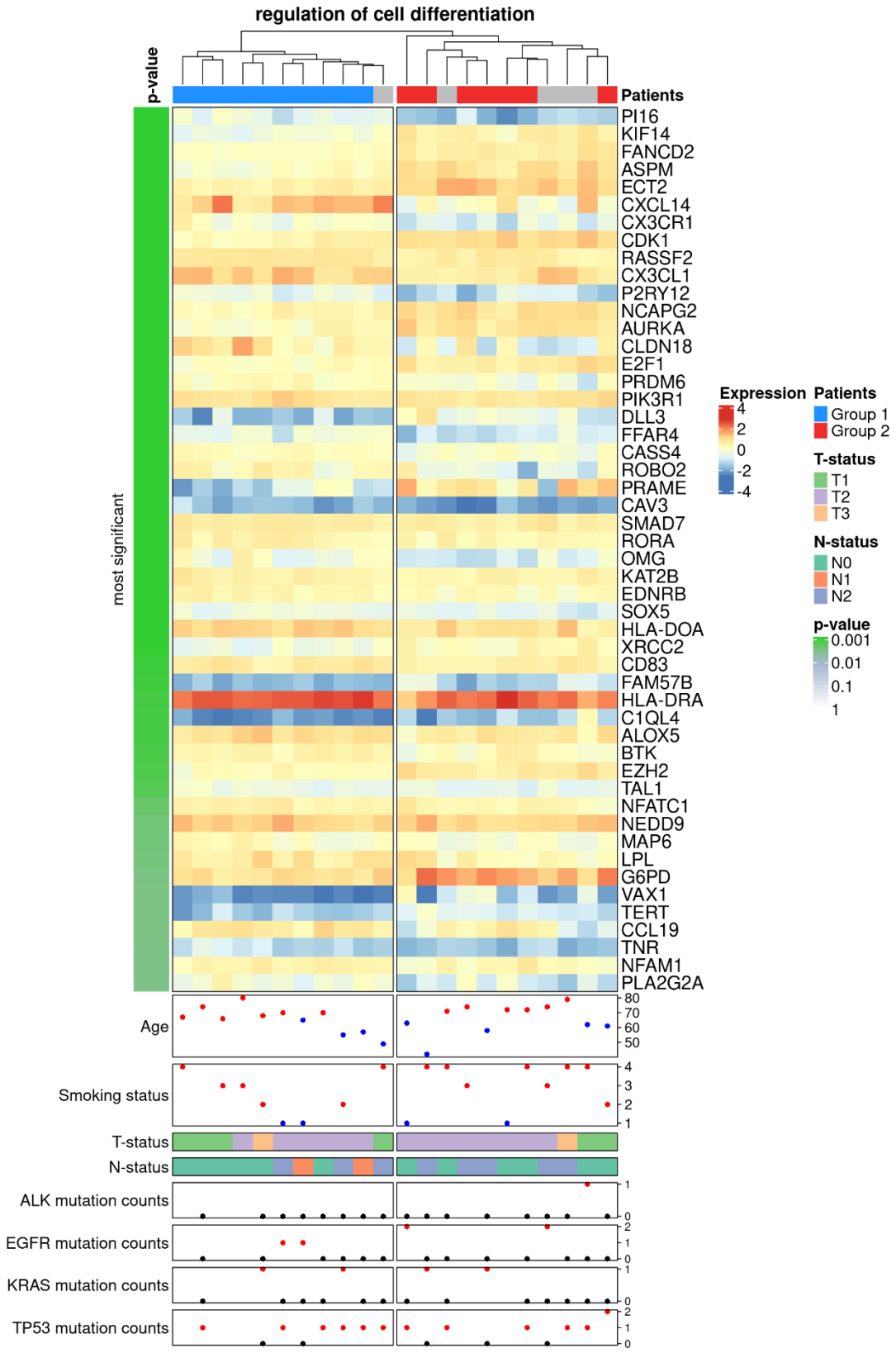


Figure 30 RAM for regulation of cell differentiation based on mRNA expression clustering
Taken from Spath and Fischer et al. [43]

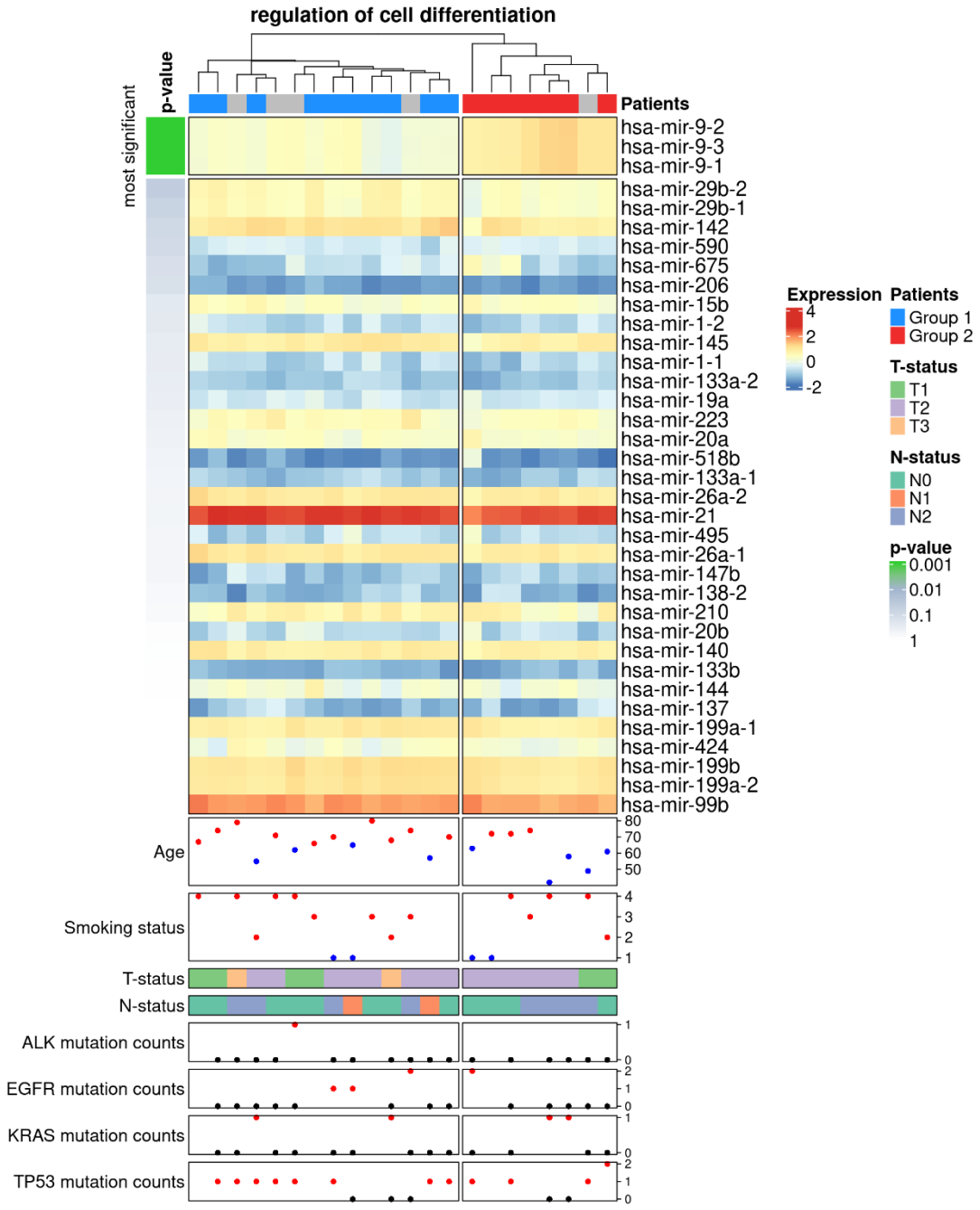


Figure 31 RAM for regulation of cell differentiation based on miRNA expression clustering
Taken from Spath and Fischer et al. [43]

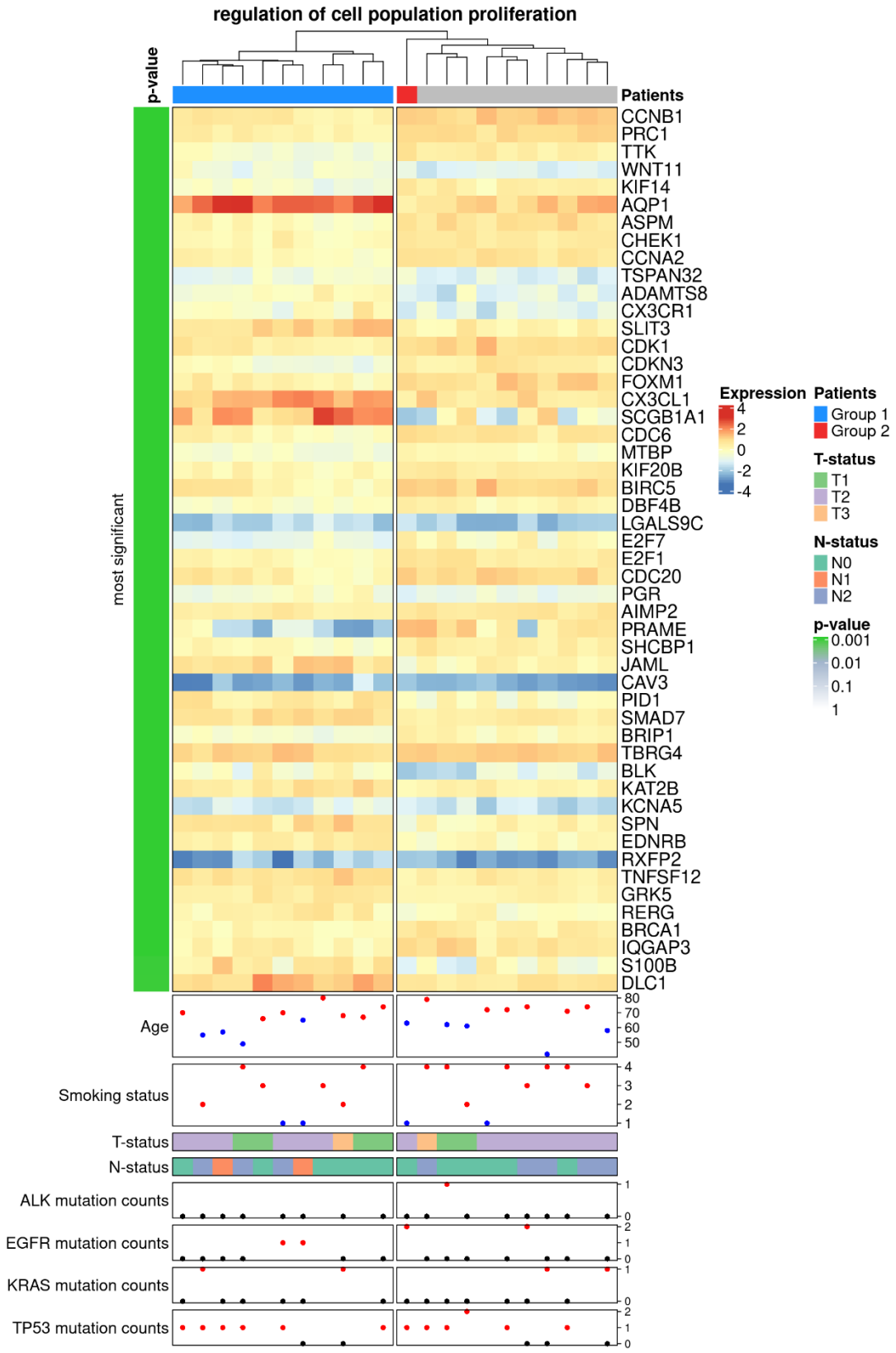


Figure 32 RAM for regulation of cell population proliferation based on mRNA expression clustering
Taken from Spath and Fischer et al. [43]

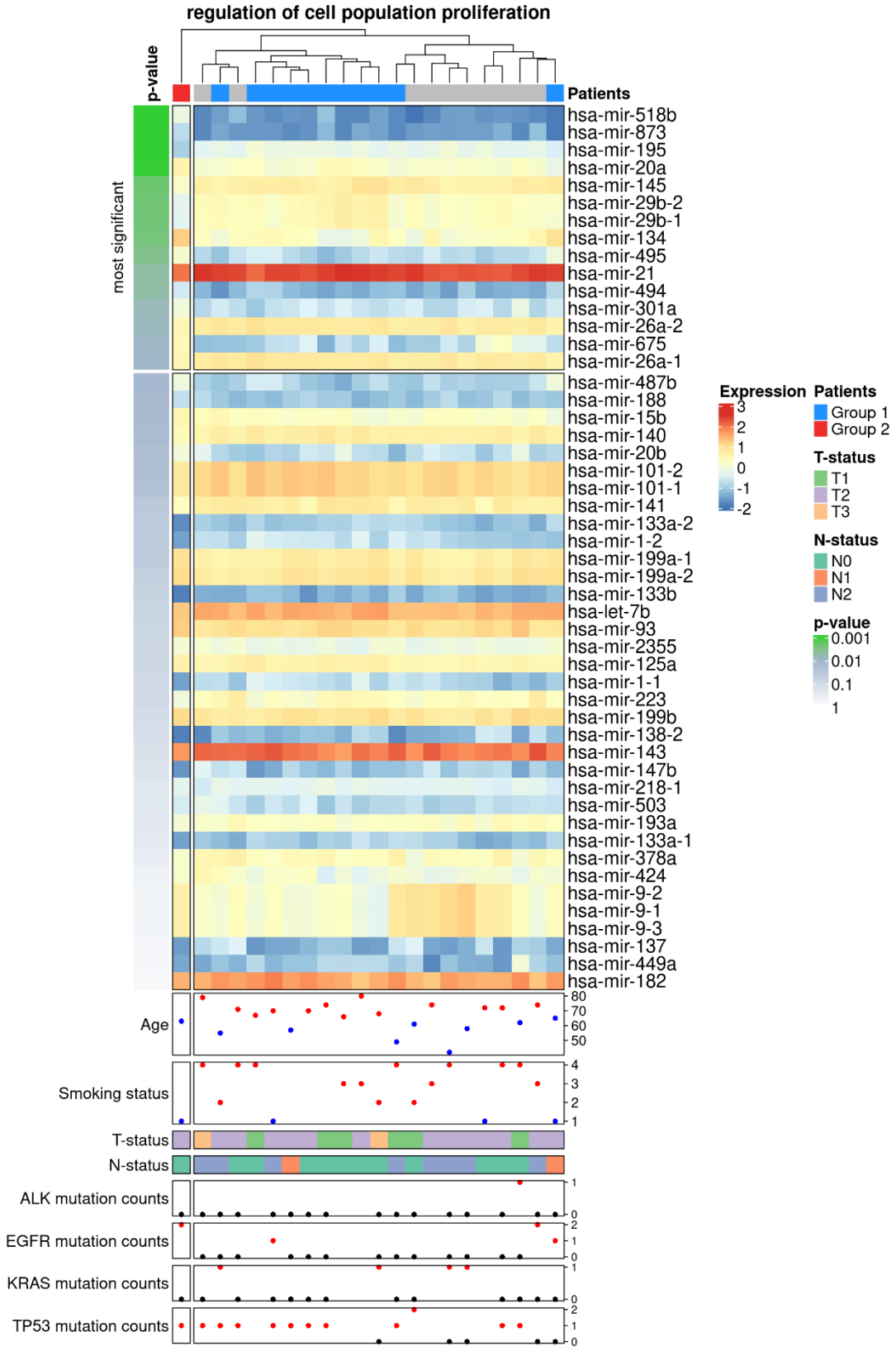


Figure 33 RAM for regulation of cell population proliferation based on miRNA expression clustering
Taken from Spath and Fischer et al. [43]

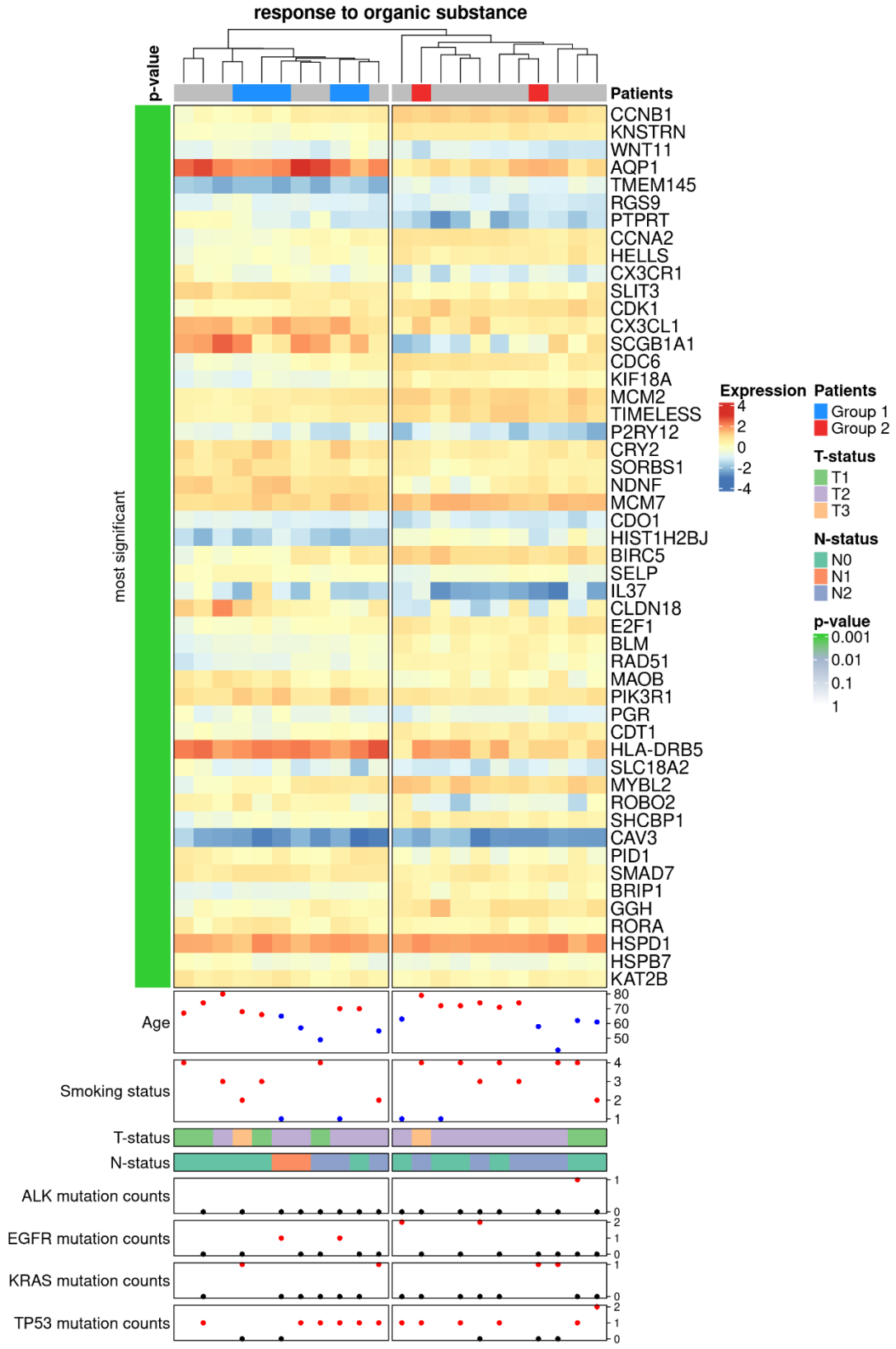


Figure 34 RAM for response to organic substance based on mRNA expression clustering
Taken from Spath and Fischer et al. [43]

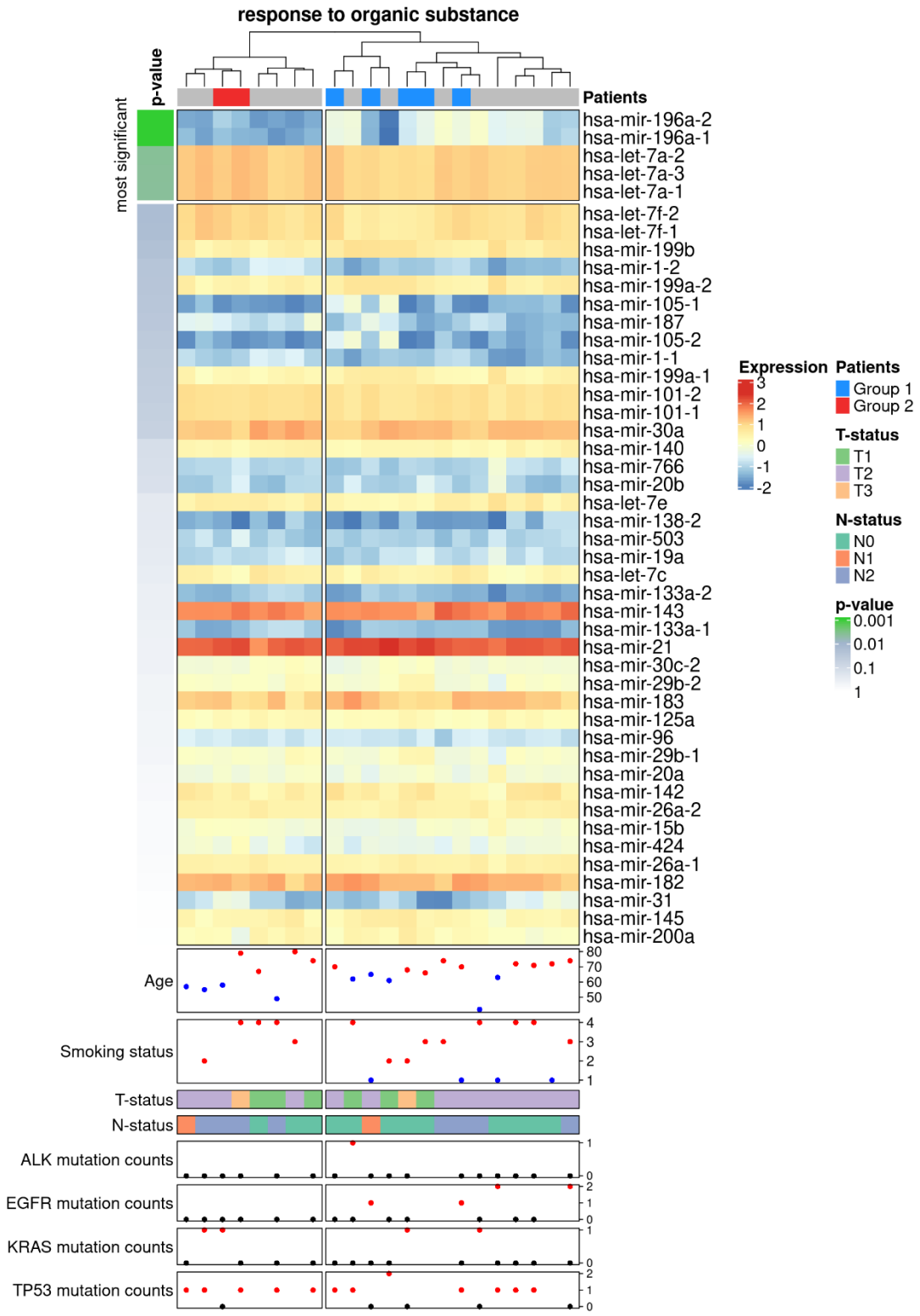


Figure 35 RAM for response to organic substance based on miRNA expression clustering
Taken from Spath and Fischer et al. [43]

9 CURRICULUM VITAE

PERSONAL REFERENCES

Name, Surname	Spath, Nicolas
Date of Birth	15.03.1996
Nationality	German
Address	Weißenburgstr. 16, 24116 Kiel
Email	nicolas.spath@web.de

CARRER SUMMARY

Physician in Oncology and Hematology 2021-2024	UKSH, Campus Kiel
Speaker at Radiogenomics Workshop 2019	Nile University, Kairo Supported by the DAAD

EDUCATION

Medical Approbation 2020	University of Rostock
Erasmus 2017/18	Centre Hospitalier Universitaire Grenoble Alpes
Abitur 2014	Don-Bosco School Rostock

PUBLICATIONS

1. Fischer S, Spath N, Hamed M. Data-Driven Radiogenomic Approach for Deciphering Molecular Mechanisms Underlying Imaging Phenotypes in Lung Adenocarcinoma: A Pilot Study. International Journal of Molecular Sciences. 2023
2. A Radiogenomic Approach in Lung Adenocarcinoma to Unravel Methylation Patterns Underlying CT Imaging Phenotypes. Spath N, Fischer S, Letsch A, Baldus C, Hövener J, Hamed M. DGHO 2023 Poster Abstract P1087

10 AFFIDAVIT

I hereby confirm that this thesis is the result of my own work. I did not receive any help or support from commercial consultants. All sources or materials applied are listed and specified in the thesis.
Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process, neither in an identical nor similar form.

Place, Date Signature

Ich versichere eidesstattlich durch eigenhändige Unterschrift, dass ich die Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen sind, habe ich als solche kenntlich gemacht.

Die Arbeit ist noch nicht veröffentlicht und ist in gleicher oder ähnlicher Weise noch nicht als Studienleistung zur Anerkennung oder Bewertung vorgelegt worden. Ich weiß, dass bei Abgabe einer falschen Versicherung die Prüfung als nicht bestanden zu gelten hat.

Ort, Datum Unterschrift