The Journal of Molecular Diagnostics, Vol. 26, No. 1, January 2024

ELSEVIER

the Journal of Molecular Diagnostics
jmdjournal.org

miR-122 and miR-21 are Stable Components of miRNA Signatures of Early Lung Cancer after Validation in Three Independent Cohorts

Joanna Zyla,* Rafal Dziadziuszko,† Michal Marczyk,*‡ Magdalena Sitkiewicz,† Magdalena Szczepanowska,† Edoardo Bottoni,§ Giulia Veronesi,¶‖ Witold Rzyman,† Joanna Polanska,* and Piotr Widlak†

From the Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland; the Medical University of Gdansk,† Gdansk, Poland; the Yale Cancer Center,‡ Yale School of Medicine, New Haven, Connecticut; Humanitas Research Hospital,§ Milan, Italy; the School of Medicine and Surgery,¶ Vita-Salute San Raffaele University, Milan, Italy; and the Department of Thoracic Surgery,‖ IRCCS San Raffaele Scientific Institute, Milan, Italy*

Accepted for publication September 28, 2023.

Address correspondence to Joanna Polanska, Ph.D., Department of Data Science and Engineering, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland; or Piotr Widlak, Ph.D., Medical University of Gdansk, M. Skłodowskiej-Curie 3a, 80-210 Gdansk, Poland. E-mail: joanna.polanska@polsl.pl or piotr.widlak@gumed.edu.pl.

Several panels of circulating miRNAs have been reported as potential biomarkers of early lung cancer, yet the overlap of components between different panels is limited, and the universality of proposed biomarkers has been minimal across proposed panels. To assess the stability of the diagnostic potential of plasma miRNA signature of early lung cancer among different cohorts, a panel of 24 miRNAs tested in the frame of one lung cancer screening study (MOLTEST-2013, Poland) was validated with material collected in the frame of two other screening studies (MOLTEST-BIS, Poland; and SMAC, Italy) using the same standardized analytical platform (the miRCURY LNA miRNA PCR assay). On analysis of selected miRNAs, two associated with lung cancer development, miR-122 and miR-21, repetitively differentiated healthy participants from individuals with lung cancer. Additionally, miR-144 differentiated controls from cases specifically in subcohorts with adenocarcinoma. Other tested miRNAs did not overlap in the three cohorts. Classification models based on neither a single miRNA nor multicomponent miRNA panels (24-mer and 7-mer) showed classification performance sufficient for a standalone diagnostic biomarker (AUC, 75%, 71%, and 53% in MOLTEST-2013, SMAC, and MOLTEST-BIS, respectively, in the 7-mer model). The performance of classification in the MOLTEST-BIS cohort with the lowest contribution of adenocarcinomas was increased when only this cancer type was considered (AUC, 60% in 7-mer model). *(J Mol Diagn 2024, 26: 37–48; https://doi.org/10.1016/j.jmoldx.2023.09.010)*

Lung cancer is the leading cause of cancer-related death worldwide.[1] Late diagnosis is a crucial issue related to mortality from this malignancy: About 15% of cases are diagnosed at an early stage (I or II), in which the likelihood of 5-year survival exceeds 60%, while almost 60% of cases are diagnosed in the metastatic stage (IV), in which the prognosis is very poor, with a likelihood of 5-year survival of <5%.[2,3] The main etiologic factor of lung cancer is cigarette smoking, with 85% of all cases being attributable to this exposure.[4] In addition to the efforts to reduce nicotine use (primary prevention), secondary prevention with lung cancer screening in high-risk groups is crucial for a reduction in the risk for lung cancer–related mortality. At present, screening using low-dose computed tomography (LDCT) in the high-risk smoking population is the only effective tool. In 2011 and 2020, the results from two randomized, controlled studies of screening were published, NLST (National Lung Screening Trial)[5] and NELSON (NEderlands Leuvens Screening ONderzoek trial),[6] which showed 20% and 26% reductions, respectively, in the rates of lung cancer–related mortality in a high-risk group undergoing three rounds of LDCT screening. In the United

Supported by National Science Centre (Poland) grant 2017/27/B/NZ7/01833; clinical material was collected and initially characterized in the frame of MOLTEST-BIS project DZP/PBS3/247184/2014 and CLEARLY project ERA-NET TRANSCAN/04/2018; partially supported by Silesian University of Technology grant 02/070/BK_23/0043 for Support and Development of Research Potential. All research performed by J.Z., M.M., and J.P. is within the scope of the Technical Informatic and Telecommunication discipline recognized by the Polish Ministry of Science.

https://doi.org/10.1016/j.jmoldx.2023.09.010

States, Croatia, Australia, and a few Chinese cantons, lung cancer screening has already been introduced as a population-based study, and pilot programs are ongoing in several European countries. However, the disadvantage of the LDCT-based test is a large number of indeterminant lung nodules, that is, in high-risk patients, lung nodules are detected in 30% to 50% of patients, while the actual malignancy is confirmed with further diagnostics in 1% to 2% of patients. It has been reported that a variable but non-negligible number of screened patients with lung lesions detected on LDCT are subjected to unnecessary diagnostic procedures, depending on the experience of the screening team.[7] This problem, referred to as false-positive diagnosis, is, apart from economic issues, the main source of doubt as to the feasibility and legitimacy of the widespread introduction of this screening method. Therefore, it is generally assumed that combining LDCT with an additional test would be a beneficial strategy for increasing the effectiveness and lowering the cost of lung cancer screening programs.

An obvious candidate for the LDCT's support test is a molecular diagnostic test, and for over a decade, intensive research to identify biomarkers of early lung cancer has been conducted. These studies have been focused on various components of the blood, including circulating tumor cells, tumor-informed platelets, circulating free DNA, and autoantibodies, and on components of the proteome, peptidome, and transcriptome of serum/plasma.[8,9] As a result, several biomarkers with potential in identifying lung cancer have been proposed. The most advanced clinical tests of the two most frequently studied types of putative biomarkers—panels of serum/plasma proteins and multicomponent signatures of miRNA—are currently in progress. miRNAs are a class of short (18 to 24 nucleotides), noncoding RNAs involved in regulating the expression of specific target genes. To date, 2500 human miRNAs that regulate the expression of thousands of genes have been described, including >500 miRNA types detected in the blood. The repertoire of miRNAs is affected by a variety of pathologic conditions; hence, serum/plasma miRNA content is an attractive source of biomarkers for many diseases, including cancer.[10] The literature on the subject has reported >20 multicomponent serum/plasma miRNA panels as potential biomarkers of lung cancer (with sensitivity and specificity of 70% to 90%).[11,12] However, a few panels have been clinically verified in validation studies. Only three registered clinical studies to validate early lung cancer serum/plasma miRNA signatures are currently underway: NCT02247453 (*https://clinicaltrials.gov/study/NCT02247453*, last accessed August 22, 2023), NCT01248806 (*https://clinicaltrials.gov/study/NCT01248806*, last accessed August 22, 2023), and NCT03452514 (*https://clinicaltrials.gov/study/NCT03452514*, last accessed August 22, 2023). None of these clinical trials have yet been concluded, and no properly validated diagnostic tests for early lung cancer based on the miRNA signature are currently on the market. Importantly, of all proposed lung cancer signatures involving >100 miRNA species, only four (namely, miR-21, miR-148b, miR-126, miR-486-5p) recurred in more than five signatures.[12] Hence, the overlap among different signatures is limited, which putatively reflects the differing clinical characteristics of lung cancer patients and their ethnic/genetic backgrounds, as well as differing analytical approaches and statistical methods used in various studies. Moreover, the repeatability of the diagnostic performance of specific miRNAs among cohorts has not been verified.

The present retrospective analysis aimed to validate a specific panel of miRNA and the diagnostic stability of its components, using blood components collected from three large-scale, independent, LDCT-based lung cancer screening studies. The diagnostic performance of the signature of 24 miRNAs established and used for the classification of cancer cases diagnosed in the frame of the Pomeranian Lung Cancer Screening Program (2009 to 2010) was validated in the two independent lung screening cohorts from programs performed recently in Poland and Italy.

Materials and Methods

Study Subjects

The biological material included in this study was collected during three independent lung cancer screening programs that offered LDCT to current or former smokers: i) the Pomeranian Lung Cancer Screening Program (MOLTEST-2013; 2009 to 2010); ii) MOLTEST-BIS, performed by the Medical University of Gdansk (Gdansk, Poland; 2016 to 2018); and iii) the Smokers Health Multiple Action (SMAC) program, performed by the Humanitas Clinical and Research Center (Milan, Italy; 2018 to 2021). These programs enrolled over 3600, 6000, and 2000 participants, respectively, from whom blood samples were collected together with LDCT scans. Blood samples from participants who were ultimately diagnosed with lung cancer and participants with no CT-detected lung nodules and no other cancer-related health problems were matched according to age, sex, and smoking history in the present study. The characteristics of all groups are presented in Table 1.

Due to the low number of available screening-detected cancer cases, a few patients with asymptomatic, low-advanced lung cancer detected occasionally not in the frame of the screening, and who fit the screening inclusion criteria, were included (69 and 13 patients in MOLTEST-2013 and SMAC, respectively; all cancer cases in MOLTEST-BIS were detected on screening). Study protocols were approved by the appropriate ethics committees (Medical University of Gdansk approval numbers NKEBN/42/2009 and NKBBN/376/2014; and Humanitas Clinical and Research Center approval number CE Humanitas ex DM 390/18), and all participants provided informed consent indicating their voluntary participation in the project and provision of blood samples for future research. The

Table 1 Characteristics of Donor Cohorts

Cohort/parameter	MOLTEST-2013	MOLTEST-BIS	SMAC	Population equality and independency testing *P* value
Healthy controls				
n	291	296	88	
Sex: male/female	154/137 (53%/47%)	171/125 (62%/38%)	55/33 (62%/38%)	$P = 0.2248$
Age range, years (median)	50–77 (63)	52–78 (67)	54–91 (67)	$P < 0.0001$
Smoking habit, packs/year (median)	10–90 (30)	26–132 (47)	1–100 (44)	$P < 0.0001$
Lung cancer cases				
n	102	99	32	
Sex: male/female	56/46 (55%/45%)	54/45 (55%/45%)	19/13 (59%/41%)	$P = 0.8852$
Age range, years (median)	49–77 (63)	53–78 (67)	61–88 (75)	$P < 0.0001$
Smoking habit, packs/year (median)	15–80 (30)	29–138 (48)	0–96 (49)	$P < 0.0001$
Cancer type, *n* (%)				
Adenocarcinoma	62 (61)	58 (59)	24 (75)	$P = 0.0191$
Squamous cell carcinoma	37 (36)	31 (31)	4 (12.5)	
Other types of lung cancer	3 (3)	10 (10)	4 (12.5)	
Cancer stage, *n* (%)				
Stage I	64 (63)	48 (49)	17 (53)	$P < 0.0001$
Stage II	32 (31)	19 (19)	7 (22)	
Stage III	6 (6)	18 (18)	6 (19)	
Stage IV	0 (0)	14 (14)	2 (6)	

distributions of age and smoking status were compared between cohorts by analysis of variance with the Tukey *post-hoc* procedure, while the χ^2 test was used to test for the independence between cohort, sex, and group variables. $P < 0.05$ was assumed as significant.

Analysis of Plasma miRNA

Plasma specimens were purified from blood samples collected in EDTA-containing Vacutainer tubes (Becton Dickinson, Franklin Lakes, NJ) using a standardized protocol based on two-step centrifugation. Briefly, blood was centrifuged at 600 × *g* for 20 minutes then at 1500 × *g* for 15 minutes (both steps at 4°C). Plasma specimens were apportioned into 0.5-mL aliquots and stored at −80°C; all samples were prepared within 1 hour of blood collection. Plasma miRNA was purified using the miRNeasy MicroKit (Qiagen, Germantown, MD) according to the manufacturer's protocol. The miRCURY LNA miRNA PCR assay (Exiqon/Qiagen) was applied to analyze miRNA levels quantitatively; the Exiqon/Qiagen Service was used each time. The automated PCR data quality control pipeline included standardized reverse-transcription and amplification of cDNA products with SYBR Green using selected panels of quantitative real-time PCR primers (the identity of miR sequences among panels was verified). An RNA spike-in (Sp6) and a DNA spike-in (Sp3) were used as controls for reverse transcription and quantitative real-time PCR controls, respectively. Each sample was analyzed in duplicate. In a pilot experiment (performed in June 2012), a set of 20 samples was analyzed using the miRCURY LNA Universal RT miRNA PCR Human Panel I and II version 2 with 742 target miRNA species; 117 miRs were detected in all samples (mean miRNAs per sample, 260). Based on that pilot study, miRCURY LNA miRNA Custom Panel A was designed, which targeted 168 miRNA species commonly expressed in human plasma samples. Plasma samples from 393 MOLTEST-2013 participants were analyzed using Panel A (performed in September 2012); the resulting data set was used to identify the initial miRNA signature of lung cancer composed of 24 species (the patent signature; Patent Office of the Republic of Poland patent PL_230661_B, November 30, 2018) and to perform the present study (MOLTEST-2013 set). A subset of 48 miRNA species included in Panel A was selected to construct the miRCURY LNA miRNA Custom Panel B. Plasma samples from 395 MOLTEST-BIS participants were analyzed using Panel B (performed in December 2019); the resulting data set was used as the MOLTEST-BIS set. A subset of 47 miRNA species included in Panel A was selected to construct the miRCURY LNA miRNA Custom Panel C. Plasma samples from 120 SMAC participants were analyzed using Panel C (performed in January 2021); the resulting data set was used as the SMAC set. A total of 30 miRNA species (not counting quality controls) were common in all three panels, including all 24 miRNA species present in the patent signature (miRNA species listed in Table 2).

Data Preprocessing

Missing values in each data set (15.53%, 3.09%, and 8.55% in the MOLTEST-2103, MOLTEST-BIS, and SMAC data sets, respectively) were filled using the k-nearest neighbors technique, with the median of the $k = 10$ nearest neighbors.

Table 2 Sequences of 24 miRNA Species That Were Analyzed in the Current Study

Target miRNA	Primer sequence	Primer catalog number
hsa-let-7a-5p	5′-UGAGGUAGUAGGUUGUAUAGUU-3′	YP00205727
hsa-let-7f-5p	5′-UGAGGUAGUAGAUUGUAUAGUU-3′	YP00204359
hsa-miR-103a-3p	5′-AGCAGCAUUGUACAGGGCUAUGA-3′	YP00204063
hsa-miR-107	5′-AGCAGCAUUGUACAGGGCUAUCA-3′	YP00204468
hsa-miR-122-5p	5′-UGGAGUGUGACAAUGGUGUUUG-3′	YP00205664
hsa-miR-142-3p	5′-UGUAGUGUUUCCUACUUUAUGGA-3′	YP00204291
hsa-miR-142-5p	5′-CAUAAAGUAGAAAGCACUACU-3′	YP00204722
hsa-miR-144-3p	5′-UACAGUAUAGAUGAUGUACU-3′	YP00204754
hsa-miR-148b-3p	5′-UCAGUGCAUCACAGAACUUUGU-3′	YP00204047
hsa-miR-17-5p	5′-CAAAGUGCUUACAGUGCAGGUAG-3′	YP02119304
hsa-miR-181a-5p	5′-AACAUUCAACGCUGUCGGUGAGU-3′	YP00206081
hsa-miR-199a-3p	5′-ACAGUAGUCUGCACAUUGGUUA-3′	YP00204536
hsa-miR-21-5p	5′-UAGCUUAUCAGACUGAUGUUGA-3′	YP00204230
hsa-miR-23b-3p	5′-AUCACAUUGCCAGGGAUUACC-3′	YP00204790
hsa-miR-27a-3p	5′-UUCACAGUGGCUAAGUUCCGC-3′	YP00206038
hsa-miR-27b-3p	5′-UUCACAGUGGCUAAGUUCUGC-3′	YP00205915
hsa-miR-29c-3p	5′-UAGCACCAUUUGAAAUCGGUUA-3′	YP00204729
hsa-miR-30b-5p	5′-UGUAAACAUCCUACACUCAGCU-3′	YP00204765
hsa-miR-339-5p	5′-UCCCUGUCCUCCAGGAGCUCACG-3′	YP00206007
hsa-miR-33a-5p	5′-GUGCAUUGUAGUUGCAUUGCA-3′	YP00205690
hsa-miR-374a-5p	5′-UUAUAAUACAACCUGAUAAGUG-3′	YP00204758
hsa-miR-374b-5p	5′-AUAUAAUACAACCUGCUAAGUG-3′	YP00204608
hsa-miR-376c-3p	5′-AACAUAGAGGAAAUUCCACGU-3′	YP00204442
hsa-miR-942-5p	5′-UCUUCUCUGUUUUGGCCAUGUG-3′	YP00204440

In the MOLTEST-2013 set, the bi-clustering technique was applied to select 78 miRNAs with strong signals for further analysis [miRNAs with crosspoint (Cp) = 40 in most samples were removed]. All 78 miRNAs were used to normalize the MOLTEST-2013 set. In the MOLTEST-BIS and SMAC data sets, not all 78 miRNAs were measured; thus, to keep a similar normalization, the linear model to reconstruct the median shift in normalization was built on common miRNAs. This process allowed the same normalization process between different data sets to be kept. Finally, using the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction technique,[13] the batch effect related to the different data sets was visualized. The internal standardization to relative values was performed within each data set to converge data into a similar range and remove the batch effect. For this purpose, nonparametric standardization was performed, in which the medians and interquartile ranges were calculated from control samples within each set.

Development of Machine-Learning Models

Logistic regression (LR) was used as a classification model to distinguish healthy individuals from lung cancer patients. The multiple random cross-validation procedure was run on all measurements obtained from the MOLTEST-2013 set to select the most important miRNAs. In each of the 500 iterations of multiple random cross-validation, the stratified sampling method was used, in which 50% of data were randomly selected for model training, and the remaining samples were used for internal model validation. In model training, P values from U-testing were used to rank 78 miRNAs. The final number of miRNAs was selected by maximizing the mean F1 score calculated from 500 internal validation sets. This process resulted in the establishment of the 24 most influential miRNAs. On the other hand, a stepwise regression procedure with Bayesian information criterion[14] was run on the 29 miRNAs left after filtering using P values from the U-test ($P < 0.2$) to obtain a simpler model. In both models, the classification probability cutoff was tuned using maximization of negative predictive value while keeping the positive predictive value at >30%.

External Validation of Machine-Learning Models

The developed models were validated using the MOLTEST-BIS and SMAC data sets. For each of the most influential features (24 miRNAs), single-input/single-output (SISO) LR models were built. The β-regression coefficient of each investigated influential miRNA was tested against equality to zero. Odds ratios were calculated based on the LR model to assess the risk for lung cancer with the level change of a particular miRNA. Results from coefficient testing were integrated across investigated data sets using weighted z-transformation, in which the square root of sample sizes in each data set was taken as weight.[15,16] Finally, the mean

odds ratio (95% CI) from LR models in each set was calculated. $P < 0.05$ was considered significant.

Results

The miRNA signature that discriminated between patients with lung cancer diagnosed during the lung cancer screening and screening participants with no pulmonary nodules was tested. Three independent cohorts that were recruited during different screening programs performed in Poland (MOLTEST-2013 and MOLTEST-BIS) and Italy (SMAC) were included (Table 1). Cancer cases were matched with controls on smoking habit, age, and sex within each cohort, yet differences were observed between cohorts regarding age and smoking. Participants in MOLTEST-2013 were younger and characterized by a lower number of pack-years compared to participants in MOLTEST-BIS and SMAC, which reflected differences in inclusion criteria. Differences in the percentages of patients at each cancer stage and histologic type were noted among cohorts. The percentage of patients with confirmed adenocarcinoma was highest in SMAC and lowest in MOLTEST-BIS (61%, 59%, and 75% in MOLTEST-2013, MOLTEST-BIS, and SMAC, respectively). The percentage of more advanced cases was greater in MOLTEST-BIS than in the two other cohorts (stage III/IV, 6%, 32%, and 25%, respectively). All analyzed cancer cases were detected on screening only in the MOLTEST-BIS cohort.

Initially, the data set that included information on the expression of 168 miRNA species in 393 participants in MOLTEST-2013 (102 lung cancer cases and 291 controls) was used to establish the signature of early lung cancer that was patented by the authors. This signature included 24 miRNA species (Table 2) and a simpler model with 7 components that classified lung cancer cases and controls, with AUCs of 80.0% and 75.5% in the 24- and 7-mer variants, respectively. The information on the expression of the 24 miRNA species that comprised the patent signature was extracted from data sets obtained from MOLTEST-2013, MOLTEST-BIS, and SMAC. The analysis of original (raw) data revealed a large heterogeneity in the complete data set (908 samples), with three batches of samples that corresponded to three independent cohorts (Figure 1A). Therefore, an original method of data transformation was implemented that corrected original data based on the distribution of expression in controls from each cohort (the corrected abundance was expressed as a relative score); Figure 1B illustrates the results of such correction to relative score. On analysis of the global structure of the corrected data set, implemented nonparametric transformation successfully removed the batch effect visible in the raw data set (Figure 1C); hence, transformed data were used in further analyses aimed at addressing the stability of lung cancer signatures among the three cohorts. Transformed abundances of 24 analyzed miRNA species are presented in Supplemental Figure S1.

In the first step, a multivariate LR model with miRNA signatures based on a complete 24-mer set and its 7-mer variant (which appeared the most promising as the patent signature variant) was addressed. Model parameters are presented in Supplemental Tables S1 and S2. Classification models were tested using the MOLTEST-2013 data set and validated using the MOLTEST-BIS and SMAC data sets. The classifier was tuned to maximize its negative predictive value while keeping its positive predictive value at >30%. The indices of such classification models in training and validation cohorts, presented in Table 3 and Figure 2, show the receiver operating characteristic curves of classifiers. As expected, indices of tested classifiers were reduced in validation cohorts, particularly in the MOLTEST-BIS cohort (AUC, <50%). Interestingly, a 7-mer variant of the classifier, which involved miR-17, miR-21, miR-23b, miR-33a, miR-122, miR-144, and miR-148b, performed better than a complete 24-mer variant. Nonetheless, although both classification models showed satisfying accuracy in the validation SMAC cohort (AUC, ~70%), insufficient potential applicability of the tested multicomponent signatures was assumed.

In the second step, the performance and diagnostic stability of each miRNA species were analyzed using the SISO classification models. Table 4 shows the significance of differences between lung cancer cases and controls with each of the 24 miRNA species obtained in each cohort separately and after the integration of data. Eleven miRNA species significantly differentiated between lung cancer cases and controls after data integration (all, $P < 0.05$), including miR-21, miR-142-5p, miR-339, miR-107, miR-103a, miR-374b, miR-23b, miR-27a (up-regulated in cancer) and miR-122, miR-17, miR-942 (down-regulated in cancer). However, only miR-122 showed statistically significant down-regulation ($P < 0.05$) in each cohort separately. Only miR-122 showed significant differences between cases and controls when odds ratios were analyzed (Figure 3A). On the other hand, miR-21 also showed a clear tendency toward significant up-regulation in the analyzed groups (particularly in the SMAC cohort). Nonetheless, the intercohort repeatability of differences between cases and controls was rather low in the remaining miRNA species. Figure 3B shows the abundance of miR-122 and miR-21 in plasma samples from each cohort, illustrating the down-regulation of miR-122 and the up-regulation of miR-21 in cancer cases. The performance of the SISO classification models was tested in each of the 24 miRNA species. Figure 3C shows the receiver operating characteristic curves of the two most promising ones. The performance of miR-122 was better in the MOLTEST-2013 and MOLTEST-BIS cohorts, while the performance of miR-21 was better in the SMAC cohort. Nonetheless, the performance of the two single-component models was comparable to (or better than) that of the two multicomponent models (24-mer and 7-

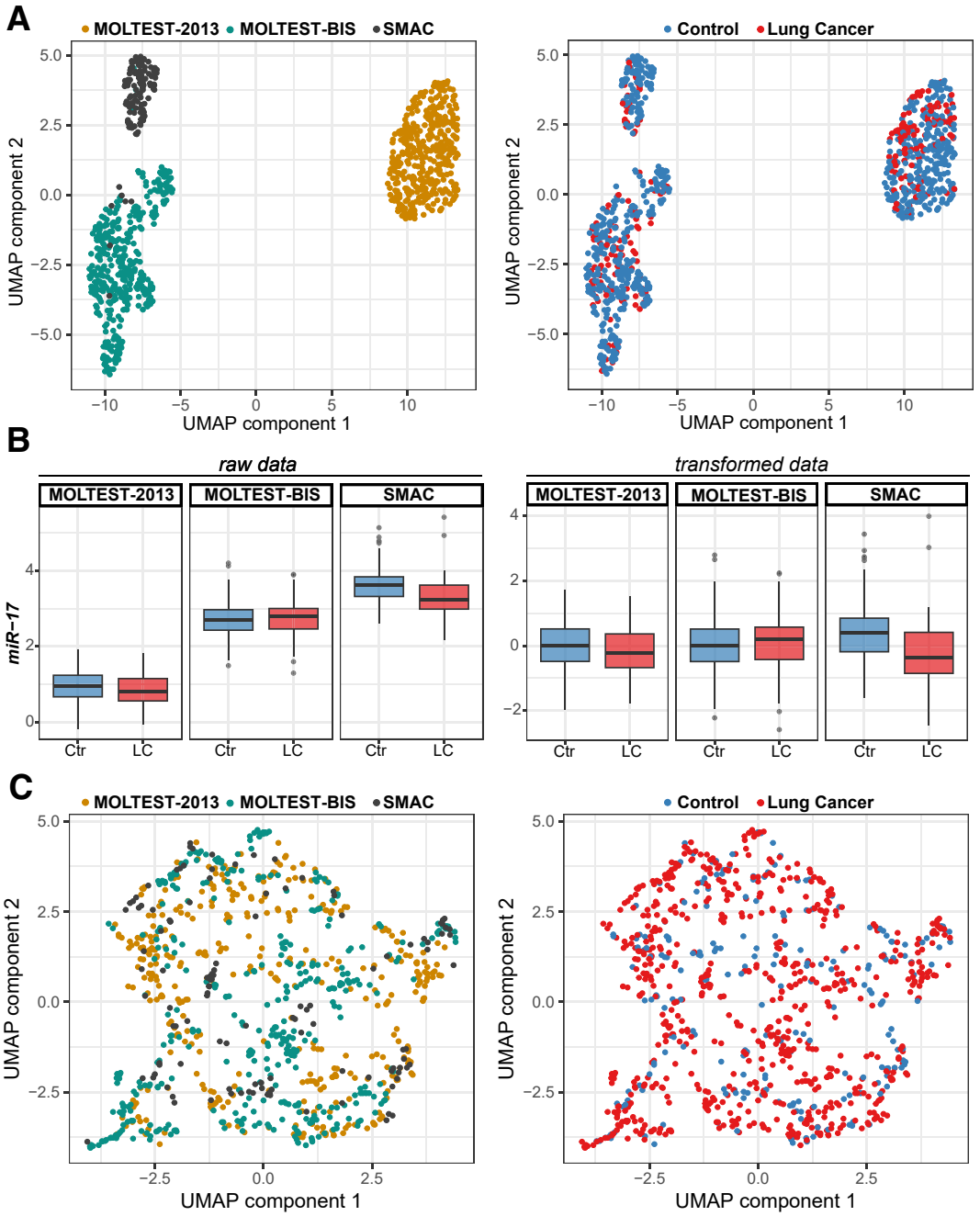


Figure 1 Transformation of miRNA levels among three analyzed cohorts. **A:** The global structure of the raw data set; spatial visualization was generated using the UMAP data transformation from 24-dimensional miRNA space to 2-dimensional view, preserving the structure of the high-dimensional data to explore the potential sample clustering. Samples from the MOLTEST-2013, MOLTEST-BIS, and SMAC cohorts (left) and samples from lung cancer cases and controls (right) are marked using separate colors. **B:** Abundances of exemplary miRNA (miR-17) before (raw data) and after transformation in lung cancer and control groups. **C:** The global structure of the corrected data set after transformation (descriptions as in **A**). Data in boxplots are expressed as lower whisker, lower quartile, median, upper quartile, upper whisker (whiskers were calculated using the Tukey method), and outliers (**dots**); arbitrary units are used. UMAP, Uniform Manifold Approximation and Projection.

mer) on validation of the MOLTEST-BIS and SMAC cohorts.

Given that differences in gene expression patterns between histologic types of lung cancer might be expected, relevant analyses were repeated on samples from patients with adenocarcinoma only, which was the predominant cancer type in all three cohorts (abundances of 24 analyzed miRNA species in control and adenocarcinoma samples are presented in Supplemental Figure S2). A multivariate LR model with miRNA signatures based on a complete 24-mer set and its 7-mer variant was addressed (model parameters are presented in Supplemental Tables S3 and S4). The

Table 3 Results of Classification Assessment for Multivariate Logistic Regression Models of 24 and 7 miRNA Species

Cohort/indices	PPV, %	NPV, %	Sensitivity, %	Specificity, %	AUC, %
Classification model based on 24 components					
MOLTEST-2013 (training)	33.2 (31.8–34.6)	89.1 (85.0–93.3)	85.4 (78.7–92.1)	38.6 (32.8–44.5)	80.0
MOLTEST-BIS (validation)	23.0	70.4	62.6	29.7	48.2
SMAC (validation)	32.1	86.1	84.4	32.2	67.1
Classification model based on 7 components					
MOLTEST-2013 (training)	31.2 (28.8–33.6)	87.6 (84.0–91.1)	88.5 (83.6–93.3)	29.3 (20.4–39.0)	75.5
MOLTEST-BIS (validation)	25.2	75.3	76.8	23.6	53.3
SMAC (validation)	30.5	88.0	90.6	25.0	70.7

indices of resulting classification models in training and validation cohorts are presented in Supplemental Table S5. The performance of the 24-mer variant of the classification model was comparable in adenocarcinoma-only and all-cancer groups in all three cohorts. However, the performance of the 7-mer variant in the validation MOLTEST-BIS subcohort of adenocarcinomas was better when compared to that in the all-cancer cohort (AUC, 53.3% and 60.1% in the all-cancer and adenocarcinoma-only subgroups, respectively) (Figure 4A). Finally, the SISO classification models were tested for each of the 24 miRNA species (Supplemental Table S6). The analysis revealed significant down-regulation of miR-122 in all three subcohorts of adenocarcinoma. However, miR-21 remained significantly up-regulated only in the SMAC adenocarcinoma subcohort (yet a similar trend remained in the other two groups). On the other hand, miR-144, which did not differentiate any all-cancer group from relevant controls, showed significant up-regulation in all three subcohorts of adenocarcinomas (Figure 4, B and C). Nonetheless, the

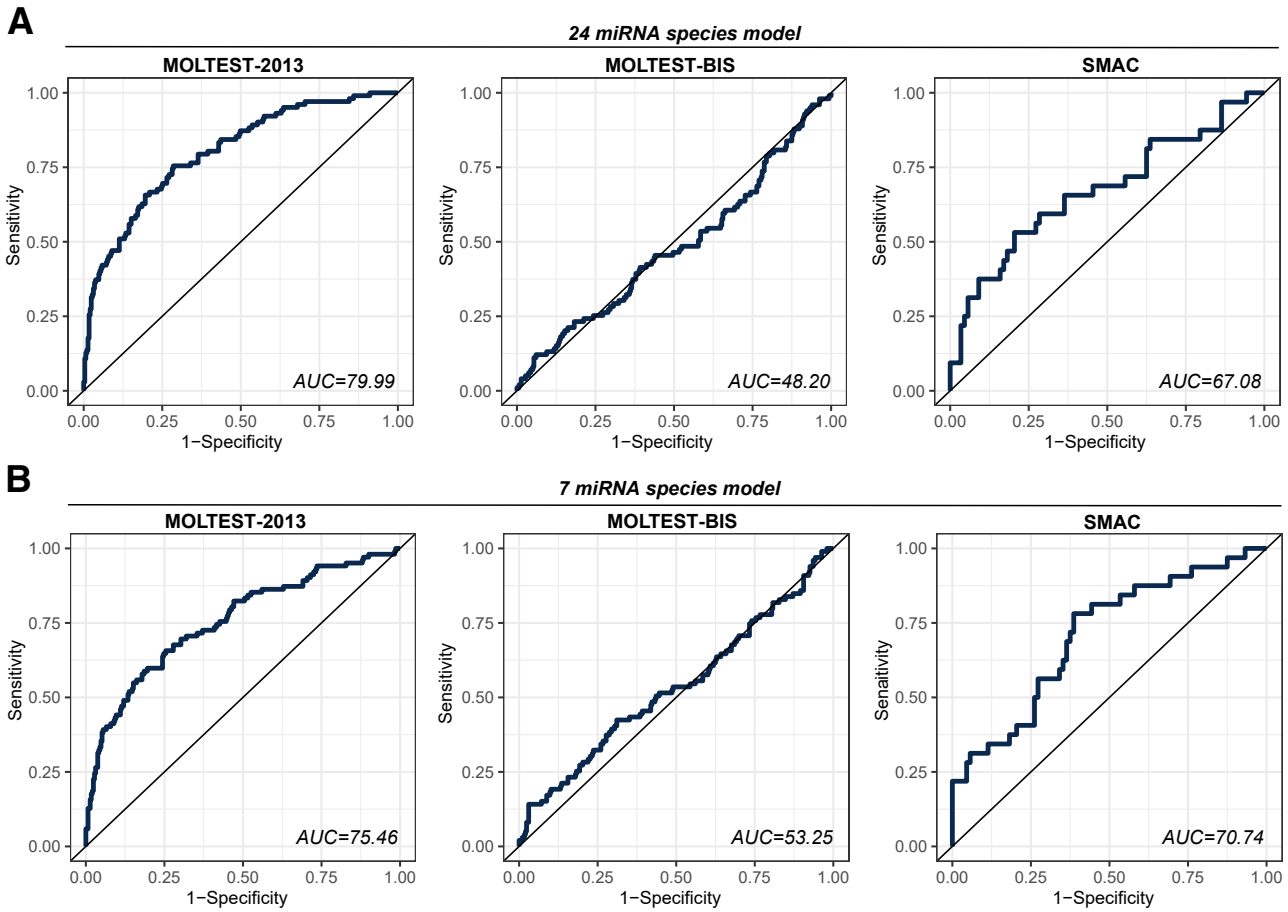


Figure 2 Receiver operating characteristic (ROC) curves of the multivariate cancer classifier. **A** and **B:** ROC curves of the lung cancer classifier based on 24 (**A**) and 7 (**B**) miRNA species (namely miR-17, miR-21, miR-23b, miR-33a, miR-122, miR-144, and miR-148b).

Table 4 Results of Single-Input—Single-Output Logistic Regression Models for Control versus Lung Cancer Classifiers

Cohort	MOLTEST-2013		MOLTEST-BIS		SMAC		Data integration			
								OR (95% CI)		
miRNA/index	*P* value	OR	*P* value	OR	*P* value	OR	*P* value	Low	Mean	High
miR-122-5p	<0.001	0.58	0.014	0.70	0.003	0.40	<0.001	0.19	0.56	0.93
miR-21-5p	0.068	1.28	0.157	1.18	0.001	2.19	0.003	0.17	1.55	2.93
miR-142-5p	0.005	0.66	0.115	1.23	0.472	1.19	0.006	0.24	1.03	1.81
miR-17-5p	0.017	0.67	0.343	1.15	0.010	0.56	0.006	0.02	0.79	1.57
miR-339-5p	0.015	1.40	0.345	1.14	0.029	0.53	0.009	−0.08	1.03	2.13
miR-942-5p	0.039	1.31	0.598	1.09	0.001	0.44	0.014	−0.19	0.95	2.08
miR-107	0.024	1.43	0.426	1.11	0.021	0.57	0.015	−0.05	1.04	2.12
miR-103a-3p	0.006	1.59	0.686	1.06	0.012	0.64	0.016	−0.09	1.09	2.28
miR-374b-5p	0.029	1.41	0.115	1.32	0.622	0.90	0.027	0.53	1.21	1.88
miR-23b-3p	0.032	1.35	0.452	1.11	0.050	0.62	0.029	0.11	1.03	1.94
miR-27a-3p	0.346	0.87	0.031	1.31	0.141	1.49	0.030	0.43	1.22	2.01
miR-374a-5p	0.229	1.20	0.406	1.14	0.003	2.13	0.051	0.12	1.49	2.86
miR-30b-5p	0.145	1.29	0.247	1.21	0.112	1.50	0.056	0.96	1.33	1.71
miR-27b-3p	0.130	0.79	0.173	1.24	0.352	1.23	0.067	0.45	1.09	1.73
miR-33a-5p	0.012	0.72	0.423	1.12	0.673	0.92	0.073	0.42	0.92	1.42
miR-148b-3p	0.084	0.75	0.472	1.11	0.096	1.58	0.076	0.10	1.15	2.19
let-7f-5p	0.001	1.68	0.697	1.06	0.798	1.07	0.082	0.39	1.27	2.15
miR-142-3p	0.005	1.55	0.957	1.01	0.014	1.82	0.083	0.43	1.46	2.49
miR-144-3p	0.037	1.38	0.624	0.94	0.279	1.30	0.118	0.61	1.21	1.80
miR-199a-3p	0.353	0.87	0.159	1.22	0.414	0.83	0.162	0.43	0.97	1.51
miR-181a-5p	0.082	1.33	0.978	1.00	0.000	0.37	0.214	−0.30	0.90	2.10
let-7a-5p	0.064	1.31	0.988	1.00	0.002	0.48	0.291	−0.10	0.93	1.97
miR-29c-3p	0.173	0.82	0.280	1.18	0.930	1.02	0.320	0.55	1.01	1.46
miR-376c-3p	0.159	0.80	0.530	1.10	0.756	0.89	0.361	0.55	0.93	1.32

performance of the SISO classification models based on miR-122 and -21 was comparable in the all-cancer group and adenocarcinoma-only group in all three cohorts. The SISO model built for miR-144 performed similarly in all three cohorts of adenocarcinoma (AUC, approximately 58%) (Figure 4D).

Discussion

The miRNA profile assessed in serum or plasma is a possible candidate for a biomarker of early lung cancer that could be applied to support lung cancer screening. The diagnostic potential of miRNA signatures has been tested in several studies,[10–12] and the clinical applicability of a few signatures[17,18] is being validated in ongoing clinical trials. However, although >100 miRNA species were included in proposed lung cancer signatures, the overlap among signatures is rather low, and only a few miRNA species recurred in multiple signatures.[12] The discrepancies among proposed miRNA signatures are related either to differing clinical and demographic characteristics of the included cohorts or to differing methodologic approaches implemented in such studies. Therefore, to assess the stability of the diagnostic potential of specific miRNAs among different cohorts, a miRNA signature identified and optimized in the frame of one lung cancer screening study (MOLTEST-2013, Poland) was validated using material collected in two independent lung cancer screening studies (MOLTEST-BIS, Poland; and SMAC, Italy) using the same standardized analytical platform.

In all three studies, quantitative analysis of plasma miRNA was applied based on the commercially available technological platform (ie, miRCURY LNA miRNA PCR). The miRNA from plasma was purified in the same laboratory while miRNA was processed by the external service center (Exiqon/Qiagen) using the same standardized analytical, which would be expected to provide repeatable results. However, the unsupervised analysis of samples revealed strong clustering, indicating significant differences between compared cohorts. However, after the transformation of raw data, classification models tested in the MOLTEST-2013 cohort could be validated in the MOLTEST-BIS and SMAC cohorts. The performance of classification models was comparable between the test cohort (MOLTEST-2013) and one validation cohort (SMAC), reaching a negative predictive value close to 90% and a positive predictive value of >30%. The diagnostic accuracy of the classification (AUC, <80%) appeared insufficient for a standalone biomarker. The performance of analyzed models in another validation cohort (MOLTEST-BIS) was markedly lower. In that cohort, the percentage of

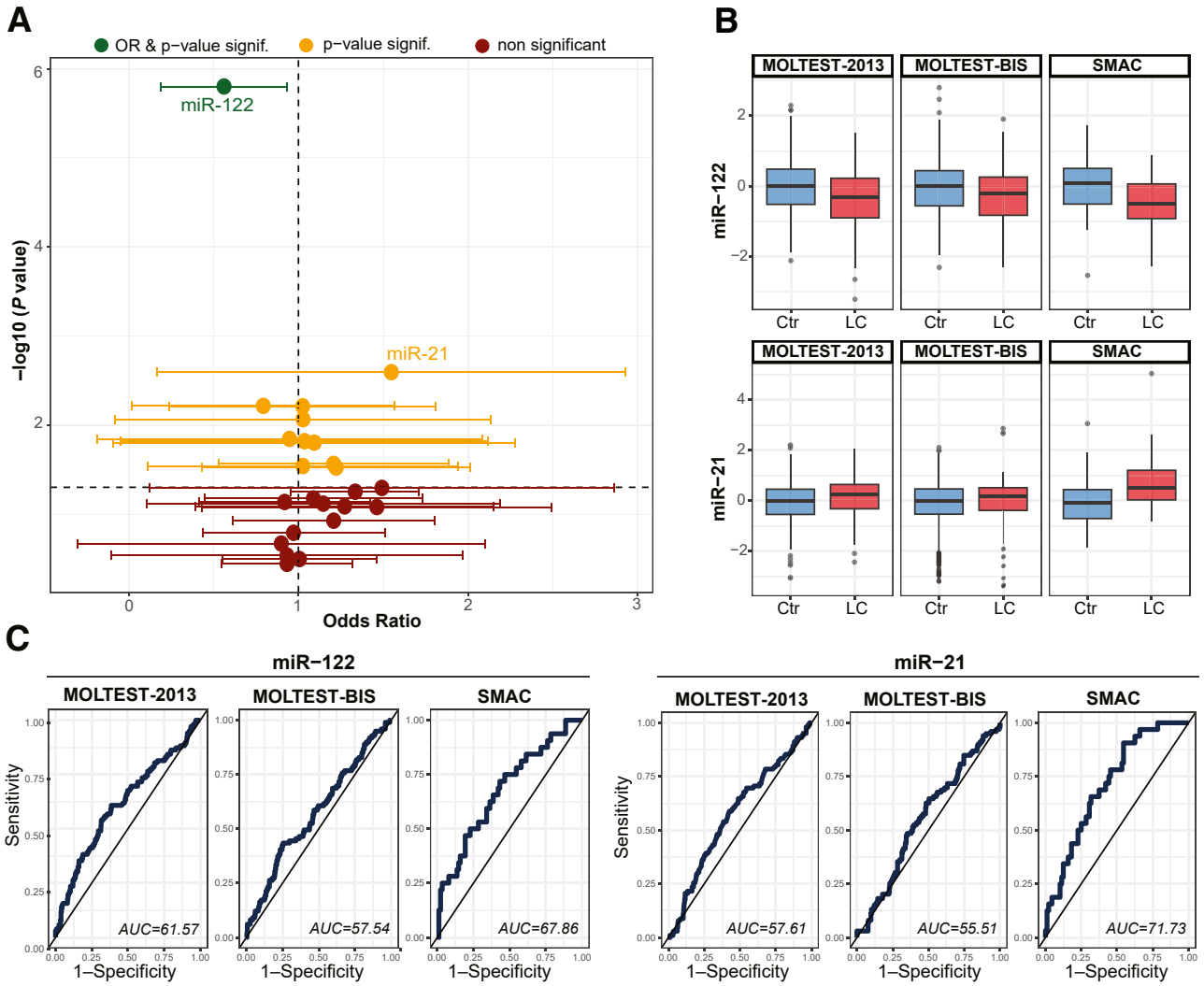
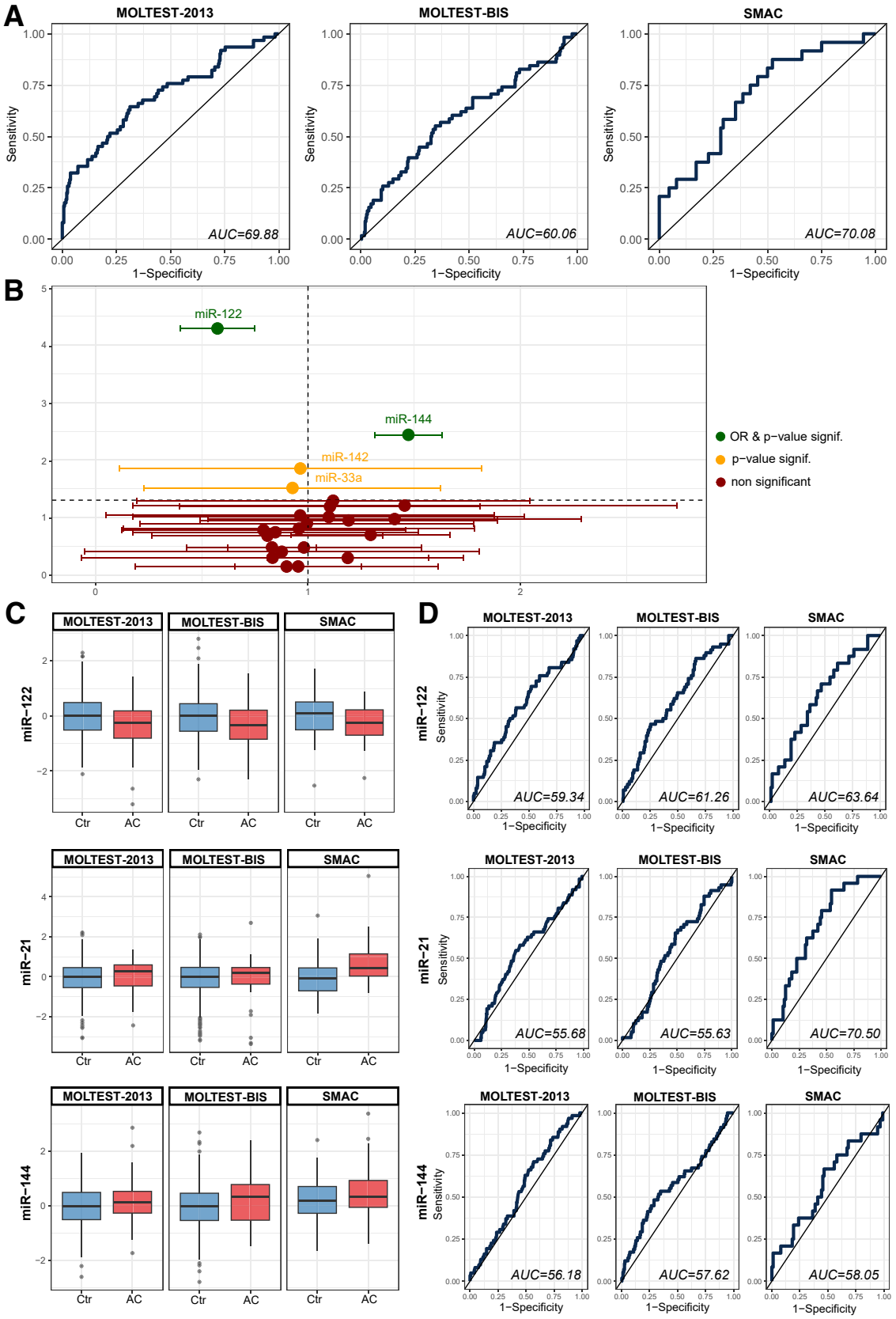


Figure 3 Levels of miR-21 and miR-122 differ between lung cancer cases and controls. **A:** The significance of differences in levels of each of 24 miRNA species between lung cancer cases and controls. The **vertical dashed line** represents a nonsignificant odds ratio (OR) equal to 1; the **horizontal dashed line** represents $-\log_{10}$ of significance level 5%. **B:** Corrected abundances of miR-122 and -21 (in arbitrary units) in lung cancer and control groups. **C:** ROC curves of the lung cancer classifier based on either miR-122 (left) or miR-21 (right) alone. Data are expressed as integrated *P* values and mean ORs (95% CIs) (**A**); data in boxplots (**B**) are expressed as lower whisker, lower quartile, median, upper quartile, upper whisker (whiskers were calculated using the Tukey method), and outliers (**dots**); arbitrary units are used.

adenocarcinomas was less than that in the MOLTEST-2013 and SMAC cohorts. Differences in gene expression patterns between lung adenocarcinomas and squamous cell carcinomas, two major histologic types of lung cancer, are well documented,[19] including specific features of miRNA profiles in tumor tissue.[20] Two plasma miRNA species (miR-944 and miR-3662) were shown to discriminate controls from lung cancer cases in a histologic type−specific pattern.[21] Hence, specific features of plasma miRNA profiles linked to different histologic types of lung cancer could contribute to the poorer performance of the classification in a group with the lowest percentage of adenocarcinomas. It is noteworthy that, when the analysis was limited to adenocarcinomas only, the performance of cancer classification in MOLTEST-2013 was increased, which supports a potential influence of cancer type−related specificities. Differences among cohorts included the percentage of screening-detected and clinic-detected cancer cases, and MOLTEST-BIS was the only cohort in which all cancer cases included in the study were detected during screening. Even more surprising, in that cohort, the percentage of more advanced cases was greater than those in the two other cohorts. Intuitively, this finding would suggest easier cancer detection, which was not the case. Nonetheless, although the existence of molecular differences between cancer cases diagnosed during screening programs and symptomatic cancers detected during other clinical procedures is still under debate (eg, Cheasley et al[22]), the present results may indicate potential limitations of clinic-detected lung cancer cases for the construction of molecular biomarkers to be applied to the detection of cancer within screening programs.


A
MOLTEST-2013
MOLTEST-BIS
SMAC
Sensitivity
1−Specificity
AUC=69.88
AUC=60.06
AUC=70.08
B
miR-122
miR-144
miR-142
miR-33a
OR & p−value signif.
p−value signif.
non significant
C
MOLTEST-2013
MOLTEST-BIS
SMAC
miR−122
miR−21
miR−144
Ctr
AC
D
MOLTEST-2013
MOLTEST-BIS
SMAC
miR−122
miR−21
miR−144
Sensitivity
1−Specificity
AUC=59.34
AUC=61.26
AUC=63.64
AUC=55.68
AUC=55.63
AUC=70.50
AUC=56.18
AUC=57.62
AUC=58.05

In addition to the assessment of the classification power of multicomponent panels, the ability of a single miRNA to differentiate controls from cancer cases was compared among cohorts. Eleven miRNA species significantly differentiated between controls and cancer cases after data integration. Two of them, miR-17 and -21, are known oncomirs that appeared in multiple plasma/serum miRNA signatures of lung cancer.[12] Five other miRNAs also appeared in lung cancer signatures; miR-103a, miR-142-5p, and miR-374b were present in serum miRNA signatures,[23] while miR-23b and miR-122 were present in circulating exosome miRNA signatures.[24,25] However, in the present study, these miRNA species, except miR-122 and to some extent miR-21, did not discriminate between controls and tumors reproducibly in all three cohorts, indicating their limited use as universal biomarkers of early lung cancer. miR-122 and miR-21 repetitively discriminated between controls and cases in each cohort and were components of miRNA signatures of early lung cancer (although the significance of miR-21 was limited in the MOLTEST-BIS cohort). miR-21 is frequently up-regulated in all types of cancers and down-regulates several tumor-suppressor genes (including *PTEN* and *TP63*)[26] and was the most frequent component of circulating miRNA signatures of lung cancer. According to a recent meta-analysis,[27] the capacity of miR-21 in diagnosing lung cancer has been addressed in 31 papers (part of them including low-advanced cases), with an integrated AUC of 0.87 (95% CI, 0.84−0.90). Although less recognized, miR-122 has also been suggested, in several cancer-related functions, as having either tumor-promoting or tumor-suppressing roles. miR-122 is aberrantly expressed in various tumors, including lung cancer.[28] It has been found that miR-122 inhibits proliferation and radiosensitizes lung cancer cells by lowering the expression of *BCLW* and *IGF1R*[29] and that this miRNA is down-regulated in non−small-cell lung cancer tissues.[30] Hence, the reduced level of circulating miR-122 in the plasma of lung cancer patients observed in the present study fits the known mechanism of its association with lung cancer progression. In addition to miR-122 and miR-21, an increased plasma concentration of miR-144 was noted in adenocarcinomas from all three cohorts included in the study. miR-144 is another oncomir whose aberrant expression was noted in different hematologic malignancies and solid cancers.[31,32] Interestingly, the expression of miR-144 was reported to be reduced in squamous non−small-cell lung cancer compared to that in normal lung tissue,[33] which suggested differences between lung cancer types.

To further confirm the potential diagnostic capacities of miRNA species included in the present signature, information on their expression levels in the setting of 100 lung cancers and 100 controls was extracted from the data set that was made publicly available by Wozniak et al[34] (Gene Express Omnibus; *https://www.ncbi.nlm.nih.gov/geo*; accession number GSE64591, last accessed August 22, 2023). Unfortunately, a multivariate LR model built for the present signature (7-mer version) did not classify these samples (AUC, 50%). miRNA species tested in the present study did not show statistically significant differences between cancer and control samples from that data set. Notably, however, those data were obtained using the TaqMan miRNA arrays (Thermo Fisher Scientific, Waltham, MA). Hence, the observed lack of consistency was probably the result of differences between analytical platforms used in both studies, which is among the potential confounding factors that impair the capacity to cross-validate miRNA signatures.

Conclusions

On analysis of selected miRNA expression patterns in plasma from participants in three independent lung cancer screening studies, two miRNA species known to be associated with lung cancer, miR-122 and miR-21, repetitively differentiated between healthy participants of the screening and individuals with lung cancer. miR-144 was up-regulated specifically in adenocarcinomas from all of three cohorts. However, with other tested miRNAs, there were significant differences between cohorts. In classification models based on neither a single miRNA nor multicomponent miRNA panels, classification performance was sufficient for a standalone diagnostic biomarker. The poorer classification power noted in a cohort with a relatively low contribution of adenocarcinomas among cancer cases may suggest limited applicability of cancer type−independent biomarkers of early lung cancer.

Disclosure Statement

None declared.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2023.09.010*.

Figure 4 Differences in plasma miRNA levels between adenocarcinoma cases and controls. **A:** Receiver operating characteristic (ROC) curves of the lung cancer classifier based on 7 miRNA species tested for adenocarcinoma. **B:** The significance of differences in levels of each of 24 miRNA species between adenocarcinomas and controls; the **vertical line** represents a nonsignificant OR equal to 1; the **horizontal line** represents $-\log_{10}$ of significance level 5%. **C:** Corrected abundances of miR-122, miR-21, and miR-144 in adenocarcinoma (AC) and control (Ctr) groups. **D:** ROC curves of the lung cancer classifier based on miR-122, miR-21, or miR-144 (right) alone tested for adenocarcinoma only. Data are expressed as integrated *P* values and mean ORs (95% CIs) (**B**); data in boxplots (**C**) are expressed as lower whisker, lower quartile, median, upper quartile, upper whisker (whiskers were calculated using the Tukey method), and outliers (**dots**); arbitrary units are used.

References

1. Wong MCS, Lao XQ, Ho KF, Goggins WB, Tse SL: Incidence and mortality of lung cancer: global trends and association with socioeconomic status. Sci Rep 2017, 7:14300
2. Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell CD: Progress and prospects of early detection in lung cancer. Open Biol 2017, 7:170070
3. Eggert JA, Palavanzadeh M, Blanton A: Screening and early detection of lung cancer. Semin Oncol Nurs 2017, 33:129−140
4. Hecht SS: Cigarette smoking and lung cancer: chemical mechanisms and approaches to prevention. Lancet Oncol 2002, 3:461−469
5. National Lung Screening Trial Research Team: Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011, 365:395−409
6. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, Lammers JW, Weenink C, Yousaf-Khan U, Horeweg N, van't Westeinde S, Prokop M, Mali WP, Mohamed Hoesein FAA, van Ooijen PMA, Aerts JGJV, den Bakker MA, Thunnissen E, Verschakelen J, Vliegenthart R, Walter JE, ten Haaf K, Groen HJM, Oudkerk M: Reduced lung-cancer mortality with volume CT screening in a randomized trial. N Engl J Med 2020, 382:503−513
7. Rzyman W, Jelitto-Gorska M, Dziedzic R, Biadacz I, Ksiazek J, Chwirot P, Marjanski T: Diagnostic work-up and surgery in participants of the Gdansk lung cancer screening programme: the incidence of surgery for non-malignant conditions. Interact Cardiovasc Thorac Surg 2013, 17:969−973
8. Chu GCW, Lazare K, Sullivan F: Serum and blood based biomarkers for lung cancer screening: a systematic review. BMC Cancer 2018, 18:1−6
9. Ostrin E, Sidransky D, Spira A, Hanash SM: Biomarkers for lung cancer screening and detection. Cancer Epidemiol Biomarkers Prev 2020, 29:2411−2415
10. Montani F, Bianchi F: Circulating cancer biomarkers: the macro-revolution of the micro-RNA. EBioMedicine 2016, 5:4−6
11. Han Y, Li H: miRNAs as biomarkers and for the early detection of non-small cell lung cancer (NSCLC). J Thorac Dis 2018, 10: 3119−3131
12. Smolarz M, Widlak P: Serum exosomes and their miRNA load–a potential biomarker of lung cancer. Cancers 2021, 13:1373
13. McInnes L, Healy J, Saul N, Grossberger L: UMAP: uniform manifold approximation and projection. J Open Source Softw 2018, 3:861
14. Schwarz G: Estimating the dimension of a model. Ann Stat 1978, 6: 461−464
15. Liptak T: On the combination of independent tests. Magyar Tud Akad Mat Kutato Int Kozl 1958, 3:171−197
16. Whitlock MC: Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol 2005, 18:1368−1373
17. Sozzi G, Boeri M, Rossi M, Verri C, Suatoni P, Bravi F, Roz L, Conte D, Grassi M, Sverzellati N, Marchiano A, Negri E, La Vecchia C, Pastorino U: Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening, a correlative MILD trial study. J Clin Oncol 2014, 32:768−773
18. Montani F, Marzi MJ, Dezi F, Dama E, Carletti RM, Bonizzi G, Bertolotti R, Bellomi M, Rampinelli C, Maisonneuve P, Spaggiari L, Veronesi G, Nicassio F, Di Fiore P, Bianchi F: miR-Test, a blood test for lung cancer early detection. J Natl Cancer Inst 2015, 107:djv063
19. Faruki H, Mayhew GM, Serody JS, Hayes DN, Perou CM, Lai-Goldman M: Lung adenocarcinoma and squamous cell carcinoma gene expression subtypes demonstrate significant differences in tumor immune landscape. J Thorac Oncol 2017, 12:943−953
20. Petkova V, Marinova D, Kyurkchiyan S, Stancheva G, Mekov E, Kachakova-Yordanova D, Slavova Y, Kostadinov D, Mitev V, Kaneva R: MiRNA expression profiling in adenocarcinoma and squamous cell lung carcinoma reveals both common and specific deregulated microRNAs. Medicine (Baltimore) 2022, 101:e30027
21. Powrozek T, Krawczyk P, Kowalski DM, Winiarczyk K, Olszyna-Serementa M, Milanowski J: Plasma circulating microRNA-944 and microRNA-3662 as potential histologic type-specific early lung cancer biomarkers. Transl Res 2015, 166:315−323
22. Cheasley D, Li N, Rowley SM, Elder K, Mann GB, Loi S, Savas P, Goode DL, Kader T, Zethoven M, Semple T, Fox SB, Pang JM, Byrne D, Devereux L, Nickson C, Procopio P, Lee G, Hughes S, Saunders H, Fujihara KM, Kuykhoven K, Connaughton J, James PA, Gorringe KL, Campbell IG: Molecular comparison of interval and screen-detected breast cancers. J Pathol 2019, 248:243−252
23. Bianchi F, Nicassio F, Marzi M, Belloni E, Dall'olio V, Bernard L, Pelos G, Maisonneuve P, Veronesi G, Di Fiore PP: A serum circulating miRNA diagnostic test to identify asymptomatic high-risk individuals with early stage lung cancer. EMBO Mol Med 2011, 3: 495−503
24. Giallombardo M, Chacartegui BJ, Castiglia M, Van Der Steen N, Mertens I, Pauwels P, Peeters M, Rolfo C: Exosomal miRNA analysis in non-small cell lung cancer (NSCLC) patients' plasma through qPCR, a feasible liquid biopsy tool. J Vis Exp 2016, 111:53900
25. Liu Q, Yu Z, Yuan S, Xie W, Li C, Hu Z, Xiang Y, Wu N, Wu L, Bai L, Yafei L: Circulating exosomal microRNAs as prognostic biomarkers for non-small-cell lung cancer. Oncotarget 2017, 8:13048−13058
26. Feng YH, Tsao CJ: Emerging role of microRNA-21 in cancer. Biomed Rep 2016, 5:395−402
27. Wang W, Li X, Liu C, Zhang X, Wu Y, Diao M, Tan S, Huang S, Cheng Y, You T: MicroRNA-21 as a diagnostic and prognostic biomarker of lung cancer: a systematic review and meta-analysis. Biosci Rep 2022, 42:BSR20211653
28. Faramin Lashkarian M, Hashemipour N, Niaraki N, Soghala S, Moradi A, Sarhangi S, Hatami M, Aghaei-Zarch F, Khosravifar M, Mohammadzadeh A, Najafi S, Majidpoor J, Farnia P, Aghaei-Zarch SM: MicroRNA-122 in human cancers: from mechanistic to clinical perspectives. Cancer Cell Int 2023, 23:29
29. Ma D, Jia H, Qin M, Dai W, Wang T, Liang E, Dong G, Wang Z, Zhang Z, Feng F: MiR-122 induces radiosensitization in non-small cell lung cancer cell line. Int J Mol Sci 2015, 16:22137−22150
30. Gao L, Chen X, Wang Y, Zhang J: Up-regulation of FSTL3, regulated by lncRNA DSCAM-AS1/miR-122-5p axis, promotes proliferation and migration of non-small cell lung cancer cells. Onco Targets Ther 2020, 13:2725
31. Kooshkaki O, Rezaei Z, Rahmati M, Vahedi P, Derakhshani A, Brunetti O, Baghbanzadeh A, Mansoori B, Silvestris N, Baradaran B: MiR-144: a new possible therapeutic target and diagnostic/prognostic tool in cancers. Int J Mol Sci 2020, 21:2578
32. Zhou M, Wu Y, Li H, Zha X: MicroRNA-144: a novel biological marker and potential therapeutic target in human solid cancers. J Cancer 2020, 11:6716−6726
33. Uchida A, Seki N, Mizuno K, Misono S, Yamada Y, Kikkawa N, Sanada H, Kumamoto T, Suetsugu T, Inoue H: Involvement of dual-strand of the miR-144 duplex and their targets in the pathogenesis of lung squamous cell carcinoma. Cancer Sci 2019, 110:420−432
34. Wozniak MB, Scelo G, Muller DC, Mukeria A, Zaridze D, Brennan P: Circulating microRNAs as non-invasive biomarkers for early detection of non-small-cell lung cancer. PLoS One 2015, 10: e0125026