Received: 14 April 2022 | Revised: 23 August 2022 | Accepted: 12 September 2022

DOI: 10.1111/cas.15591

ORIGINAL ARTICLE

Cancer Science WILEY

Genetic predisposition to lung adenocarcinoma outcome is a feature already present in patients' noninvolved lung tissue

Francesca Minnai[1] | Sara Noci[2] | Marco Chierici[3] | Chiara Elisabetta Cotroneo[2] | Barbara Bartolini[2] | Matteo Incarbone[4] | Davide Tosi[5] | Giovanni Mattioni[5] | Giuseppe Jurman[3] | Tommaso A. Dragani[2] | Francesca Colombo[1]

[1]Institute for Biomedical Technologies, National Research Council, Segrate, Italy

[2]Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy

[3]Data Science for Health Research Unit, Bruno Kessler Foundation, Trento, Italy

[4]Department of Surgery, IRCCS Multimedica, Milan, Italy

[5]Thoracic Surgery and Lung Transplantation Unit, Fondazione IRCCS Cà Granda Ospedale Maggiore Policlinico, Milan, Italy

Correspondence
Tommaso A. Dragani, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori, Via G.A. Amadeo 42, I-20133 Milan, Italy.
Email: tommaso.dragani@istitutotumori.mi.it

Funding information
Associazione Italiana per la Ricerca sul Cancro, Grant/Award Number: IG 20226

Abstract

Emerging evidence suggests that the prognosis of patients with lung adenocarcinoma can be determined from germline variants and transcript levels in nontumoral lung tissue. Gene expression data from noninvolved lung tissue of 483 lung adenocarcinoma patients were tested for correlation with overall survival using multivariable Cox proportional hazard and multivariate machine learning models. For genes whose transcript levels are associated with survival, we used genotype data from 414 patients to identify germline variants acting as *cis*-expression quantitative trait loci (eQTLs). Associations of eQTL variant genotypes with gene expression and survival were tested. Levels of four transcripts were inversely associated with survival by Cox analysis (*CLCF1*, hazard ratio [HR] = 1.53; *CNTNAP1*, HR = 2.17; *DUSP14*, HR = 1.78; and *MT1F*: HR = 1.40). Machine learning analysis identified a signature of transcripts associated with lung adenocarcinoma outcome that was largely overlapping with the transcripts identified by Cox analysis, including the three most significant genes (*CLCF1*, *CNTNAP1*, and *DUSP14*). Pathway analysis indicated that the signature is enriched for ECM components. We identified 32 *cis*-eQTLs for *CNTNAP1*, including 6 with an inverse correlation and 26 with a direct correlation between the number of minor alleles and transcript levels. Of these, all but one were prognostic: the six with an inverse correlation were associated with better prognosis (HR < 1) while the others were associated with worse prognosis. Our findings provide supportive evidence that genetic predisposition to lung adenocarcinoma outcome is a feature already present in patients' noninvolved lung tissue.

KEYWORDS
gene expression, lung neoplasm, machine learning, prognosis, quantitative trait locus

1 | INTRODUCTION

Lung adenocarcinoma is the most common type of non-small-cell lung cancer. It is characterized by poor prognosis, with a median 5-year survival rate of 45%.[1] High variability in lung adenocarcinoma outcome has been repeatedly observed, especially among patients diagnosed with different pathological stages. Such variability has often been ascribed to a high tumor genetic heterogeneity among patients. Indeed, beside the well-described cancer-driving mutations in the *EGFR*, *KRAS*, *ALK*, *BRAF*, and *ROS1* genes,[2,3] several other somatic mutations are present in lung adenocarcinomas. Lung tumor profiling studies have documented this genetic variability[4–6] and associated these profiles with different outcomes.[7]

In addition to somatic mutations, other prognostic factors in lung adenocarcinoma include germline variants,[8–12] tumor gene expression levels[13,14] and gene methylation profiles,[15,16] the immunophenotype of the tumor microenvironment,[17] patients' clinical traits,[18,19] and the combination of these factors.[20,21] We previously reported that gene expression levels in the noninvolved lung tissue of lung adenocarcinoma patients also have prognostic value.[22] In that study, we used transcriptome analysis of noninvolved, apparently normal lung tissue (in a discovery series of 204 samples and a validation series of 78) to identify a signature of 10 genes (*CNTNAP1*, *PKNOX1*, *FAM156A*, *FRMD8*, *GALNTL1*, *TXNDC12*, *SNTB1*, *PPP3R1*, *SNX10*, and *SERPINH1*) associated with survival. This finding supports the hypothesis that noninvolved tissue already possesses the genetic signature predisposing to a different cancer outcome.

Two limitations of that study[22] were the rather small patient series and use of the standard Cox proportional hazard model to assess associations between gene expression and prognosis. Indeed, with the Cox model, the effect on overall survival of expression levels of each transcript is evaluated one at a time. Thus, this approach does not take into account possible interactions between different transcripts in modulating the lung adenocarcinoma outcome. Machine learning algorithms are multivariable methods that overcome this limitation.[23] They model the relationship between gene expression and outcome by accounting for the simultaneous interaction of several genes or related molecular pathways. Thus, they consider the complexity of the overall system.

Here, we expanded the transcriptome analysis to a larger patient series. We used a machine learning algorithm, the Random Forest, in addition to the standard Cox proportional hazard model to assess associations between gene expression in noninvolved tissue and prognosis of patients with lung adenocarcinoma. For the top-ranked genes whose levels were associated with lung adenocarcinoma outcome in both the Cox and Random Forest approaches, we also tested, in an independent series, whether their expression levels in lung adenocarcinoma tissue associated with survival. Additionally, we evaluated whether the expression in noninvolved lung tissue of the same transcripts was regulated by germline variants, that is, we looked for *cis*-expression quantitative trait loci (eQTLs) of these genes. This investigation was undertaken with the aim of understanding whether the observed interindividual variability in noninvolved lung tissue gene expression levels was due to host genetics. Finally, for the identified regulatory variants, we tested their association with lung adenocarcinoma patient survival.

2 | MATERIALS AND METHODS

2.1 | Ethics statement

The ethics committees of the recruiting hospitals approved the protocol for tissue collection, and patients provided written informed consent to the use of their biological samples and data for research purposes, as already described.[24]

2.2 | Study population, samples, and clinical information

This study analyzed transcriptome and genotype data from noninvolved (apparently normal) lung parenchyma tissues excised from 483 patients who underwent lobectomy for lung adenocarcinoma in the authors' institutes in the area around Milan, Italy, between 1992 and 2017. These tissues and the extracted nucleic acids are stored in a biobank of the Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy.

Clinical data for each patient were collected regarding sex, age at diagnosis, pathological stage, self-reported habit regarding the smoking of tobacco-containing cigarettes (recorded as ever or never if they had ever or never smoked in their life, respectively), and survival status (limited to 60 months after surgery). Methods for tissue collection and nucleic acid extraction have already been described.[25] Briefly, noninvolved (apparently normal) lung parenchyma specimens had been placed in RNAlater solution (Thermo Fisher Scientific) and transported to the laboratory at 4°C. DNA had been extracted from a portion of tissue using the DNeasy Blood & Tissue Kit (Qiagen), and RNA had been extracted from another portion using Qiazol Lysis Reagent and purified with RNeasy silica membrane columns (Qiagen). DNA was stored at −20°C and RNA at −80°C.

2.3 | Gene expression analysis

We used our already available transcriptome data (GSE71181 and GSE123352)[22,24] from noninvolved lung tissue of lung adenocarcinoma patients. Gene expression profiles had been obtained using HumanHT-12 v4 Expression BeadChips (Illumina) in different batches. Microarray data had been log-transformed and normalized as described,[22] and lumiBatch objects (i.e., a specific class of Illumina microarray data) had been created with the package lumi in R.

For each batch, we loaded the lumiBatch objects into the R environment and applied a variance-stabilizing transformation to the expression levels (function lumi::lumiT). Next, we filtered the probes based on the detection call p-value and kept only those with $p < 0.01$. Probes belonging to the same gene (according to Illumina annotation) were mean-summarized. We then corrected for known batch effects with the ComBat function (sva library[25]) and also adjusted for the following variables of interest: sex, stage, age at diagnosis, smoking habit, and 60-month follow-up status. The adjustment was required to prevent the suppression of potential differences due to these phenotypes. Finally, the gene expression table was filtered to keep most variant transcripts, based on the interquartile range (IQR) of normalized probe intensity, using a threshold of IQR > 0.75.

2.4 | Survival analyses

We first did a preliminary multivariable Cox proportional hazard analysis to identify prognostic factors 60 months after lung adenocarcinoma surgical resection; this analysis considered the clinical variables sex, age at diagnosis (a continuous variable), pathological stage (I vs. >I) and smoking habit. Patients for whom stage information was lacking were excluded from this and subsequent analyses.

The relationships between gene expression levels (normalized probe intensities; continuous variables) and patient overall survival were analyzed using a multivariable Cox proportional hazard model considering sex, age at diagnosis, pathological stage (I vs. >I), and smoking habit as covariates. Data were censored at 60 months of follow-up. Each gene was tested individually. False discovery rate (FDR) was calculated using the Benjamini–Hochberg method,[26] and an FDR < 0.05 was set as a stringent significance threshold. Another, more permissive threshold at nominal $p < 0.01$ was used in the comparison between Cox and Random Forest analysis results. Kaplan–Meier curves were drawn to visualize the differences in survival between patients expressing high or low levels (above or below the median value of $\log_2$-transformed probe intensities, respectively) of the genes of interest, and log–rank p values were calculated. Analyses were carried out using the survival package of R.

For genes selected in the Cox proportional hazard analysis, we used the online tool Kaplan–Meier Plotter[27] to look for associations with survival in published gene expression data from lung adenocarcinoma tumor tissue. We selected the following parameters (different from default settings): follow-up threshold of 60 months; all probe sets per gene; adenocarcinoma histology; and multivariable Cox regression, with stage, sex, and smoking habit as covariates. With these settings, the tool analyzed data from 387 patients. Gene expression data were dichotomized at the median value of $\log_2$-transformed probe intensities to form high and low expression groups. For comparison to the noninvolved lung tissue analyzed in this study, we used a multivariable Cox model to calculate hazard ratios between samples with high and low levels (above and below the median of $\log_2$-transformed values, respectively) of the same genes, with sex, age, and stage as covariates.

2.5 | Machine learning analysis

The association between patient outcome (60-month survival status) and gene expression was assessed using the Random Forest classifier.[28] To limit overfitting effects that could prevent generalization from the model, the machine learning framework was built according to guidelines by the US FDA MicroArray/Sequencing Quality Control (MAQC/SEQC) initiatives[29,30] regarding the development of predictive models for the analysis of high-throughput biomedical data. The data were randomly split into 20 train and test partitions with an 80%/20% train/test proportion, preserving the original class stratification. Each of the 20 train partitions underwent 10 iterations of a stratified 5-fold cross-validation. A Random Forest model with 501 trees was used as the classifier, ranking genes according to the ANOVA F-value. At each cross-validation iteration, 10 Random Forest models were built using an increasing number of ranked genes (namely 5, 10, 15, 20, 50, and 100 genes, then 5%, 25%, 50%, and 100% of the total number of genes). Predictive performance of the 10 models was evaluated in terms of sensitivity, specificity, positive predictive value, negative predictive value, and the Matthews correlation coefficient (MCC).[31] The MCC is a classification metric that combines precision and accuracy in a single value ranging from −1 (inverse prediction) to +1 (perfect prediction), with 0 meaning random guess or, in general, no correlation between the predictions and the actual class labels. It has been shown that MCC is a more reliable metric than accuracy and F1 score.[32]

The ranked gene lists generated by the cross-validation procedure were aggregated into a single ranked list with the Borda method.[33,34] The top-ranking genes (highest MCC) were taken as the optimal gene list for the classification task. The length of this optimal gene list was termed “signature size”. A Random Forest model was finally refitted on the whole train partition restricted to the optimal genes; this model was validated on the test partition. To ensure that the analysis was not affected by systematic error, the whole predictive modeling pipeline was also run after randomly scrambling the patient outcome labels: the performance of the classifier was close to that of a random one (i.e., MCC near 0), verifying the absence of bias. The performance of the pipeline run with true patient outcome labels was compared to that of the pipeline run with random labels by a Wilcoxon rank sum test. The similarity between two gene lists, indicated by A and B, was assessed by the Jaccard similarity coefficient, which is defined as the ratio of intersection over union: $J(A, B) = |A \cap B| / |A \cup B|$.

2.6 | Functional annotation and xCell analysis

Reactome pathway analyses were performed using the ReactomePA R library.[35] Pathway significance was assessed using a one-sided hypergeometric test, adjusting p values by the Benjamini–Hochberg method (i.e., calculating FDR). Pathways with an FDR < 0.05 were deemed significant.

The cellular composition of lung tissue samples was estimated from gene expression data using the xCell package in R, with default parameters.[36] For each patient's transcriptome, we calculated an immune cell enrichment score (considering B cells, CD4+ T cells, CD8+ T cells, dendritic cells, eosinophils, macrophages, monocytes, mast cells, neutrophils, and natural killer cells) and a stromal cell enrichment score (considering adipocytes, endothelial cells, and fibroblasts). The association between rank-transformed scores and the expression of selected genes was tested in a generalized linear model, with age at diagnosis, sex, stage and smoking status as covariates.

2.7 | Expression quantitative trait loci analysis and association between eQTL SNPs and survival

Genome-wide genotype data from DNA of noninvolved lung tissues of 414 of the 483 lung adenocarcinoma patients were available in our laboratory.[37] A principal component analysis was carried out using PLINK software.[38] Using a custom script in R, we projected our patients into the 1000 Genomes principal component space to evaluate deviations from European ethnicity.

To evaluate whether the expression of a gene associated with patient survival was regulated in *cis* by an eQTL, we selected all SNPs mapping in a 1 Mb window spanning each gene and tested the association between SNP genotypes and gene expression levels (normalized probe intensities). This analysis was carried out by linear regression, with age, sex, pathological stage, and smoking habit as covariates, using PLINK software. An FDR <0.05 was set as the significance threshold. The SNP with the smallest p value in each locus was termed the "lead variant". Linkage disequilibrium (LD) between the lead variant and the other SNPs in each locus was assessed by calculating the correlation coefficient r^2 for each pair using PLINK software. The eQTL p values and r^2 values for all the SNPs in each locus were plotted in a regional association plot using the LocusZooms function.[39] All coding genes annotated in the displayed region according to the UCSC database were shown. To validate the identified eQTLs, we searched in the Genotype-Tissue Expression (GTEx) project database version 7[40] for prior reports of the variants in lung tissue.

Differences in gene expression levels between the three genotype groups of most significant eQTL SNPs were analyzed by one-way ANOVA followed by Tukey's test for multiple comparisons, in R environment. Genotypes of eQTL SNPs were also tested for association with patient survival using a multivariable Cox model, with age at diagnosis and stage as covariates, in the R environment.

3 | RESULTS

This study considered 483 surgically treated lung adenocarcinoma patients (Table 1), including 284 individuals whose noninvolved lung transcriptome data had been investigated in our previous study.[22] In this expanded series, approximately two-thirds of the patients (64%) were men and most were ever smokers (84%). More patients had pathological stage I (64%) than higher stages of lung adenocarcinoma. Information about ethnicity was not available, but according to genotype principal components all patients but two were European (Figure S1).

TABLE 1 Clinical characteristics of lung adenocarcinoma patients

Characteristic	All cases (N = 483)
Sex, *n* (%)	
Female	174 (36)
Male	309 (64)
Age at diagnosis, median (range), years	66 (36–85)
Smoking habit, *n* (%)	
Ever	404 (84)
Never	59 (12)
Unknown	20 (4)
Pathological stage, *n* (%)	
I	312 (65)
II	56 (12)
III	98 (20)
IV	14 (2.9)
Unknown	3 (0.6)
Dead at the 60-month follow-up, *n* (%)	
Yes	161 (33.3)
No	322 (66.7)

A preliminary analysis between overall survival and clinical characteristics did not identify significant associations with sex or smoking habit ($p > 0.05$). As expected, survival was inversely associated with pathological stage (stage >I vs. stage I, hazard ratio [HR] = 3.47; 95% confidence interval [CI], 2.50–4.82; $p = 9.67 \times 10^{-14}$). A less significant effect was also observed for age at diagnosis (HR = 1.02; 95% CI, 1.00–1.03; $p = 0.040$).

3.1 | Prognostic gene expression levels in noninvolved lung tissue

We started with an expression dataset regarding 14,481 genes and, after filtering, there were 3621 genes for survival analysis. A multivariable Cox proportional model identified 174 transcripts whose levels (considered as continuous variables) in noninvolved lung tissue associated with overall survival of lung adenocarcinoma patients at a nominal $p < 0.01$ (Table S1). Of these, four transcripts (i.e., *CLCF1*, *CNTNAP1*, *DUSP14*, and *MT1F*) were significantly associated with overall survival at FDR <0.05. High levels of these four transcripts were associated with a higher risk of death (HR >1).

Figure 1 shows Kaplan–Meier curves and log–rank test p values for *CLCF1*, *CNTNAP1*, *DUSP14*, and *MT1F*, in noninvolved tissue. Patients were divided into two groups according to the

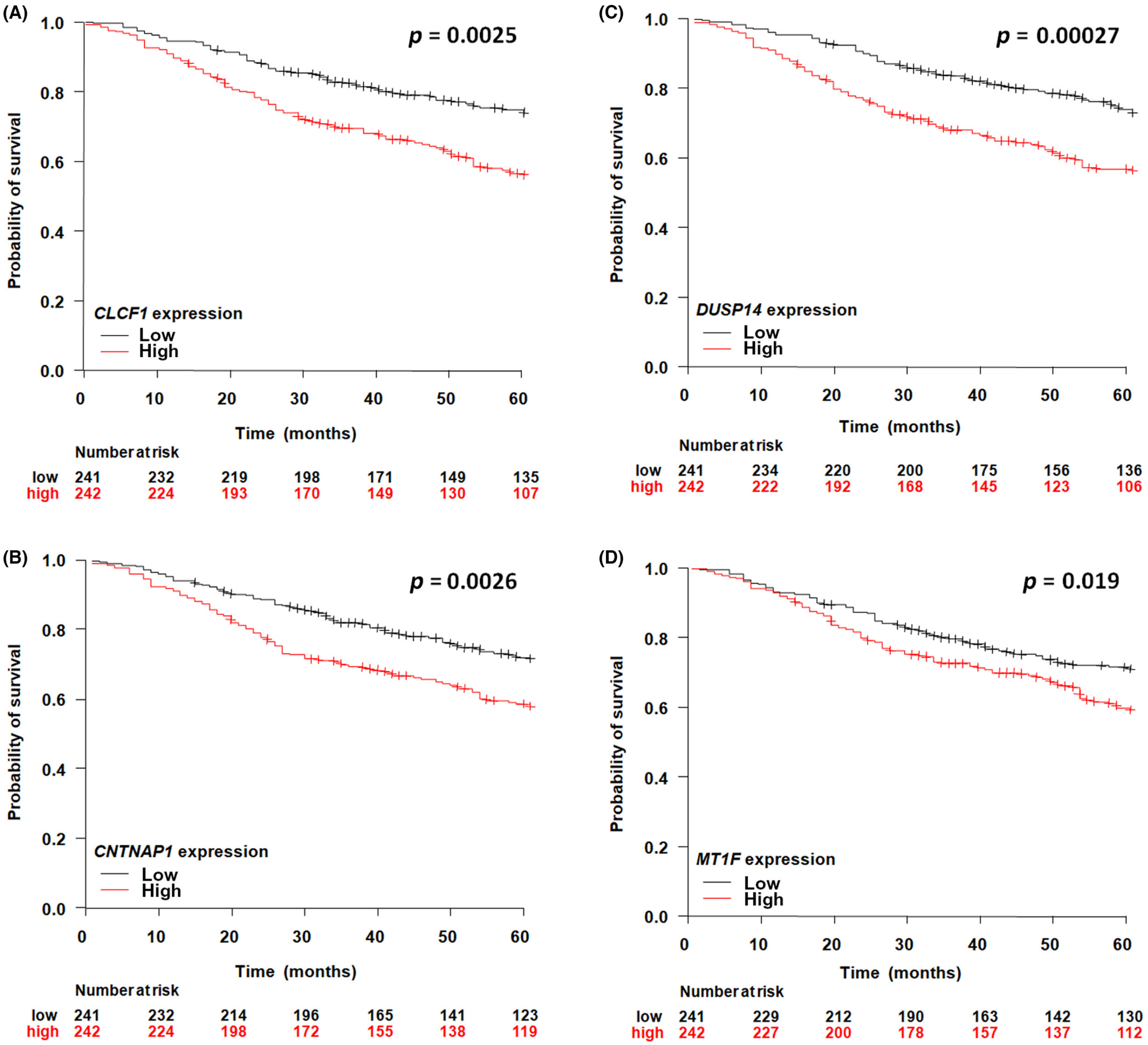

FIGURE 1 Kaplan–Meier survival curves for lung adenocarcinoma patients according to the expression levels of (A) *CLCF1*, (B) *CNTNAP1*, (C) *DUSP14*, and (D) *MT1F* genes in noninvolved lung tissue. Red and black lines represent overall survival probability of patients expressing high and low levels, respectively, of each gene. Crosses denote censored samples. Below each plot are indicated the number of patients at risk in the two groups. Log–rank *p* values are shown

expression level of each gene (above or below the median value of $\log_2$-transformed probe intensities). Patients with high levels of *CLCF1*, *DUSP14*, and *MT1F* genes in noninvolved tissues had a lower probability of survival than patients expressing low levels of these genes.

3.2 | Machine learning analysis of gene expression data and functional annotation

We tested the association between lung adenocarcinoma outcome and gene expression in noninvolved lung tissue with the Random Forest algorithm in a machine learning analysis. This computational approach has the limitation of not performing a time-to-event analysis, as, instead, the Cox model does. However, it allows taking into consideration the simultaneous effects of multiple genes, expressed in the analyzed tissue, on patient outcome. On the filtered dataset consisting of 3621 genes, the machine learning analysis achieved a relatively compact median signature size of 100 genes, with an average cross-validation MCC of 0.118 (95% CI, 0.111–0.125; Figure S2a). Sensitivity was low (<25%) but specificity was high (87%). On the test set, the model achieved lower metrics than on the cross-validation set. Although the predictive performances of these models are quite low, it is important to note that several genes found by the multivariable Cox proportional model (i.e., 24 of the 31 genes identified at FDR < 0.1) were also independently identified by the machine learning analysis within the first 100 positions (e.g., the three most significant genes in the Cox analysis, *CLCF1*, *CNTNAP1*, and *DUSP14*, were at positions 11, 27, and 44, respectively; Table S1). These findings indicate that the two independent analyses (i.e., Cox proportional hazard and Random Forest models) generate partially overlapping results. The lists of 100 and 174 genes resulting from the machine learning and Cox analyses, respectively, have 86 common genes (Table S2), with a Jaccard similarity coefficient of 0.46.

We also carried out a machine learning analysis on the top 100 FDR-ranked genes from the multivariable Cox model, to understand whether the Random Forest algorithm could perform better when starting from a preselected set of transcripts than from the total 3621 genes. The analysis in the training set achieved a median signature size of 100 genes, with a cross-validation MCC of 0.153 (95% CI, 0.147–0.160) (Figure S2b). On the test set, the model achieved an average MCC of 0.125 (95% CI, 0.07–0.182). Albeit poor, this performance is still superior to that of a random classifier, as assessed by a Wilcoxon rank-sum test for both the cross-validation ($p < 1.0 \times 10^{-16}$) and test set ($p = 6.5 \times 10^{-3}$) results. Additionally, the performance of these models was superior to that obtained starting with the entire set of 3621 transcripts. Overall, these results indicate an association, although weakly predictive, between noninvolved lung transcripts and adenocarcinoma patients' outcome.

We then explored the functions of genes associated with overall survival (according to Cox or Random Forest model) by Reactome pathway analysis (Figure 2 and Table S3). In a first analysis, we used the 100 top-ranked genes of the best-performing machine learning model, built starting from the 3621 transcripts, and identified nine pathways with an adjusted $p < 0.05$ (Figure 2A). A second functional analysis was undertaken on the 174 top-ranked genes of the multivariable Cox analysis with nominal $p < 0.01$, identifying 13 pathways (Figure 2B). Eight of the pathways identified in the two analyses were the same. These eight pathways included 12 genes coding for collagens and components of the ECM (*COMP*, *COL14A1*, *COL15A1*, *SERPINH1*, *COL7A1*, *COL16A1*, *FBLN2*, *COL18A1*, *ITGA5*, *COL1A1*, *TIMP1*, and *SERPINE1*). These results suggest that differences in the expression of some collagen genes in the noninvolved lung tissue of lung adenocarcinoma patients predispose them to a different outcome. The other transcripts associated with lung adenocarcinoma outcome according to Cox or Random Forest, but not belonging to the identified pathways, might be involved in lung adenocarcinoma progression by acting in different molecular processes and interacting with each other in a way to be further investigated.

We also predicted which cell types contributed to the expression of the identified genes (i.e., *CLCF1*, *CNTNAP1*, *DUSP14*, *MT1F*, *COMP*, *COL14A1*, *COL15A1*, *SERPINH1*, *COL7A1*, *COL16A1*, *FBLN2*, *COL18A1*, *ITGA5*, *COL1A1*, *TIMP1*, and *SERPINE1)*. First, the lung transcriptome data of each patient was analyzed to infer the cellular composition and calculate stromal and immune scores. Then we looked for associations between these scores and the expression levels of the above-listed transcripts. All genes but *CLCF1* were significantly more expressed in lung samples enriched with stromal components (i.e., high stromal enrichment scores directly correlated with high gene expression; Table S4). High levels of *COL14A1* and *MT1F* also significantly associated with high immune enrichment score.

3.3 | *CLCF1*, *DUSP14*, and ECM transcript levels associate with survival in lung adenocarcinoma tissue

Analyses thus far identified 15 genes whose levels in noninvolved lung tissue associated with survival: three genes by Cox and Random Forest models (i.e., *CLCF1*, *CNTNAP1*, and *DUSP14*) and 12 genes by pathway analysis. To determine whether levels of these transcripts in lung adenocarcinoma tissue also associated with overall survival, we used the Kaplan–Meier Plotter to analyze already published data. This analysis showed that the expression levels of *CLCF1*, *DUSP14*, *COL1A1*, *COL7A1*, *COL14A1*, *COL15A1*, and *SERPINH1* genes in lung adenocarcinoma tissue from 387 independent patients were significantly associated with overall survival (log–rank $p < 0.05$; Table 2). Apart from *COL14A1*, the effects for the other six transcripts were concordant with those observed in noninvolved tissue. In particular, high levels of these six genes were poor prognostic factors in both tissue types. Therefore, at least for these six genes identified by both statistical methods, the association with lung adenocarcinoma prognosis is an intrinsic feature of the noninvolved tissue that is conserved in the tumor. These results strongly support a prognostic role of these genes.

3.4 | Expression quantitative trait loci analysis SNPs for *CNTNAP1*, *DUSP14*, *COMP*, and *FBLN2* associate with gene expression

We next looked for germline variants associated with expression levels of the 15 selected genes by analyzing 1 Mb regions around each gene (ranging from 229 variants for *COL7A1* to 787 for *FBLN2* gene). This analysis identified 32 *cis*-eQTLs of *CNTNAP1* at FDR < 0.05 (Table S5). For six eQTL SNPs (i.e., rs12451036, rs679, rs8079855, rs55985470, rs4793253, and rs200190875), there was an inverse correlation between the number of minor alleles and levels of the transcript. For the remaining 26 SNPs, an increasing number of minor alleles was associated with higher levels of transcript. For the *DUSP14* gene, 11 *cis*-eQTLs were identified. In five cases the number of minor alleles was inversely associated with transcript levels and in six cases there was a direct association. For the *COMP* gene, three *cis*-eQTLs were identified (two with a direct and one with an inverse correlation). For the *FBLN2* gene, five *cis*-eQTLs were identified (four directly and one inversely correlated).

To validate our results, we searched the GTEx database for prior reports of these eQTLs in lung. Interestingly, for the *CNTNAP1* gene, we found that 28 of 32 eQTLs identified in our study had already been reported, and all but one showed concordant effects of the respective minor alleles on *CNTNAP1* gene expression (Table S5). Regarding the

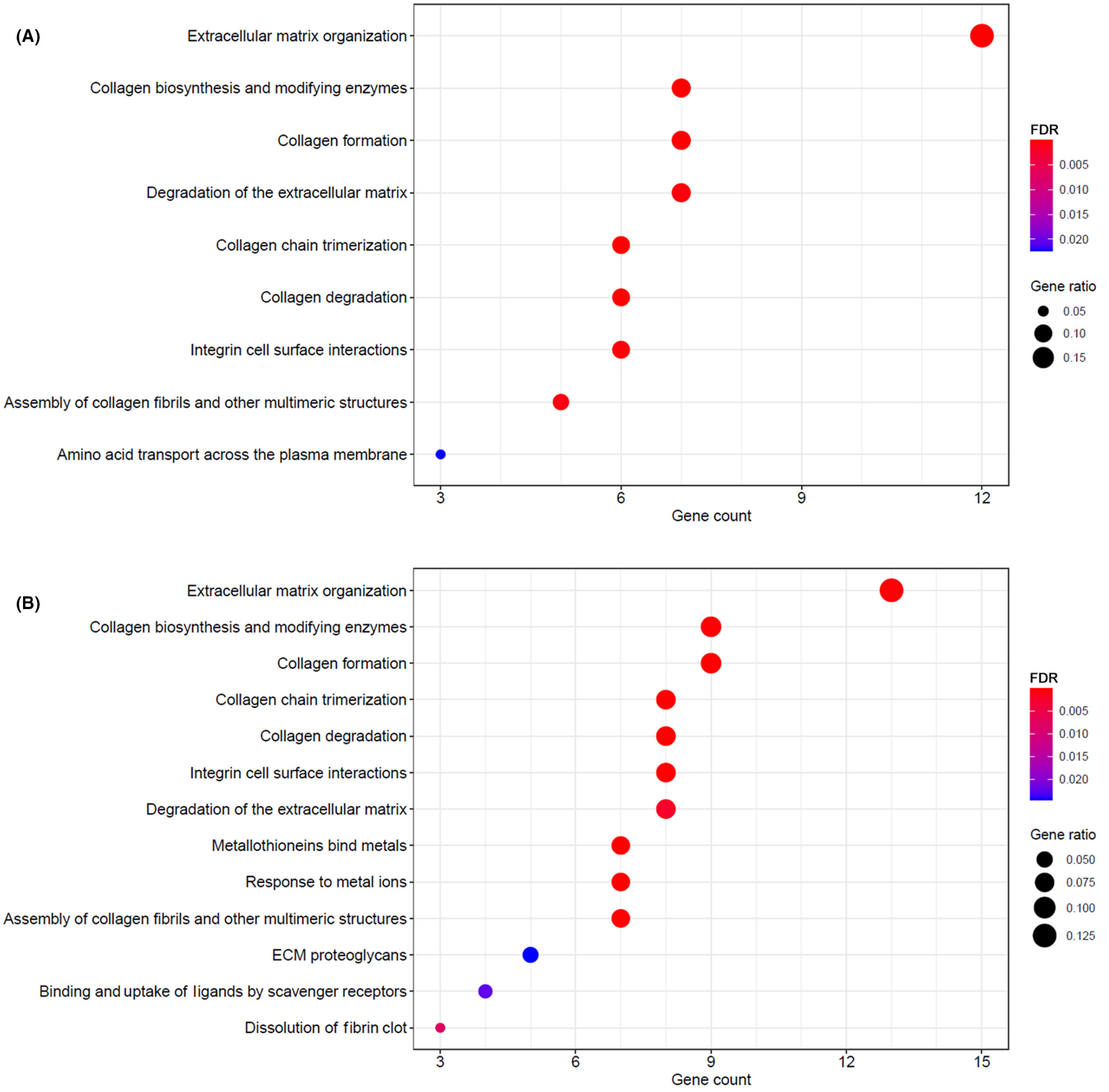

FIGURE 2 Reactome pathways enriched in the genes identified by the (A) Cox and (B) machine learning analyses. “Gene count” is the number of genes enriched in a pathway; “Gene ratio” is the percentage of input genes that are annotated in a pathway; FDR is the false discovery rate calculated to adjust nominal *p* values by the Benjamini–Hochberg method. Pathways are visualized in dotplot enrichment maps: dot size represents the gene ratio; color represents the Benjamini–Hochberg adjusted *p* value (red < purple < blue)

variant for which an opposite effect had been reported, in GTEx there was an inverted annotation of the major and minor alleles compared to our series. Of the five *FBLN2* eQTLs identified in this study, four were listed in GTEx with effects of the minor alleles concordant with what we observed. Of the 11 eQTLs identified for *DUSP14*, only one variant (rs1051849) had previously been reported. Finally, the three *COMP* eQTLs were not validated by GTEx data.

Figure 3 shows regional association plots for the two genes with the higher numbers of eQTLs. Among the 32 *CNTNAP1 cis*-eQTLs (Figure 3A), the lead variant rs2271028 is intronic. The other three most significant eQTL SNPs (i.e., rs9766, rs200701491, and rs2089115; Figure 3A, red dots) map from 830 bp to 6.8 kbp downstream of the gene and are all in strong LD ($r^2 > 0.8$) with the lead variant, rs2271028. For *DUSP14*, the lead variant rs1051849 maps in the 3′-UTR of the gene. rs1051849 is in strong LD ($r^2 > 0.8$) only with rs2074411 (Figure 3B, red dot), which maps in an open chromatin region (according to Variant Effect Predictor of Ensembl). The other variants mapping in or around *DUSP14* (rs853214, intronic; rs853195,

TABLE 2 Transcripts whose levels in lung tissue associated with overall survival of lung adenocarcinoma patients

		Noninvolved tissue[a]		Tumor tissue[b]	
Gene symbol	Gene name	*p* value	HR (95% CI)	*p* value[c]	HR (95% CI)
CLCF1	Cardiotrophin like cytokine factor 1	3.28×10^{-3}	1.62 (1.18–2.24)	0.0150	1.70 (1.10–2.62)
DUSP14	Dual specificity phosphatase 14	3.20×10^{-5}	1.98 (1.44–2.73)	0.0070	1.82 (1.17–2.82)
COL1A1	Collagen type I alpha 1 chain	2.53×10^{-4}	1.89 (1.34–2.65)	0.0013	2.04 (1.31–3.19)
COL7A1	Collagen type VII alpha 1 chain	6.80×10^{-4}	1.79 (1.28–2.49)	0.0032	1.91 (1.23–2.97)
COL14A1	Collagen type XIV alpha 1 chain	5.84×10^{-5}	2.02 (1.44–2.85)	4.8×10^{-7}	0.30 (0.18–0.49)
COL15A1	Collagen type XV alpha 1 chain	7.95×10^{-5}	2.02 (1.42–2.86)	0.0440	1.55 (1.01–2.39)
SERPINH1	Serpin family H member 1	1.04×10^{-3}	1.75 (1.25–2.43)	0.0410	1.57 (1.01–2.42)

Abbreviations: CI, confidence interval; HR, hazard ratio.

[a] Cox model results from analyses comparing patients with high and low transcript levels (above and below the median of $\log_2$-transformed values, respectively), with sex, age, and stage as covariates.

[b] Data from Kaplan–Meier Plotter for lung adenocarcinoma ($n = 387$); sex, age, and stage were used as covariates in the multivariable analysis.[27]

[c] Log–rank *p* value.

1.7 kbp upstream; and rs1063215, 1.3 kbp downstream) are in weaker LD with the lead variant ($0.6 < r^2 \leq 0.8$; Figure 3B, orange dots).

Figure 4 reports the effects of genotype of the most significant eQTL of *CNTNAP1* and *DUSP14* on transcript levels. For *CNTNAP1* (Figure 4A), patients homozygous for the major allele of rs2271028 (G) expressed lower transcript levels than either heterozygotes ($p = 0.040$, ANOVA followed by Tukey's test for multiple comparisons) or homozygotes ($p < 0.001$) for the minor allele A. In contrast, patients homozygous for the major allele of rs1051849 (A) expressed higher levels of *DUSP14* transcript than heterozygotes ($p < 0.001$). *DUSP14* levels in homozygotes for the minor allele G (only seven patients) were not significantly different from those of the AA subgroup.

3.5 | Expression quantitative trait loci SNPs for *CNTNAP1* associate with lung adenocarcinoma prognosis

To test whether the eQTL SNPs were associated themselves with patient survival, we carried out a Cox analysis using genotypes coded as in the genetic additive model, and with age and pathological stage as covariates. All *CNTNAP1* eQTL SNPs but rs11651246 were found to be prognostic variants (Table S6). Of the 26 eQTL SNPs for which there was a direct correlation between the number of minor alleles and *CNTNAP1* transcript level (Table S5), 25 were poor prognostic factors (HR > 1). The six variants for which the number of minor alleles was inversely correlated with transcript levels were associated with better prognosis (HR < 1). Among the three *COMP* eQTL SNPs, only rs12977772 was associated with survival with HR = 1.32 (95% CI, 1.01–1.71; $p = 0.041$). No significant association between *DUSP14* or *FBLN2* eQTLs and survival was observed (FDR > 0.05 for all).

Figure 5 shows Kaplan–Meier curves of survival probability for lung adenocarcinoma patients according to the genotypes of rs7359598 and rs2242461, the most significant variants in the Cox analysis (Table S6). A higher number of minor alleles associated with poorer prognosis (log–rank $p = 1.1 \times10^{-3}$ and 5.94×10^{-4}, respectively), in agreement with the HR of 1.69 and 1.67, respectively (Table S6).

4 | DISCUSSION

We analyzed the transcriptome of noninvolved lung tissue from 483 lung adenocarcinoma patients to identify genes whose expression was associated with overall survival. Cox proportional hazard analysis identified four genes whose transcript levels were associated with outcome (*CLCF1*, *CNTNAP1*, *DUSP14*, and *MT1F*; FDR < 0.05). Machine learning with a Random Forest model confirmed the association of *CLCF1*, *CNTNAP1*, and *DUSP14* and identified another 83 transcripts (with $p < 0.01$ in Cox analysis) associated with lung adenocarcinoma outcome. Additionally, for a subset of 100 genes (most significant transcripts from the Cox analysis), the Random Forest algorithm confirmed their importance in lung adenocarcinoma outcome, although the predictive performance achieved by the model was quite low. Pathway analysis of the 100 top-ranking genes identified by Random Forest (without gene preselection) and of the 174 transcripts associated with survival at $p < 0.01$ (according to Cox) indicated that the two gene lists were enriched in genes involved in

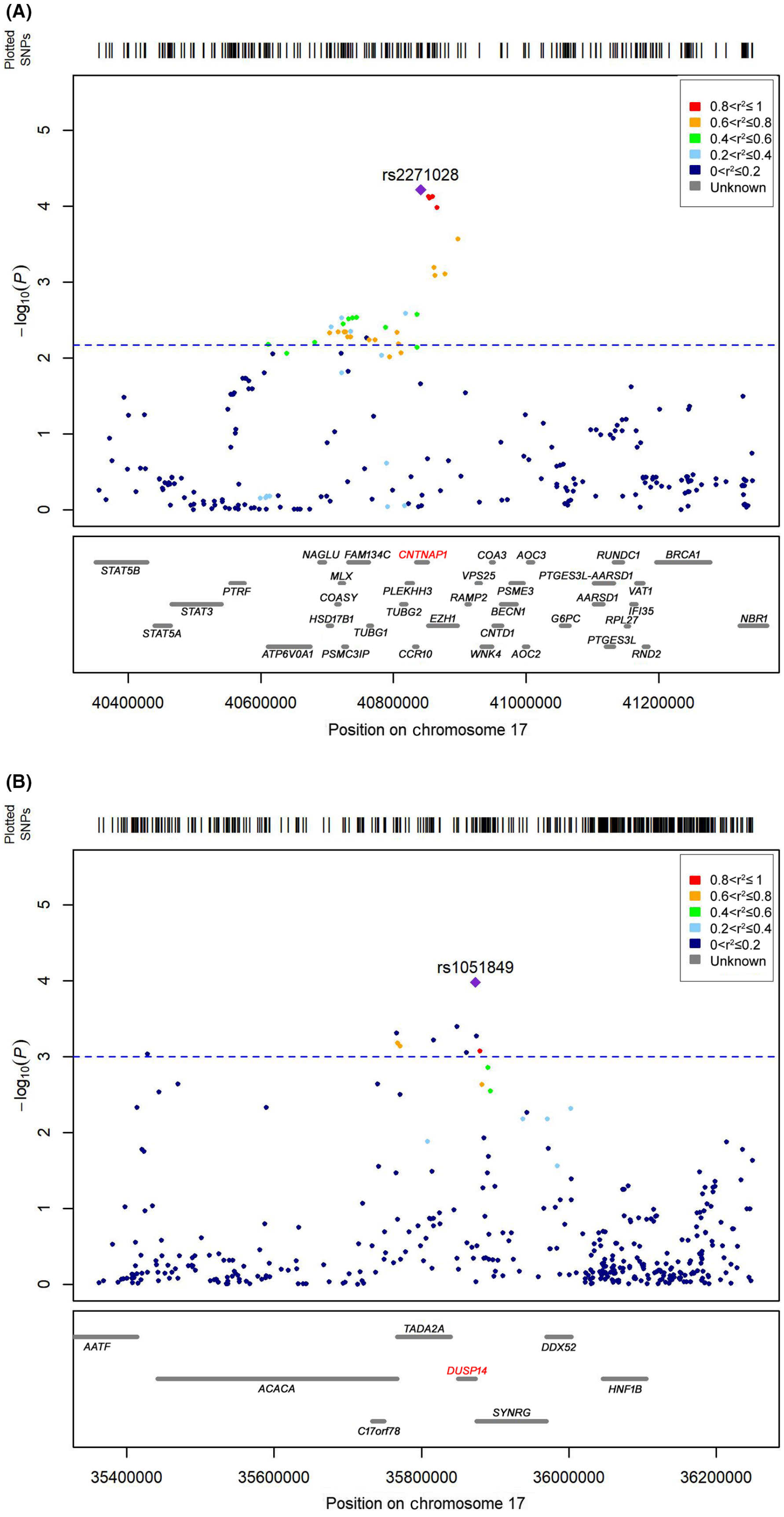

FIGURE 3 Regional association plots for expression quantitative trait loci SNPs. Plots span a 1 Mb region around (A) *CNTNAP1* and (B) *DUSP14* genes. SNPs are plotted on the *x*-axis according to their position on chromosome 17 (GChr 37, hg19 release), and *p* values (−log10[*P*]) for their association with transcript levels are plotted on the *y*-axis. Horizontal dashed blue lines represent the threshold of significance (false discovery rate < 0.05). Dot color represents the level of linkage disequilibrium, expressed as r^2 between each SNP and the lead variant (purple diamond)

ECM organization, for example, collagens. To further investigate the genes identified by the Cox and Random Forest models (3 genes) and those in the ECM pathway (12 genes), we tested whether their levels were also associated with survival in the tumor tissue of an independent series of lung adenocarcinoma patients. Of note, high expression levels of *CLCF1*, *DUSP14*, *COL1A1*, *COL7A1*, *COL14A1*,

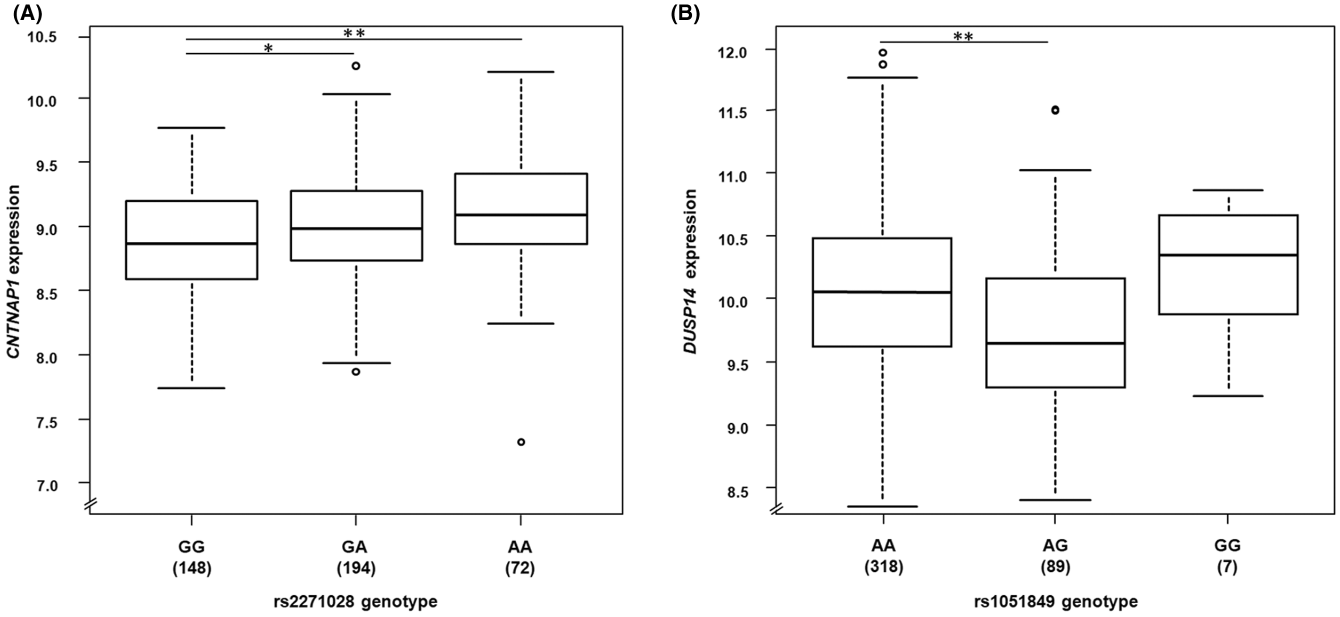

FIGURE 4 Germline control of *CNTNAP1* and *DUSP14* transcript levels in noninvolved lung tissue. (A) Gene expression levels (normalized probe intensities) of *CNTNAP1*, according to rs2271028 genotype. A is the minor allele. (B) Gene expression levels of *DUSP14*, according to rs1051849 genotype. G is the minor allele. Numbers in parentheses are the individuals carrying the indicated genotype. The line within each box represents the median normalized probe intensity; upper and lower edges of each box are the 75th and 25th percentiles, respectively; top and bottom whiskers indicate the largest and smallest value within 1.5 times the interquartile range above the upper quartile and below the lower quartile, respectively; circles denote outliers (extreme values, >1.5 times the interquartile range). $^*p = 0.040$, $^{**}p < 0.001$ vs. homozygotes for the major allele, ANOVA followed by Tukey's test for multiple comparisons

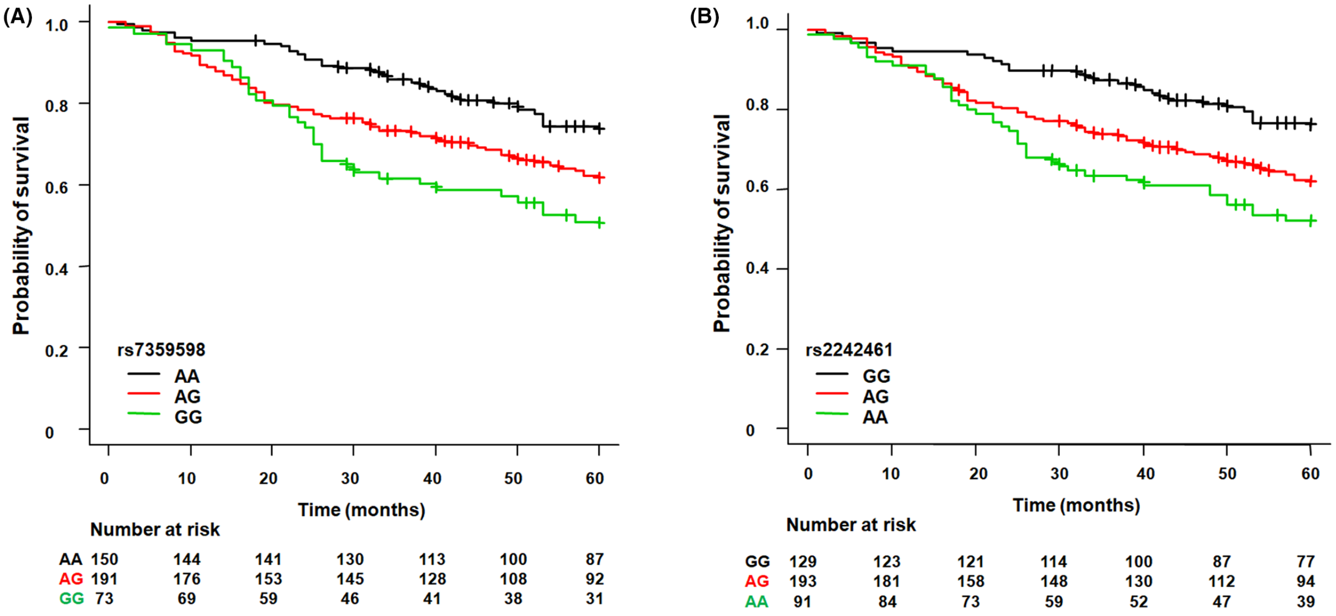

FIGURE 5 Kaplan–Meier survival curves for lung adenocarcinoma patients according to the genotype of the *CNTNAP1* expression quantitative trait loci variants (A) rs7359598 and (B) rs2242461. Green, red. and black lines represent patients homozygous for the minor allele, heterozygous, and homozygous for the major allele, respectively. Crosses denote censored samples. Log–rank *p*-value is shown

COL15A1, and *SERPINH1* were also poor prognostic factors in lung adenocarcinoma. We also observed that the expression of *CNTNAP1* and *DUSP14* genes was genetically regulated by 32 and 11 *cis*-eQTL SNPs, respectively, while that of *COMP* and *FBLN2* (belonging to the ECM pathways) correlated with the genotypes of five and three polymorphisms, respectively. Interestingly, the genotype of 31 of the 32 SNPs associated with *CNTNAP1* levels also associated with patient survival: 25 SNPs directly correlated with *CNTNAP1* transcript levels were poor predictors, while 6 SNPs inversely related to transcript levels associated with a favorable prognosis.

This study confirms, in a larger, partially overlapping patient series, the association between lung adenocarcinoma patient survival and gene expression in noninvolved lung tissue for six genes we previously identified in a 10-gene prognostic signature.[22] These genes are *CNTNAP1*, *SERPINH1*, *FRMD8*, and *SNX10* (here, all associated with survival at $p < 0.01$) and *PPP3R1* and *SNTB1* ($p < 0.05$). *CNTNAP1*, *SERPINH1*, and *FRMD8* ranked within the 100 top positions in the Random Forest model. However, this latter model was trained to identify transcripts whose levels associated with outcome without taking into account the time-to-event data, as the Cox analysis does; this could explain the lack of more extended validation of our previous results. A machine learning algorithm that can model event probabilities as a function of time (e.g., see Kvamme et al.[41]) could be useful in future studies.

Our finding of 100 transcripts whose levels in noninvolved lung tissue associated with outcome corroborates the role of the microenvironment in driving tumor progression. Some of these genes belong to pathways involved in ECM remodeling, and stromal cells were predicted here to significantly contribute to their expression. Additionally, the finding that the expression levels of some of these genes were also poor prognostic factors in lung adenocarcinoma tissue from an independent patient series, strengthen their role, and that of the stromal component, in promoting tumor progression. The ECM is an integral part of the tumor microenvironment. It influences cell aggregation and migration and growth factor presentation to cancer cells, influencing tumor growth.[42] Its composition and organization vary between tumor and normal tissue.[43–45] An overproduction of ECM molecules increases tumor mass and can mask tumor cells from immune cell surveillance and from pharmacological therapy.[46] Collagens, the fibrous proteins that are the main components of the ECM, accumulate within the tumor stroma and increase tissue stiffness, inducing novel cell–cell and cell–matrix interactions that culminate in tumor progression. Different types of collagens have been reported to be prognostic factors in breast, colorectal, and hepatocellular cancer, among others.[47] A similar role was predicted for lung adenocarcinoma, where evidence links collagen X overexpression with poor prognosis.[48] An enhanced expression of ECM genes in noninvolved tumor tissue could be the first step in the modification of a microenvironment that could, in turn, drive tumor progression. Should this close relationship between noninvolved and tumor tissue be confirmed, the genetic analysis of normal tissue could be used to predict disease progression.

This study confirms the association between survival and *CNTNAP1* levels in noninvolved tissue of lung adenocarcinoma patients, which we previously reported.[22] *CNTNAP1* encodes contactin-associated protein 1 (CNTP1), a transmembrane protein that is mainly expressed in brain. Contactin-associated protein 1 plays a role in the development of neural tissue by regulating nerve fiber myelination,[49] together with its binding protein contactin 1 (CNTN1).[50] There is currently scarce information about the functional involvement of CNTP1 in pathologies other than neurodegenerative diseases. Interestingly, CNTN1 has been found to be expressed in several tumor tissues, including lung.[51–55] To our knowledge, no other members of the CNTN1 network have been associated with lung adenocarcinoma and its prognosis. Functional studies are needed to determine whether CNTP1 upregulation in nontumor tissue can boost expression of CNTN1, predisposing to tumor progression.

The *DUSP14* gene encodes dual specificity phosphatase 14, a tyrosine phosphatase involved in signaling pathways related to cell growth, proliferation, and differentiation.[56] Given its involvement in these cellular pathways, the potential role of *DUSP14* in tumor progression is more apparent than for *CNTP1*. The gene was reported to be upregulated in a cellular model of pancreatic cancer.[57] We observed that at least one minor allele of *DUSP14* SNP rs1051849 was associated with low transcript levels. The same association of this SNP with *DUSP14* levels was previously observed in lymphoblastoid cell lines.[58] That same study also found that the minor allele of rs1051849 was a protective factor against melanoma. Although our survival analysis did not reveal an association between this SNP and survival, it is worth investigating whether this regulatory variant of *DUSP14* levels plays a role in lung cancer progression.

This study also found that transcript levels of *CLCF1* are linked to prognosis of lung adenocarcinoma. *CLCF1* encodes cardiotrophin like cytokine factor 1, a pro-inflammatory cytokine member of the interleukin-6 family.[59] A role of CLCF1 in promoting tumor progression has been described in different cancer types.[60,61] Studies in cell and animal models support its involvement in tumor progression in lung,[62] so targeting the related signaling pathway might be a possible therapy for lung adenocarcinoma. Moreover, *CLCF1* is often expressed by cancer-associated fibroblasts, which shape the tumor microenvironment and thus influence tumor cell behavior and ultimately tumor fate.[63]

A limitation of the present study is the lack of a validation cohort. Nevertheless, the machine learning analysis was carried out by repeatedly partitioning the patient series in training and test sets. The MCC values obtained in both the training and test sets were similar and superior to those of a random classifier, indicating the validity of our results. Additionally, different methodological approaches (i.e., standard Cox and machine learning analyses) gave overlapping results, strengthening the findings. The added value of machine learning in the identification of a transcriptomic prognostic signature is the possibility to overcome the main limitation of standard Cox analysis. Indeed, the Cox approach analyzes one gene at a time without considering interactions among them. Multivariate approaches such as machine learning[23] overcome this limitation. They model the relationship between gene expression and survival by accounting for the simultaneous interaction of several genes or related molecular pathways. Thus, they consider the complexity of the overall system.

Notwithstanding the low performance of machine learning in predicting the survival of patients from noninvolved tissue expression data (as evidenced by the relatively low MCC), this result was in some way expected. Indeed, we are aware that the role played by germline regulatory variants in controlling lung cancer progression could be limited. Several other factors participate in cancer progression, including somatic mutations, epigenetic alterations, and environmental factors (e.g., different pharmacological treatments). This study did not take these factors into consideration.

In conclusion, this study shows that genetic predisposition to the prognosis of lung adenocarcinoma is an individual feature already present in patients' noninvolved lung tissue, as gene expression is associated with overall survival. In addition, our findings provide supporting evidence for a role of regulatory germline variants in lung adenocarcinoma outcome.

AUTHOR CONTRIBUTIONS

TAD, FC, MC, and GJ conceived the study and were involved in experimental design. MI, DT, and GM provided biological samples from lung cancer patients and follow-up data. SN prepared DNA and RNA samples. FC, CC, and FM were involved in genotyping and gene expression data management and analysis. MC performed survival analyses and applied machine learning method. FC, MC, BB, and TAD were involved in manuscript preparation. All authors participated in critical revision of the article and approved the final manuscript.

ACKNOWLEDGMENTS

The authors acknowledge the contributions of Valerie Matarese, PhD, who provided scientific editing. This work was supported in part by a grant from the Italian Association for Cancer Research (AIRC, grant no. IG 20226 to T.A. Dragani). The funding organization had no role in design or conduct of the study, collection, management, analysis, or interpretation of the data, preparation, review, or approval of the manuscript.

DISCLOSURE

No potential conflict of interest was reported by the authors. Tommaso A. Dragani is a current Associate Editor of *Cancer Science*.

DATA AVAILABILITY STATEMENT

Expression data that support the findings of this study have been deposited at GEO (accession numbers: GSE71181 and GSE123352).

ETHICS STATEMENT

Approval of the research protocol by an institutional review board: The ethics committees of the recruiting hospitals approved the protocol for tissue collection and genetic studies.

Informed consent: Patients provided written informed consent to the use of their biological samples and data for research purposes.

Registry and the registration no. of the study/trial: N/A.

Animal studies: N/A.

ORCID

Marco Chierici https://orcid.org/0000-0001-9791-9301
Chiara Elisabetta Cotroneo https://orcid.org/0000-0001-6648-2987
Tommaso A. Dragani https://orcid.org/0000-0001-5915-4598
Francesca Colombo https://orcid.org/0000-0003-2015-4317

REFERENCES

1. Oskarsdottir GN, Bjornsson J, Jonsson S, Isaksson HJ, Gudbjartsson T. Primary adenocarcinoma of the lung--histological subtypes and outcome after surgery, using the IASLC/ATS/ERS classification of lung adenocarcinoma. *APMIS*. 2016;124:384-392.
2. Inamura K. Clinicopathological characteristics and mutations driving development of early lung adenocarcinoma: tumor initiation and progression. *Int J Mol Sci*. 2018;19:1259.
3. Zhu Q-G, Zhang S-M, Ding X-X, He B, Zhang H-Q. Driver genes in non-small cell lung cancer: characteristics, detection methods, and targeted therapies. *Oncotarget*. 2017;8:57680-57692.
4. Zhang Y, Chang L, Yang Y, et al. Intratumor heterogeneity comparison among different subtypes of non-small-cell lung cancer through multi-region tissue and matched ctDNA sequencing. *Mol Cancer*. 2019;18:7.
5. Skoulidis F, Byers LA, Diao L, et al. Co-occurring genomic alterations define major subsets of KRAS-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities. *Cancer Discov*. 2015;5:860-877.
6. Yue D, Liu W, Gao L, et al. Integrated multiomics analyses revealing different molecular profiles between early- and late-stage lung adenocarcinoma. *Front Oncol*. 2021;11:746943.
7. Menor M, Zhu Y, Wang Y, Zhang J, Jiang B, Deng Y. Development of somatic mutation signatures for risk stratification and prognosis in lung and colorectal adenocarcinomas. *BMC Med Genomics*. 2019;12:24.
8. Lee Y, Yoon KA, Joo J, et al. Prognostic implications of genetic variants in advanced non-small cell lung cancer: a genome-wide association study. *Carcinogenesis*. 2013;34:307-313.
9. Galvan A, Colombo F, Frullanti E, et al. Germline polymorphisms and survival of lung adenocarcinoma patients: a genome-wide study in two European patient series. *Int J Cancer*. 2015;136:E262-E271.
10. Pintarelli G, Cotroneo CE, Noci S, et al. Genetic susceptibility variants for lung cancer: replication study and assessment as expression quantitative trait loci. *Sci Rep*. 2017;7:42185.
11. Enguix-Riego MDV, Cacicedo J, Delgado Leon BD, et al. The single nucleotide variant rs2868371 associates with the risk of mortality in non-small cell lung cancer patients: a multicenter prospective validation. *Radiother Oncol J Eur Soc Ther Radiol Oncol*. 2019;136:29-36.
12. Liao W-Y, Ho CC, Tsai TH, Chen KY, Shih JY, Yu CJ. Combined effect of ERCC1 and ERCC2 polymorphisms on overall survival in non-squamous non-small-cell lung cancer patients treated with first-line pemetrexed/platinum. *Lung Cancer*. 2018;118:90-96.
13. Xiong Y, Feng Y, Qiao T, Han Y. Identifying prognostic biomarkers of non-small cell lung cancer by transcriptome analysis. *Cancer Biomark*. 2020;27:243-250.
14. Yu X, Zhang X, Zhang Y. Identification of a 5-gene metabolic signature for predicting prognosis based on an integrated analysis of tumor microenvironment in lung adenocarcinoma. *J Oncol*. 2020;2020:5310793.
15. Pan X, Ji P, Deng X, Chen L, Wang W, Li Z. Genome-wide analysis of methylation CpG sites in gene promoters identified four pairs of CpGs-mRNAs associated with lung adenocarcinoma prognosis. *Gene*. 2022;810:146054.
16. Wang X, Zhou B, Xia Y, et al. A methylation-based nomogram for predicting survival in patients with lung adenocarcinoma. *BMC Cancer*. 2021;21:801.
17. Li J, Li H, Zhang C, et al. Identification of a gene signature closely related with immunosuppressive tumour microenvironment predicting prognosis of patients in EGFR mutant lung adenocarcinoma. *Front Oncol*. 2021;11:732841.
18. Nie K, Nie W, Zhang Y-X, Yu H. Comparing clinicopathological features and prognosis of primary pulmonary invasive mucinous adenocarcinoma based on computed tomography findings. *Cancer Imaging off Publ Int Cancer Imaging Soc*. 2019;19:47.

19. Shim WS, Yim K, Kim T-J, et al. DeepRePath: identifying the prognostic features of early-stage lung adenocarcinoma using multi-scale pathology images and deep convolutional neural networks. *Cancers (Basel)*. 2021;13:3308.
20. Song C, Guo Z, Yu D, et al. A prognostic nomogram combining immune-related gene signature and clinical factors predicts survival in patients with lung adenocarcinoma. *Front Oncol*. 2020;10:1300.
21. Lathwal A, Kumar R, Arora C, Raghava GPS. Identification of prognostic biomarkers for major subtypes of non-small-cell lung cancer using genomic and clinical data. *J Cancer Res Clin Oncol*. 2020;146:2743-2752.
22. Galvan A, Frullanti E, Anderlini M, et al. Gene expression signature of non-involved lung tissue associated with survival in lung adenocarcinoma patients. *Carcinogenesis*. 2013;34:2767-2773.
23. Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S. Machine learning and its applications to biology. *PLoS Comput Biol*. 2007;3:e116.
24. Pintarelli G, Noci S, Maspero D, et al. Cigarette smoke alters the transcriptome of non-involved lung tissue in lung adenocarcinoma patients. *Sci Rep*. 2019;9:13039.
25. Maspero D, Dassano A, Pintarelli G, et al. Read-through transcripts in lung: germline genetic regulation and correlation with the expression of other genes. *Carcinogenesis*. 2020;41:918-926.
26. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. 1995;57:289-300.
27. Győrffy B, Surowiak P, Budczies J, Lánczky A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One*. 2013;8:e82241.
28. Breiman L. Random Forests. *Mach Learn*. 2001;45:5-32.
29. Shi L, Campbell G, Jones WD, et al. The MicroArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol*. 2010;28:827-838.
30. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat Biotechnol*. 2014;32:903-914.
31. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000;16:412-424.
32. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21:6.
33. Jurman G, Merler S, Barla A, Paoli S, Galea A, Furlanello C. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*. 2008;24:258-264.
34. Jurman G, Riccadonna S, Visintainer R, Furlanello C. Algebraic comparison of partial lists in bioinformatics. *PLoS One*. 2012;7:e36540.
35. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst*. 2016;12:477-479.
36. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol*. 2017;18:220.
37. Cotroneo CE, Mangano N, Dragani TA, Colombo F. Lung expression of genes putatively involved in SARS-CoV-2 infection is modulated in cis by germline variants. *Eur J Hum Genet*. 2021;29:1019-1026.
38. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559-575.
39. Major T, Takei R. LocusZoom-Like Plots for GWAS Results. 2021. doi:10.5281/ZENODO.5154379
40. GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat Genet*. 2013;45:580-585.
41. Kvamme H, Borgan Ø. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Anal*. 2021;27:710-736.
42. Winkler J, Abisoye-Ogunniyan A, Metcalf KJ, Werb Z. Concepts of extracellular matrix remodelling in tumour progression and metastasis. *Nat Commun*. 2020;11:5120.
43. Provenzano PP, Inman DR, Eliceiri KW, et al. Collagen density promotes mammary tumor initiation and progression. *BMC Med*. 2008;6:11.
44. Auvinen P, Tammi R, Kosma V-M, et al. Increased hyaluronan content and stromal cell CD44 associate with HER2 positivity and poor prognosis in human breast cancer. *Int J Cancer*. 2013;132:531-539.
45. Mammoto T, Jiang A, Jiang E, et al. Role of collagen matrix in tumor angiogenesis and glioblastoma multiforme progression. *Am J Pathol*. 2013;183:1293-1305.
46. Henke E, Nandigama R, Ergün S. Extracellular matrix in the tumor microenvironment and its impact on cancer therapy. *Front Mol Biosci*. 2019;6:160.
47. Xu S, Xu H, Wang W, et al. The role of collagen in cancer: from bench to bedside. *J Transl Med*. 2019;17:309.
48. Liang Y, Xia W, Zhang T, et al. Upregulated collagen COL10A1 remodels the extracellular matrix and promotes malignant progression in lung adenocarcinoma. *Front Oncol*. 2020;10:573534.
49. Lesmana H, Vawter Lee M, Hosseini SA, et al. CNTNAP1-related congenital Hypomyelinating neuropathy. *Pediatr Neurol*. 2019;93:43-49.
50. Peles E, Nativ M, Lustig M, et al. Identification of a novel contactin-associated transmembrane receptor with multiple domains implicated in protein-protein interactions. *EMBO J*. 1997;16:978-988.
51. Hung Y-H, Hung W-C. 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) enhances invasiveness of lung cancer cells by up-regulating contactin-1 via the alpha7 nicotinic acetylcholine receptor/ERK signaling pathway. *Chem Biol Interact*. 2009;179:154-159.
52. Shi K, Xu D, Yang C, et al. Contactin 1 as a potential biomarker promotes cell proliferation and invasion in thyroid cancer. *Int J Clin Exp Pathol*. 2015;8:12473-12481.
53. Xu S, Lam S-K, Cheng PN-M, Ho JC-M. Contactin 1 modulates pegylated arginase resistance in small cell lung cancer through induction of epithelial-mesenchymal transition. *Sci Rep*. 2019;9:12030.
54. Zhang D, Zhou S, Liu B. Identification and validation of an individualized EMT-related prognostic risk score formula in gastric adenocarcinoma patients. *Biomed Res Int*. 2020;2020:7082408.
55. Wang B, Yang X, Zhao T, et al. Upregulation of contactin-1 expression promotes prostate cancer progression. *Oncol Lett*. 2020;19:1611-1618.
56. Lountos GT, Tropea JE, Cherry S, Waugh DS. Overproduction, purification and structure determination of human dual-specificity phosphatase 14. *Acta Crystallogr D Biol Crystallogr*. 2009;65:1013-1020.
57. Wei Y, Wang G, Wang C, et al. Upregulation of DUSP14 affects proliferation, invasion and metastasis, potentially via epithelial-mesenchymal transition and is associated with poor prognosis in pancreatic cancer. *Cancer Manag Res*. 2020;12:2097-2108.
58. Liu H, Wang L-E, Loiu Z, et al. Association between functional polymorphisms in genes involved in the MAPK signaling pathways and cutaneous melanoma risk. *Carcinogenesis*. 2013;34:885-892.
59. Sims NA. Cardiotrophin-like cytokine factor 1 (CLCF1) and neuropoietin (NP) signalling and their roles in development, adulthood, cancer and degenerative disorders. *Cytokine Growth Factor Rev*. 2015;26:517-522.
60. Hu X, Zhao Y, He X, et al. Ciliary neurotrophic factor receptor alpha subunit-modulated multiple downstream signaling pathways in hepatic cancer cell lines and their biological implications. *Hepatology*. 2008;47:1298-1308.
61. Kober P, Bujko M, Olędzki J, Tysarowski A, Siedlecki JA. Methyl-CpG binding column-based identification of nine genes hypermethylated in colorectal cancer. *Mol Carcinog*. 2011;50:846-856.

62. Kim JW, Marquez CP, Kostyrko K, et al. Antitumor activity of an engineered decoy receptor targeting CLCF1-CNTFR signaling in lung adenocarcinoma. *Nat Med*. 2019;25:1783-1795.
63. Lei MML, Lee TKW. Cancer-associated fibroblasts: orchestrating the crosstalk between liver cancer cells and neutrophils through the Cardiotrophin-like cytokine factor 1-mediated chemokine (C-X-C motif) ligand 6/TGF-β Axis. *Hepatology (Baltimore, MD)*. 2021;73:1631-1633.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Minnai F, Noci S, Chierici M, et al. Genetic predisposition to lung adenocarcinoma outcome is a feature already present in patients' noninvolved lung tissue. *Cancer Sci*. 2023;114:281-294. doi: 10.1111/cas.15591