IJC
International Journal of Cancer

Germline polymorphisms and survival of lung adenocarcinoma patients: A genome-wide study in two European patient series

Antonella Galvan[1], Francesca Colombo[1], Elisa Frullanti[1], Alice Dassano[1], Sara Noci[1], Yufei Wang[2], Timothy Eisen[3], Athena Matakidou[4], Luisa Tomasello[5], Marzia Vezzalini[5], Claudio Sorio[5], Matteo Dugo[1], Federico Ambrogi[6], Ilaria Iacobucci[7], Giovanni Martinelli[7], Matteo Incarbone[8], Marco Alloisio[9], Mario Nosotti[10], Davide Tosi[10], Luigi Santambrogio[10], Giuseppe Pelosi[1], Ugo Pastorino[1], Richard S. Houlston[2] and Tommaso A. Dragani[1]

[1] Fondazione IRCCS, Istituto Nazionale dei Tumori, Milan, Italy
[2] Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey SM2 5NG, United Kingdom
[3] Cambridge University Health Partners, Cambridge, United Kingdom
[4] Cambridge Biomedical Centre, Cambridge, United Kingdom
[5] Department of Pathology and Diagnostics, University of Verona School of Medicine and Surgery, Verona, Italy
[6] Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy
[7] Institute of Hematology "Seragnoli", Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, Bologna, Italy
[8] Department of Surgery, San Giuseppe Hospital-MultiMedica, Milan, Italy
[9] Department of Surgery, Istituto Clinico Humanitas, Rozzano, Italy
[10] Fondazione IRCCS Ospedale Maggiore Policlinico, University of Milan, Milan, Italy

Cancer Genetics

In lung cancer, the survival of patients with the same clinical stage varies widely for unknown reasons. In this two-phase study, we examined the hypothesis that germline variations influence the survival of patients with lung adenocarcinoma. First, we analyzed existing genotype and clinical data from 289 UK-resident patients with lung adenocarcinoma, identifying 86 single nucleotide polymorphisms (SNPs) that associated with survival ($p < 0.01$). We then genotyped these candidate SNPs in a validation series of 748 patients from Italy that resulted genetically compatible with the UK series based on principal component analysis. In a Cox proportional hazard model adjusted for age, sex and clinical stage, four SNPs were confirmed on the basis of their having a hazard ratio (HR) indicating the same direction of effect in the two series and $p < 0.05$. The strongest association was provided by rs2107561, an intronic SNP of *PTPRG,* protein tyrosine phosphatase, receptor type, G; the C allele was associated with poorer survival in both patient series (pooled analysis $\log_e HR = 0.31$; 95% CI: 0.15–0.46, $p = 8.5 \times 10^{-5}$). *PTPRG* mRNA levels in 43 samples of lung adenocarcinoma were 40% of those observed in noninvolved lung tissue from the same patients. *PTPRG* overexpression significantly inhibited the clonogenicity of A549 lung carcinoma cells and the anchorage-independent growth of the NCI-H460 large cell lung cancer line. These four germline variants represent promising candidates that, with further study, may help predict clinical outcome. In addition, the *PTPRG* locus may have a role in tumor progression.

Key words: clonogenicity, clinical stage, genome-wide association, prognostic markers, PTPRG, tumor progression
Additional Supporting Information may be found in the online version of this article.
Grant sponsors: Bobby Moore Fund, Cancer Research UK; **Grant numbers:** C1298/A8780, C1298/A8362; **Grant sponsor:** Italian Association for Cancer Research; **Grant numbers:** 10323 and 12162; **Grant sponsors:** Allan J Lerner Fund, National Cancer Research Network, Helen Rollason Heal Cancer Charity and Sanofi-Aventis
DOI: 10.1002/ijc.29195
History: Received 28 Feb 2014; Accepted 5 Aug 2014; Online 6 Sep 2014
Present address: Elisa Frullanti's current address is: Medical Genetics, University of Siena, Policlinico S, Maria alle Scotte, Siena, Italy
Correspondence to: Tommaso A. Dragani, Department of Predictive and Preventive Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, Via Amadeo 42, I-20133 Milan, Italy, Tel.: +39-0223-902642, Fax: +39-0223-902764.
E-mail: tommaso.dragani@istitutotumori.mi.it

Introduction

The TNM staging system is commonly used in clinical practice for predicting prognosis and treatment requirements of patients with lung cancer.[1,2] However, the outcome of patients who apparently have the same stage of disease can vary.[3,4] As the reasons for such variability in outcome are not known, the search for prognostic factors independent of clinical stage, for example individual characteristics and biomarkers,[5] is an active field of research. A more precise assessment of prognosis would be beneficial in clinical decision-making, allowing physicians to make patient-tailored decisions on drug therapy and consequently improve patient outcome.

Most of the studies on prognostic markers for lung cancer have focused on the tumor tissue itself. Many studies have found that mutations in *KRAS*, a gene encoding a small GTPase, associate inversely with survival in patients with lung adenocarcinoma, as documented by two systematic reviews.[6,7] Differently, the presence of *EGFR* mutations

What's new?

Lung adenocarcinoma patients are at a high risk of poor prognosis in spite of surgical resection, and identifying the factors for individual predisposition to poor prognosis is needed to improve follow-up and therapy. Here, the authors discovered a novel genetic profile of four SNPs associated with survival. If the results are confirmed, the profile could help stratify lung adenocarcinoma patients into different prognostic groups. The strongest association was provided by rs2107561, an intronic SNP of *PTPRG*, whose overexpression inhibited clonogenicity and anchorage-independent growth of lung cancer cells, pointing to a functional role of *PTPRG* in the prognosis of lung adenocarcinoma.

predicts a clinical response to tyrosine kinase inhibitor therapy and longer progression-free survival, as shown in a meta-analysis.[8] Thus, there is an increasing trend to test tumoral tissue for the presence of known mutations, particularly in *EGFR*, to assess a patient's treatment schedule and improve patients' prognosis.[9] Additionally, in the attempt to develop a more sensitive tool that considers multiple genes, numerous studies have described prognostic gene expression profiles of lung cancer tissue, although in a critical analysis none was judged to be ready for clinical application.[10]

A recent trend in cancer research is to shift the focus of the search for prognostic markers from the tumoral DNA to the germline DNA, based on the realization that genetic polymorphisms can modulate the risk of developing certain types of cancer and even influence the prognosis.[11,12] In lung cancer, several studies have tested candidate single nucleotide polymorphisms (SNPs) in association with disease prognosis, but only a few studies used a genome-wide design.[13,14] However, these studies included patients with all forms of non-small cell lung cancer, without taking into account the different histological subtypes despite the fact that histological type is a determinant of survival.[15,16] Therefore, it could be useful to analyze a single histotype of lung cancer for genetic determinants of survival.

Working with the hypothesis that germline polymorphisms modulate overall survival in lung cancer patients, we designed a study to identify such polymorphisms in a single histological subtype, focusing on lung adenocarcinoma. Here, we report the results of a genome-wide association (GWA) study in a homogeneous UK discovery series of lung adenocarcinoma patients, followed by external validation of the best associated SNPs in an independent Italian series of patients with lung adenocarcinoma. In conducting the validation study, we were aware of the limited comparability with the discovery series regarding some demographic and clinical variables; to partially overcome these difficulties, we carried out the survival analyses by adjusting for relevant covariates in each series and we also examined the genetic comparability of these two European case series.

Subjects, Material and Methods

Ethics

Collection of blood samples and clinicopathological information from patients was undertaken with ethical review board approval and informed consent in accordance with the tenets of the Declaration of Helsinki. Specifically, in the UK ethical approval was obtained from the Medical Research Ethics Committee (MREC) while in Italy the study protocol was approved by the Committees for Ethics of the institutes involved in recruitment (Fondazione IRCCS Istituto Nazionale dei Tumori, San Giuseppe Hospital, Ospedale Maggiore Policlinico).

Discovery phase

For the discovery phase, we used data from a GWA study on lung cancer risk conducted in the UK.[17] From that work, we studied the subset of 289 cases with pathologically confirmed lung adenocarcinoma for whom survival data were available; these cases had been collected as part of the Genetic Lung Cancer Predisposition Study (GELCAPS) from oncology centres all over the UK.[18] All cases were British residents with self-reported European ancestry. Genomic DNAs from these patients had been genotyped for 30,568 SNPs using Illumina SNP-arrays (Illumina).[17] In addition to the data regarding genotypes, we obtained information regarding age at diagnosis, sex, smoking status, clinical stage, treatment and overall survival (up to 50 months after treatment).

Validation phase

For the validation phase that was conducted in Italy, we accessed a biobank containing samples of peripheral blood of Italian lung cancer patients. Inclusion criteria for this study were that the patients had received surgical treatment for lung adenocarcinoma, and that follow-up data were available. The biobank also provided clinical and personal data, such as the patients' age at diagnosis, clinical stage and smoking status. We considered a maximum follow-up of 60 months, as deaths after this period may be unrelated to cancer; therefore, survival durations longer than 60 months were censored in the analysis of survival.

For candidate gene analysis, we selected from a separate biobank 43 pairs of specimens of lung adenocarcinoma and non-involved (apparently normal) lung tissue from patients being treated in the authors' institutes around Milan, Italy. Noninvolved lung specimens had been excised during surgical resection of the cancer, with prior approval of the institutes' Committees for Ethics. In particular, specimens were

taken as far as possible from the tumor, from the part of the lobe that would otherwise have been discarded.

Genomic DNA was extracted from blood samples using the DNeasy Blood & Tissue Kit (Qiagen, Valencia, USA) and quantified by fluorimetry using the Picogreen dsDNA Quantitation Kit (Life Technologies). Top 96 SNPs, according to the *p*-values, identified during the discovery phase were genotyped in this series using customized 96.96 Dynamic Arrays (Fluidigm, South San Francisco, CA), which are able to analyze 96 samples with 96 SNPtype assays on a BioMark platform (Fluidigm). Each SNPtype assay is based on an allele-specific PCR detection system, using allele-specific primers labeled with FAM or HEX fluorescent dyes and a locus-specific primer (LSP). Before SNP genotyping, genomic DNA was amplified using specific target amplification primers and LSPs, and then diluted 1:100, according to the manufacturer's instructions; each array was loaded with 91 samples (two samples were loaded in duplicate to monitor quality control) and three nontemplate controls (NTC).

Data were analyzed using Fluidigm SNP Genotyping Analysis software (version 3.1.2). First, the FAM and HEX fluorescence intensities of the two possible alleles (called *x* and *y*) for each SNPassay were normalized using the NTC normalization method, which assigns the NTC cells to the $x = 0.1$ and $y = 0.1$ locations on the plot. To obtain genotype calls for each SNP, a k-means clustering algorithm was used to divide samples into three groups; samples having a signal intensity <0.1 for either allele were excluded from the respective SNP call. Samples having a genotype call rate <0.10 were excluded from further analysis.

Analysis of genetic comparability

To visually assess the genetic comparability of the discovery and validation series, we performed principal component analysis (PCA) and discriminant analysis of principal component (DAPC) on genotype data using the *glPca* and *dapc* functions of the *adegenet* package[19,20] for the R software. PCA and DAPC are multivariate methods that allow the investigation of the genetic structure in large datasets. PCA aims to measure the overall variability between individuals, which is composed of between-groups and within-groups variability. DAPC transforms the data using PCA to reduce the number of variables analyzed and then builds synthetic variables, the discriminant functions, as linear combinations of alleles to maximize the between-groups variability while minimizing the within-groups variation, leading to a better distinction between predefined groups of individuals. Briefly, the genotype data were pooled into a single dataset excluding patients with a SNP genotype missingness >30% ($n = 23$, all in the validation series), genotypes were expressed as 0, 1, or 2 copies of the minor allele, the patients' country of origin was used as the prior group assignment, and data were analyzed using PCA and DAPC.

Cell cultures

Three human lung cancer cell lines were used for molecular and functional assays. A549 lung carcinoma cells were cultured in Ham's F12 medium containing 10% fetal bovine serum (FBS). NCI-H520 squamous cell and NCI-H460 large cell lines were cultured in RPMI1640 with 10% FBS.

Quantitative real-time PCR

Total RNA was extracted from 43 matched pairs of noninvolved lung parenchyma and lung adenocarcinomas and from A549, NCI-H520 and NCI-H460 cell lines using Trizol (Life Technologies). RNA (1 μg) was used to synthesize cDNA using the Transcriptor First Strand cDNA Synthesis Kit (Roche, Basel, Switzerland). To quantify *PTPRG* expression, we used intron-spanning primers (forward, 5′-GGCAGGAGGTTTCCTGTTGA-3; reverse, 5′-GCAGAATTGTCCCTCGGACT-3′) in quantitative PCR. Each reaction comprised 12 ng cDNA template diluted in RNase-free dH_2O, 5 μl 2× Fast SYBR Green Master Mix (Life Technologies), and 0.3 μ*M* PCR primers in a final volume of 10 μl. The human hypoxanthine phosphoribosyltransferase 1 (*HPRT1*; forward, 5′-GACTTTGCTTTCCTTGGTCAGG-3′; reverse, 5′-TCCTTTTCACCAGCAAGCTTG-3′) gene was used to normalize expression data. Reactions were run in duplicate on an ABI 7900HT platform (Life Technologies). Relative quantities of *PTPRG* mRNA levels in each pair of adenocarcinoma tissue and noninvolved lung tissue, as well as in cell lines were assessed using the comparative cycle threshold (Ct) method and calculated with respect to a pool of 64 RNA samples from noninvolved lung tissue (different from the previous 43 samples), used as calibrator.

Immunohistochemistry

Sections of neutral buffered formalin-fixed paraffin-embedded tissues from lung adenocarcinoma and paired normal lung tissue were analyzed by immunohistochemistry for detection of PTPRG protein. Antigen retrieval was performed in 10 m*M* citrate buffer (pH 6.0) heated in a microwave at 360 W for 20 min. Sections were incubated with anti-PTPRG primary chicken antibody (Aves Labs, Oregon, OR) (6.25 μg/ml) for 90 min at room temperature in PBS/0.05% Tween® 20/1% BSA/4 m*M* NaN_3, then washed and incubated with rabbit antichicken IgY (IgG) (whole molecule)-HRP (A9046 Sigma-Aldrich, 1:1000). The immunoreaction was visualized using 3,3′-diaminobenzidine (DAB) staining (K3648, Dako) and sections were counterstained with hematoxylin and dehydrated. For negative controls, the primary antibody was replaced with the preimmune IgY.

Transfections

A549, NCI-H520 and NCI-H460 cell lines were transfected with the pCR3.1 expression vector containing the *PTPRG* coding sequence (ENST00000295874)[21] or with the empty plasmid (pCR3.1, Life Technologies). Briefly, cells growing in 12-well culture plates were treated with 2.5 μl FuGene HD Transfection Reagent (Roche, Basel, Switzerland) and 1 μg (0.163 pmol) *PTPRG*-containing plasmid or an equimolar amount of empty pCR3.1 (544 ng, plus 456 ng carrier

Table 1. SNPs associated with overall survival in the UK discovery series and confirmed in the Italian validation series on the basis of having a hazard ratio indicating the same direction of effect, with $p < 0.05$

					Discovery series (UK)		Validation series (Italy)	
Chr.	Position[1]	SNP[2]	Gene[3]	Minor allele	$\log_e$HR (95% CI)[4]	*p*	$\log_e$HR (95% CI)	*p*
3	61.994	rs2107561	*PTPRG*	C	0.50 (0.21 to 0.80)	8.9E-04	0.24 (0.06 to 0.42)	8.2E-03
5	116.664	rs6882451		T	−0.38 (−0.65 to −0.12)	4.7E-03	−0.21 (−0.37 to −0.04)	1.3E-02
5	116.625	rs1826692		G	−0.38 (−0.63 to −0.12)	3.8E-03	−0.18 (−0.32 to −0.03)	1.7E-02
5	116.608	rs6595026		C	−0.38 (−0.63 to −0.13)	2.6E-03	−0.15 (−0.30 to −0.01)	3.9E-02

[1]Position in Mb, based on Assembly GRCh37.p5, Genome Build 37.3.
[2]Listed in order of increasing *p*-value in the validation series.
[3]Gene containing the SNP in its gene region (*PTPRG*).
[4]Natural logarithm of the hazard ratio, obtained by Cox regression adjusted for sex, age at diagnosis and clinical stage (I *vs.* >I) in both series, and additionally for surgery and radiotherapy in the UK series.

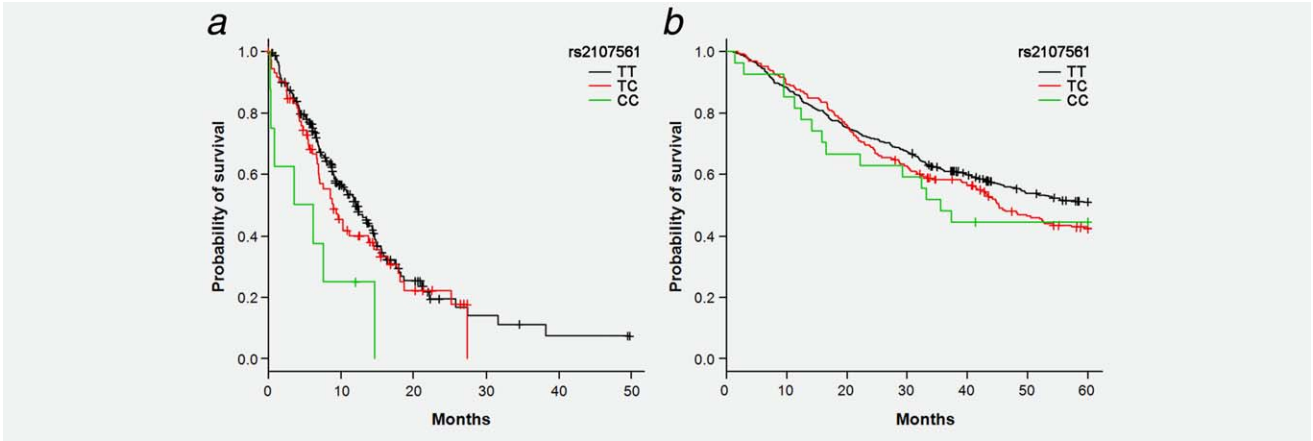


Figure 1. Kaplan–Meier survival curves for the discovery series (panel *a*) and the validation series (panel *b*) patients by rs2107561 genotype (TT, black lines and crosses; T/C, red lines and crosses; CC, green lines or crosses). Crosses denote censored samples.

pUC19 plasmid DNA, Life Technologies). PTPRG protein expression after transient transfection was confirmed by Western blot analysis. Briefly, total proteins were extracted, 48 h after transfection, from NCI-H460 cells transfected with *PTPRG* containing plasmid or with the empty vector, using RIPA buffer containing protease inhibitor cocktail (diluted 1:100). Total protein content was assessed using Bradford assay (Sigma, MO). Twenty-five microgram of protein extracts were loaded and run on a 6% polyacrylamide, 0.1% SDS gel. Gel was electro-blotted on polyvinylidene difluoride membrane (Sigma Aldrich, Milan, Italy). After saturation for 1 h in 0.05% TBS/Tween® 20/5% BSA and overnight incubation with rabbit anti-PTPRG (P4) antibody (1 μg/ml), the membrane was washed and incubated for 1 h with ECL Rabbit IgG, HRP-linked F(ab)2 fragment from donkey (GE Healthcare Life Sciences, Buckinghamshire, UK), further washed and assayed with ECL (Millipore, Billerica, MA).

Transfected cells were used in colony forming and *in vitro* migration assays.

Colony forming assays

Anchorage-dependent growth was tested in a clonogenic assay. Cells were exposed to the transfection reagents (as above) for 24 h prior to being detached and seeded at low density in 6-well plates (∼10,000 cells/well). The selection of stable transfectant clones was started 24 h later by adding G418 (Geneticin; Invitrogen Life Technologies) to the culture medium at 0.5 mg/ml (A549) or 1 mg/ml (NCI-H460, NCI-H520 cells). After 2 weeks, clones were methanol-fixed and stained with 10% Giemsa. Photographs of clone-containing wells were acquired with UVIdoc HD2 instrument, and colonies were counted with UVIdoc software (UVItec Limited, Cambridge, UK). In the first experiment, each cell line was transfected with each plasmid in triplicate, with each transfection assayed for colony formation in three wells. Then, the experiment in A549 cells was repeated twice.

Anchorage-independent growth was tested in a soft agar colony formation assay (CytoSelect 96-well cell

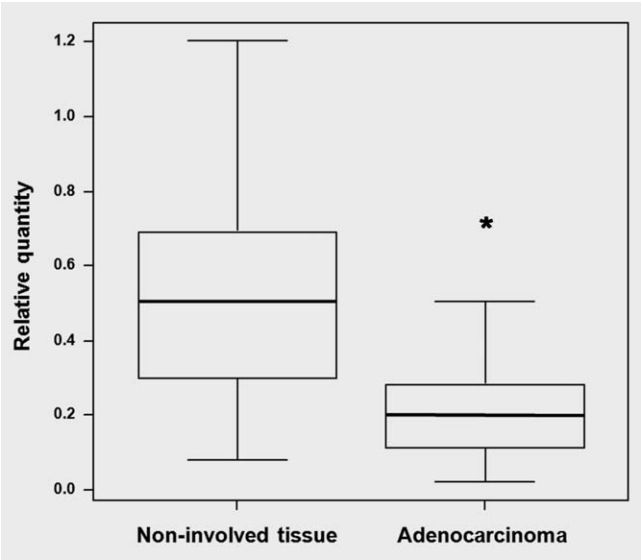


Figure 2. Relative expression levels of *PTPRG* mRNA in 43 pairs of noninvolved lung tissue and lung adenocarcinoma taken from the same patients. The line within each box represents the median relative quantity obtained by qPCR analysis; upper and lower edges of each box represent the 75th and 25th percentile, respectively; upper and lower bars indicate the highest and lowest values less than one interquartile range from the extremes of the box. *$p < 0.0001$, paired *t*-test.

transformation assay; Cell Biolabs, San Diego, CA). Briefly, 96-well plates were prepared with a base agar layer consisting of 0.6% agar in culture medium, 10% FBS and Geneticin at the cell-specific concentrations indicated above (complete medium). Forty-eight hour after the beginning of transfection, cells were detached, suspended in 0.4% agar in complete medium and seeded in the prepared plates at 5000 cells/well. Complete medium was added over the soft agar layers and renewed every two days. On the seventh day, the agar was solubilized, the cells were lysed and DNA was labeled with the CyQuant GR dye. Lysates were transferred to 96-well black plates (three wells for each soft agar well) and fluorescence (excitation, 485 nm; emission, 520 nm) was measured on a microplate reader (Tecan ULTRA). Each cell line was transfected with each plasmid in duplicate (two independent experiments), each transfection was assayed for colony formation in triplicate wells, and each well was measured in triplicate with the fluorimeter. Data were normalized to the mean fluorescence value of each independent experiment.

In vitro cell migration

After 24 h of exposure to the transfection reagents, the cell culture medium was replaced with serum-free medium and the cells were incubated for 24 h. Then the cells were harvested and plated (150,000 cells/well) in polycarbonate Transwell inserts (with 8.0 μm pores; Corning B.V. Life Sciences, Amsterdam, The Netherlands) in serum-free medium, whereas medium containing 10% FBS was added to the bottom wells. Cells were incubated for 24 h prior to measuring migration through the Transwell. Briefly, medium was removed and wells were washed twice with phosphate-buffered saline. Cells that had migrated through the Transwell were detached from the bottom side of the membrane using the Cell Dissociation Solution (Cultrex, Trevigen, Gaithersburg, MD) and labeled with calcein AM (Life Technologies) for 30 min at 37°C in a 5% CO_2 atmosphere. Fluorescence (excitation, 485 nm; emission, 520 nm) was measured in black 96-well plates using a microplate reader (Tecan ULTRA). A549 and NCI-H460 cells were transfected with each plasmid in triplicate, each transfection was seeded onto three Transwell inserts, and fluorescence of the migrated cells was measured in triplicate. Transfected NCI-H520 cells were not assayed because preliminary experiments using untransfected cells revealed a relatively low motility.

Statistical analyses

The association between SNP genotype and survival in the discovery series was evaluated in a Cox proportional hazard model. In the validation series, a numerical value (0, 1, 2) was assigned to each patient according to the number of minor alleles of the SNP in that patient's genotype. The statistical model used is based on the additive effects of minor allele number on the log of the instantaneous risk of dying. Data were adjusted for the phenotypic variables that were found to impact upon survival, using a multivariable Cox analysis. Log_e-transformed hazard ratios (HRs) and *p*-values were computed for each SNP.[19] Deviations of allelic frequencies from the Hardy-Weinberg equilibrium were assessed using the PLINK software.[22]

Candidate SNPs were defined as having a $p < 0.05$ and an HR with the same sign (indicating the same direction of effect) in the discovery and validation series.

Survival curves were generated using the Kaplan–Meier method and differences between groups were assessed with the log-rank test. The extent of linkage disequilibrium (LD) between SNPs was assessed using PLINK software.[22] Differences between quantitative variables were tested for significance using analysis of variance (ANOVA), with $p < 0.05$ (two-sided) indicating statistical significance.

Results

SNP genotyping and association with overall survival

To identify germline variations that influence the survival of patients with lung adenocarcinoma, we designed a two-stage association study. In the discovery phase, we accessed genotype and clinical data for 289 patients with lung adenocarcinoma from a larger UK study on lung cancer risk (Supporting Information Table S1).[17] Multivariable Cox analysis showed that age, sex, smoking status and a clinical record of chemotherapy were not associated with survival, whereas clinical stage (Stage I *vs.* >I) and a record of radiotherapy or surgery were significantly associated. Using a Cox

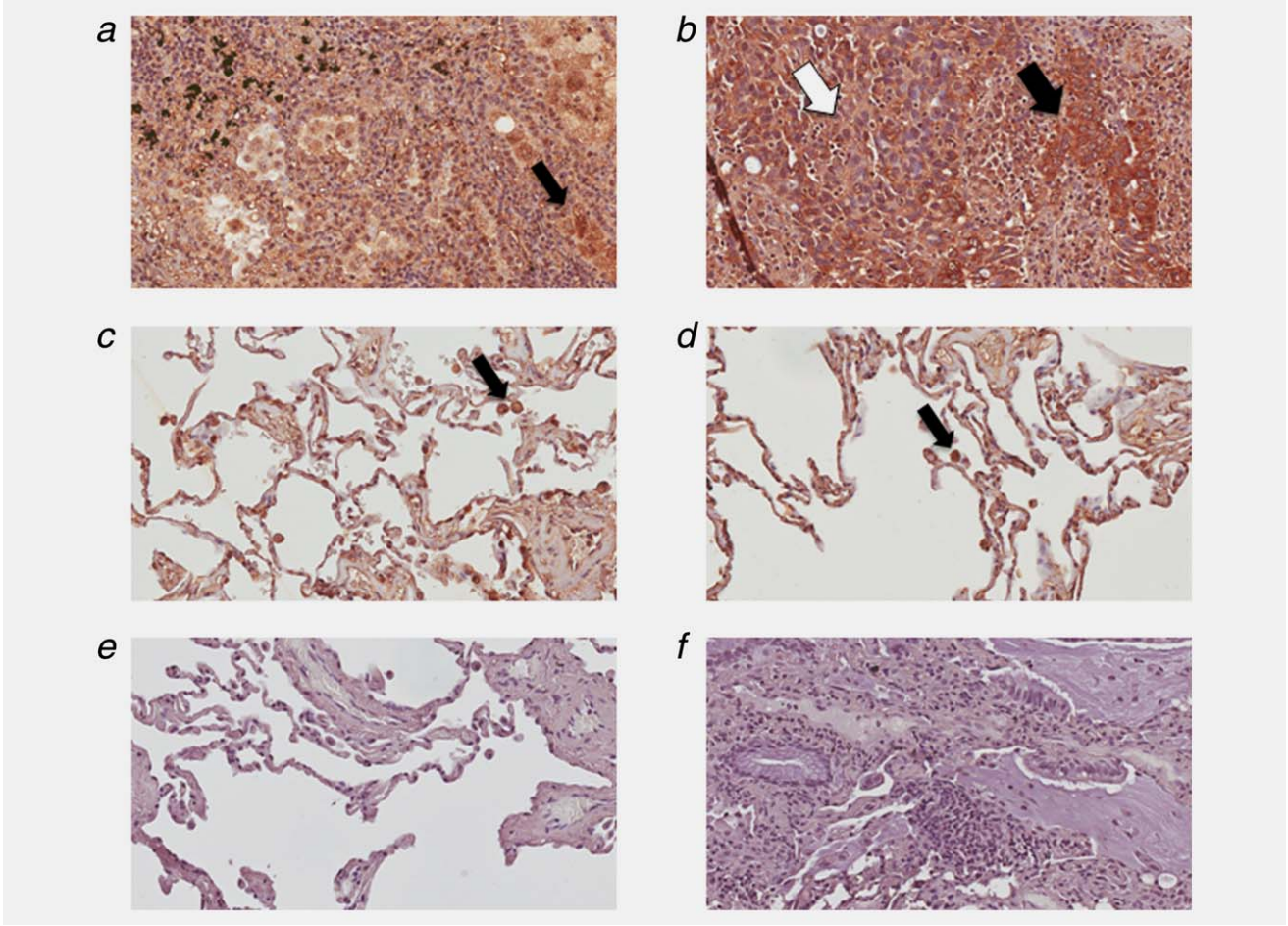


Figure 3. Panel *a*: lung adenocarcinoma. Arrow indicates macrophages, while carbon particles are visible as dark spots on the upper-left. Panel *b*: Lung adenocarcinoma from another subject where differential level of PTPRG expression is detectable: white arrow indicates low PTPRG-expressing cells while black arrow indicates high expressors. Panel *c* and *d*: non-neoplastic lung in the same histologic section from the respective patients (*c* correspond to the patient shown in *a* and panel *d* correspond to the patient shown in panel *b*). Black arrows indicate lung macrophages that are positive for PTPRG expression. Panels *e* and *f*: Total lack of staining with preimmune IgY of the same samples demonstrates the specificity of the signal detected.

proportional hazard model, we determined the association of 30,568 SNPs with survival, resulting in 202 SNPs associated at $p < 0.005$.

To confirm and further investigate these findings, we examined the association of the top 96 SNPs with survival in an Italian validation series comprising patients with lung adenocarcinoma, all of whom had been treated surgically. Genotyping was performed on customized 96.96 Dynamic Arrays. However, ten SNPs failed genotyping. Initially, DNA samples and clinical data were available for 889 patients, but during the process of filtering the raw genotype data, 152 samples (17%) were discarded since they had more than 90% of missing genotypes, leaving 748 samples (Supporting Information Table S2) for which the individual genotyping rate (average per-patient SNP call rate) was 0.96. Multivariable Cox analysis showed that smoking status was not associated with survival whereas age, sex and stage were all significantly associated.

Cox's survival analysis of the 152 samples with missing genotypes *versus* the 748 genotyped samples, after adjusting for age, sex and clinical stage, showed no significant difference in survival (data not shown; $p = 0.57$), suggesting that the exclusion of these samples did not alter the group's survival phenotype.

The validation series was similar to the discovery series in terms of age and smoking status, but there were noticeable differences in gender profile (the discovery series had a high proportion of women), clinical stage (the discovery series had fewer patients with Stage I disease), and treatment (only 33% of the discovery series had surgery). Overall survival was poorer in the discovery series ($p < 2.0 \times 10^{-16}$, log-rank test; Supporting Information Fig. S1), reflecting the higher proportion of stage >I patients in that group.

In the association analysis for the validation series where survival data were adjusted for age, sex and clinical stage, four SNPs associated with survival at $p < 0.05$ and the HR

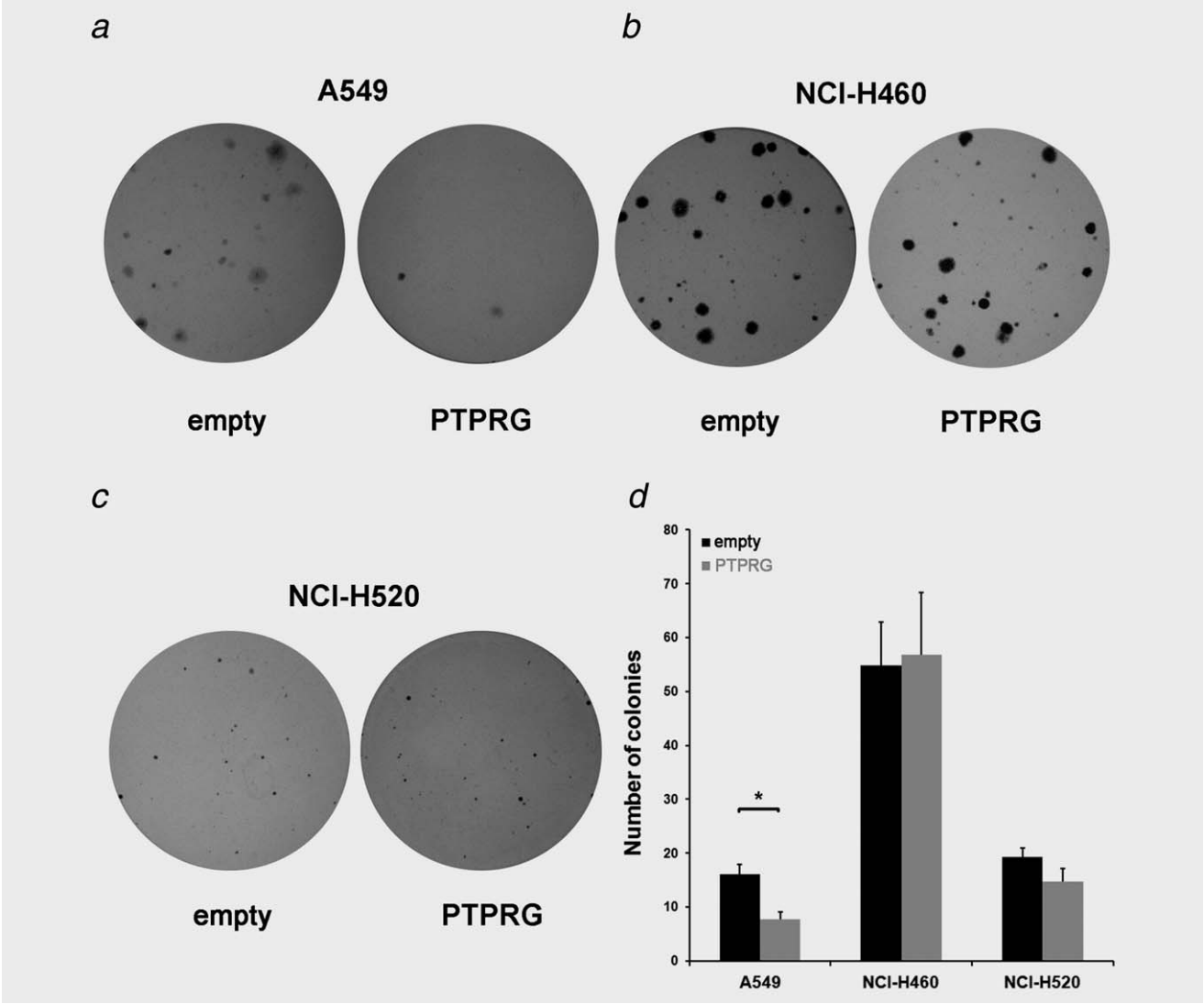
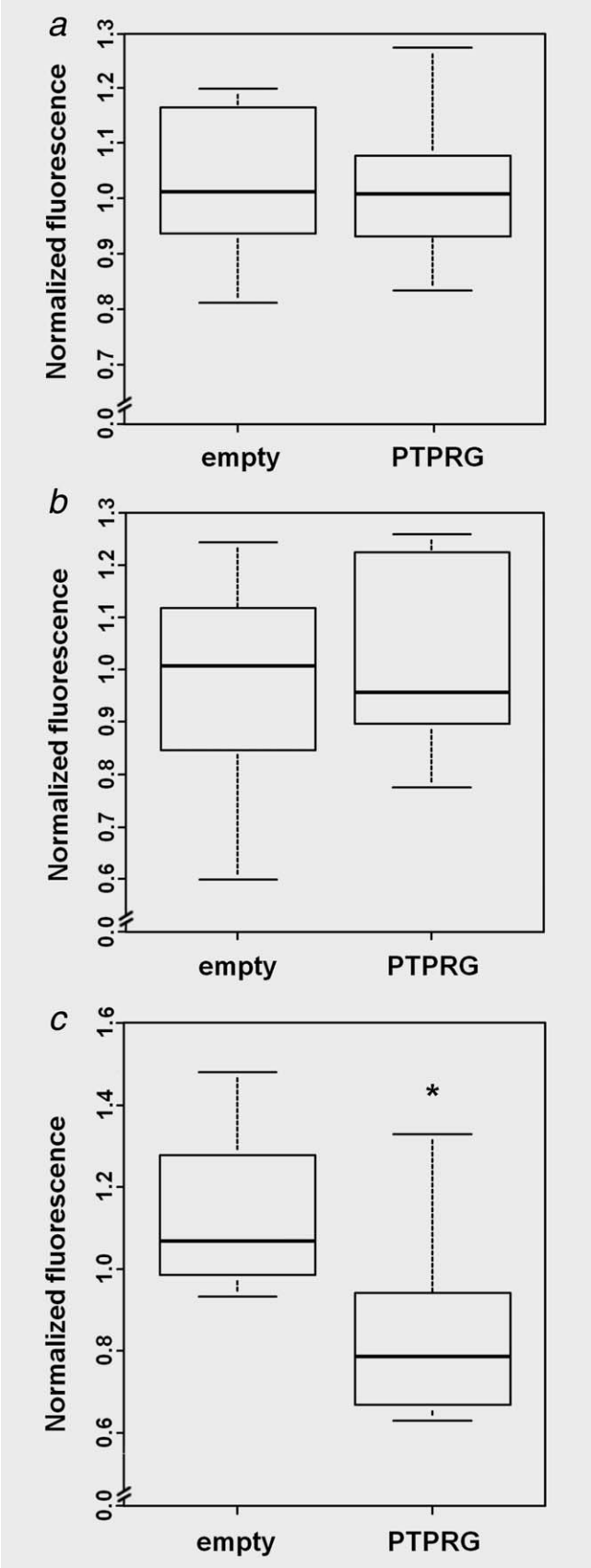


Figure 4. Influence of the *PTPRG* gene on the ability of human lung cancer cell lines to form adherent colonies. (*a*) A549 cells stably transfected with either the empty plasmid or the *PTPRG* expression plasmid. Shown are representative culture dishes stained with Giemsa. (*b*) Transfected NCI-H460 cells. (*c*) Transfected NCI-H520 cells. (*d*) Number of colonies of transfected A549 (18 replicas), NCI-H460 (9 replicas) and NCI-H520 (9 replicas) cells. Values are mean and SE. $*p < 0.05$, ANOVA.

had the same sign, indicating the same direction of effect, as it did in the UK series, and thus these markers were considered to be validated (Table 1). The SNP with the strongest association, rs2107561, mapped to intron 4 of the protein tyrosine phosphatase, receptor type, G (*PTPRG*) gene, on 3p21-p14 ($p = 0.0089$); carrier status for the minor C allele was associated with a poorer survival in both series (Fig. 1). The other three SNPs mapped to a ~56 kb region of chromosome 5q23.1, in a region containing no known transcripts; all showed evidence of LD (pairwise r^2, 0.51–0.96).

Genetic comparability of the two study series

The two study series had demographic and clinical differences that could limit our ability to identify and validate SNPs associated with survival in lung adenocarcinoma. To partially overcome these differences, in the association analyses we adjusted the Cox proportional hazards models for the variables that influenced survival. Because the patients were from different European countries, with different ethnic backgrounds, it was also important to examine their genetic comparability; we pooled the genotype data into a single dataset and analyzed the data using PCA and DAPC (Supporting Information Fig. S2). Visual inspection of the PCA scatter plot showing the proportions of variance explained by the first two principal components (Supporting Information Fig. S2A) revealed that the patients did not cluster separately according their geographic origins. A similar result was obtained with DAPC (Supporting Information Fig. S2B), where the curves for the two discriminant functions



overlapped substantially. These results suggest that the two series are not genetically distinct, justifying their use in the present study with its two-stage discovery-validation design.

Molecular and functional analyses of rs2107561 and the PTPRG gene

To begin to investigate the biological relevance of the four validated SNPs to lung adenocarcinoma, we focused on the best associated SNP, rs2107561. This marker is located in intron 4 of *PTPRG*, which encodes the receptor-type tyrosine-protein phosphatase gamma. First, we tested whether or not *PTPRG* expression is similar in noninvolved lung tissue and in lung adenocarcinoma tissue, by comparing mRNA levels in pairs of surgical specimens from 43 patients. By quantitative RT-PCR, *PTPRG* mRNA was less abundant in the tumor tissue of 36 of the 43 pairs. The mean mRNA relative quantity in lung adenocarcinoma was 40% of that in the noninvolved lung tissue counterpart ($p = 1.33 \times 10^{-6}$, ANOVA; Fig. 2). Additionally, we confirmed these expression data at protein level in lung adenocarcinoma and corresponding normal tissue sections by immunohistochemistry, using an anti-PTPRG primary chicken antibody already tested in immunofluorescence and Western blot[23,24] (Fig. 3).

PTPRG mRNA levels were also lower in mRNA from three human lung cancer cell lines (A549, NCI-H460 and NCI-H520) than in a pool of mRNA from 64 additional specimens of noninvolved lung tissues (Supporting Information Fig. S3). In particular, the mean relative quantities in A549, NCI-H460 and NCI-H520 cells were 2.5%, 7% and 10%, respectively, of that in noninvolved lung tissues.

The lower expression of *PTPRG* mRNA in lung adenocarcinoma and in three lung cancer cell lines as compared to noninvolved lung tissue motivated us to over-express this gene in these cell lines to examine its impact on tumor cell growth and aggressiveness. First, we tested the ability of transfected cells to form adherent colonies (Fig. 4). A549 cells stably transfected with the *PTPRG* expression vector formed 48% fewer colonies than those transfected with the empty vector ($p = 0.0008$, ANOVA; Figs. 4*a* and 4*d*). In contrast, stable expression of *PTPRG* had no effect on colony formation of NCI-H460 cells (Figs. 4*b* and 4*d*) or NCI-H520 cells (Figs. 4*c* and 4*d*).

We also tested whether overexpression of *PTPRG* influenced anchorage-independent growth in soft agar colony formation assays (Fig. 5). Transiently transfected A549 and NCI-H520 cells grew in soft agar to similar extents whether

Figure 5. Influence of the *PTPRG* gene on anchorage-independent colony formation. Cells transiently transfected with either the empty plasmid or the *PTPRG* expression plasmid were grown in soft agar; growth was quantified by fluorescent staining with CyQuant GR dye. (*a*) A549 cells. (*b*) NCI-H520 cells. (*c*) NCI-H460 cells. Data are relative fluorescence units normalized to the mean value of each experiment. The line within each box represents the median; upper and lower edges of each box are 75th and 25th percentiles, respectively; upper and lower bars indicate the highest and lowest values less than one interquartile range from the extremes of the box, $*p < 0.05$, ANOVA.

they were transfected with the control vector or the *PTPRG* expression vector (Figs. 5*a* and 5*b*). In contrast, NCI-H460 cells transiently expressing *PTPRG* made fewer colonies than the mock-transfected cells, as shown by the lower DNA-associated fluorescence ($p = 0.0001$, ANOVA; Fig. 5*c*). Finally, A549 and NCI-H460 cells transiently transfected with the control vector or the *PTPRG* expression vector migrated to the same extent in Transwell assays (data not shown). PTPRG protein expression after transfection was confirmed by Western blot analysis (Supporting Information Fig. S4). Overall, the results obtained with these assays suggest that *PTPRG* might be implied in anchorage-dependent or -independent growth of specific types of lung tumor cell.

Discussion

Here, we sought to examine the role of inherited genetic variation as a determinant of outcome from adenocarcinoma of the lung. Our two-stage study design enabled us to identify four SNPs potentially modulating overall survival. We identified a promising relationship between variation at 5q23.1 and outcome; interestingly, deletions of chromosome 5q are common in myelodysplastic syndrome[25] and in several types of cancer, including lung adenocarcinoma,[26] pointing on a functional role in cancer of genes mapping in this region.

Of the four validated SNPs, the strongest association was provided by rs2107561, an intron variant of *PTPRG*. The minor allele of this marker was associated with poorer survival in both the discovery and validation series (Fig. 1); such allele showed a similar frequency in both series (0.16 and 0.19 in the UK and Italian series, respectively), and the observed frequencies agreed with the reported frequency (MAF = 0.18) in the population with European ancestry (http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=rs2107561). *PTPRG is* a member of the family of protein tyrosine phosphatase receptor genes. It is ubiquitously expressed, acts as a tumor suppressor,[21] and is down-regulated in several neoplasms, including chronic myeloid leukaemia and ovarian, breast and lung cancers.[21,27,28] *PTPRG* expression, at mRNA and protein level, was lower in lung adenocarcinoma tissue specimens than in noninvolved lung tissue from the same patients; it was also relatively low in three human lung cancer cell lines. Overexpression of *PTPRG* inhibited the anchorage-dependent growth of A549 human lung carcinoma cells and the anchorage-independent growth of the NCI-H5460 human large cell lung cancer line; at the moment, we do not have any obvious explanation of the cell-specific effects exerted by *PTPRG* overexpression, although contrasting effects of candidate cancer-modulating genes overexpressed in different cell types have been reported and may perhaps be attributed to the complex signalling pathways regulating cell growth and differentiation.[29,30] Testing the effects of *PTPRG* overexpression in additional cancer cell lines and analysis of the signalling pathways involved may allow to establish the reasons underlying the cell-type effects of *PTPRG* overexpression that we have observed.

Altogether, these results support a potential role of genetic variation in *PTPRG* in determining lung cancer prognosis; however, rs2107561 does not seem to be the causal variation, since it does not map to an evolutionary conserved region of the genome or to a known regulatory region of the gene. Indeed, in a preliminary analysis in normal lung tissue,[31] we have not found a statistically significant association between rs2107561 genotype and *PTPRG* mRNA levels (not shown); therefore, it is likely that its association with survival is mediated by a causal variant yet to be identified. Possible functional candidate variants may be represented by missense variants affecting the biochemical activity of the PTPRG protein by mimicking the variation of activity associated with different expression levels of the gene, or by variants modulating the amount of specific isoforms. In the latter case, constructs can be generated to test whether alleles of these variants, including rs2107561, affect splicing of *PTPRG*.

One issue in the design of this study was the different proportion of males and females in the discovery and validation series and the larger number of clinical Stage III and IV patients in the UK discovery series, resulting in a significantly higher mortality than in the validation series. The comparison of two groups of patients with a large difference in prognosis could have confounded the identification of individual genetic variations modulating survival, since advanced diseases may be refractory to any kind of modulation.[32–34] However, despite the differences in survival between the two series in this study, they were nonetheless genetically comparable since PCA and DAPC were not able to distinguish the individuals according to their countries of origin. Although we cannot exclude the possibility that a repeat analysis with a larger number of SNPs may have identified genetic clusters according to ancestry in the two series, we think that the 86 SNPs used here suffice for a rough estimation of genetic homogeneity, since comparable numbers of SNPs (*e.g.*, 92 and 93) have been proposed to be adequate for controlling for admixture in association studies.[35,36]

In conclusion, this study provides evidence supporting the hypothesis that germline variations may influence the survival of patients with lung adenocarcinoma, as we identified four genetic variants that associated with clinical outcome in both the discovery and the validation series. The best associated SNP is in the *PTPRG* gene, for which this study provides preliminary evidence associating reduced expression levels with lung adenocarcinoma. Validation of these findings in independent case series is required to establish the robustness of the association with survival, while deciphering the biological basis of the association requires fine genetic mapping and additional functional analyses.

Acknowledgements

The authors thank Dr. Valerie Matarese for scientific editing. They would like to thank all the individuals that participated in the research[17,18] leading up to this study and the clinicians who took part in the GELCAPS consortium. The funders had no role in the design or conduct of the study, in the collection, analysis or interpretation of the data, or in the preparation, review or approval of the manuscript.

References

1. Marshall HM, Leong SC, Bowman RV, et al. The science behind the 7th edition tumour, node, metastasis staging system for lung cancer. *Respirology* 2012;17:247–60.
2. Goldstraw P, Crowley J, Chansky K, et al., International Association for the Study of Lung Cancer International Staging Committee, Participating Institutions. The IASLC lung cancer staging project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM classification of malignant tumours. *J Thorac Oncol* 2007;2:706–14.
3. Yoshizawa A, Motoi N, Riely GJ, et al. Impact of proposed IASLC/ATS/ERS classification of lung adenocarcinoma: prognostic subgroups and implications for further revision of staging based on analysis of 514 stage I cases. *Mod Pathol* 2011;24:653–64.
4. Zhang J, Wu J, Tan Q, et al. Why do pathological stage IA lung adenocarcinomas vary from prognosis?: a clinicopathologic study of 176 patients with pathological stage IA lung adenocarcinoma based on the IASLC/ATS/ERS classification. *J Thorac Oncol* 2013;8:1196–202.
5. Coate LE, John T, Tsao MS, et al. Molecular predictive and prognostic markers in non-small-cell lung cancer. *Lancet Oncol* 2009;10:1001–10.
6. Mascaux C, Iannino N, Martin B, et al. The role of RAS oncogene in survival of patients with lung cancer: a systematic review of the literature with meta-analysis. *Br J Cancer* 2005;92:131–9.
7. Meng D, Yuan M, Li X, et al. Prognostic value of K-RAS mutations in patients with non-small cell lung cancer: a systematic review with meta-analysis. *Lung Cancer* 2013;81:1–10.
8. Lee CK, Brown C, Gralla RJ, et al. Impact of EGFR inhibitor in non-small cell lung cancer on progression-free and overall survival: a meta-analysis. *J Natl Cancer Inst* 2013;105:595–605.
9. Robinson KW, Sandler AB. EGFR tyrosine kinase inhibitors: difference in efficacy and resistance. *Curr Oncol Rep* 2013;15:396–404.
10. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst* 2010;102:464–74.
11. Egan KM, Nabors LB, Olson JJ, et al. Rare TP53 genetic variant associated with glioma risk and outcome. *J Med Genet* 2012;49:420–1.
12. Lin WY, Camp NJ, Cannon-Albright LA, et al. A role for XRCC2 gene polymorphisms in breast cancer risk and survival. *J Med Genet* 2011;48:477–84.
13. Wu X, Wang L, Ye Y, et al. Genome-wide association study of genetic predictors of overall survival for non-small cell lung cancer in never smokers. *Cancer Res* 2013;73:4028–38.
14. 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;34:307–13.
15. Blanchon F, Grivaux M, Asselain B, et al. 4-year mortality in patients with non-small-cell lung cancer: development and validation of a prognostic index. *Lancet Oncol* 2006;7:829–36.
16. Kogure Y, Ando M, Saka H, et al. Histology and smoking status predict survival of patients with advanced non-small-cell lung cancer. results of west japan oncology group (WJOG) study 3906L. *J Thorac Oncol* 2013;8:753–8.
17. Broderick P, Wang Y, Vijayakrishnan J, et al. Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res* 2009;69:6633–41.
18. Eisen T, Matakidou A, Houlston R, GELCAPS Consortium. Identification of low penetrance alleles for lung cancer: the GEnetic lung CAncer predisposition study (GELCAPS). *BMC Cancer* 2008;8:244, 2407–8–244.
19. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* 2010;11:94, 2156–11–94.
20. Jombart T, Ahmed I. Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 2011;27:3070–1.
21. Della Peruta M, Martinelli G, Moratti E, et al. Protein tyrosine phosphatase receptor type {gamma} is a functional tumor suppressor gene specifically downregulated in chronic myeloid leukemia. *Cancer Res* 2010;70:8896–906.
22. 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;81:559–75.
23. Lorenzetto E, Moratti E, Vezzalini M, et al. Distribution of different isoforms of receptor protein tyrosine phosphatase gamma (ptprg-RPTP gamma) in adult mouse brain: upregulation during neuroinflammation. *Brain Struct Funct* 2014;219:875–90.
24. Mafficini A, Vezzalini M, Zamai L, et al. Protein tyrosine phosphatase gamma (PTPgamma) is a novel leukocyte marker highly expressed by CD34 precursors. *Biomark Insights* 2007;2:218–25.
25. Ebert BL. Molecular dissection of the 5q deletion in myelodysplastic syndrome. *Semin Oncol* 2011;38:621–6.
26. Petersen I, Bujard M, Petersen S, et al. Patterns of chromosomal imbalances in adenocarcinoma and squamous cell carcinoma of the lung. *Cancer Res* 1997;57:2331–5.
27. LaForgia S, Morse B, Levy J, et al. Receptor protein-tyrosine phosphatase gamma is a candidate tumor suppressor gene at human chromosome region 3p21. *Proc Natl Acad Sci U S A* 1991;88:5036–40.
28. Vezzalini M, Mombello A, Menestrina F, et al. Expression of transmembrane protein tyrosine phosphatase gamma (PTPgamma) in normal and neoplastic human tissues. *Histopathology* 2007;50:615–28.
29. Choi PM, Tchou-Wong KM, Weinstein IB. Overexpression of protein kinase C in HT29 colon cancer cells causes growth inhibition and tumor suppression. *Mol Cell Biol* 1990;10:4650–7.
30. Housey GM, Johnson MD, Hsiao WL, et al. Overproduction of protein kinase C causes disordered growth control in rat fibroblasts. *Cell* 1988;52:343–54.
31. 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;34:2767–73.
32. Bidard FC, Pierga JY, Soria JC, et al. Translating metastasis-related biomarkers to the clinic—progress and pitfalls. *Nat Rev Clin Oncol* 2013;10:169–79.
33. Laird BJ, Kaasa S, McMillan DC, et al. Prognostic factors in patients with advanced cancer: a comparison of clinicopathological factors and the development of an inflammation-based prognostic system. *Clin Cancer Res* 2013;19:5456–64.
34. Sylvester RJ. Combining a molecular profile with a clinical and pathological profile: biostatistical considerations. *Scand J Urol Nephrol Suppl* 2008;218:185–90. doi:185-90.
35. Pakstis AJ, Speed WC, Fang R, et al. SNPs for a universal individual identification panel. *Hum Genet* 2010;127:315–24.
36. Nassir R, Kosoy R, Tian C, et al. An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genet* 2009;10:39, 2156–10–39.