Variation rs2235503 C>A within the promoter of *MSLN* affects transcriptional rate of mesothelin and plasmatic levels of the soluble mesothelin-related peptide

Roberto Silvestri[1], Perla Pucci[2], Chiara De Santi[3], Irene Dell'Anno[1], Simona Miglietta[4], Alda Corrado[5], Vanessa Nicolì[1], Daniela Marolda[1], Monica Cipollini[1], Enrica Pellegrino[1], Monica Evangelista[6], Alessandra Bonotti[7], Rudy Foddis[1], Alfonso Cristaudo[1], Stefano Landi[1,1], and Federica Gemignani[1]

[1]University of Pisa
[2]The Open University
[3]Royal College of Surgeons in Ireland
[4]IRCCS San Raffaele Scientific Institute
[5]University of Milan
[6]CNR di Pisa
[7]Pisa University Hospital

April 29, 2020

Abstract

Soluble mesothelin-related peptide (SMRP) is a promising biomarker for malignant pleural mesothelioma (MPM), but several confounding factors can reduce SMRP-based tests accuracy. The identification of these confounders could improve the diagnostic performance of SMRP. In this study, we evaluated the sequence of 1000 base pairs encompassing the minimal promoter region of *MSLN* gene to identify expression quantitative trait loci (eQTL) that can affect SMRP. We assessed the association between four *MSLN* promoter variants and SMRP levels in a cohort of 72 MPM and 677 non-MPM subjects, and we carried out in vitro assays to investigate their functional role. Our results show that rs2235503 is an eQTL for *MSLN* associated with increased levels of SMRP in non-MPM subjects. Furthermore, we show that this polymorphic site affects the accuracy of SMRP, highlighting the importance of evaluating the individual's genetic background and giving novel insights to refine SMRP specificity as a diagnostic biomarker.

Introduction

Malignant pleural mesothelioma (MPM) is a highly aggressive and rare cancer of the pleura triggered by exposure to asbestos and associated with a poor prognosis (Bianchi et al., 2017). The long latency between the exposure to asbestos and the onset of the disease, the unspecific symptoms at the presentation, and the lack of accurate and non-invasive diagnostic tools make the diagnosis quite challenging (Bianco, Valente, De Rimini, Sica, & Fiorelli, 2018; Rusch et al., 2012). Thus, MPM is often diagnosed at advanced stages, thereby further reducing the limited therapeutic options available and resulting in poor prognostic outcomes. The identification of accurate diagnostic biomarkers, or the optimization of those identified so far, could help to overcome this problem (Sun, Vaynblat, & Pass, 2017). To date, one of the most promising biomarkers for the diagnosis of MPM is “soluble mesothelin-related peptide” (SMRP), which is released in the extracellular environment following the proteolytic cleavage of mesothelin (a membrane protein encoded by the *MSLN* gene) (Sapede et al., 2008). The plasmatic concentration of SMRP is easily measurable from blood samples, and it is usually higher in MPM than in non-MPM subjects (either healthy or affected by other respiratory conditions), allowing fair discrimination between these two groups (Gao et al., 2019; Gillezeau et al., 2019). Despite these promising features, several confounding factors can reduce the accuracy of SMRP as a diagnostic biomarker, thereby preventing its employment in the clinical practice. While many studies extensively documented the impact of age, body mass index, glomerular filtration rate (Casjens et al., 2017), and tumor histology on SMRP (Scherpereel et al., 2006), none of them explored the role of the genetic background. However, we believe that the genetic background could play a prominent role in this context. An increasing number of studies, by showing the importance of genetic variants in altering the concentration and the accuracy of several biomarkers, strongly support this idea (Cramer et al., 2003; Enroth, Johansson, Enroth, & Gyllensten, 2014; Gudmundsson et al., 2010; Ruggiero et al., 2015; Wang et al., 2018). Furthermore, we showed for the first time that single nucleotide polymorphisms (SNPs) located within the 3'-UTR of the*MSLN* gene could affect the SMRP levels through a miRNA-mediated mechanism (Garritano et al., 2014). Recently, we reported similar results for other variants lying within the *MSLN* promoter (De Santi et al., 2017). Based on these observations, we hypothesized that expression quantitative trait loci (eQTL) for *MSLN* could affect the plasmatic concentration and the accuracy of SMRP. Thus, we assessed the association between SNPs located within the promoter region of*MSLN* and the plasmatic levels of SMRP in a cohort of 72 MPM and 677 non-MPM subjects. We carried out *in vitro* assays to investigate the functional role of these SNPs, and highlighted the relationship between the transcriptional rate of *MSLN* gene and serum concentrations of SMRP. Our results show that evaluating the individual's genotype to establish personalized cut-off values can increase the diagnostic accuracy of SMRP, thereby helping to promote its translation into clinical practice.

Materials and methods

Editorial Policies and Ethical Consideration

This study was approved by the institutional ethical committee of the University Hospital of Pisa. All subjects gave written informed consent.

Selection of the SNPs

Following the hypothesis that the transcriptional rate of the*MSLN* gene could affect the plasmatic concentration of SMRP, we extracted a window of 1000 base pairs upstream of the transcription starting site (TSS), encompassing the minimal promoter region, using the UCSC genome browser (*https://genome.ucsc.edu/*). Within this sequence, we selected those SNPs showing a minor allele frequency (MAF) > 0.05 and a significant association with*MSLN* mRNA levels in lung tissues (according to the GTEX Portal;*https://gtexportal.org/*). This selection led to the identification of four SNPs worth of further investigations: rs3764247 A>C (NC_000016.10:g.760039A>C), rs3764246 A>G (NC_000016.10:g.760143A>G), rs2235503 C>A (NC_000016.10:g.760593C>A) and rs2235504 A>G (NC_000016.10:g.760655A>G).

According to 1000 Genomes (*https://www.internationalgenome.org/1000-genomes-browsers/*), their combination elicits four haplotypes (Table 1), accounting for almost the totality (98.5%) of the genetic variability of the promoter region within Caucasians, and Tuscans.

Genotyping, haplotype reconstruction, SMRP measurements, and association study

Detailed information about the studied population, DNA extraction, genotyping, and measurement of SMRP levels, can be found elsewhere (De Santi et al., 2017). Briefly, we analyzed a total of 677 non-MPM and 72 MPM volunteers. Among the 677 non-MPM, 371 were healthy people, while 306 were patients affected by benign respiratory diseases (BRD). All the subjects were recruited at the University Hospital of Pisa as part of an occupational surveillance program on workers previously exposed to asbestos, as described by *Garritano et al.* (Garritano et al., 2014). Blood and serum samples were obtained by venipuncture and store at -80°C. To measure the plasmatic concentration of the SMRP, we employed the Mesomark enzyme-linked immunosorbent assay (Fujirebio Diagnostics, Japan) following the manufacturer's instructions. We carried out DNA extraction and genotyping using the EuroGOLD Blood DNA Mini Kit (EuroClone, Pero, Italy) and the KASPar PCR SNP genotyping system (LGC Genomics Ltd, Teìddington, Middlesex, UK) respectively. For this study, we reconstructed the individual haplotypes and diplotypes using PHASE 2.1.1 (M Stephens, Smith, & Donnelly, 2001; Matthew Stephens & Scheet, 2005) and stratified our cohort according to health status and diplotypes. To precisely assess the effect of each haplotype, we restricted the association study to carriers of the haplotype #1 (e.g., H1H2, H1H3, and H1H4), using the H1H1 homozygotes as reference. The number of subjects excluded from the analysis was minimal (19 non-MPM and 6 MPM) and did not hamper the statistical power of the study.

Cell lines

For the functional study, we employed a non-malignant SV40-immortalized epithelial cell line (MeT-5A) and an epithelioid malignant mesothelioma cell line (Mero-14). MeT-5A were purchased from ATCC (American Type Culture Collection, CRL-9444) and cultured in Medium-199 (Gibco, Life Technologies, Monza, Italy) supplemented with 10% fetal bovine serum (FBS), 1% penicillin/streptomycin, 3nM epidermal growth factor (EGF), 400nM hydrocortisone and 870nM insulin. Mero-14 were kindly donated by Istituto tumori of Genova (National Research Council, Genova, Italy) and cultured in Dulbecco's modified Eagle's medium (DMEM; EuroClone, Pero, Italy) supplemented with 10% FBS and 1% penicillin/streptomycin. Both cell lines were maintained at 37°C and 5% CO_2.

Construction of plasmids

In the first part of this work, we aimed to evaluate the role of the four haplotypes herein identified in affecting the activity of the*MSLN* promoter. Thus, we constructed four vectors to employ in the functional study. Each of these vectors harbored one of the haplotypic variants immediately upstream of a green fluorescent protein (GFP) coding sequence. The same vectors also harbored a red fluorescent protein (RFP) coding sequence controlled by the constitutive eukaryotic promoter EF1 (Figure 1A). For the cloning procedure, we employed the CloneEZ PCR cloning kit (GenScript, Piscataway, U.S.A.). We used the HR220PA-1 vector (System Bioscience, Palo Alto, U.S.A.) linearized with BstBI (New England BioLabs, Ipswich, U.S.A.) and DNA fragments representing the four haplotypic variants of the *MSLN* promoter. Each of these fragments was obtained by PCR using as template the genomic DNA of four subjects homozygote for one of the four selected haplotypes (now on abbreviated as H1-H4 when referring to the human genomic DNA, or HAP1-HAP4 when referring to the cloned fragment). The resulting vectors were named HR_HAP1 to HR_HAP4. In the second part of this work, we aimed to ascertain the individual role of each of the selected SNPs. To this end, we created three additional vectors, each harboring the uncommon variant of only one of the four SNPs. These vectors were named HR_246, HR_503, and HR_504 and were obtained from HR_HAP1 by site-directed mutagenesis using the Quick Change Lightning Site-Directed Mutagenesis Kit (Agilent). Consistently with the terminology employed in this second phase of the study, since the HR_HAP4 differs from the HR_HAP1 only for the rs3764247, bearing the C-allele instead of the A-allele, we renamed it "HR_247". Figure 1A shows the details of the changes introduced by the site-directed mutagenesis to generate these additional vectors.

Fluorescent reporter assay

To evaluate the effect of the selected haplotypes and SNPs on the transcriptional activity of the *MSLN* promoter, we employed a fluorescent reporter assay. To this end, 2x10^5cells/ml were electroporated with 10µg of plasmid DNA, seeded into a six-well plate, and incubated at 37°C and 5% CO_2 for 72hours. After the incubation, we harvested the cells by trypsinization and measured the GFP/RFP fluorescence intensity using the BD FACSJazz System (BD Biosciences, Franklin Lakes, U.S.A.). We used the untreated sample to set the RFP intensity threshold for the selection of efficiently transfected cells (Figure 1B). Similarly, we used cells transfected with the "empty" HR220PA-1 vector (expressing the RFP but not the GFP reporter gene) to establish the average background signal detected by the FITC/GFP channel (Figure 1C). These two steps allowed us to restrict the statistical analysis only to the cells that were efficiently transfected, and that showed a GFP signal above the average background (Figure 1D; P3). All the electroporation steps were carried out using the Neon Transfection System (Thermo Fisher Scientific, Monza, Italy), and the following parameters: 1230V, 30ms, 2pulses for MeT-5A and 1130V, 30ms, 2pulses for Mero-14.

Statistical analyses

For the statistical analyses, we employed GraphPad PRISM 7.0 software. We carried out the analysis of variance (ANOVA) for the association study and the multifactor ANOVA (mANOVA) for the functional studies, both followed by Dunnett's multiple comparison test. To evaluate the diagnostic performance of SMRP along with the optimal cut-off values, we calculated the ROC curves employing the same software.

Results

Haplotype #2 and haplotype #3 are associated with SMRP levels*in vivo*.

Among the non-MPM volunteers, we found a statistically significant difference between the plasmatic levels of SMRP in heterozygotes H1H2, H1H3 and H1H4 compared with those of the reference group H1H1 (p-value ANOVA <10^{-5}). The Dunnett's post-test showed a statistically significant difference between carriers of haplotype #2 or haplotype #3, but not haplotype #4, and the reference H1H1 (p-value <10^{-5} and 0.0047 respectively) (Figure. 2A). On average, carriers of haplotype #2 showed the highest SMRP level (average ± standard error: 1.30 ± 0.046 nM) followed by carriers of haplotype #3, #1 and #4 (0.98 ± 0.060 nM, 0.80 ± 0.022 nM and 0.78 ± 0.047 respectively). We did not observe any significant difference within the group of MPM patients (Figure. 2B).

In vitro studies confirmed the functional role of haplotype #2 and identify rs2235503 C>A as the most likely causative SNP

To evaluate whether the haplotype #2 and #3 could enhance the activity of the *MSLN* promoter, we carried out *in vitro* experiments using a fluorescence reporter assay. We transfected MeT-5A and Mero-14 cells with HR vectors carrying the GFP reporter gene under the control of the different haplotypic variants of *MSLN* promoter, named HR_HAP1, HR_HAP2, HR_HAP3 and HR_HAP4 (Figure 1A). Results showed a statistically significant difference among haplotypes but not between cell lines (mANOVA p-value=0.0016 and 0.67). When compared to cells transfected with HR_HAP1, cells transfected with HR_HAP2 showed a 1.41-fold increased RFU (± 0.267; p-value = 6x10^{-4}). HR_HAP3 conferred a slight increased expression but not statistically significant (1.07 ± 0.131; p-value = 0.45). HR_HAP4 showed a promoter activity similar to HR_HAP1 (0.97 ± 0.05; p-value = 0.98) (Figure. 2C). Since haplotype #2 bears the uncommon variants of the four SNPs, we investigated the functional role of each of them, by employing four more constructs: HR_247, HR_246, HR_503, and HR_504 (Figure 1A). Each of these vectors bore the uncommon variant of rs3764247, rs3764246, rs2235503 or rs2235504, thus differing from HR_HAP1 only for one SNP. Again, mANOVA showed a statistically significant difference among genotypes but not between cell lines (p-value<10^{-5} and 0.15). When compared with HR_HAP1, only cells transfected with HR_503 showed significant RFU increase (fold change 1.35 ± 0.164; p-value<10^{-5}) (Figure. 2D).

Rs2235503 C>A significantly affects the performance of SMRP as diagnostic biomarker for MPM

Since the results indicated that the uncommon variant A-rs2235503 caused an increase of *MSLN* expression and SMRP levels in non-MPM subjects, we sought to determine whether this effect could significantly affect the accuracy of SMRP as a diagnostic biomarker for MPM. To verify this aspect, we stratified our cohort in two subgroups, the former containing the carriers of A-rs2235503 (rs2235503_C/A + A/A) and the latter containing all the other subjects (rs2235503_C/C). We then used the SMRP levels to calculate the "Receiver Operating Characteristic" (ROC) curve for each of the two subgroups. Results suggested a strong influence of the genotype on the performance of SMRP. In fact, the ROC curve for the C/A + A/A group resulted in an area under the curve (AUC) of only 0.798 ± 0.055 compared with an AUC of 0.915 ± 0.018 for the C/C group (Figure. 3A). Interestingly, the AUC for the C/C group was also higher than that for the overall population

(0.877 ± 0.019). Moreover, the Yuden index pointed at an optimal cut-off value of 1.118 (sensitivity = 84.31, specificity = 82.86) for the C/C group, and 3.092 (sensitivity = 52.58, specificity = 96.69) for the C/A + A/A group. The optimal cut-off value for the overall population was 1.28, with a sensitivity of 77.78 and a specificity of 79.91. Notably, using a cut-off of 1.28 nM, the specificity for the C/C group raised to 88.71, but the sensitivity dropped to 76.47. Similarly, the specificity for the C/A + AA group raised to 55.80, but the sensitivity dropped to 80.95. Supplementary table 1, 2 and 3 report the complete list of the possible cut-off values along with the associated sensitivity and specificity for each group. These results were not surprising as the difference in the average SMRP concentration between non-MPM and MPM subjects, although still significant (p-value < 0.0001), dropped from 2.734 ± 0.174 nM for the C/C group to only 1.706 ± 0.388 nM for the A/A group (Figure. 3B). Intriguingly, data from GTEx Portal showed a similar trend for the tagging SNP rs12597489 C>T, in linkage disequilibrium (LD) with rs2235503 ($r^2 = 0.8$ according to HaploReg v4.1), suggesting a strong relationship between the mRNA levels of *MSLN* and the serum concentration of SMRP (Figure. 3C).

Discussion

SMRP is one of the most promising biomarkers for MPM, but its sensitivity and specificity are suboptimal for the use in the clinical practice (Casjens et al., 2017; Scherpereel et al., 2006). Our previous data showed that SNPs within regulatory regions of *MSLN* gene were associated with SMRP levels, raising the question of whether the evaluation of these variants could improve the performance of SMRP in the diagnosis of MPM (De Santi et al., 2017; Garritano et al., 2014). Herein, we explored the possibility of exploiting eQTL for *MSLN*to increase the diagnostic accuracy of SMRP. Firstly, we carried out an association study in a cohort of 677 non-MPM subjects and 72 MPM patients. We found that two haplotypes (#2 and #3) of the minimal promoter region of *MSLN* gene were associated with increased serum levels of SMRP in non-MPM subjects. In agreement with our previous studies, we did not observe any association in the cohort of MPM patients (De Santi et al., 2017; Garritano et al., 2014). This could be ascribed to a low statistical power of this sample setting or to other factors affecting SMRP in the context of the malignant transformation (Pass et al., 2008). Nonetheless, the *in vitro* assays confirmed the functional role of haplotype #2, able to increase the expression of a reporter gene in a chimerized vector of about 42%. Moreover, cells transfected with a vector harboring the A-rs2235503 showed increased promoter activity of a similar extent, compared to C-rs2235503. This result straightened the role of haplotype #2 and pinpointed rs2235503 as the SNP responsible for the enhancement of *MSLN* gene expression. Since SMRP derives from the proteolytic cleavage of mesothelin (Sapede et al., 2008), it is conceivable that its plasmatic levels can be influenced by factors such as the eQTL for *MSLN* . Our work supports the rationale that the polymorphism rs2235503 is among these factors. Conversely, our *in vitro* studies ruled out the possibility of a functional role for the SNPs that characterize the haplotype #3 (i.e. rs3764246 and rs2235504). Thus, likely the association observed *in vivo* should be ascribed to other SNPs in LD with haplotype #3 but residing outside the 1000-bps promoter region herein considered. Then, we assessed whether the performance of SMRP could be improved when taking into account the subjects' genetic background, in agreement with previous findings for other biomarkers (Cramer et al., 2003; Enroth et al., 2014; Gudmundsson et al., 2010; Ruggiero et al., 2015; Wang et al., 2018). Notably, we found that the AUC for individuals carrying the rs2235503-C/C genotype was higher (0.915) than that of carriers of the A-allele (0.798) and higher as compared to the whole population (0.877). Moreover, the Youden index pointed at significantly different cut-off values according to the rs2235503 genotypes, suggesting that the stratification of subjects based on this SNP could improve the accuracy of SMRP. In conclusion, our work shows that rs2235503 affects *MSLN* gene transcription and represents an eQTL for SMRP levels. Moreover, we highlighted for the first time that the rs2235503 can affect the diagnostic performance of SMRP, reinforcing the importance of considering the individual genetic background to improve the accuracy of cancer biomarkers (Cramer et al., 2003; Enroth et al., 2014; Gudmundsson et al., 2010; Wang et al., 2018). Therefore, it is conceivable that the characterization of further eQTL could help translating SMRP into the clinical practice and improve the efficiency of this biomarker.

CONFLICT OF INTERESTS

The authors declared no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Bianchi, C., Bianchi, T., Pass, H. I., Wali, A., Tang, N., Ivanova, A., … Shi, H. H.-Z. (2017). Global mesothelioma epidemic: Trend and features. *Journal of Thoracic Oncology : Official Publication of the International Association for the Study of Lung Cancer* ,*9* (11), 82–88. https://doi.org/10.1016/j.athoracsur.2007.07.042

Bianco, A., Valente, T., De Rimini, M. L., Sica, G., & Fiorelli, A. (2018). Clinical diagnosis of malignant pleural mesothelioma.*Journal of Thoracic Disease* , *10* (Suppl 2), S253–S261. https://doi.org/10.21037/jtd.2017.10.09

Casjens, S., Weber, D. G., Johnen, G., Raiko, I., Taeger, D., Meinig, C., … Pesch, B. (2017). Assessment of potential predictors of calretinin and mesothelin to improve the diagnostic performance to detect malignant mesothelioma: results from a population-based cohort study. *BMJ Open* , *7* (10), e017104. https://doi.org/10.1136/bmjopen-2017-017104

Cramer, S. D., Chang, B.-L., Rao, A., Hawkins, G. A., Zheng, S. L., Wade, W. N., … Xu, J. (2003). Association between genetic polymorphisms in the prostate-specific antigen gene promoter and serum prostate-specific antigen levels. *Journal of the National Cancer Institute* , *95* (14), 1044–1053. https://doi.org/10.1093/jnci/95.14.1044

De Santi, C., Pucci, P., Bonotti, A., Melaiu, O., Cipollini, M., Silvestri, R., … Landi, S. (2017). Mesothelin promoter variants are associated with increased soluble mesothelin-related peptide levels in asbestos-exposed individuals. *Occupational and Environmental Medicine* , *74* (6), 456–463. https://doi.org/10.1136/oemed-2016-104024

Enroth, S., Johansson, A., Enroth, S. B., & Gyllensten, U. (2014). Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nature Communications* , *5* , 4684. https://doi.org/10.1038/ncomm

Gao, R., Wang, F., Wang, Z., Wu, Y., Xu, L., Qin, Y., … Tong, Z. (2019). Diagnostic value of soluble mesothelin-related peptides in pleural effusion for malignant pleural mesothelioma: An updated meta-analysis. *Medicine* , *98* (14), e14979. https://doi.org/10.1097/MD.0000000000014979

Garritano, S., De Santi, C., Silvestri, R., Melaiu, O., Cipollini, M., Barone, E., … Landi, S. (2014). A common polymorphism within MSLN affects miR-611 binding site and soluble mesothelin levels in healthy people. *Journal of Thoracic Oncology : Official Publication of the International Association for the Study of Lung Cancer* ,*9* (11), 1662–1668. https://doi.org/10.1097/JTO.0000000000000322

Gillezeau, C., van Gerwen, M., Ramos, J., Liu, B., Flores, R., & Taioli, E. (2019). Biomarkers for malignant pleural mesothelioma: a meta-analysis. *Carcinogenesis* . https://doi.org/10.1093/carcin/bgz103

Gudmundsson, J., Besenbacher, S., Sulem, P., Gudbjartsson, D. F., Olafsson, I., Arinbjarnarson, S., … Stefansson, K. (2010). Genetic correction of PSA values using sequence variants associated with PSA levels. *Science Translational Medicine* , *2* (62), 62ra92. https://doi.org/10.1126/scitranslmed.3001513

Pass, H. I., Wali, A., Tang, N., Ivanova, A., Ivanov, S., Harbut, M., … Allard, J. (2008). Soluble mesothelin-related peptide level elevation in mesothelioma serum and pleural effusions. *The Annals of Thoracic Surgery* , *85* (1), 265–272; discussion 272. https://doi.org/10.1016/j.athoracsur.2007.07.042

Ruggiero, D., Nappo, S., Nutile, T., Sorice, R., Talotta, F., Giorgio, E., … Ciullo, M. (2015). Genetic variants modulating CRIPTO serum levels identified by genome-wide association study in Cilento isolates.*PLoS Genetics* , *11* (1), e1004976. https://doi.org/10.1371/journal.pgen.1004976

Rusch, V. W., Giroux, D., Kennedy, C., Ruffini, E., Cangir, A. K., Rice, D., … van Meerbeeck, J. P. (2012). Initial analysis of the international association for the study of lung cancer mesothelioma database. *Journal* *of Thoracic Oncology : Official Publication of the International Association for the Study of Lung Cancer* ,*7* (11), 1631–1639. https://doi.org/10.1097/JTO.0b013e31826915f1

Sapede, C., Gauvrit, A., Barbieux, I., Padieu, M., Cellerin, L., Sagan, C., . . . Gregoire, M. (2008). Aberrant splicing and protease involvement in mesothelin release from epithelioid mesothelioma cells.*Cancer Science* , *99* (3), 590–594. https://doi.org/10.1111/j.1349-7006.2007.00715.x

Scherpereel, A., Grigoriu, B., Conti, M., Gey, T., Gregoire, M., Copin, M.-C., . . . Lassalle, P. (2006). Soluble mesothelin-related peptides in the diagnosis of malignant pleural mesothelioma. *American Journal of Respiratory and Critical Care Medicine* , *173* (10), 1155–1160. https://doi.org/10.1164/rccm.200511-1789OC

Stephens, M, Smith, N. J., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* , *68* (4), 978–989. https://doi.org/10.1086/319501

Stephens, Matthew, & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation.*American Journal of Human Genetics* , *76* (3), 449–462. https://doi.org/10.1086/428594

Sun, H. H., Vaynblat, A., & Pass, H. I. (2017). Diagnosis and prognosis-review of biomarkers for mesothelioma. *Annals of Translational Medicine* , *5* (11), 244. https://doi.org/10.21037/atm.2017.06.60

Wang, K., Bai, Y., Chen, S., Huang, J., Yuan, J., Chen, W., . . . Wei, S. (2018). Genetic correction improves prediction efficiency of serum tumor biomarkers on digestive cancer risk in the elderly Chinese cohort study. *Oncotarget* , *9* (7), 7389–7397. https://doi.org/10.18632/oncotarget.23205

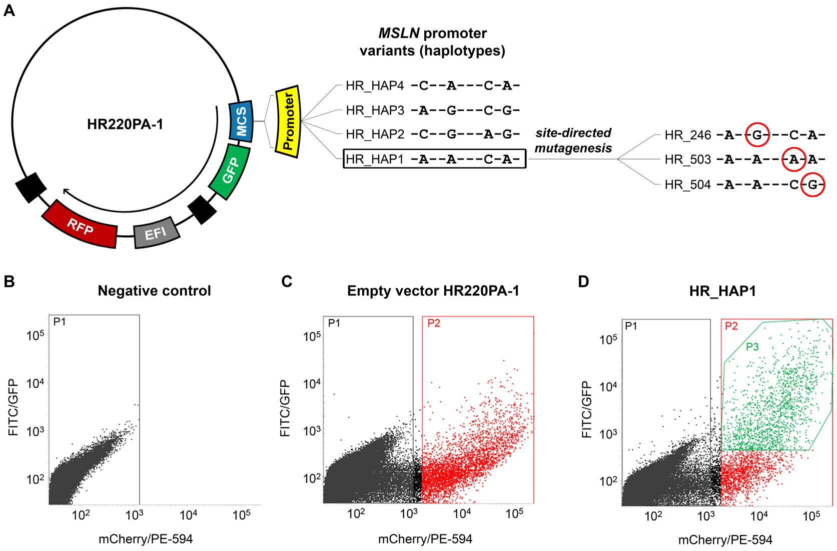
Figure Legends

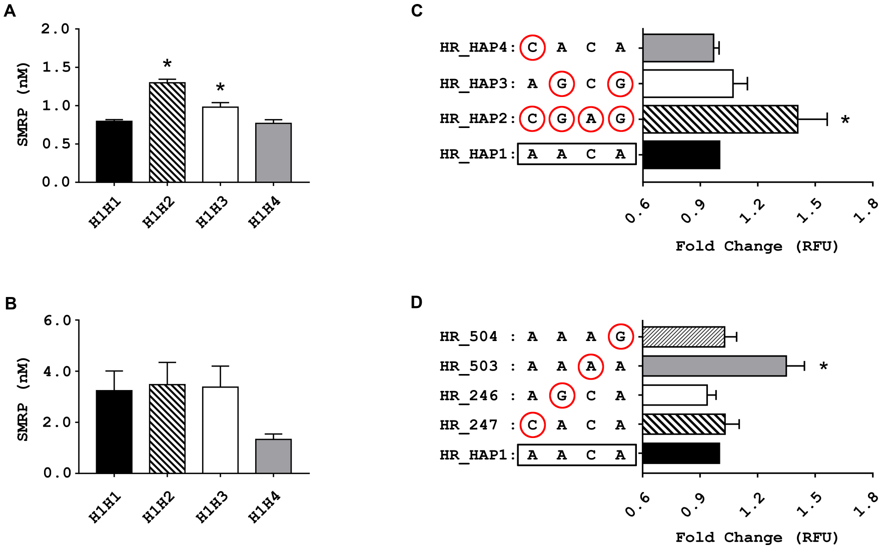
Figure 1: (A) scheme of the vectors harboring the four haplotypic variants of MLSN promoter (HR_HAP1 to HR_HAP4) upstream of a GFP reporter gene employed in the functional study. The same vectors also harbor an internal control represented by an RFP reporter gene under the control of the elongation factor 1 (EF1) constitutive promoter. HR_HAP1 has been used as a template for a site-directed mutagenesis procedure, to obtain the vectors HR_246/503/504 that have been employed to assess the effect of the single SNPs in altering the GFP expression. The SNPs variants are circled in red. **(B, C, D)**Graphs showing the fluorescence intensity, at single cell level, measured using FITC/GFP (GFP fluorescence) and mCherry/PE-594 (RFP fluorescence) filters with BD FACSJazz System. **(B)**Untransfected cells (P1) have been used as a control to establish the RFP intensity threshold for the successfully transfected cells.**(C)** Cells successfully transfected (P2) with the empty vector HR220PA-1, expressing only the RFP reporter gene, have been used to determine the average GFP background level. **(D)** The statistical analysis has been restricted to successfully transfected cells that showed a GFP signal above the average GFP background (P3) (HR_HAP1 is reported as an example).

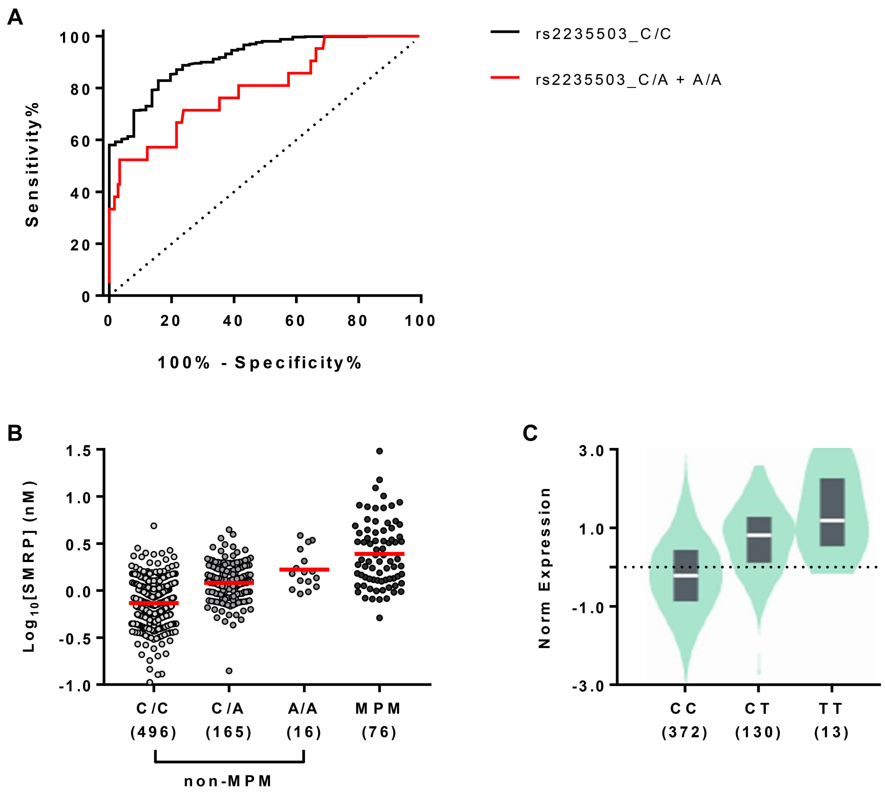
Figure 2: (A, B) bar charts showing the association between the genetic variants of MSLN promoter and the levels of soluble mesothelin-related peptide (SMRP) measured in blood samples from non-MPM subjects **(A)** and form patients affected by pleural mesothelioma (MPM) **(B)** stratified according to their individual diplotypes. The bar charts show the concentration of SMRP expressed in nM, along with the standard error of the mean (SEM). (*p<0.05, one-way ANOVA and Dunnett's post-test). **(C, D)** bar charts showing the relative fluorescence units (RFU), along with SEM, of cells 72h after being electroporated with vectors harboring the four haplotypic variants of MSLN promoter **(C)** or the rare variant for each of the considered SNPs (circled) **(D)** upstream of a GFP reporter. Since no statistically significant difference between the RFU of the two cell lines emerged from the mANOVA, the data from MeT-5A and Mero-14 have been combined. (*p<0.05, one-way ANOVA and Dunnett's post-test).

Figure 3: (A) ROC curves showing the performance of SMRP as diagnostic biomarker for MPM in the subset of individuals carrying the common variant of the rs2235503 C>A (rs2235503_C/C) and in the subset of subjects carrying at least one rare variant of the same SNP (rs2235503 C/A + A/A). **(B)**Comparison between the plasmatic concentrations of SMRP in non-MPM subjects stratified based on their rs2235503

alleles, and in MPM patients. The dots represent the $\log_{10}$ [SMRP] (nM) of each subject. The red lines represent the median of the $\log_{10}$ [SMRP]. **(C)** Violin plot showing the normalized levels of MSLN mRNA in lung samples from 515 subjects stratified based on their rs12597489 C>T alleles, according to the GTEx Portal. Given the good linkage disequilibrium between the rs2235503 C>A and the rs12597489 C>T ($r^2 = 0.8$), we considered this latter as a "tagging SNP" for rs2235503.






A
Sensitivity%
100% - Specificity%
0
20
40
60
80
100
rs2235503_C/C
rs2235503_C/A + A/A
B
Log10[SMRP] (nM)
1.5
1.0
0.5
0.0
-0.5
-1.0
C/C (496)
C/A (165)
A/A (16)
MPM (76)
non-MPM
C
Norm Expression
3.0
1.0
-1.0
-3.0
CC (372)
CT (130)
TT (13)