

Whole exome sequencing and machine learning germline analysis of individuals presenting with phenotypes of extreme high and low risk of developing tobacco-induced lung adenocarcinoma.

Jose Luis Perez-Gracia, Elisabet Guruceaga, Maria Pilar Andueza, Marimar Ocon, Nicolas de Villalonga Zornoza, Jafait Junior Fodop Sokoudjou, Gorka Alkorta-Aranburu, Carlos Camps, Eloisa Jantus-Lewintre, Maria Navamuel Andueza, Miguel F. Sanmamed, Alfonso Gulpide, Ignacio Melero Bermejo, Mohamed Elgendi, Juan Alberto Perez Valencia, Mikel Hernaez, Ruben Pio Osés, Luis M. Montuenga, Idoia Ochoa, Ana Patiño-García; Department of Oncology, Clínica Universidad de Navarra, Pamplona, Spain; Center for Applied Medical Research (CIMA), Pamplona, Spain; Pulmonary Department, Clínica Universidad de Navarra, Pamplona, Spain; Electrical and Electronics Department, Tecnun, University of Navarra, San Sebastian, Spain; Universidad de Navarra - CIMA LAB Diagnostics, Pamplona, Spain; Hospital General Universitario Valencia, Valencia, Spain; Department of Biotechnology, Universitat Politècnica de València, Valencia, Spain & Mixed Unit TRIAL (Príncipe Felipe Research Centre & General University Hospital Of Valencia Research Foundation), Valencia, Spain; Department of Medical Oncology, Clínica Universidad de Navarra, Pamplona, Spain; Clínica Universidad de Navarra, Pamplona, NAVARRA, Spain; Universidad de Navarra, Center for Applied Medical Research (CIMA), Pamplona, Spain; Metabolic Alterations in Cancer/University Hospital Carl Gustav Carus, Medical Clinic I, Technical University Dresden, Dresden, Germany; Computational Biology Program, CIMA, University of Navarra, Pamplona, Spain; Program in Solid Tumors, Center for Applied Medical Research (CIMA), University of Navarra, Pamplona, Spain; Department of Pediatrics and Clinical Genetics, Clínica Universidad de Navarra, Pamplona, Spain

Background: Tobacco is the main risk factor for developing lung cancer. Yet, while some heavy smokers develop lung cancer at young age others never develop it, even at advanced age. This suggests a remarkable variability in the individual susceptibility to the carcinogenic effects of tobacco. We characterized the germline profile of subjects presenting these extreme phenotypes with Whole Exome Sequencing (WES) and Machine Learning (ML). **Methods:** We sequenced germline DNA from heavy smokers who either developed lung adenocarcinoma at early age (*extreme cases*) or did not develop it at advanced age (*extreme controls*). The discovery and validation cohorts included respectively 50 and 66 extreme cases and 50 and 83 extreme controls, selected from databases including > 6,000 subjects. We selected individual coding variants and variant-rich genes showing a significantly different distribution between extreme cases and controls. We trained ML models (Logistic Regression, Random Forest, Support Vector machine Classifier (SVC)) on the discovery cohort to classify subjects into their respective phenotypes and tested them on the validation cohort. **Results:** Mean age for extreme cases and controls in both cohorts was 50.2 and 78.4 years. Mean tobacco consumption was 38.1 and 59.1 pack-years. We validated 16 significant individual variants. The most significant variants were in *ADAMTS7* (2 variants) in cases and *TMEM191B* (1) in controls. We validated 33 genes enriched with significant variants. The genes harboring more variants were *HLA-A* (4 variants) and *ADAMTS7* (2) in cases; and *PLIN4* (2) in controls (Table). We trained several ML models on the discovery cohort using as input the 16 significant individual variants and the number of variants in the 33 enriched genes. We tested them in the validation cohort obtaining accuracy of 72% and AUC-ROC of 87.4% with the best model (SVC), using 16 variants as input, confirming their association with the phenotypes. Functions of validated genes included oncogenes, tumor-suppressors, DNA repair, maintenance of genomic stability, HLA mediated antigen presentation and regulation of proliferation, migration, apoptosis and inflammatory pathways. **Conclusions:** Individuals presenting phenotypes of extreme high and low risk of developing tobacco-induced lung adenocarcinoma have different germline profiles. Our strategy may allow to identify high-risk subjects and to develop new therapeutic approaches. Research Sponsor: Spanish Society of Medical Oncology, Fundación SEOM and Fundación Salud 2000.

	Discovery		Validation		Function
	N.variants	p	N.variants	p	
Genes(cases)					
<i>HLA-A</i>	4	3.5 E-07	4	1.7 E-06	Antigen presentation
<i>ADAMTS7</i>	2	0.02	2	3.1 E-05	Metalloproteinase
<i>SPINK5</i>	2	0.002	1	0.044	Tumor suppressor
<i>REX04</i>	2	0.0002	1	0.046	DNA repair
Genes(controls)					
<i>PLIN4</i>	2	0.0006	2	0.002	Lipid droplet metabolism
<i>ZNF214</i>	2	6.7 E-05	1	0.03	Transcriptional regulation
<i>KRT18</i>	1	0.03	2	0.003	Ras activated oncogene