Cancer Treatment and Research Communications 15 (2018) 21–31

Contents lists available at ScienceDirect

Cancer Treatment and Research Communications

journal homepage: www.elsevier.com/locate/ctarc

Genomic profiling in advanced stage non-small-cell lung cancer patients with platinum-based chemotherapy identifies germline variants with prognostic value in *SMYD2*

Iván Galván-Femenía[a], Marta Guindo[b], Xavier Duran[a], Sílvia Calabuig-Fariñas[c,d,e], Josep Maria Mercader[b,1,2], Jose Luis Ramirez[f], Rafael Rosell[f], David Torrents[b,g], Anna Carreras[a], Takashi Kohno[h], Eloisa Jantus-Lewintre[d,e,i], Carlos Camps[c,d,j,k], Manuel Perucho[f], Lauro Sumoy[l], Jun Yokota[f], Rafael de Cid[a,*]

[a] *Genomes For life-GCAT Lab. Program of Predictive and Personalized Medicine of Cancer (PMPPC), Institute for Health Science Research Germans Trias i Pujol (IGTP), Can Ruti Biomedical Campus, Crta de Can Ruti, Camí de les Escoles S/N, 08916 Badalona, Barcelona, Spain*
[b] *Barcelona Supercomputing Center (BSC-CNS), Joint BSC-CRG-IRB Research Program in Computational Biology, Carrer de Jordi Girona, 29-31, 08034 Barcelona, Spain*
[c] *Department of Medical Oncology, Hospital General Universitario de Valencia, Avenida Tres Cruces, 2, 46014, València, Spain*
[d] *Molecular Oncology Laboratory, Fundación Hospital General Universitario de Valencia, Avda. Tres Cruces s/n 46014 València, Spain*
[e] *Department of Pathology, Universitat de València, Av. de Blasco Ibáñez, 13, 46010 València, Spain*
[f] *Program of Predictive and Personalized Medicine of Cancer (PMPPC), Institute for Health Science Research Germans Trias i Pujol (IGTP), Can Ruti Biomedical Campus, Crta de Can Ruti, Camí de les Escoles S/N, 08916 Badalona, Barcelona, Spain*
[g] *ICREA, Catalan Institution for Research and Advanced Studies, Spain*
[h] *Division of Genome Biology, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan*
[i] *Molecular Oncology Laboratory, Fundación Hospital General Universitario de València, Avda. Tres Cruces s/n, 46014 València*
[j] *Department of Biotechnology, Universitat Politècnica de València, Camí de Vera, s/n, 46022 València, Spain*
[k] *Department of Medicine, Universitat de València, Av. de Blasco Ibáñez, 13, 46010 València, Spain*
[l] *Genomics and Bioinformatics. Program of Predictive and Personalized Medicine of Cancer (PMPPC), Institute for Health Science Research Germans Trias i Pujol (IGTP), Can Ruti Biomedical Campus, Crta de Can Ruti, Camí de les Escoles S/N, 08916 Badalona, Barcelona, Spain*

ARTICLE INFO

Keywords:
Lung cancer
NSCLC
Advanced stage
Prognostic factors
Genome-Wide-Association Studies

ABSTRACT

Objective: The aim of the study was to investigate the relationship between germline variations as a prognosis biomarker in patients with advanced Non-Small-Cell-Lung-Cancer (NSCLC) subjected to first-line platinum-based treatment.

Materials and Methods: We carried out a two-stage genome-wide-association study in non-small-cell lung cancer patients with platinum-based chemotherapy in an exploratory sample of 181 NSCLC patients from Caucasian origin, followed by a validation on 356 NSCLC patients from the same ancestry (Valencia, Spain).

Results: We identified germline variants in *SMYD2* as a prognostic factor for survival in patients with advanced NSCLC receiving chemotherapy. *SMYD2* alleles are associated to a decreased overall survival and with a reduced Time to Progression. In addition, enrichment pathway analysis identified 361 variants in 40 genes to be involved in poorer outcome in advanced-stage NSCLC patients.

Conclusion: Germline *SMYD2* alleles are associated with bad clinical outcome of first-line platinum-based treatment in advanced NSCLC patients. This result supports the role of *SMYD2* in the carcinogenic process, and might be used as prognostic signature directing patient stratification and the choice of therapy.

Microabstract: A two-Stage Genome wide association study in Caucasian population reveals germline genetic variation in *SMYD2* associated to progression disease in first-line platinum-based treatment in advanced NSCLC

* Corresponding author.
E-mail addresses: igalvan@igtp.cat (I. Galván-Femenía), marta.guindomartinez@bsc.es (M. Guindo), xduran@igtp.cat (X. Duran), calabuix_sil@gva.es (S. Calabuig-Fariñas), mercader@broadinstitute.org (J.M. Mercader), jramirez@iconcologia.net (J.L. Ramirez), rrosell@iconcologia.net (R. Rosell), david.torrents@bsc.es (D. Torrents), acarreras@igtp.cat (A. Carreras), tkkohno@ncc.go.jp (T. Kohno), jantus_elo@gva.es (E. Jantus-Lewintre), camps_car@gva.es (C. Camps), mperucho@igtp.cat (M. Perucho), lsumoy@igtp.cat (L. Sumoy), jyokota@igtp.cat (J. Yokota), rdecid@igtp.cat (R. de Cid).
[1] Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.
[2] Diabetes Unit and Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA.

https://doi.org/10.1016/j.ctarc.2018.02.003
Received 24 July 2017; Received in revised form 26 January 2018; Accepted 19 February 2018
2468-2942/ © 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

Table 1
Clinical and pathological characteristics of the discovery (BREC, n = 178) and validation sample (n = 323).

		BREC		Disease progression					Validation Sample		Disease progression					BREC + Validation Sample		
				NO		YES					NO		YES					
		N	%	N	%	N	%	p-value$_1$	N	%	N	%	N	%	p-value$_2$	NBREC	NValidation	p-value$_3$
Gender								0.29							0.55			0.16
	Male	139	78	104	76	35	85		270	83	171	83	99	85		139	270	
	Female	39	22	33	24	6	15		53	17	36	17	17	15		39	53	
Smoker								0.16							-			
	Yes	167	94	126	92	41	100											
	No	10	6	10	7	0	0											
	NA	1	0	1	1	0	0											
ECOG								0.73							**0.002**			0.23
	0	59	33	45	33	14	34		95	29	74	36	21	18		59	95	
	1	114	64	88	65	26	64		220	68	128	61	92	80		114	220	
	2	2	1	2	1	0	0		7	2	4	2	3	2		2	7	
	NA	3	2	2	1	1	2		1	1	1	1	0	0		3	1	
Histology								**0.006**							0.83			**0.002**
	ADCA	99	56	83	61	16	39		164	51	105	51	59	51		99	164	
	SCC	64	36	44	32	20	49		101	31	67	32	34	29		64	101	
	LCC	6	3	6	4	0	0		45	14	28	13	17	15		6	45	
	Others	9	5	4	3	5	12		13	4	7	4	6	5		9	13	
Treatment								0.66							1			-
	doce/cis	123	69	93	68	30	73		323	100	207	100	116	100				
	gemci/cis	44	25	36	26	8	20											
	doce	11	6	8	6	3	7											
Stage								1							0.70			**0.0001**
	III	7	4	5	4	2	5		52	15	35	17	17	15		7	52	
	IV	164	92	127	92	37	90		271	85	172	83	99	85		164	271	
	NA	7	4	5	4	2	5											
RECIST																		**6.6×10^{-9}**
	PD	41	23						123	37						41	123	
	SD	56	31						113	34						56	113	
	PR	58	32						93	28						58	93	
	CR	23	14						3	1						23	3	

ECOG, performance status; doce, docetaxel; cis, cisplatin; gemci, gemcitabine; RECIST, response evaluation criteria in solid tumors; PD, progression disease; SD, stable disease; PR, partial response; CR, complete response; NA, not available.
The **p-value$_1$** and **p-value$_2$** columns show the difference between progression disease regarding each clinical variable in BREC and validation sample respectively.
The **p-value$_3$** column show the difference between the number of patients in BREC and validation sample for each clinical variable.

patients. *SMYD2* profiling might have prognostic / predictive value directing choice of therapy and enlighten current knowledge on pathways involved in human carcinogenesis as well in resistance to chemotherapy.

1. Introduction

Lung cancer is the most common cancer in the world, and the leading cause of mortality among cancer-related deaths. The Non-Small-Cell-Lung-Cancer (NSCLC), being the most common form, has an overall 5-years survival of less than 15% [15]. NSCLC is a histological diverse group of tumors, with major classes being squamous (SCC), adenocarcinoma (ADC), and large cell carcinoma (LCC). Despite the enormous heterogeneity, these tumors have been treated homogeneously for a long time with cytotoxic chemotherapy as the choice treatment [17].

Platinum-based chemotherapy is still widely used for treatment of the vast majority of NSCLC patients with advanced-stage disease, with the exception of cases bearing E*GFR, BRAF*, *ROS1* or *EML4-ALK tumor* mutations. The latter have greatly benefited from the advances achieved in the last ten years in targeted therapy based on somatic genetic/molecular profiling. While chemotherapy provides palliation, advanced NSCLC remains incurable in most cases since acquired resistance is common, response rates are only 15%–30%, and median OS is less than 12 months. Resistance can arise from a several causes (drug delivery, altered target, tolerance to damage, etc…) [16] and differences observed in therapy efficacy could be explained by the impact of host genotype variants on target/resistance factors.

Customization of chemotherapy has relied on tumor cell expression profiles of specific genes. Candidate gene studies have indicated possible association to response of genetic variants in genes of the platinum pathway (reviewed by Hildebrandt et al. [20]) and DNA-repair genes [32]. Genome-wide association studies (GWAS) have been used successfully to identify germline genetic variants associated with an increased risk of developing lung cancer including NSCLC, and have been applied to identify prognostic biomarkers [22,27,49,50,54,55] as well as to identify genetic variability associated to adverse effects to chemotherapy [6,7]. Furthermore, re-positioning of GWAS-derived germline predisposition markers as prognostic markers, have been successfully reported in other cancer forms.

The aim of this study was to investigate the relationship between germline variants to identify prognosis biomarkers for clinical outcome in patients with advanced NSCLC subjected to first-line platinum-based treatment. In this study we report a genome-wide scan study in two independent samples from the same ancestry (Spain).

We present evidence of germline variation in the *SMYD2* affecting the clinical outcome of advanced NSCLC patients. *SMYD2* profiling might have prognostic/predictive value directing choice of therapy and enlighten current knowledge on pathways involved in human carcinogenesis as well in resistance to chemotherapy.

2. Material and Methods

2.1. Patients

This study was approved by the institutional review board of the IGTP. The recruitment of NSCLC patients in the discovery phase and validation phase was approved by the institutional review board of each participating institution.

2.2. Discovery sample

Patients included in the study were selected from the BREC clinical trial study (Multicenter, Predictive, Prospective, Phase III, Open, Randomized, Pharmacogenomic Study in Patients with Advanced Lung Carcinoma (BREC)) [35]. BREC patients with advanced NSCLC who had not received treatment for the disease at the time they entered the study and had a good performance status (ECOG 0–1) and measurable disease (at least one target lesion according to the RECIST (response evaluation criteria in solid tumors), received six to eight chemotherapy cycles. The 94% received cisplatin 75 mg/m2 combined with Docetaxel 75 mg/m2 (73%) (group 1) or Gemcitabine 1250 mg/m2 (27%) (group 2), both at day 1, in 21-day cycles. Remaining 6% received Docetaxel 75 mg/m2 (group 3), on day 1 every 3 weeks. See complete description at clinicaltrials.gov/show/NCT00617656.

A total of 178 patients were included in the genetic analysis. All considered patients were *EGFR*-WT. The median age was 62 years (range: 39–82), 78% males, and 92% stage IV of the disease. ADC was the most common histological subtype (56% of patients) of NSCLC, followed by squamous cell carcinoma (SCC) (36%) and large cell carcinoma (LCC) (3%), 5% were grouped in other categories. The most frequent ECOG score was 1 (64%) (33% and 1% for 0 and 2 status). Overall progression free survival (PFS) (calculated from the date of randomization to progression or death) was 5.3 months (95% CI 4.71–5.88), and survival time (Overall Survival OS; calculated from the date of randomization to death) was 10.16 months (95% CI 8.32–12.01).

2.3. Validation sample

Patients included in the validation cohort were from a Multicenter study coordinated by the Spanish Lung Cancer Group. All patients were with advanced NSCLC, from Caucasian ancestry and the same geographical region (Valencia, Spain) [24,25,47]. Blood samples were recollected from 356 NSCLC stage IIIB with pleural effusion or stage IV, who received cisplatin (75 mg/m^2) and docetaxel (75 mg/m^2) on day 1 every 3 weeks. Among 356 patients, 323 with fulfilled response data were considered for the analysis.

The median age was 59 years (range: 31–80), 83% males. 15% of patients had stage III and 85% stage IV of the disease. Like in BREC patients, ADC was the most common histological subtype (51%) of NSCLC, followed by SCC (31%) and LCC (14%), 4% were grouped in other categories. The most frequent ECOG score was 1 (68%) (29% and 2% for 0 and 2 status). TTP was 5.53 months (95% CI 4.93–6.33) and OS 9.9 months (95% CI 9.17–11.07).

According to the study objectives, the clinical outcomes were diagnosis of NSCLC and response to treatment (according to the criteria established in the RECIST). Patients were categorized as progressing disease if its RECIST was assessed as PD (PD) (23% BREC, 37% validation sample) and as non-progressing if their RECIST was complete (CR) or partial response (PR) (14%, 1% and 32%, 28%) or stable disease (31%, 34%)(SD), in both exploratory and replication cohorts, respectively. The main clinical and pathological characteristics of the discovery and validation samples are shown in Table 1.

2.4. Genome scan

2.4.1. Genotyping

Genome-wide genotypes were generated for the discovery sample using SNP-array technology. The Infinium® HTS Assay automated protocol, was used on HumanCoreExome-24v1-0 BeadChips scanned with a HiScan confocal scanner (ILLUMINA, San Diego, CA). Genotyping was performed at the Genomic Units of the PMPPC-IGTP. Genome Studio version 2011.1 was used for raw data analysis. All illumina internal system controls were fulfilled. Before the genetic

association analysis, we conducted systematic quality control on the raw genotyping data to filter both unqualified samples and SNPs. Overall call rate was 99.89%. Samples were excluded if they failed genotyping in more than 10% of variants. Variants were excluded if they failed genotyping in more than 10% of samples, were non-polymorphic, or showed departure from Hardy-Weinberg Equilibrium (HWE) (p value > 0.0001). 40% of genotyped markers were monomorphic in our sample. Gender control detected a mismatch in one sample that was included in the study after database consultation. After these quality control steps, 181 cases with 325,762 SNPs were considered. PLINK 1.9 version [9,43] was used to perform the quality control analysis. Genotyping of candidate SNPS in the replication sample was done at the Spanish National Centre of Genotyping (ISCIII-CEGEN-Santiago Node) facility by using the iPLEX Sequenom MassARRAY platform (Sequenom Inc., San Diego, CA, USA) and at PMPPC-IGTP by Real-Time PCR, using TaqMan™ (ILLUMINA, San Diego, CA) when do not fit Sequenom's basics.

2.4.2. *In silico genotyping*

IMPUTE2 [21] was used to impute untyped SNPs from sequence-based reference panels (1000Genomes, UK10K, GoNL). SHAPEIT [11] was used for haplotype estimation prior to imputation procedures. Imputed genotypes with IMPUTE2info lower than 0.7 were discarded for association analysis. The best IMPUTE information score was used for those SNPs present in more than one reference panel. Finally, from 24,873,940 imputed SNPs we considered 10,307,177 unique SNPs for association analysis.

2.5. *Population structure and relatedness*

All patients in the discovery sample were used to detect population substructure and independence. Principal Component Analysis (PCA) was applied to detect any hidden substructure, and method of moments (MoM) for the estimation of identity by descent probabilities was applied to exclude cases with cryptic relatedness. A Spanish population based cohort GCAT (genomesforlife.com) and public databases (HapMap) were used to test ancestry homogeneity before imputation analysis [2].

2.6. *Statistical analysis*

We perform a multivariate logistic regression model, under an additive model, adjusted by gender, smoking (yes/no), tumor histology (i.e. ADCA, SCC, LCC, other), pretreatment performance status (ECOG score) (0, 1, > 2), chemotherapy treatment group (docetaxel/cisplatin, gemcitabine/cisplatin, docetaxel) and the first seven principal components (PC) as covariates. Genomic control inflation for the association results was calculated from observed and imputed data ($\lambda = 1.12$). All p-values were corrected for genomic inflation factor.

For the replication analysis we considered those SNPs with corrected p-values $< 1 \times 10^{-5}$ (Fig. 1). For validation purposes, due to the relative small sample size and the inflated or deflated size effect for SNPs with MAF < 0.01 generated from imputation methods, we considered those with an effect size (OR) in the [0.05–20] range. From suggestive signals, alternative SNPs were selected with LDlink [30] and FINEMAP software [3]. Both tools were used for exploring possible functional variants via linkage disequilibrium and a shotgun stochastic search algorithm. Selected candidates are shown in Fig. 2.

We analyze all candidates SNPs in the validation sample under the same model assumptions, but excluding smoking, since was not relevant in the BREC analysis, and was not available in the replication sample. Then a joint analysis was performed. Since differences were evident among cohorts (Table 1), a heterogeneity analysis was performed and I^2 measure was estimated. Heterogeneity source was investigated by a multiple correspondence analysis [26] to detect any data structuring within BREC and Valencia sample regarding gender, histology, stage and ECOG categorical variables (Supplementary Fig. 1). For replication, we performed a matched analysis with resampling. Each Valencia's individual was matched with BREC cohort by disease progression and stage to preserve the same clinical features before association analysis. Then, we resampled 10,000 times and p-values were derived and ranked. A p-value representing the 5% percentile of the p-values distribution [13] was considered for each SNP.

Cox proportional hazard regression models were used to evaluate survival outcomes (TTP and OS) in the validation cohort, and multivariate analysis was performed adjusting the Cox models by age, gender, ECOG and disease progression status. No individual data was available from the BREC cohort. Survival curves were computed with the Kaplan–Meier estimator. Hazard ratios (HRs) and their 95% confidence intervals were assessed to evaluate the risk of death.

We used SNPtest software [31] for association analysis in the discovery sample, and PLINK 1.9 version for association analysis in the validation sample. SNPtest allows worked seamlessly with imputation data. R software (version 3.3.1, R Core Team, 2016) was used for data visualization (Manhattan plot, QQ plot, Kaplan-Meier and ROC curves) and statistical analyses. Data visualization was made with LocusZoom, for plotting chromosomal regions.

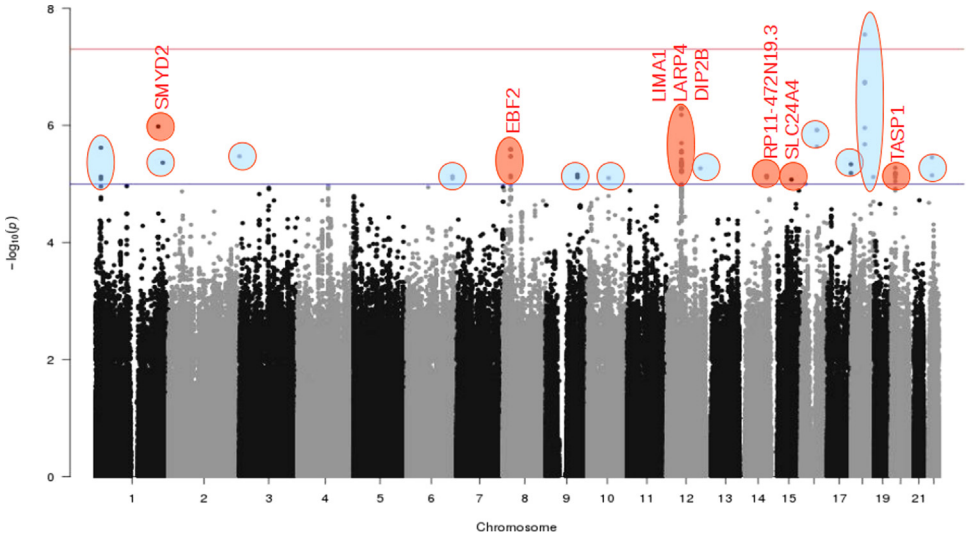

Fig. 1. Manhattan plot for genome-wide association results in the BREC discovery sample. Association p-values are expressed as -log10(p). P-values comes from multivariate models accounting for gender, smoking status, histology, ECOG performance status, chemotherapy treatment and the first seven principal components. Red circles are the selected peaks used for replication purposes (MAF > 0.01 and 0.05 > OR < 20). The blue and red lines indicate the p-value threshold for the candidate genes at -log10(10^5) and -log10(5×10^8) respectively.

CHR:BP	Alleles	Freq.	Info	Proxy	R2	P−val	OR	CI95% OR
1:214481630	G/A*	0.47	1.00	FINEMAP	0.59	8×10^{-5}	0.23	[0.11−0.5]
1:214495703	C/T*	0.47	1.00	BI_p	0.59	1.3×10^{-4}	0.25	[0.12−0.52]
1:214502898	C/G*	0.31	0.99	BI_p	0.65	1.9×10^{-3}	3.30	[1.52−7.15]
1:214503489	T*/G	0.41	0.99	BI	–	1×10^{-6}	7.02	[2.88−17.12]
8:25643577	G/A*	0.50	1.00	FINEMAP		8×10^{-5}	0.22	[0.1−0.49]
8:25647711	A/G*	0.32	1.00	BI_p		1.3×10^{-5}	5.81	[2.43−13.86]
8:25652367	T/C*	0.31	1.00	BI_p		2.6×10^{-6}	7.40	[2.89−18.9]
8:25662655	G/T*	0.31	0.99	FINEMAP		3.3×10^{-6}	7.36	[2.87−18.89]
12:50589836	A/G*	0.20	1.4×10^{-4}	BO_p	0.77	4.6×10^{-4}	5.25	[1.92−14.32]
12:50645471	A/G*	0.21	1.00	BO	–	5.1×10^{-7}	17.10	[4.23−69.14]
12:50824232	G/A*	0.20	0.97	BI	–	3.8×10^{-6}	14.42	[3.55−58.58]
12:50881148	C/A*	0.26	0.99	BI_p	0.61	1.2×10^{-5}	6.96	[2.62−18.49]
12:50947729	G*/T	0.29	1.00	FINEMAP	0.85	3×10^{-5}	5.30	[2.22−12.64]
12:50959426	G/A*	0.25	1.00	BO	–	4.6×10^{-6}	8.30	[2.93−23.55]
12:51015509	A/G*	0.26	1.00	BO_p	0.97	1.3×10^{-5}	6.94	[2.6−18.54]
12:51026878	G/T*	0.17	1.00	FINEMAP	0.07	3.7×10^{-5}	0.20	[0.09−0.44]
14:92726738	A*/G	0.30	1.00	BO_p	0.37	6.9×10^{-3}	0.39	[0.2−0.76]
14:92726813	C*/G	0.22	0.99	FINEMAP	0.61	1.4×10^{-4}	0.23	[0.11−0.5]
14:92729757	C/A*	0.27	1.00	BO_p	1	7.3×10^{-6}	0.19	[0.09−0.41]
14:92729907	G/T*	0.27	1.00	BO	–	7.2×10^{-6}	0.19	[0.09−0.41]
15:70089618	A*/G	0.47	0.96	FINEMAP	0.56	3.3×10^{-4}	0.30	[0.15−0.59]
15:70095755	T/C*	0.21	1.00	BO_p	0.27	6.2×10^{-3}	0.38	[0.2−0.74]
15:70098476	A/G*	0.48	0.96	BO_p	0.68	4.2×10^{-5}	0.24	[0.12−0.5]
15:70104713	C*/T	0.46	1.00	BO	–	8.4×10^{-6}	0.23	[0.11−0.46]
18:46538022	A*/G	0.12	0.76	BI	–	2.8×10^{-8}	0.01	[0.002−0.11]
20:13582476	C/A*	0.35	1.00	BO	–	1.2×10^{-5}	5.15	[2.36−11.26]
20:13582879	T/G*	0.33	1.00	BO_p	0.94	6.9×10^{-6}	5.52	[2.48−12.32]
20:13612157	G/T*	0.40	1.00	BO_p	0.73	1×10^{-3}	3.19	[1.58−6.43]

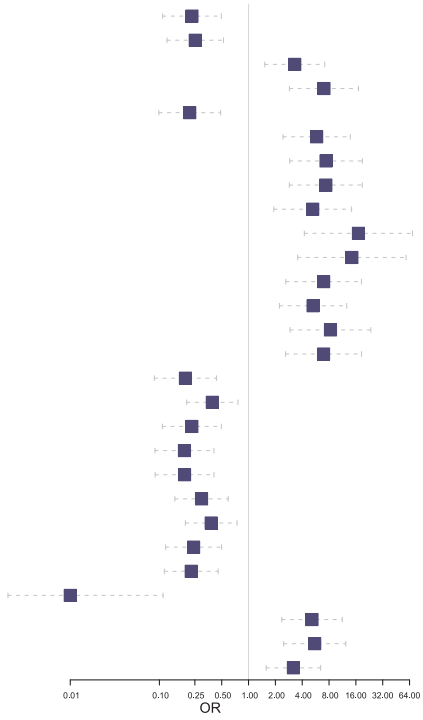

Fig. 2. Forest plot diagram of the association results of the BREC discovery dataset used for replication analysis. The variants are listed by chromosome and position (CHR:BP) showing the IMPUTE information measure (Info) and the effect size (OR) regarding the first allele of the Alleles column. BO, best observed; BI, best imputed; BO_p, best observed proxy; BI_p, best imputed proxy.

2.7. Pathway enrichment analysis

In order to provide biological hypotheses from our GWAS results we performed a pathway analysis to highlight enriched pathways based on genes in associated loci. All genes with at least one variant at p-value $< 1 \times 10^{-4}$ were included in the analysis. We used the seq2-pathway R package [51] to select the subset of the most significant genes within a search radius of 150 kbps from the SNPs with an association p-value below 1×10^{-4}. Pathway enrichment analysis of the 889 selected genes was performed against Gene Ontology and Reactome annotation data with both seq2pathway and PANTHER Overrepresentation Test tool (release 20160715) [34]. The significance of the GO terms was estimated through the adjusted p-values based on the binomial testing with Bonferroni correction for multiple hypotheses.

2.8. Fine-mapping and functional annotation

Variant Effect Predictor (VEP) tool [33] was used for the functional characterization of identified variants (hg19). The VEP application determines the effect of variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, protein and regulatory regions.

3. Results

3.1. Clinical and pathological characteristics of the two-stage used cohorts

Bivariate analysis of the clinical and pathological characteristics shows differences in tumor histological type ($p = 0.002$), stage ($p = 0.0001$) and progression disease ($p = 6.6 \times 10^{-9}$), with more cases of LC, stage III and PD in the validation sample than in BREC. No other differences in gender, age, and pretreatment performance status were statistically significant. Regarding PD, we observed differences in tumor histological type, slightly different in the discovery sample, but not in the replication sample, and ECOG, related to PD in the Valencia sample but not in BREC. Concerning chemotherapy treatment group, no significant differences were observed in BREC. All statistically significant differences were considered as covariates in further analyses. The clinical and pathological characteristics of the study population are shown in Table 1.

3.2. Twenty genomic regions show association with disease progression outcome in the discovery sample

PCA analysis indicated that the BREC as an ethnically homogenous Caucasian. All patients except three overlapped with the CEU ancestry reference panel from HapMap and with the geographically matched sample from the Spanish GCAT cohort (genomesforlife.com). The three genetically distant patients were discarded for the genomic analysis. The first seven PCA dimensions were incorporated in the association analysis as covariates. No cryptic relatedness was found by estimating identity by descent (IBD) probabilities.

Association analysis was made with observed and imputed data recovered from three public reference panels. In the discovery phase we observed one SNP with p-value $< 1 \times 10^{-8}$, two SNPs with p-value $< 1 \times 10^{-7}$, 22 SNPs with p-value $< 1 \times 10^{-6}$, 147 SNPs with p-value $< 1 \times 10^{-5}$, 864 SNPs with $< 1 \times 10^{-4}$, 8,674 SNPS with $< 1 \times 10^{-3}$ and 90,826 SNPs with p-value $< 1 \times 10^{-2}$, associated with PD. Resulting genome-wide association results are shown by the Manhattan plot in Fig. 1. Top hits with a p-value $< 1 \times 10^{-5}$, and (OR) [0.05–20] were selected for replication in the Valencia sample. None of the retained SNPs reached the genome-wide threshold. Further, as single SNP analysis results could be misleading, we plotted genotypes 500Kb around the peak together with along additional annotation from the GWAS catalogue, recombination rates, LD measures with genotyped or imputed SNPs in the region, and functional annotation for each SNP. After visualization, eight regions were retained (Supplementary Fig. 2). Observed genotype was preferentially retained when imputed signals were also present; three derived from *in silico* genotyping (imputation). We selected additional SNPs as proxies ($r^2 = 1$–0.6 on average) for

individual genotyping, by using FINEMAP and LDlink tools. In addition to this selection, the genome-wide associated SNP (p-value = 2.8×10^{-8}) at Chr18, was included in the replication step (Supplementary Fig. 2). A total of 28 SNPs in nine chromosomal regions were chosen for replication testing in the Valencia cohort. All of them were in Hardy-Weinberg equilibrium (p-value > 0.001). Results of the association analysis and minor allele effect sizes for selected SNPs are shown in Fig. 2.

3.3. SMYD2 replicated the association in an independent sample

Five variants out of 28 analyzed were associated with a PD in the validation cohort, overlapping with the *SMYD2*, *LARP4*, *RP11–472N19.3*. The observed variant effect size was in the same direction in both discovery and validation samples, except for the variant in *LARP4* (Fig. 3). Variants in *SMYD2* and *RP11-472N19.3* were statistically significant. *SMYD2* carry two variants, chr1:214502898-rs4655246 and chr1:214503489-rs2291830, associated with a poor outcome in both cohorts. Minor alleles at two positions (p-value = 1.9×10^{-3}, freq. = 0.31 and p-value = 1.0×10^{-6}, freq. = 0.41 for BREC; p-value = 0.016, freq. = 0.32 and p-value = 0.038, freq. = 0.42 for validation sample) were associated with PD in BREC and validation cohort. The rs4655246-C allele variant, and the rs2291830-T allele variant showed a strong effect towards progressing disease; OR = 3.33 and OR = 1.47 for C-allele, and OR = 7.02 and OR = 1.26 for T-allele, for BREC and the validation cohort respectively. In *RP11–472N19.3*, two variants, chr14:92726738-rs7142050 and chr14:92726813-rs4904853, show a protective effect (i.e. favoring non progressing disease) by the minor allele for both cohorts (p-value = 6.9×10^{-3}, freq. = 0.3 and p-value = 1.4×10^{-4}, freq. = 0.22 for BREC; p-value = 0.045, freq. = 0.34 and p-value = 0.035, freq. = 0.23 for validation sample); for the rs7142050-A allele variant the effect size was OR = 0.39 and OR = 0.81, and for rs4904853-C, OR = 0.23 and OR = 0.75, for BREC and the validation cohort respectively.

Table 2
Results from survival analysis for overall survival (OS) and time to progression (TTP) of significant variants in the validation sample.

	Gene	Variant	HR (95% CI)	p-value
(OS)	*SMYD2*	chr1:214481630-rs6665343-A	1.370 (1.050, 1.788	0.020
		chr1:214495703-rs11120295-T	1.368 (1.047, 1.787)	0.022
		chr1:214503489-rs2291830-T	1.289 (1.017, 1.633)	0.036
(TTP)		chr1:214495703-rs11120295-T	1.331 (1.020, 1.737)	0.035

Variant, chromosome position in GRch37/hg19, rs identifier and the allele effect; HR (95% CI), hazard ratio; 95% confidence interval of the hazard ratio; p-value of the variant calculated from the Cox regression model with gender, age, ECOG, progression disease and stage as covariates.

3.4. SMYD2 variants have an impact on survival endpoints in the validation cohort

Survival analysis was assessed in the replicated regions, and only reach statistical significance for *SMYD2*. Impact on survival outcomes was analyzed for overall changes in survival (OS) as well as in time to progression (TTP). We then stratified survival analysis by outcome (i.e. disease progression) to test the impact on other aspects of survival outcomes. Median OS and TTP was lower in the PD patients from those with response and stable disease (non PD); OS = 6 months (CI95% = [5.1,7.1]) and 13.4 (CI95% = [12.1,15.7]) and TTP = 2.8 months (CI95% = [2.6,3.1]) and 7.9 months (CI95% = [7.5,8.4]). Summary results for *SMYD2* variants are presented in Table 2.

In *SMYD2*, three analyzed variants (rs6665343 G/A, rs11120295 C/T, rs2291830 T/G) were associated with a reduced survival time. OS was shorter for the rs6665343-A, rs11120295-T, rs2291830-T allele carriers, showing a dominant effect. Allele variant rs6665343-A carriers had a shorter OS; 12.8 to 9.7 months (p-value = 0.020, HR = 1.370 95% (1.047–1.787)), allele variant rs11120295-T shows similar reduction of OS (12.5 to 9.7 months, p-value = 0.022, HR = 1.368 95% (1.047–1.787)), and rs2291830-T shows the lower effect, with a slight reduction (10.4 to 9.8 months, p-value = 0.036, HR = 1.289 95%

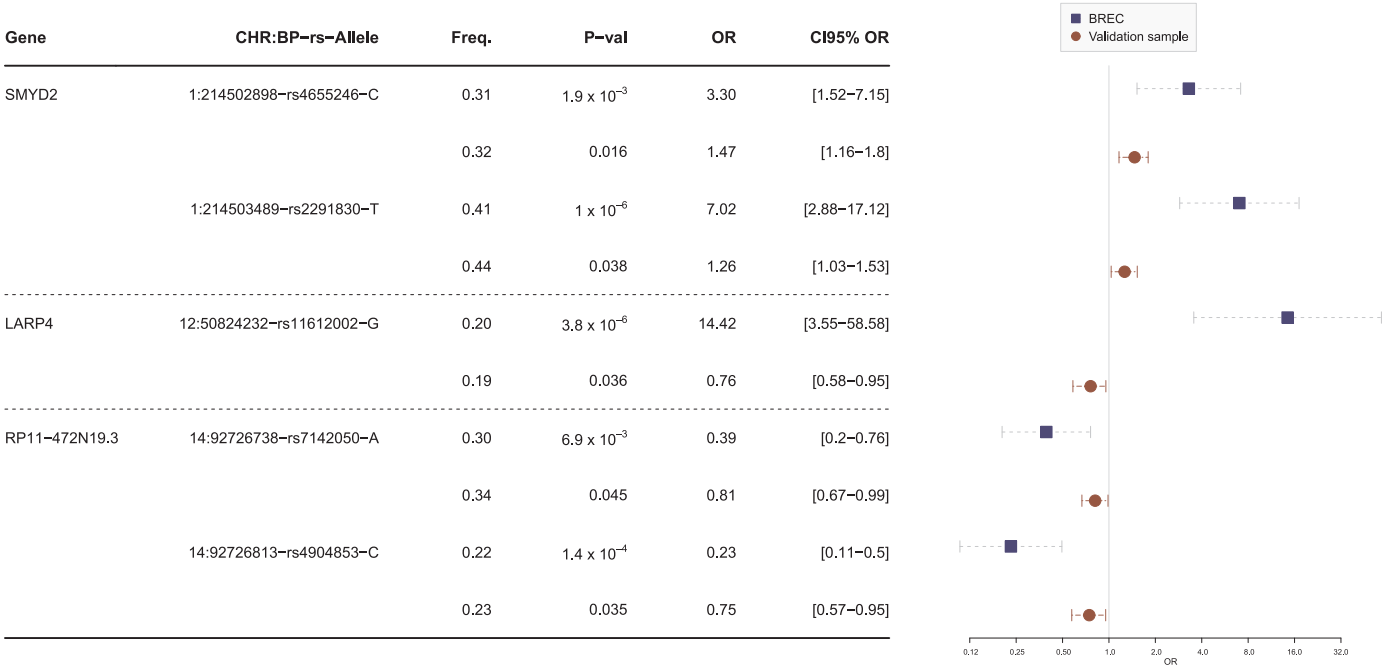

Fig. 3. Forest plot diagram of the replicated variants in the discovery and validation sample. Variants in *SMYD2* (chr1:214502898; chr1: 214503489), *RP11–472N19.3* (chr14:92726738; chr1492726813) show the same effect direction, but it is discordant in LARP4.

(1.017–1.633). When considered survival, only rs11120295-T allele was associated with shorter TTP with a dominant effect for the common allele (freq. = 0.53) with a reduction in 1.6 months (from 6.9 months to 5.3 months, p-value = 0.035, HR = 1.331 95% (1.020–1.737)) (Fig. 4). In the BREC cohort, individual survival data was not available but in concordance, rs11120295-T allele was associated with a PD outcome (OR = 0.25, CI95% = [0.12, 0.52], p-value = 1.3×10^{-4}).

3.5. Pathway analysis

From the filtered raw association signals shown in Fig. 1, we performed the functional characterization of the 889 selected genes overlapping with genome scan signals with a nominal p-value $< 1 \times 10^{-4}$. Nine GO pathways were significantly enriched with the overlapping genes (Table 3). The sequence-specific DNA binding pathway (GO: 0043565) (OR = 5.32, p value = 0.0050) with nominal values overlapping *ATF1*, *PAX7*, *TBX3*, *IRX5*, *IRX3* and *CERS5,* and the cAMP-mediated signaling pathway (GO:0019933) (OR = 13.60, p value = 0.0054) highlighted by the *ADM*, *EIF4, EBP2*, *PDE4D*, *RAPGEF2*, *PCLO* genes were the most significant ones. None of the Reactome pathways reached significant level after multiple-testing correction.

4. Discussion

To better understand the germline genetic factors modulating disease progression in advanced NSCLC with first-line platinum-based treatment we performed a genome wide analysis in a two-stage approach, including two independent populations with the same ethnic ancestry. Our results provide evidence for implication in disease progression and overall survival of germline genetic variants in *SMYD2*.

In our study, the *SMYD2* variant chr1:214503489-rs2291830 T/G, is associated with poor clinical outcome for treated patients. The effect size observed for the rs2291830-T allele is the highest *SMYD2* signal observed in our study; OR = 7.02, CI 95% = [2.88 − 17.12]. Furthermore, survival analysis shows that rs2291830-T carriers have a reduction in the survival time (10.4 to 9.8 months, p-value = 0.036) in the validation cohort. *SMYD2* (SET and MYND domain containing 2) encode for one of the SMYD methyltransferase family proteins (SMYD1–5) [18], some of which have already been reported as candidate targets for anticancer drugs [48]. *SMYD2* is overexpressed in multiple cancer cells [10], and in addition to histones, methylates other protein substrates, including RB1 and p53, leading to loss of its tumor suppressive function [23]. There are also interesting observations, showing that depletion of SMYD2 is linked to cancer chemotherapy improvement, through the reduction of PARP1 activity, which is involved in DNA repair, chromatin modification, transcriptional regulation and genomic stability [40]. Concordantly, genetic variants in *PARP1* have been associated to a better response to platinum-based chemotherapy in NSCLC [46]. Furthermore, a prognostic value has been proposed for this protein, but there is contradictory data on functionality, while SMYD2 overexpression has been reported as a bad prognostic factor in leukemia, esophageal squamous cell carcinoma and gastric carcinoma, low expression levels in renal tumors have been associated with worse disease-specific survival and disease-free survival [41]. Supporting the role in the carcinogenic process, Nakamura's Group recently reported SMYD2-mediated ALK methylation as a new mechanisms regulating cell growth in NSCLC ALK-fused gene cell lines [52].

The other SMYD2 variants in close LD (rs6665343, rs4655246, rs11120295, rs2291830, $r^2 > 0.60$) were concordant with the observed SMYD2 association (Fig. 2), however, none of the variants had any clinical significance. No variation effect on protein function was observed using SIFT and Polyphen analysis. All variants were intronic. Expression quantitative trait loci (eQTL) analysis was performed, a significant cis-eQTL, on SMYD2 expression for rs2291830-T allele *(pvalue = 7.30 x10^{-7}, eQTL effect size (es) = −0.31), as well for* rs6665343-A, rs4655246-C and rs11120295-T alleles. *(pvalue = 3.3 × 10^{-5}, es-0.16, pvalue = 3.7 pvalue = 3,7 × 10^{6}, es = 0.17, and pvalue = 4.3 × 10^{-5}, es = −0.17)* was present in transformed samples (fibroblasts) on the GTEx database (Release V6p (dbGaP Accession phs000424.v6.p1), and non-transformed samples (peripheral blood cells) (*pvalue* = 3.4 10^{-6}, *pvalue* = 9.2 × 10^{-11}, *pvalue* = 2.11 × 10^{-9}, *pvalue* = 2.6 × 10^{-11}) from Westra et al. [53] but not in lung tissues. However, a trans-eQTL, was observed when consider lung tissue samples on KCNK2 *(potassium two pore domain channel subfamily K member 2)* expression; rs6665343-A, rs11120295-T, and rs2291830-T alleles were correlated with a higher expression of *KCNK2 (pvalue = 1.1 × 10^{-2}, es = 0.18)*. KCNK2 belongs to the two-pore-domain background potassium channel protein family, and interestingly overexpression of the channel protein, in prostate cancer, has been

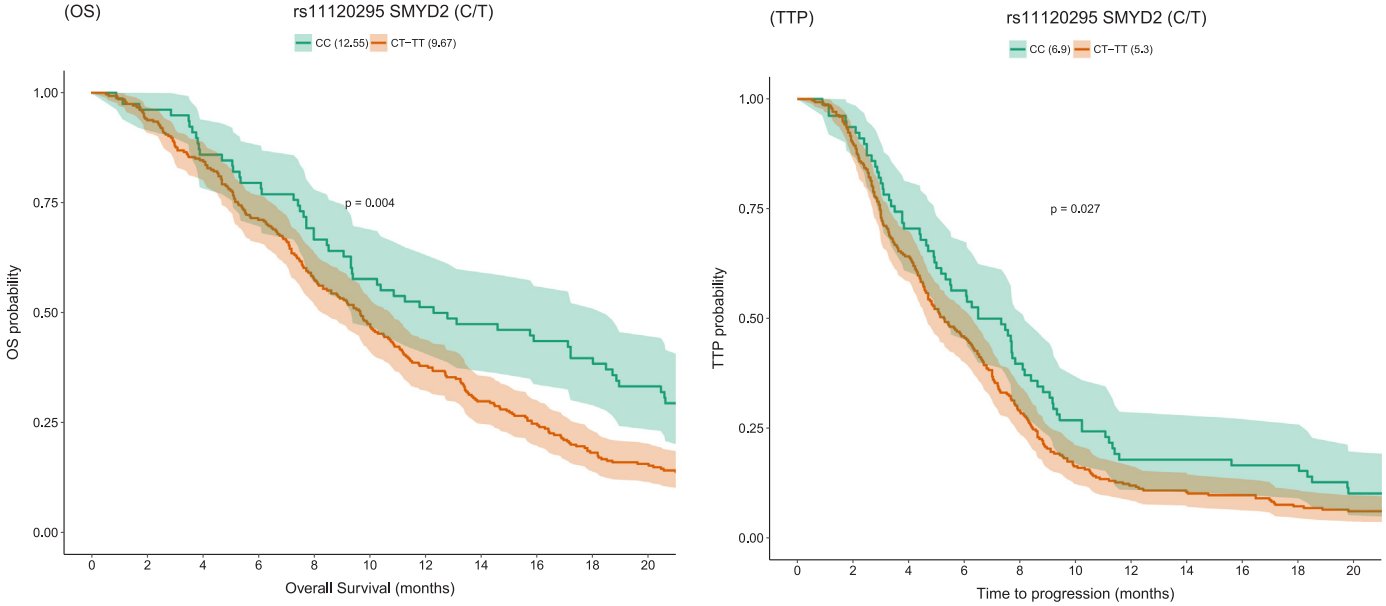

Fig. 4. Kaplan Meier plot for the validation sample: overall survival (OS) and time to progression (TTP) of patients with risk (CT-TT) and non-risk genotypes (CC) for the rs11120295 SMYD2 variant.

Table 3
Summary of the pathway enrichment analysis results in the discovery sample.

Method	GWAs	GO:ID	Description	Corrected p-value	OR	Intersect count	GO count	Intersect genes
PANTHER	−4	GO:0019864	IgG binding	0.016	20.46	5	12	*FCGR2A FCGR3B FCGR2C FCGR2B FCGR3A*
PANTHER	−4	GO:0060986	Endocrine hormone secretion	0.032	22.32	5	11	*GATA3 CGA GHRL TBX3* ***FZD4***
seq2pathway	−4	GO:0019933	cAMP-mediated signaling	0.005	13.60	5	22	*ADM EIF4* ***EBP2 PDE4D*** *RAPGEF2 PCLO*
seq2pathway	−4	GO:0010595	Positive Regulation Of Endothelial Cell Migration	0.026	7.70	5	35	*AGT ANGPT1 GATA3* ***PROX1*** *NRP1*
seq2pathway	−5	GO:0006351	Transcription, DNA-templated	0.036	2.41	10	1766	***PTPN14*** *TBX3 IRX5 IRX3 SALL3 ESF1* ***SMYD2 EBF2*** *ING5 TLE3*
seq2pathway	−5	GO:0045893	Positive regulation of transcription, DNA-templated	0.03	4.15	5	487	***PROX1*** *TBX3* ***TASP1 EBF2*** *ING5*
seq2pathway	−4	GO:0016055	Wnt signaling pathway	0.03	3.35	10	150	*HHEX PITX2 TLE3 TLE4* ***FZD4*** *PYGO1 WWOX CXXC4 NKD2 RSPO2*
seq2pathway	−5	GO:0003700	Sequence-specific DNA binding transcription factor activity	0.018	3.11	7	990	***ATF1*** *PAX7 TBX3 IRX5 IRX3 CERS5* ***PROX1***
seq2pathway	−5	GO:0043565	Sequence-specific DNA binding	0.005	5.32	6	500	***ATF1*** *PAX7 TBX3 IRX5 IRX3 CERS5*

Methods, PANTHER and seq. 2pathway overrepresentation methods; GWAS, p-value below 10^{-4} and 10^{-5} threshold for SNP inclusión; Corrected p-values on seq. 2pathway correspond to FDR while corrected p-values on PANTHER overrepresentation tests are adjusted with Bonferroni correction.

associated with a reduced survival, while knockdown inhibits cell proliferation in vivo [56].

In order to identify possible functional haplotypes, we estimated genewide haplotype structure of *SMYD2*, and haploblocks were inferred with the CI method as implemented in Haploview. All four variants were in the same block, the largest conserved block in the 3′-terminal region, but interestingly rs4655246-C/G differentiate two different haplotypes; ATCT (freq = 0.315) and ATGT (freq = 0.093), suggesting a functional role for ATCG / ATCT haplotype carriers.

Genes frequently methylated in lung cancer cells and associated with oncogenic growth of cancer cells could be targets of SMYD2, which is over-expressed in most cancer types. All of the validated methylated substrates of SMYD2 are implicated in stress responses and cellular checkpoints, it is possible that overexpression and dysregulated methylation activity could lead to compromised chemotherapy response and reduced overall survival [12,44]. Nowadays, 20 published non-histone proteins have been reported as validated targets of SMYD2 [1]. In concrete, some authors have reported that SMYD2-methylation mediated of RB1, HSP90, PTEN, PARP1 has a critical roles in tumorigenesis [10,19,36,40], and confirm, as a possible common mechanism for SMYD2 cancer progression, a SMYD2-mediated methylation causing the nuclear translocation of b-catenin and activation of Wnt/b catenin signaling pathway [12], a hallmark of a large proportion of human cancers. A higher methylation activity leads to an increased nuclear translocation activity for b-catenin, then to a high activation of the Wnt/b-catenin pathway and cancer cell progression. However lower activity could produce the contrary effect, leading to cancer cell death apoptosis, hence a higher resistance to the cisplatine action.

Identified genetic polymorphism show neighborhood enrichments of chromatin functional annotations in rs4655246 with enhancer and promoter functions (i.e., 11_TxEnh3, H3K4me3_Pro, H3K27ac_Enh) (Roadmap Epigenomics Consortium, 2015). Even out of the promoter regions, this could suggest a cryptic promoter region modulating the expression of alternative regulatory transcripts, but to date only one alternative transcript has been described in placenta tissues.

We do not have any available data for somatic mutations and methylation in cancer cells of those patients, and further studies will be needed to clarify the significance of SMYD2 polymorphisms.

Another interesting finding from our study is *RP11–472N19.3.*, a long non-coding RNA (lncRNA) locus. LncRNAs are normally found as endogenous cellular RNAs, larger than 200nt, and lacking an open reading frame of significant length. They are functional RNA elements, expressed at low levels in a tissue-specific and time-restricted manner. *RP11–472N19.3*. is transcribed in several tissues, including lung, but to date no phenotype, functional annotation or eQTL have been reported in this locus. The uncommon rs7142050-C allele was associated to a better prognosis, suggesting RP11–472N19.3 as a possible new candidate therapeutic target for lung cancer treatment. Based on several evidences (score 2b RegulomeDB, Version 1.1.) the variant rs7142050, is likely to affect binding of several transcription factors such as IRF4 (*Interferon Regulatory Factor 4*), SPI1 (*Spi-1* Proto-Oncogene), and ATF2 (*Activating transcription factor 2*). ATF2 is a transcription factor involved in stress and DNA damage which has been recently involved in cisplatin resistance in non-small cell lung cancer. LncRNAs are regarded with increasing interest as new targets for cancer therapy. Dysregulation of lncRNA expression has been implicated in lung cancer etiology, oncogenic or tumor suppressive. Zhou et al., proposed a eight-lncRNA signature as an effective independent prognostic molecular biomarker in the prediction of NSCLC patient survival [57]. Recent studies, using RNAi experiments to inhibit *HOTAIR* (Hot Transcription Antisense RNA), have reported a decreased migration, invasion and metastasis in NSCLC cells along with reduced expression of genes involving and antisense RNA inhibitory process. Similar results were reported for *MALAT1 (Metastasis Associated Lung Adenocarcinoma Transcript 1)* in mouse lung cancer models [14].

In addition to the single analysis, we performed a pathway enrichment analysis to analyze all excluded signals (pvalue $> 1 \times 10^{-5}$) from the replication phase. With this analysis we highlighted several pathways involved in differential clinical outcome. Some of the identified signals were in primary retained regions with a suggestive profile but were discarded prior to the replication phase (*PAX7*, *IRX5*, or *ATF1)*. *SMYD2* has been identified in the pathway enrichment analysis belonging to one of the statistically significant overrepresented pathways; GO:0006351 (OR = 2.4, p-value = 0.036), a wide functional category that includes transcription regulator activity genes. Furthermore, it is interesting to note two of the enriched pathways. The cAMP-mediated signaling pathway (GO:0019933) is the second most significantly associated term (OR = 13.60, p value = 0.0054), with 5 genes out 22 associated to clinical outcome (*ADM*, *EIF4, EBP2*, *PDE4D*, *RAPGEF2*, *PCLO)*. Among them, EBP2 (EBNA1-binding protein (homolog)) and PDE4D (Phosphodiesterase 4) are relevant as therapeutic targets for lung cancer therapy. EBP2 has been reported as a novel binding partner of c-Myc, regulating the function of nucleolar c-Myc, cell proliferation and tumorigenesis [28], and PDE4D has been reported as a promoter of proliferation and angiogenesis of lung cancer [42]. Moreover, the Wnt signaling pathway (GO:0016055) was overrepresented, with 10 out 150 genes (*HHEX, PITX2, TLE3, TLE4, FZD4, PYGO1, WWOX, CXXC4, NKD2, RSPO2H*)(OR = 3.35 p value = 0.03). In NSCLC it has been reported that Wnt ligand and Fzd are overexpressed and that Wnt antagonists are downregulated [37]. The same authors suggest that

elevation of the β-catenin pathway is a common mechanism for conferring resistance to cancer treatment, not only to EGFR tyrosine kinase inhibitors (TKIs), but also to other types of treatment, including chemotherapy and radiotherapy. In NSCLC, a study reported inherited genetic variation in the Wnt signaling pathway contributing to variable clinical outcomes for patients with early-stage disease [8]. The involvement in NSCLC, but in different stage could indicate a common mechanism related to resistance in both phases of the disease.

In the last years, genetic analysis of somatic variation has yielded valuable profiles for lung cancer subtype classification and prediction of response to treatment [4,24]. Individual germline genetic configuration could help to improve disease management and guide treatment choice decisions. GWAS has been used successfully to identify susceptibility genes to lung cancer, has also been used to identify prognostic and predictive biomarkers to response in early [50,55] or advanced NSCLC patients [22,27,45,54], as well as to analyze adverse effects of drug treatment [6,7,49]. Any of the genes uncovered in our study has been previously reported in advanced NSCLC patients treated with chemotherapy. Most of the reported GWAS (GWAS catalog) are from Asian ancestry populations (11 out 12), and, until now, only one study is from European ancestry patients [54]. Other study using mixed ancestry data come from a different approach using cell lines in the discovery phase. As seen for susceptibility factors, ethnic differences could account for these inconsistencies.

Here, using a genomewide screening approach, we have identified a gene with potential clinical value in advanced NSCLC patients treated with chemotherapy. It is noteworthy that our approach takes advantage of massive variation information collected in deep sequencing derived public panels to empower the study. Identified SMYD2 variant have been genotyped by imputation, and inferred genotypes predicted by IMPUTE2 info shown a highly concordance (average for all inferred variants 96.9%) with genotyping. As widely reported elsewhere, these results corroborate the power of SNP imputation using sequencing derived panels for improving genome scanning results.

We identified 20 regions in the exploratory sample, and even if those signals did not reach genomewide significance, we have replicated one region in an independent sample. All signals remained significant in the joint analysis, however, heterogeneity analysis for replicated variants precluded any joint meta-analysis interpretation (median, mean $I^2 = 92.2\%$, 84.9%). We can discard a genetic bias from different ethnicity since both cohorts are from the same wide-geographical area (Spain), and share the same ancestry; or from genotyping platform, or imputation, since a high correlation was observed in our study among imputation panels. But slight differences were present regarding stage, histological type, and ECOG status that could account for these heterogeneous values. Even if clinical regimens are standardized we cannot underscore the effect of these differences between cohorts. Moreover, the treatment choices in both cohorts were slightly different, and therefore even if we account for these differences in the analysis, in the BREC cohort, we cannot overcome if present the distinct effect of cisplatin and the other dual combination chemotherapies (cisplatin-gemcitabine, cisplatin-docetaxel) in the genetic variant effects. Cisplatin enters cells via multiple pathways, and forms DNA-platinum adducts initiating a cellular self-defense system resulting in cancer cell destruction. Since resistance is supposed to be pleiotropic, these differences do not invalidate the identified signals. In the same way, a pleiotropy of alterations could be related to natural or acquired resistance [16].

Data dimensionality in genome wide analyses is a major concern when applied to clinical cohort series, generally composed by a small number of patients. In order to increase the robustness of the results, our study only considered signals with a reasonable effect in a two-stage design. The large effect size observed for *SMYD2* alleles in the BREC cohort should be considered with caution, since overestimation of the initial effect size could be present. In addition, we cannot discard that other genomic mutations, further than EGFR mutations, could be confounding the results.

5. Conclusion

In conclusion, our study identified germline genetic variation in *SMYD2* associated to bad clinical outcome (PD) in first-line platinum-based treatment in advanced NSCLC patients. These results support the biological significance of methylation process in human carcinogenesis, and open up new drug targeting possibilities and patient stratification in lung cancer therapy based on germline profiling. *SMYD2* profiling could represent an additional prognostic biomarker to better tailor multidisciplinary treatment of patients.

Clinical practice points

- What is already known about this subject?
  Tumor genomic profiling of advanced NSCLC patients determines an increase in the overall survival rates when matched therapies are provided compared with cytotoxic chemotherapy
  In advanced NSCLC patients under first-line cytotoxic chemotherapy, tumor profiling is always a tardy option.
  Furthermore, repeat tissue biopsies should be avoided and sometimes genomic profiling is precluded due to exhausted sample.
  Alternative, germline variants are identified as a valuable prognostic marker in those patients (e.g. DNA-repair genes, *CTNNB1* or *CMKLR1*).
- What are the new findings?
  In this article, we have show that genetic variation on SMYD2 is a biomarker for a bad outcome and reduced overall survival of advanced NSCLC patients when risk alleles are carried at germinal level.
  Multivariate survival analysis showed that genetic variants were independent prognostic factors.
  We report evidences of SMYD2 genetic variation impact on its own expression, and support the biological significance of methylation process of SMYD2 in human carcinogenesis.
- How might it impact on clinical practice in foreseeable future?

Evidences for SMYD2 genetic variation lead to new drug targeting possibilities.

SMYD2 alleles could be used as a biomarker for patient stratification in lung cancer therapy prior to tumor genomic profiling.

Acknowledgement

This study makes use of data generated by the GCAT |Genomes for Life Project. Cohort study of the Genomes of Catalonia", PMPPC-IGTP. *Obón-Santacana et al 2018 (BMJ Open2018;0:e018324. doi:10.1136/bmjopen-2017-018324). A full list of the investigators who contributed to the generation of the data is available from www.genomesforlife.com.* We are grateful to F. J. Gómez-Santonja for giving help on the statistical analysis. The authors also thank Laia Ramos and Raquel Pluvinet from Genomics and Bioinformatics Unit for genotyping support.

Authors' contributions

RdC, JY, and IGF conceived and designed the study. RdC, IGF, JY, SC, EJ, CC, AC, JLR, RR and LS contributed to the generation, collection and assembly. RdC, MG, JMM, XD, SC and IGF contributed to the analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis). RdC, JY, LS, JMM, MP, TK, SC, EJ and IGF contributed to writing, review, and/or revision of the manuscript. All authors approved the final version of the manuscript.

Funding

This study was funded, by the Ministerio de Economía y Competitividad (ADE10/00026) and Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) (SGR 1269). R de Cid is granted by the Ramón y Cajal (RYC) Program (RYC-2011–07822). IGTP is part of the CERCA Program / Generalitat de Catalunya.

Conflicts of interest

None.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at https://dx.doi.org/10.1016/j.ctarc.2018.02.003.

References

[1] H. Ahmed, S. Duan, C.H. Arrowsmith, D. Barsyte-Lovejoy, M. Schapira, An integrative proteomic approach identifies novel cellular SMYD2 substrates, J Proteome Res 15 (2016) 2052–2059.
[2] A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature 526 (2015) 68–74.
[3] C. Benner, C.C. Spencer, A.S. Havulinna, V. Salomaa, S. Ripatti, M. Pirinen, FINEMAP: efficient variable selection using summary data from genome-wide association studies, Bioinformatics 32 (2016) 1493–1501.
[4] L. Bonanno, C. Costa, M. Majem, J.J. Sanchez, A. Gimenez-Capitan, I. Rodriguez, A. Vergnenegre, B. Massuti, A. Favaretto, M. Rugge, et al., The predictive value of 53BP1 and BRCA1 mRNA expression in advanced non-small-cell lung cancer patients treated with first-line platinum-based chemotherapy, Oncotarget 4 (2013) 1572–1581.
[6] S. Cao, C. Wang, H. Ma, R. Yin, M. Zhu, W. Shen, J. Dai, Y. Shu, L. Xu, Z. Hu, H. Shen, Genome-wide association study on platinum-induced hepatotoxicity in non-small cell lung cancer patients, Sci Rep 5 (2015) 11556.
[7] S. Cao, S. Wang, H. Ma, S. Tang, C. Sun, J. Dai, C. Wang, Y. Shu, L. Xu, R. Yin, et al., Genome-wide association study of myelosuppression in non-small-cell lung cancer patients with platinum-based chemotherapy, Pharmacogenomics J 16 (2016) 41–46.
[8] A. Coscio, D.W. Chang, J.A. Roth, Y. Ye, J. Gu, P. Yang, X. Wu, Genetic variants of the Wnt signaling pathway as predictors of recurrence and survival in early-stage non-small cell lung cancer patients, Carcinogenesis 35 (2014) 1284–1291.
[9] C.C. Chang, C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, Gigascience 4 (2015) 7.
[10] H.S. Cho, S. Hayami, G. Toyokawa, K. Maejima, Y. Yamane, T. Suzuki, N. Dohmae, M. Kogure, D. Kang, D.E. Neal, et al., RB1 methylation by SMYD2 enhances cell cycle progression through an increase of RB1 phosphorylation, Neoplasia 14 (2012) 476–486.
[11] O. Delaneau, B. Howie, A.J. Cox, J.F. Zagury, J. Marchini, Haplotype estimation using sequencing reads, Am J Hum Genet 93 (2013) 687–696.
[12] X. Deng, R. Hamamoto, T. Vougiouklakis, R. Wang, Y. Yoshioka, T. Suzuki, N. Dohmae, Y. Matsuo, J.H. Park, Y. Nakamura, Critical roles of SMYD2-mediated beta-catenin methylation for nuclear translocation and activation of Wnt signaling, Oncotarget 8 (2017) 55837–55847.
[13] F. Dudbridge, A. Gusnanto, Estimation of significance thresholds for genomewide association scans, Genet Epidemiol 32 (2008) 227–234.
[14] M. Eissmann, T. Gutschner, M. Hammerle, S. Gunther, M. Caudron-Herger, M. Gross, P. Schirmacher, K. Rippe, T. Braun, M. Zornig, S. Diederichs, Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development, RNA Biol 9 (2012) 1076–1087.
[15] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D.M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012, Int J Cancer 136 (2015) E359–E386.
[16] L. Galluzzi, L. Senovilla, I. Vitale, J. Michels, I. Martins, O. Kepp, M. Castedo, G. Kroemer, Molecular mechanisms of cisplatin resistance, Oncogene 31 (2012) 1869–1883.
[17] J. Goffin, C. Lacchetti, P.M. Ellis, Y.C. Ung, W.K. Evans, First-line systemic chemotherapy in the treatment of advanced non-small cell lung cancer: a systematic review, J Thorac Oncol 5 (2010) 260–274.
[18] P.D. Gottlieb, S.A. Pierce, R.J. Sims, H. Yamagishi, E.K. Weihe, J.V. Harriss, S.D. Maika, W.A. Kuziel, H.L. King, E.N. Olson, et al., Bop encodes a muscle-restricted protein containing MYND and SET domains and is essential for cardiac differentiation and morphogenesis, Nat Genet 31 (2002) 25–32.
[19] R. Hamamoto, G. Toyokawa, M. Nakakido, K. Ueda, Y. Nakamura, SMYD2-dependent HSP90 methylation promotes cancer cell proliferation by regulating the chaperone complex formation, Cancer Lett 351 (2014) 126–133.
[20] M.A. Hildebrandt, J. Gu, X. Wu, Pharmacogenomics of platinum-based chemotherapy in NSCLC, Expert Opin Drug Metab Toxicol 5 (2009) 745–755.
[21] B.N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, PLoS Genet 5 (2009) e1000529.
[22] L. Hu, C. Wu, X. Zhao, R. Heist, L. Su, Y. Zhao, B. Han, S. Cao, M. Chu, J. Dai, et al., Genome-wide association study of prognosis in advanced non-small cell lung cancer patients receiving platinum-based chemotherapy, Clin Cancer Res 18 (2012) 5507–5514.
[23] J. Huang, L. Perez-Burgos, B.J. Placek, R. Sengupta, M. Richter, J.A. Dorsey, S. Kubicek, S. Opravil, T. Jenuwein, S.L. Berger, Repression of p53 activity by Smyd2-mediated methylation, Nature 444 (2006) 629–632.
[24] E. Jantus-Lewintre, E. Sanmartin, R. Sirera, A. Blasco, J.J. Sanchez, M. Taron, R. Rosell, C. Camps, Combined VEGF-A and VEGFR-2 concentrations in plasma: diagnostic and prognostic implications in patients with advanced NSCLC, Lung Cancer 74 (2011) 326–331.
[25] E. Jantus-Lewintre, R. Sirera, A. Cabrera, A. Blasco, C. Caballero, V. Iranzo, R. Rosell, C. Camps, Analysis of the prognostic value of soluble epidermal growth factor receptor plasma concentration in advanced non-small-cell lung cancer patients, Clin Lung Cancer 12 (2011) 320–327.
[26] S. Lê, J. Josse, F. Husson, FactoMineR: an R Package for Muktivariate Analysis, Journal of Statistical Software 25 (2008) 1–18.
[27] Y. Lee, K.A. Yoon, J. Joo, D. Lee, K. Bae, J.Y. Han, J.S. Lee, Prognostic implications of genetic variants in advanced non-small cell lung cancer: a genome-wide association study, Carcinogenesis 34 (2013) 307–313.
[28] P. Liao, W. Wang, M. Shen, W. Pan, K. Zhang, R. Wang, T. Chen, Y. Chen, H. Chen, P. Wang, A positive feedback loop between EBP2 and c-Myc regulates rDNA transcription, cell proliferation, and tumorigenesis, Cell Death Dis 5 (2014) e1032.
[30] M.J. Machiela, S.J. Chanock, LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants, Bioinformatics 31 (2015) 3555–3557.
[31] J. Marchini, B. Howie, Genotype imputation for genome-wide association studies, Nat Rev Genet 11 (2010) 499–511.
[32] A. Matakidou, R. el Galta, E.L. Webb, M.F. Rudd, H. Bridle, T. Eisen, R.S. Houlston, Genetic variation in the DNA repair genes is predictive of outcome in lung cancer, Hum Mol Genet 16 (2007) 2333–2340.
[33] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor, Genome Biol 17 (2016) 122.
[34] H. Mi, A. Muruganujan, J.T. Casagrande, P.D. Thomas, Large-scale gene function analysis with the PANTHER classification system, Nat Protoc 8 (2013) 1551–1566.
[35] T. Moran, J. Wei, M. Cobo, X. Qian, M. Domine, Z. Zou, I. Bover, L. Wang, M. Provencio, L. Yu, et al., Two biomarker-directed randomized trials in European and Chinese patients with nonsmall-cell lung cancer: the BRCA1-RAP80 Expression Customization (BREC) studies, Ann Oncol 25 (2014) 2147–2155.
[36] M. Nakakido, Z. Deng, T. Suzuki, N. Dohmae, Y. Nakamura, R. Hamamoto, Dysregulation of AKT pathway by SMYD2-mediated lysine methylation on PTEN, Neoplasia 17 (2015) 367–373.
[37] A. Nakata, R. Yoshida, R. Yamaguchi, M. Yamauchi, Y. Tamada, A. Fujita, T. Shimamura, S. Imoto, T. Higuchi, M. Nomura, et al., Elevated β-catenin pathway as a novel target for patients with resistance to EGF receptor targeting drugs, Scientific Reports 5 (2015) 13076.
[40] L. Piao, D. Kang, T. Suzuki, A. Masuda, N. Dohmae, Y. Nakamura, R. Hamamoto, The histone methyltransferase SMYD2 methylates PARP1 and promotes poly(ADP-ribosyl)ation activity in cancer cells, Neoplasia 16 (2014) 257–264 (264 e252).
[41] A.S. Pires-Luis, M. Vieira-Coimbra, F.Q. Vieira, P. Costa-Pinheiro, R. Silva-Santos, P.C. Dias, L. Antunes, F. Lobo, J. Oliveira, C.S. Goncalves, et al., Expression of histone methyltransferases as novel biomarkers for renal cell tumor diagnosis and prognostication, Epigenetics 10 (2015) 1033–1043.
[42] S.S. Pullamsetti, G.A. Banat, A. Schmall, M. Szibor, D. Pomagruk, J. Hanze, E. Kolosionek, J. Wilhelm, T. Braun, F. Grimminger, et al., Phosphodiesterase-4 promotes proliferation and angiogenesis of lung cancer by crosstalk with HIF, Oncogene 32 (2013) 1121–1134.
[43] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, P.C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses, Am J Hum Genet 81 (2007) 559–575.
[44] N. Reynoird, P.K. Mazur, T. Stellfeld, N.M. Flores, S.M. Lofgren, S.M. Carlson, E. Brambilla, P. Hainaut, E.B. Kaznowska, C.H. Arrowsmith, et al., Coordination of stress signals by the lysine methyltransferase SMYD2 promotes pancreatic cancer, Genes Dev 30 (2016) 772–785.
[45] Y. Sato, N. Yamamoto, H. Kunitoh, Y. Ohe, H. Minami, N.M. Laird, N. Katori, Y. Saito, S. Ohnami, H. Sakamoto, et al., Genome-wide association study on overall survival of advanced non-small cell lung cancer patients treated with carboplatin and paclitaxel, J Thorac Oncol 6 (2011) 132–138.
[46] K. Shiraishi, T. Kohno, C. Tanai, Y. Goto, A. Kuchiba, S. Yamamoto, K. Tsuta, H. Nokihara, N. Yamamoto, I. Sekine, et al., Association of DNA repair gene polymorphisms with response to platinum-based doublet chemotherapy in patients with non-small-cell lung cancer, J Clin Oncol 28 (2010) 4945–4952.
[47] R. Sirera, R.M. Bremnes, A. Cabrera, E. Jantus-Lewintre, E. Sanmartin, A. Blasco, N. Del Pozo, R. Rosell, R. Guijarro, J. Galbis, et al., Circulating DNA is a useful prognostic factor in patients with advanced non-small cell lung cancer, J Thorac Oncol 6 (2011) 286–290.
[48] N. Spellmon, J. Holcomb, L. Trescott, N. Sirinupong, Z. Yang, Structure and function of SET and MYND domain-containing proteins, Int J Mol Sci 16 (2015) 1406–1428.
[49] X.L. Tan, A.M. Moyer, B.L. Fridley, D.J. Schaid, N. Niu, A.J. Batzler, G.D. Jenkins, R.P. Abo, L. Li, J.M. Cunningham, et al., Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving

platinum-based chemotherapy, Clin Cancer Res 17 (2011) 5801–5811.
[50] S. Tang, Y. Pan, Y. Wang, L. Hu, S. Cao, M. Chu, J. Dai, Y. Shu, L. Xu, J. Chen, et al., Genome-wide association study of survival in early-stage non-small cell lung cancer, Ann Surg Oncol 22 (2015) 630–635.
[51] B. Wang, J.M. Cunningham, X.H. Yang, Seq. 2pathway: an R/Bioconductor package for pathway analysis of next-generation sequencing data, Bioinformatics 31 (2015) 3043–3045.
[52] R. Wang, X. Deng, Y. Yoshioka, T. Vougiouklakis, J.H. Park, T. Suzuki, N. Dohmae, K. Ueda, R. Hamamoto, Y. Nakamura, Effects of SMYD2-mediated EML4-ALK methylation on the signaling pathway and growth in non-small cell lung cancer cells, Cancer Sci. (2017).
[53] H.J. Westra, M.J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M.W. Christiansen, B.P. Fairfax, K. Schramm, J.E. Powell, et al., Systematic identification of trans eQTLs as putative drivers of known disease associations, Nat Genet 45 (2013) 1238–1243.
[54] X. Wu, Y. Ye, R. Rosell, C.I. Amos, D.J. Stewart, M.A. Hildebrandt, J.A. Roth, J.D. Minna, J. Gu, J. Lin, et al., Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy, J Natl Cancer Inst 103 (2011) 817–825.
[55] K.A. Yoon, M.K. Jung, D. Lee, K. Bae, J.N. Joo, G.K. Lee, H.S. Lee, J.S. Lee, Genetic variations associated with postoperative recurrence in stage I non-small cell lung cancer, Clin Cancer Res 20 (2014) 3272–3279.
[56] G.M. Zhang, F.N. Wan, X.J. Qin, D.L. Cao, H.L. Zhang, Y. Zhu, B. Dai, G.H. Shi, D.W. Ye, Prognostic significance of the TREK-1 K2P potassium channels in prostate cancer, Oncotarget 6 (2015) 18460–18468.
[57] M. Zhou, Y. Sun, W. Xu, Z. Zhang, H. Zhao, Z. Zhong, J. Sun, Comprehensive analysis of lncRNA expression profiles reveals a novel lncRNA signature to discriminate nonequivalent outcomes in patients with ovarian cancer, Oncotarget 7 (2016) 32433–32448.