Gene-gene Interaction of *AhR* With and Within the *Wnt* Cascade Affects Susceptibility to Lung Cancer

Albert Rosenberger (arosenb@gwdg.de)
University Medical Center, Georg-August-University Göttingen

Nils Muttray
University Medical Center, Georg-August-University Göttingen

Rayjean J Hung
Lunenfeld-Tanenbaum Research Institute, University of Toronto

David C Christiani
Harvard T.H. Chan School of Public Health and Massachusetts General Hospital, Harvard Medical School

Neil E Caporaso
National Cancer Institute

Geoffrey Liu
Princess Margaret Cancer Centre

Stig E Bojesen
Herlev and Gentofte Hospital

Loic Le Marchand
University of Hawaii Cancer Center

Demetrios Albanes
National Cancer Institute

Melinda C Aldrich
Vanderbilt University Medical Center

Adonina Tardon
University of Oviedo

Guillermo Fernández-Tardón
University of Oviedo

Gad Rennert
Clalit National Cancer Control Center at Carmel Medical Center and Technion Faculty of Medicine

John K Field
The University of Liverpool

Michael P.A. Davies
The University of Liverpool

Triantafillos Liloglou
The University of Liverpool

Lambertus A Kiemeney

Radboud University Medical Center

Philip Lazarus

Washington State University

Bernadette Wendel

University Medical Center, Georg-August-University Göttingen

Aage Haugen

National Institute of Occupational Health

Shanbeh Zienolddiny

National Institute of Occupational Health

Stephen Lam

British Columbia Cancer Agency

Matthew B Schabath

H. Lee Moffitt Cancer Center and Research Institute

Angeline S Andrew

Geisel School of Medicine

Eric J Duell

Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL)

Susanne M Arnold

University of Kentucky

Gary E Goodman

Swedish Medical Group

Chu Chen

Fred Hutchinson Cancer Research Center

Jennifer A Doherty

Huntsman Cancer Institute, University of Utah

Fiona Taylor

University of Sheffield

Angela Cox

University of Sheffield

Penella J Woll

University of Sheffield

Angela Risch

University of Salzburg and Cancer Cluster Salzburg

Thomas R Muley

University Hospital Heidelberg

Mikael Johansson

Umeå University

Paul Brennan

International Agency for Research on Cancer, World Health Organization

Maria Teresa Landi

National Cancer Institute

Sanjay S Shete

The University of Texas MD Anderson Cancer Center

Christopher I Amos

Baylor College of Medicine

Heike Bickeböller

University Medical Center, Georg-August-University Göttingen

Research Article

Keywords: susceptibility, association, prediction, polygenic risk score, decision trees

Posted Date: June 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-658360/v1

Abstract

Introduction: Aberrant *Wnt* signalling, regulating cell development and stemness, is observed in many cancer entities. Aryl hydrocarbon receptor (*AhR*) mediates tumorigenesis of environmental pollutants. Complex interaction patterns of genes assigned to *AhR/Wnt*-signalling were recently associated to lung cancer susceptibility. Aim: To assess the association and predictive ability of *AhR/Wnt*-genes with lung cancer in cases and controls of European descent.

Methods: Odds ratios (OR) were estimated for genomic variants assigned to the genes *DKK2*, *DKK3*, *DKK4*, *FRZB*, *SFRP4* and *Axin2* and other lung cancer-related genes. Logistic regression models with variable selection were trained, validated and tested to predict lung cancer. Further, decision trees were created to investigate variant x variant interaction. All analyses were performed for overall lung cancer and for subgroups.

Results: No association with overall lung cancer was observed, but within the subgroups of ever smokers (e.g. maker rs2722278 SFRP4; OR=1.20; 95%-CI: 1.13-1.27; p=5.6 10^{-10}) and never smokers. Although predictability is poor, *AhR/Wnt-variants* are unexpected overrepresented in optimized prediction scores for overall lung cancer and for small cell lung cancer. Remarkable, the score for never-smokers contained solely two *AhR/Wnt-variants*. The optimal decision tree for never smokers consists of 7 *AhR/Wnt-variants* and only two lung cancer variants, no assigned to any CHRN gene.

Conclusions: The role of variants belonging to *Wnt*/ *AhR*-pathways in lung cancer susceptibility may be underrated in main-effects association analysis. Complex interaction patterns in individuals of European descent have moderate predictive capacity for lung cancer or subgroups thereof, especially in never smokers.

1 Introduction

Genome-wide association studies (GWAS) have identified dozens of susceptibility loci throughout the genome that are associated with lung cancer (LC) or one of its histological subtypes [1–8]. Genes related to *Wnt* signalling, one of the key pathway regulating cell development and stemness, were not detected as being associated to LC susceptibility in individuals of European descent so far, unlike *TERT* (5p15.33), which was among the first [9]. However, aberrant *Wnt* signalling is observed in many cancer entities, but related susceptibility genes were found for none [10–12]. Administration of RNAi against *Wnt* was shown to reduce tumour burden in lung adenocarcinoma (adenoLC) [13]. In non-small cell lung cancer (NSCLC), overexpressed *miR-582-3p* maintains stemness features by negatively targeting the regulators of *Wnt* signalling *Axin2, DKK3* and *SRP1* for degradation, thereby increasing β-catenin mediated *Wnt* activity [14]. *TERT* expression was found to be directly enhanced by binding of β-catenin to its promoter region and thereby links telomerase activity to *Wnt* signalling [10]. This is inasmuch important, as *TERT* is one of the first susceptibility genes for LC identified by GWAS [15, 16]. The tight regulatory machinery of the *Wnt* pathway has several major antagonists, such as Secreted Frizzled related protein (*sFRP*), Dickopff 5 (*DKK*) protein and *Axin2* protein [17]. Aryl hydrocarbon receptor (7p21.1; *AhR*) is a ligand induced transcription

factor, which is translocated into the nucleus. It is known to mediate the toxicity and tumorigenesis of a variety of environmental pollutants, including for NSCLC. *AhR* upregulates the enzyme CYP1A1 when cells are exposed to carcinogenic metabolism, such as some polycyclic aromatic hydrocarbons (PAHs) found in cigarette smoke. The CYP1A1 coding gene is discussed as susceptibility gene for LC. *AhR* is a major determinant in the process of smoking driven LC [18–20]. The complexity of both the *AhR* signalling pathway and the Wnt signalling cascade is reflected by interaction effects of genomic variants within genes, which control their function [21].

Recently, the association of the *Wnt*-genes *DKK4* (8p11.21), *DKK3* (11p15.3), *DKK2* (4q25), *FRZB* (2q32.1, also known as *sFRP3*), *SFRP4* (7p14.1), *Axin2* (17q24.1) and a potential interaction with *AhR* was investigated with respect to the susceptibility to LC in a sample of 600 subjects from North India [21, 22]. A notable association with LC, e.g. for the *SFRP4* variant rs1802073 (OR = 3.19; 95%-CI 1.81–5.63), was reported. Classification And Regression Tree (CART) analysis revealed an interaction of *DKK2* and *SFRP4* polymorphisms to be the best (off all investigated) predictors for LC; especially within smokers. They also reported to have identified several high-risk subgroups in smokers, e.g. characterised by *DKK2* (rs17037102 / rs419558) and *Axin2* (rs9915936). A similar picture was observed in a sample of 270 subjects from Istanbul, Turkey [23]. A two-way interaction between DKK3 (rs3206824) and SFRP4 (rs1802074) was found to be predictive of LC.

We aimed to assess a possible association of *AhR* pathway and *Wnt* signalling cascade with LC within the large-scale series of cases and controls of European descent hold by the International Lung Cancer Consortium (ILCCO) / Integrative analysis of Lung Cancer Etiology and Risk (INTEGRAL). We also evaluated the predictive ability of these genes as a complement to known LC-related markers.

2 Materials And Methods

The work presented has been reviewed and approved by the ILCCO Steering Committee.

2.1 Cases and Controls

Phenotype and genotype data of 58,181 entries of the data repository of ILCCO were extracted. Details of the repository is described previously.[1, 24] QC control samples, individuals without information on smoking status or age, and samples of poor genotyping quality or sex discrepancies, were excluded, To avoid population stratification, this analysis is focused on European-ancestry population (defined as more than 95% probability of being of European descent). 14,068 LC-cases and 12,390 cancer-free controls of European descent remained for analysis. Those genotyped with other genome-wide array in addition to OncoArray were separated to form an independent validation set (extra test set) of size (n = 4,359, including 2,360 LC-cases and 1,999 controls).

2.2 Selected Markers

For this investigation we extracted the genotypes of 113 genomic variants (markers) assigned to 58 genes, previously reported as associated to LC or one of its histological subtypes [1–8] or proxies thereof (called

LC-marker), and 296 markers assigned to 7 genes involved in *Wnt* signalling and listed in Bahl et al. [21, 22] and Yilmaz et al. [23] (called *AhR/Wnt-marker*). Thus, we focused this analysis to genes previously investigated with respect to LC. Fifty of these 409 markers were eliminated before analysis due to a MAF < 1% (minor allele frequency), or departure from HWE (Hardy–Weinberg equilibrium) in genotypes (unaffected $p < 10^{-7}$, affected $p < 10^{-12}$), or low imputation accuracy (info < 0.8). Seventy-eight of the remaining *LC-markers* were genotyped with the OncoArray (44 thereof are proxy SNPs identified using LDlink [25]); 32 needed to be imputed. Two-hundred twenty-one of the remaining *AhR/Wnt-markers* were genotyped; 28 have been imputed. A list of these markers extracted from ILCCO OncoArray repository is given in the appendix.

2.3 Association analysis

We first performed association analysis for each marker separately using the program PLINK [26, 27]. Crude (model 1) and adjusted odds ratios (ORs) were estimated along with 95%-confidence intervals within log-additive models. Sex, age and smoking status and the first 3 principal components (PCs) to adjust for population stratification (model 2); and in addition the 6 most significant *LC-markers* (rs55781567, 15q25.1 *CHRNA5;* rs11780471, 8p21.2 *CHRNA2;* rs7705526, 5p15.33 *TERT;* rs56113850, 19q13.2 *CYP2A6;* rs71658797, 1p31.1 AK5; rs11571833, 13q13.1 *BRCA2*) (model 3) were included in adjusted models. ORs were estimated for overall LC, small cell LC (SCLC), squamous cell LC (SqCLC), adenocarcinoma LC (adenoLC), ever smokers, never smokers and young individuals (21 to 55 years of age) as subgroups. We generated QQ-plots for the *AhR/Wnt-markers* and estimated the genomic inflation factor λ.

2.4 Logistic Regression - Predicting models with model selection

We fitted logistic regression models with variable selection to find appropriate polygenic risk scores (PRS) in order to predict the disease (LC) status (affected or unaffected). To avoid multi-collinearity we removed SNPs in LD to another (R^2>0.8, pruning). The remaining entered the models as potential predictors. We performed forward selection until the Bayesian information criterion (BIC, most stringed selection), the *Akaike* information criterion (AIC, less stringent selection, contains in general more predictors) or the sample size corrected AIC (AICC) indicate a best solution (and 10 more selection steps). The resulting PRSs are called BIC-, AIC- and AICC-scores. Note, that for the purpose of model building, the AIC-selection is asymptotically equivalent to cross-validation (CV).[28, 29] To avoid overfitting, we assigned individuals to a training, validation, or testing-set (sub-sample) with a 1/3 probability each. For comparison, we also generated a BIC^{LC}-score with at least one marker, only allowing *LC-markers* to enter the model building. To compare the importance for LC prediction of the sets g of *LC-maker*s and *AhR/Wnt-markers, respectively*, we contrasted the importance-values defined as $I_g = \sum_{m \in g} |\beta_m| \cdot MAF_m$ for each score (MAF_m the minor allele frequency and β_m the logistic regression coefficient of marker m). The superiority of the AIC-scores over the BIC^{LC}-score and the BIC-score was tested applying the nonparametric test of DeLong, DeLong, and Clarke-Pearson 1988 (1-sided) on AUCs of ROC (area under the receiver operation characteristic curve) [30].

2.5 Decision trees

Decision trees were created to examine marker x marker interaction and to predict the LC-status. This was accomplished in the entire sample and in all subgroups defined above. The R packages *rpart* and *DescTools* were used [31, 32]. Since overfitting is a point of concern when building decision trees, the complexity parameter was first optimized applying 10-fold cross-validation, grading the performance on the validation set by Somers' D (concordance of true and predicted LC-status). The ability of the optimal trees to predict the LC-status was then tested within the independent sample of 4,359 cases and controls. True positive (TP) and true negative (TN) rates are given.

All statistical analyses were performed with SAS® 9.4 (RRID: SCR_008567), PLINK 1.90 and 2.0 (RRID:SCR_001757) or R 4.0.2. (RRID:SCR_001905).

2.6 Gene Expression

We extracted information on gene expression from the *Human Protein Atlas* [33, 34] and *LungGENS* [35, 36].

3 Results

3.1 Sample description

The analysed sample consists of 14,068 LC-cases and 12,390 controls with median age of 63. Sixty-three percent were male, 52% of cases and 28% of controls were current smokers. The most frequent histological subtype is adenocarcinoma (38%), followed by squamous cell carcinoma (SqCLC) (26%) and small cell lung cancer (SCLC) (10%). The proportion of never-smokers was largest within the subgroup of adenocarcinoma cases (14%), but almost the same between younger (< 55 years; 10%) and older (9%) cases. Details on smoking status and histological subtypes are presented in Table 1.

Table 1
Smoking by LC status and subgroups

		Never smoker			Ever smoker					
		total	never		former		current		ever[$]	
		n	n	%	n	%	n	%	n	%
control	teenager	6	3	50%	–	–	3	50%	–	–
	young	2,756	948	34%	698	25%	893	32%	217	8%
	old	9,628	2,960	31%	3,572	37%	2,568	27%	528	5%
	all	12,390	3,911	32%	4,270	34%	3,464	28%	745	6%
case	SqCLC	3,692	138	4%	1,257	34%	2,158	58%	139	4%
	SCLC	1,450	48	3%	383	26%	965	67%	54	4%
	other LC	3,629	405	11%	1,200	33%	1,820	50%	204	6%
	AdenoLC	5,297	740	14%	1,989	38%	2,401	45%	167	3%
	young	2,765	281	10%	452	16%	1,945	70%	87	3%
	old	11,303	1,050	9%	4,377	39%	5,399	48%	477	4%
	all	14,068	1,331	9%	4,829	34%	7,344	52%	564	4%
total		26,458	5,242	20%	9,099	34%	10,808	41%	1,309	5%

[$] ..as recorded; SCLC: small cell lung cancer, SqCLC : squamous cell lung cancer, AdenoLC: adenocarcinoma of the lung, other LC: other histological subtypes; teenager: <21 young: years of age, 21 to 55 years of age, old: >55 years of age;

3.2 Association analysis

The p-values for an association of *AhR/Wnt-markers* with LC range from 0.005 (rs12115174; 8p11.21 *DKK4*; OR = 0.9211) to 1 (model 2); with a negligible genomic inflation ($\lambda = 1.02$); a nominally significant association ($10^5 < p \leq 0.05$) was observed for only 8 of the 249 markers (~ 3%). The corresponding point estimates of OR range from 0.88 (rs1053070054; 8p11.21 *DKK4*; p = 0.007) to 1.12 (rs74596148; 7p14.1 *SFRP4*; p = 0.25). A QQ-plot indicates that achieved p-values almost perfectly agree with the expectation of no associated marker (see Fig. 1). P-values and OR are in moderate agreement between the models (e.g. model 2 to model 3: Kendall's rho_p=0.75, rho_{OR}=0.78).

Subgroup analysis: When dividing the cases according to histological subtypes (SCLC; SqCLC and adenoLC) the observation of no detectable association remains. Merely the number of nominally significant association ($10^5 < p \leq 0.05$) increases to 12 (5%) or 21 (8%) of the 249 markers for SqCLC and SCLC, respectively, hence close to the expected type 1 error (Online Resource S-Table 1). When dividing the

cases and controls according to their smoking behaviour (ever and never smokers), genome-wide significance ($p \leq 10^7$) was achieved for 7 and 8 markers, respectively. Another 12 and 3 marker, respectively, were found suggestive significant ($10^7 < p \leq 10^5$) (see Online Resource S-Figure 1). Those markers found associated among ever smokers have mainly been directly genotyped and are assigned to *SFRP4* (but also *DKK4*). E.g. for marker rs2722278 we estimated an OR = 1.20 (95%-CI: 1.13–1.27), yielding a p-value of $5.6\ 10^{-10}$. Those markers found associated among never smokers have mainly been imputed and are assigned to *AXIN2*, but also to *AHR*, *FRZB* and *DKK2*. Marker rs17037102, assigned to DKK2, was the only one found associated with LC by Bahl et al. and in this analysis (see Table 2 and Online Resource S-Table 2). Interestingly, the ORs of these markers estimated by model 3 (additionally adjusted for selected *LC-marker*) differ from that estimated by model 2. They are closer to one and no more significant. E.g. for rs1133683 (*AXIN2*) we observe an OR = 1.27 (95%-CI: 1.19–1.35, $p = 1x10^{-12}$) fitting model 2, but OR = 0.95 (95%-CI: 0.86–1.06, p = 0.3586) fitting model 3.

Table 2
Significantly associated *AhR/Wnt-markers* within never and ever smokers

SNP		Cyto band	MAF	gene	model 2			model 1	model 3
					p-value	OR	95%-CI	OR	OR
never smoker									
imputed	rs202198518	7p21.1	14%	*AHR*	$3.4\ 10^{-13}$	0.72	0.66–0.79	0.71	0.90 $^{n.s.}$
imputed	rs2237297		14%		$9.9\ 10^{-14}$	0.71	0.65–0.78	0.71	0.90 $^{n.s.}$
imputed	rs1133683	17q24.1	42%	*AXIN2*	$1.0\ 10^{-12}$	1.27	1.19–1.35	1.27	0.95$^{n.s.}$
imputed	rs2240307		5%		$7.7\ 10^{-24}$	0.41	0.34–0.49	0.40	0.62 $^{n.s.}$
imputed	rs35285779		9%		$3.2\ 10^{-22}$	0.58	0.52–0.65	0.58	1.10 $^{n.s.}$
imputed	rs35415678		9%		$3.7\ 10^{-19}$	0.62	0.56–0.69	0.62	1.10 $^{n.s.}$
imputed	rs288326	2q32.1	10%	*FRZB*	$2.5\ 10^{-8}$	1.42	1.25–1.60	1.41	0.98 $^{n.s.}$
imputed	rs17037102	4q25	15%	*DKK2*	$7.4\ 10^{-15}$	0.69	0.63–0.76	0.69	1.09 $^{n.s.}$
ever smoker									
genotyped	rs12532321	7p14.1	45%	*SFRP4*	$1.3\ 10^{-9}$	1.14	1.09–1.19	1.15	1.13 $^{s.s.}$

MAF: minor allele frequency; model 1: crude odds ratio (OR); model 2: adjusted for sex, age and smoking status and the first three principal components; model 3: OR additional adjusted for 6 selected *LC-markers*. $^{gw.s.}$ genome-wide significant; $^{s.s.}$ suggestive significant; $^{n.s.}$ not significant ($p > 0.05$).

SNP		Cyto band	MAF	gene	model 2			model 1	model 3
genotyped	rs7811872		36%		$1.3\ 10^{-8}$	0.88	0.84–0.92	0.88	0.88 gw.s.
genotyped	rs10226308		42%		$1.8\ 10^{-8}$	0.88	0.85–0.92	0.89	0.89 gw.s.
genotyped	rs10488617		42%		$1.6\ 10^{-8}$	0.88	0.85–0.92	0.89	0.89 gw.s.
genotyped	rs2722278		16%		$5.6\ 10^{-10}$	1.20	1.13–1.27	1.16	1.20 gw.s.
genotyped	rs2722279		11%		$9.0\ 10^{-9}$	1.22	1.14–1.31	1.17	1.23 gw.s.
genotyped	rs7811420		43%		$7.9\ 10^{-8}$	0.89	0.85–0.93	0.89	0.89 gw.s.
imputed	rs2073664	8p11.21	9%	*DKK4*	$9.4\ 10^{-11}$	1.20	1.14–1.27	1.15	1.08 s.s.

MAF: minor allele frequency; model 1: crude odds ratio (OR); model 2: adjusted for sex, age and smoking status and the first three principal components; model 3: OR additional adjusted for 6 selected *LC-markers*. gw.s. genome-wide significant; s.s. suggestive significant; n.s. not significant ($p > 0.05$).

3.3 Logistic Regression - Predicting models with model selection

Eight *LC-markers* from only eight LC-Genes (*CYP2A6*, *CHRNA5*, *TERT*, *AMICA1*, *CHRNA3*, COPS2, *HCG4* and *CHRNA2*) were selected for the BIC-score (most stringent selection) to predict overall LC. Hence, the BIC-score and the BIC^{LC}-score are identical. In contrast, the AIC-score (identical to the AICC-score) includes 20 *LC-markers* and remarkable 17 *AhR/Wnt-markers*, with *LC-markers* being more important than the *AhR/Wnt-markers* (importance ratio 0.56: 0.34) (see Fig. 2, Online Resource S-Figure 3 and S-Table 3). The ability to distinguish cases and controls from susceptibility genes only was, as expected, poor for each of the scores. (see Online Resource S-Table 4). Only in the training set the performance of the AIC/AICC-score (AUC = 0.607) exceeded those of the BIC/BIC^{LC}-score (AUC = 0.582) significantly ($p < 0.001$). Within the test set (AUCs: 0.577 and 0.576) and the extra test set (AUCs: 0.553 and 0.548), the higher complexity with additional *AhR/Wnt-markers* did not improve discriminability for overall LC ($p = 0.87$ and $p = 0.35$).

Similar score composition and performance was observed for most subgroups. The BIC-scores in the subgroups adenoLC (involved marker LC:*AhR/Wnt* = 6:–), SCLC (3:–) and smokers (7:–) contained *LC-markers* only, whereas *AhR/Wnt-markers* are included even under this stringent variable selection in the

subgroups SqCLC (5:1) and Young (2:2). Between 14 and 31 *AhR/Wnt-markers* entered the subgroup's AIC-scores. We observed a significant higher discriminability (larger AUCs) of the *AhR/Wnt-markers* enriched AIC-scores compared to BIC^{LC} in the subgroup of **SCLC** patients (p = 0.019; AUC_{AIC}=0.577 AUC_{BIC}=0.546) within the test set (see Online Resource S-Figure 4). The selected *AhR/Wnt-markers* contribute to the AIC-score more than twice as much as the *LC-markers* (importance ration 0.60: 1.49). In the extra test set the score-specific AUCs were similar but no more significantly different (AUC_{AIC}=0.564 vs. AUC_{BIC}=0.531; p = 0.08). The AIC-score of this SCLC-subgroup is composed of 12 *LC-markers* (assigned to *CHRNA5*, *HCG4*, *DNAJB4* (4x each), *CYP2A6*, *CHRNA3*, *CHRNA2*, *AMICA1*, *KCNJ4*, *AS1*, *BRCA2*, *EGFL8* and *WNK1* (2x each)) and 27 *AhR/Wnt-markers* (assigned to all *AhR/Wnt*-genes except *DKK3*). However, only one LC patient in the test set (n = 434) and one in the extra test set (n = 164) was recognized as a patient at a threshold of 50% case probability.

Interestingly the BIC-score for **never smokers** was built by only two *AhR/Wnt-markers* (assigned to *AXIN2* and *SFRP4*) but not a single LC-marker. Further, the *LC-markers* are the minority in the composite of the AIC-score (15:23). They also contribute less to the AIC-score than the *AhR/Wnt-markers* (importance ratio of 0.96 : 1.46). The median predicted case probability, in the test set (24.8%) and extra test set (25.6%), exceeds that of controls by 1%- to 2%-points. However, AUC differed neither in the test set (p = 0.13) nor in the extra test set (p = 0.36) significantly. Nevertheless, this observation highlights the value of the *AhR/Wnt-markers* in the subgroup of never smokers.

3.4 Decision trees

The decision tree for overall LC (whole sample) consists off solely a single decision node (rs55781567 assigned to *CHRNA5*), achieving a concordance of Somers' D = 0.0565 in the extra test set (see Online Resource S-Table 5 and S-Figure 2). A single-node decision-tree was also found optimal for young participants (split: rs1051730 assigned to *CHRNA3*), achieving a concordance of Somers' D = 0.096. These two, unsophisticated trees are characterised by balanced TP- (about 62%) and TN-rates (about 44%).

The decision trees for ever smokers, SCLC and SqCLC were more complex achieving Somers' D of 0.007, -0.0005 and 0.0126, respectively. The trees for SCLC and SqCLC are characterised by an extreme TP-rate < 5% and TN-rate > 99%; the tree for Ever Smokers by a TP-rate > 99% and TN-rate < 5%. Remarkable, a marker assigned to *CHRNA5* was always chosen as first and most important split. However, markers assigned to *AhR/Wnt-genes* (*smoker*: *DKK2*; *SCLC*: *FRZB*; *SqCLC*; *DKK2* and *DKK3*) appear at lower-level decision-nodes (Online Resource S-Figures 5, 6, 7 and 8). With the same program settings, no decision tree could be created for adenocarcinoma.

Most notable is the optimal decision tree for the 5,242 **never smokers** (75% LC-cases, 25% controls), the only one that does not contain a marker belonging to the CHRN (*Cholinergic receptors nicotinic subunits*) gene group (see Fig. 3). The tree is built from only two *LC-markers* but 7 *AhR/Wnt-markers*, achieving a concordance of Somers' D=-0.002. One can make out three branches of this tree. Branch I covers two of three individuals (n = 754, 66% of 1141 in the extra test set): All of these are graded as "unaffected" based on only the two *LC-markers*: first decision node (rs885518 assigned to MTAP) and second decision node

(rs7705526 assigned to *TERT*). For branch II an additional node (rs17214897 assigned to *DKK2*) is taken into account, covering a further tenth (9.9%) of never smokers. In this branch, very few subjects of the training set (1.7% within branch II Eq. 0.17% of all never smokers) are graded “affected”. However, one in four individuals of the extra test set belonging to both branches, I and II, is truly “affected” but has not been detected (TP-rate = 0%, TN-rate = 100%). Rated as “affected” appears in the test set only in the third branch III, covering the remaining fourth of never smokers (n = 284 of the extra test set). This third branch requires genotypes of several *AhR/Wnt-markers* assigned to *AHR*, *Axin2*, *DKK2* and/or *SFRP4*. Herein, one in three (n = 97 of the extra test set) is truly “affected” and is given a chance to be correctly identified, which appears in 8 LC-cases (TP-rate = 9%, TN-rate = 88%). We also noted that the histological subtypes are equally distributed between the branches (see Online Resource S-Table 6).

3.5 Gene Expression

AHR, *AXIN2*, *DKK3* are ubiquitously expressed, with RNA expression detected in many tissue and evidence for protein existence. *AXIN2* and *DKK3* are moderate to highly expressed in normal lung tissues according to the Human Protein Atlas [33]. *AhR* is expressed at low levels in macrophage cells of the lung. No expression is reported for other *Wnt/AhR*-genes. (see Online Resource S-Figure 9 and S-Table 8). Significant differential expression is listed in *LungGENS* for *AhR*, *AXIN2 DKK2*, *DKK3* and *SFRP4* [35] (see Online Resource S-Table 7). Further, AhR is reported to be abundantly expressed in solid lung tumours, especially in adenocarcinomas. AhR overexpression was associated with upregulation of IL-6 secretion, which is critical for lung cancer initiation [37]. Detailed information on gene expression is given in the Appendix. In addition, the *DKK*1 serum level was seen as significantly lower in NSCLC and SCLC patients compared to healthy controls [38]. Significant upregulation of *DKK2* expression was found in *APC (*adenomatous polyposis coli*)*-mutated non-SCLC lung cancers [39].

4 Discussion

This investigation was intended to discover association of the *Wnt*-genes *DKK4* (8p11.21), *DKK3* (11p15.3), *DKK2* (4q25), *FRZB* (2q32.1, also known as *sFRP*3), *SFRP4* (7p14.1), *Axin2* (17q24.1) and a potential interaction with *AhR-genes*, to LC in a large sample of 26,458 individuals of European descent. No marginal association of *AhR/Wnt-markers* with overall LC was observed. Interestingly, an accumulation of associated markers was observed splitting the sample by smoking status, where respective markers in ever smokers are assigned to *SFRP4*. The association analysis within never-smokers may reflect complex gene-gene interactions, since single marker association disappears when adjusted by known LC-markers.

Recently, marginal associations of the *AhR/Wnt-markers* were reported for subjects from North India, although in a much smaller sample of about 600 individuals [21, 22]. A notable association with LC, e.g. for the *SFRP4* variant rs1802073 (OR = 3.19; 95%-CI 1.81 – 5.63), was observed. Classification and Regression Tree (CART) analysis revealed an interaction of *DKK2* and *SFRP4* polymorphisms to be the best (off all investigated) predictors for LC; especially within smokers. They also reported to have identified several high-risk subgroups in smokers, e.g. characterised by *DKK2* (rs17037102 / rs419558) and *Axin2* (rs9915936). A similar picture was observed in a sample of 270 subjects from Istanbul, Turkey [23].

We failed to directly replicate the single marker associations reported by Bahl et al. [21, 22] (North India) and Yilmaz et al. [23] (Turkey). The Indian population is known to be a mixture of several subpopulations [40], which can result in spurious associations. E.g. for rs7396187 assigned to DKK3 Bahl et al. reported a protective effect ($OR_{GC+CC\ vs\ CC}$=0.63, 95%-CI: 0.44–0.91, p = 0.01); however, along with a significant departure from HWE in controls (χ^2=15.11, df = 2, p = 0.001). In contrast, the analysed sample was carefully examined for ethnic homogeneity and principal components were used to adjust for population stratification. The reports by Bahl et al. and Yilmaz et al. are themselves contradictory in some details.

Yilmaz at al. reported a two-way interaction between DKK3 (rs3206824) and SFRP4 (rs1802074) to be predictive of LC. Among other constellations, Bahl et al. reported that DKK3 and SFRP4 were placed closely to each other by a Multifactor dimensionality reduction (MDR) for overall LC, while two markers of *SFRP4* were closely placed within smokers. In contrast, markers assigned to *AXIN2*, but also to *AHR*, *FRZB* and *DKK2* were observed as associated within never smokers. According to Bahl et al. markers of AXIN2 and DKK2 were in never smokers closely placed by a MDR, too. The discrepancy between the total sample and the subsample association estimates point to smoking mediated associations.

We agreed with both previous studies in that complex interaction patterns between the investigated genes contribute to LC susceptibility as entirety or within specific subgroups. To discover patterns of *Ahr/Wnt-genes* involved in LC genesis we further changed the focus from significance of association to inclusion in prediction models, and followed two approaches: First, we searched for polygenic risk scores (PRS). Doing so, we add up marker main effects to construct multidimensional scores, optimising model fit (instead of marker preselection by p-value below some threshold), in order to discriminate cases from controls in a somehow ideal way. Complex gene x gene (GxG) interactions are not modelled.

Nevertheless, the proportion of *Ahr/Wnt-genes* entering the scoring models was remarkable large, given that these markers are not, all other candidates however genome-wide significantly associated to LC. This was particularly noticeable for SCLC, since *AhR/Wnt-markers* contribute more than twice as much to the score as *LC-markers*. It is known, that within current smokers, tobacco consumption is strongest associated to SCLC [41]. Moreover, within never smokers, a stringed defined score is made up from only two *AhR/Wnt-markers*, assigned to AXIN2 and SFRP4. However, the discriminative ability of PRSs for LC, contributing markers with significance for main effect at different levels, is in general poor. The AUC of the BIC^{LC} score for overall LC (0.58 in the test set and 0.55 in the extra test set) corresponds to the AUC = 0.54 based on four top LC-genes in a simulated population, as given by the GWAS-ROCS Database (https://gwasrocs.ca/). This may be due to other overpowering risk factors, since models including e.g. age, sex and smoking variables achieve higher AUCs (0.62 to 0.79) [42].

Recently two polygenic risk scores (PRSs) for overall-LC had been developed, validated and assessed with respect to improving eligibility to low-dose computed tomography (LDCT) as the only recommended screening test for lung cancer. Jia et al. [43, 44] build a PRS on 19 genome-wide associated SNPs ($p < 0.5\ 10^{-8}$). Hung et al. [45], integrated their PRS on 128 SNPs, including established LC-related loci and suggestive associated loci selected by LASSO-regression model, into the $PLCO_{all2014}$ risk model. Both approaches have been validated using data from the UK Biobank. While no substantial increase in

discriminability was reported for both set of PRS, both studies were able to show that the age at which a smoker crosses the recommended screening threshold of 1.5% for the 5 or 6-year LC risk depends on the genetic background, which is sufficiently quantified by the PRS examined. Some smokers will be eligible by < 50 years of age, others by > 60 years of age. Hence, constructing reliable PRS, even with small discriminability, may help to improve the performance of LDCT.

Two- and multiway GxG interaction can also contribute to LC susceptibility, rather than just markers with observed (marginal) main effects. GxG interaction is in general less commonly investigated, not only because this requires much larger samples. However, Li et al. [46] found RGL1:RAD51B in overall LC and non-SCLC, SYNE1:RNF43 in adenocarcinoma and FHIT:TSPAN8 in SqCLC to interactively contribute to LC susceptibility. As in the presented data analysis, the impact of these genes would also have been overlooked considering main effects only. Another reason could be that LC itself is just a generic term of several subcategories that differ in terms of LC initiation and require separate PRSs [42, 47]. A third reason of the poor performance may be due to the exclusively concentration on genetic effects, rather than modelling lifelong interaction with the environment as well. E.g. GxE interaction effects for LC have been observed smoking [48], exposure to asbestos fibres [49, 50] and exposure to radon [51, 52].

With this in mind, the data analysis presented shows that the complex interaction of Wnt-related genes has the potential to be part of an adequate risk assessment for never-smokers or in relation to certain histological subtypes of LC.

As a second approach, we constructed decision trees, which mainly depict GxG interaction patterns. Although, the ability to discriminate cases from controls is again poor, CHRNA5 was in general the most important first node for overall LC as in many subgroups. *Ahr/Wnt-genes* are the one that play a complex but important role in at least one quarter of never smokers, as seen before. Remarkable, TERT was central in that branch important for the remaining three quarter of never smoker. This corresponds to a concentration of relevant genes for this subgroup in the *CLPTM1L-TERT* region on chromosome 5, as previously reported by Hung et al. [53]. Out observations confirm the suspicion, that LC in never smokers is a different entity, justified beforehand on differences in epidemiological, clinical and molecular characteristics [47].

We would like to emphasize that this study was not intended to provide a definitive and reliable risk assessment, but rather aimed to examine in depth the LC-relevant complex interaction pattern of *AhR/Wnt-genes* hypnotized by Bahl et al.. Indeed, considering prediction instead of association provides weaker evidence for this, but is valid in view of the large amount of external evidence. The importance of the *Wnt*-signalling pathway and its antagonist's sFRP, DKKs and Axin2 for cancer is outlined in the introduction. One can also assume a connection with the molecular functionality, since involved genes are expressed ubiquitously or in lung tissues. In summary, we were unable to replicate previously reported associations of *Wnt/AhR-markers* with LC. However, we observed a small but significant impact of these genomic variants on PRSs or decision trees to predict LC.

5 Conclusion

The role of markers belonging to *Wnt* signalling and the *AhR* pathway in LC susceptibility may be underrated in main-effects association analysis. Complex interaction patterns in individuals of European decent have moderate predictive capacity for LC or subsets thereof, especially in never smokers.

6 Declarations

6.1 Funding

The National Institutes of Health (7U19CA203654-02/ 397 114564-5111078 Integrative Analysis of Lung Cancer Etiology and Risk) supported this work. CARET is funded by the National Cancer Institute, National Institutes of Health through grants U01 CA063673, UM1 CA167462, R01 CA 111703, RO1 CA 151989, U01 CA167462 and funds from the Fred Hutchinson Cancer Research Center. Other individual funding for participating studies and members of INTEGRAL-ILCCO are listed elsewhere [10, 30]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

6.2 Conflicts of interest/Competing interests

The authors declare no conflict of interest.

6.3 Data availability (data transparency)

The data that support the findings of this study are available from ILCCO/INTEGRAL but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of ILCCO/INTEGRAL.

6.4 Code availability (software application or custom code)

Not applicable

6.5 Authors' contributions

DKK-Manuskript CRediT author statement.docx

6.6 Compliance with Ethical Standards

All participants in this study signed an informed consent, approved by the local internal review board or ethics committee and administered by trained personnel. All consortium research received approval from the Dartmouth Committee for Protection of Human Subjects on 7/30/2014 with id STUDY00023602. All

experimental protocols and other methods used comply with institutional, national, or international guidelines.

References

1. McKay, J. D., Hung, R. J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D. C., … Amos, C. I. (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature Genetics*, *49*(7), 1126–1132. https://doi.org/10.1038/ng.3892
2. Timofeeva, M. N., Hung, R. J., Rafnar, T., Christiani, D. C., Field, J. K., Bickeboller, H., … Landi, M. T. (2012). Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet*, *21*(22), 4980–95. https://doi.org/10.1093/hmg/dds334
3. Rosenberger, A., Sohns, M., Friedrichs, S., Hung, R. J., Fehringer, G., McLaughlin, J., … Bickeböller, H. (2017). Gene-set meta-analysis of lung cancer identifies pathway related to systemic lupus erythematosus. *PLOS ONE*, *12*(3), e0173339. https://doi.org/10.1371/journal.pone.0173339
4. Brenner, D. R., Brennan, P., Boffetta, P., Amos, C. I., Spitz, M. R., Chen, C., … Hung, R. J. (2013). Hierarchical modeling identifies novel lung cancer susceptibility variants in inflammation pathways among 10,140 cases and 11,012 controls. *Hum Genet*, *132*(5), 579–89. https://doi.org/10.1007/s00439-013-1270-y
5. Ji, X., Bossé, Y., Landi, M. T., Gui, J., Xiao, X., Qian, D., … Amos, C. I. (2018). Identification of susceptibility pathways for the role of chromosome 15q25.1 in modifying lung cancer risk. *Nature Communications*, *9*(1), 3221. https://doi.org/10.1038/s41467-018-05074-y
6. Truong, T., Sauter, W., McKay, J. D., Hosgood, H. D., III, Gallagher, C., Amos, C. I., … Hung, R. J. (2010). International Lung Cancer Consortium: coordinated association study of 10 potential lung cancer susceptibility variants. *Carcinogenesis*, *31*(4), 625–633.
7. Wang, Y., Wei, Y., Gaborieau, V., Shi, J., Han, Y., Timofeeva, M. N., … Houlston, R. S. (2015). Deciphering associations for lung cancer risk through imputation and analysis of 12 316 cases and 16 831 controls. *European Journal of Human Genetics*, *23*(12), 1723–1728. https://doi.org/10.1038/ejhg.2015.48
8. Feng, Y., Wang, Y., Liu, H., Liu, Z., Mills, C., Owzar, K., … Wei, Q. (2018). Novel genetic variants in the P38MAPK pathway gene ZAK and susceptibility to lung cancer. *Mol Carcinog*, *57*(2), 216–224. https://doi.org/10.1002/mc.22748
9. Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., … Caporaso, N. E. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet*, *85*(5), 679–91. https://doi.org/10.1016/j.ajhg.2009.09.012
10. Zhan, T., Rindtorff, N., & Boutros, M. (2016). Wnt signaling in cancer. *Oncogene*. https://doi.org/10.1038/onc.2016.304

11. Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120
12. GWAS catalog. (2020, Dezember 22). *The NHGRI-EBI Catalog of human genome-wide association studies*. publicly accessible database. Abgerufen 22. Dezember 2020, von https://www.ebi.ac.uk/gwas/
13. Kerdidani, D., Chouvardas, P., Arjo, A. R., Giopanou, I., Ntaliarda, G., Guo, Y. A., … Tsoumakidou, M. (2019). Wnt1 silences chemokine genes in dendritic cells and induces adaptive immune resistance in lung adenocarcinoma. *Nature Communications*, *10*. https://doi.org/10.1038/s41467-019-09370-z
14. Fang, L., Cai, J., Chen, B., Wu, S., Li, R., Xu, X., … Li, M. (2015). Aberrantly expressed miR-582-3p maintains lung cancer stem cell-like traits by activating Wnt/beta-catenin signalling. *Nat Commun*, *6*, 8640. https://doi.org/10.1038/ncomms9640
15. Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., … Houlston, R. S. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet*, *40*(5), 616–622.
16. Yuan, Y., Lu, C., Xue, L., & Ge, D. (2014). Association between TERT rs2736100 polymorphism and lung cancer susceptibility: evidence from 22 case–control studies. *Tumor Biology*, *35*(5), 4435–4442. https://doi.org/10.1007/s13277-013-1583-4
17. Akiyama, T. (2000). Wnt/beta-catenin signaling. *Cytokine Growth Factor Rev*, *11*(4), 273–82.
18. Chang, J. T., Chang, H., Chen, P. H., Lin, S. L., & Lin, P. (2007). Requirement of aryl hydrocarbon receptor overexpression for CYP1B1 up-regulation and cell growth in human lung adenocarcinomas. *Clin Cancer Res*, *13*(1), 38–45. https://doi.org/10.1158/1078-0432.CCR-06-1166
19. Lin, P., Chang, H., Tsai, W. T., Wu, M. H., Liao, Y. S., Chen, J. T., & Su, J. M. (2003). Overexpression of aryl hydrocarbon receptor in human lung carcinomas. *Toxicol Pathol*, *31*(1), 22–30. https://doi.org/10.1080/01926230390173824
20. Wang, C. K., Chang, H., Chen, P. H., Chang, J. T., Kuo, Y. C., Ko, J. L., & Lin, P. (2009). Aryl hydrocarbon receptor activation and overexpression upregulated fibroblast growth factor-9 in human lung adenocarcinomas. *Int J Cancer*, *125*(4), 807–15. https://doi.org/10.1002/ijc.24348
21. Bahl, C., Singh, N., Behera, D., & Sharma, S. (2017). High-order gene interactions between the genetic polymorphisms in Wnt and AhR pathway in modulating lung cancer susceptibility. *Personalized Medicine*, *14*(6). https://doi.org/10.2217/pme-2017-0018
22. Bahl, C., Singh, N., Behera, D., & Sharma, S. (2017). Association of polymorphisms in Dickopff (DKK) gene towards modulating risk for lung cancer in north Indians. *Future Oncol*, *13*(3), 213–232. https://doi.org/10.2217/fon-2016-0117
23. Yilmaz, M., Donmez, G., Kacan, T., Sari, I., Akgül Babacan, N., Sari, M., & Kilickap, S. (2015). Significant Association Between Polymorphisms of Wnt Antagonist Genes and Lung Cancer: *Journal of Investigative Medicine*, 1. https://doi.org/10.1097/JIM.0000000000000241

24. Amos, C. I., Dennis, J., Wang, Z., Byun, J., Schumacher, F. R., Gayther, S. A., … Easton, D. F. (2017). The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev*, *26*(1), 126–135. https://doi.org/10.1158/1055-9965.EPI-16-0106
25. Machiela, M. J., & Chanock, S. J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics (Oxford, England)*, *31*(21), 3555–3557. https://doi.org/10.1093/bioinformatics/btv402
26. Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 1–16. https://doi.org/10.1186/s13742-015-0047-8
27. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, *81*(3), 559–75. https://doi.org/10.1086/519795
28. Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 44–47.
29. Fang, Y. (2009). Asymptotic Equivalence between Cross-Validations and Akaike Information Criteria in Mixed-Effects Models. *Journal of Data Science*, (9), 15–21.
30. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837–845.
31. Andri, S. (2021). *DescTools: Tools for Descriptive Statistics*. Abgerufen von https://cran.r-project.org/package=DescTools.
32. Therneau, T., Atkinson, B., & Ripley, B. (2019). *rpart: Recursive partitioning for classification, regression and survival trees.* R (≥ 2.15.0). Abgerufen von https://github.com/bethatkinson/rpart, https://cran.r-project.org/package=rpart
33. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220). https://doi.org/10.1126/science.1260419
34. The Human Protein Atlas. (o. J.). Abgerufen 5. Februar 2021, von https://www.proteinatlas.org/
35. LungGENS. (o. J.). Abgerufen 5. Februar 2021, von https://research.cchmc.org/pbge/lunggens/
36. Du, Y., Guo, M., Whitsett, J. A., & Xu, Y. (2015). „LungGENS": a web-based tool for mapping single-cell gene expression in the developing lung. *Thorax*, *70*(11), 1092–1094. https://doi.org/10.1136/thoraxjnl-2015-207035
37. Xue, P., Fu, J., & Zhou, Y. (2018). The Aryl Hydrocarbon Receptor and Tumor Immunity. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.00286
38. Xu, H., Wu, J., Chen, B., Li, M., Tian, Y., He, M., … She, J.-X. (2014). Serum Dickkopf-1 (DKK1) is significantly lower in patients with lung cancer but is rapidly normalized after treatment. *American Journal of Translational Research*, *6*(6), 850–856.

39. Shen, T., Chen, Z., Qiao, J., Sun, X., & Xiao, Q. (2019). Neutralizing monoclonal antibody against Dickkopf2 impairs lung cancer progression via activating NK cells. *Cell Death Discovery*, *5*(1), 1–12. https://doi.org/10.1038/s41420-019-0204-4

40. Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian Population History. *Nature*, *461*(7263), 489–494. https://doi.org/10.1038/nature08365

41. Lee, P. N., Forey, B. A., & Coombs, K. J. (2012). Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC Cancer*, *12*, 385. https://doi.org/10.1186/1471-2407-12-385

42. Katki, H. A., Kovalchik, S. A., Petito, L. C., Cheung, L. C., Jacobs, J., Jemal, A., … Chaturvedi, A. K. (2018, März 7). Implications of Nine Risk Prediction Models for Selecting Ever-Smokers for Computed Tomography Lung Cancer Screening. *Annals of internal medicine*. Ann Intern Med. https://doi.org/10.7326/M17-2701

43. Jia, G., Lu, Y., Wen, W., Long, J., Liu, Y., Tao, R., … Zheng, W. (2020). Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. *JNCI Cancer Spectrum*, *4*(3). https://doi.org/10.1093/jncics/pkaa021

44. Jia, G., Wen, W., Massion, P. P., Shu, X.-O., & Zheng, W. (2021). Incorporating Both Genetic and Tobacco Smoking Data to Identify High-Risk Smokers for Lung Cancer Screening. *Carcinogenesis*, (bgab018). https://doi.org/10.1093/carcin/bgab018

45. Hung, R. J., Warkentin, M. T., Brhane, Y., Chatterjee, N., Christiani, D. C., Landi, M. T., … Amos, C. I. (2021). Assessing Lung Cancer Absolute Risk Trajectory Based on a Polygenic Risk Model. *Cancer Research*, *81*(6), 1607–1615. https://doi.org/10.1158/0008-5472.CAN-20-1237

46. Li, Y., Xiao, X., Bossé, Y., Gorlova, O., Gorlov, I., Han, Y., … Amos, C. I. (2019). Genetic interaction analysis among oncogenesis-related genes revealed novel genes and networks in lung cancer development. *Oncotarget*, *10*(19), 1760–1774. https://doi.org/10.18632/oncotarget.26678

47. Sun, S., Schiller, J. H., & Gazdar, A. F. (2007). Lung cancer in never smokers–a different disease. *Nat Rev Cancer*, *7*(10), 778–90. https://doi.org/10.1038/nrc2190

48. Saccone, N. L., Culverhouse, R. C., Schwantes-An, T.-H., Cannon, D. S., Chen, X., Cichon, S., … Bierut, L. J. (2010). Multiple Independent Loci at Chromosome 15q25.1 Affect Smoking Quantity: a Meta-Analysis and Comparison with Lung Cancer and COPD. *PLOS Genetics*, *6*(8), e1001053. https://doi.org/10.1371/journal.pgen.1001053

49. Liu, C. Y., Stucker, I., Chen, C., Goodman, G., McHugh, M. K., D'Amelio, A. M., Jr., … Christiani, D. C. (2015). Genome-wide Gene-Asbestos Exposure Interaction Association Study Identifies a Common Susceptibility Variant on 22q13.31 Associated with Lung Cancer Risk. *Cancer Epidemiol Biomarkers Prev*, *24*(10), 1564–73. https://doi.org/10.1158/1055-9965.EPI-15-0021

50. Wei, S., Wang, L.-E., McHugh, M. K., Han, Y., Xiong, M., Amos, C. I., … Wei, Q. W. (2012). Genome-wide gene-environment interaction analysis for asbestos exposure in lung cancer susceptibility. *Carcinogenesis*, *33*(8), 1531–1537. https://doi.org/10.1093/carcin/bgs188

51. Rosenberger, A., Hung, R. J., Christiani, D. C., Caporaso, N. E., Liu, G., Bojesen, S. E., … Gomolka, M. (2018). Genetic modifiers of radon-induced lung cancer risk: a genome-wide interaction study in former

uranium miners. *Int Arch Occup Environ Health*. https://doi.org/10.1007/s00420-018-1334-3

52. Lorenzo-González, M., Ruano-Ravina, A., Torres-Durán, M., Kelsey, K. T., Provencio, M., Parente-Lamelas, I., … Barros-Dios, J. M. (2019). Residential radon, genetic polymorphisms in DNA damage and repair-related. *Lung cancer (Amsterdam, Netherlands)*, *135*, 10–15. https://doi.org/10.1016/j.lungcan.2019.07.003
53. Hung, R. J., Spitz, M. R., Houlston, R. S., Schwartz, A. G., Field, J. K., Ying, J., … Gorlova, O. Y. (2019). Lung Cancer Risk in Never-Smokers of European Descent is Associated With Genetic Variation in the 5p15.33 TERT-CLPTM1Ll Region. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, *14*(8), 1360–1369. https://doi.org/10.1016/j.jtho.2019.04.008
54. Du, Y., Kitzmiller, J. A., Sridharan, A., Perl, A. K., Bridges, J. P., Misra, R. S., … Xu, Y. (2017). Lung Gene Expression Analysis (LGEA): an integrative web portal for comprehensive gene expression data analysis in lung development. *Thorax*, *72*(5), 481–484. https://doi.org/10.1136/thoraxjnl-2016-209598

Figures

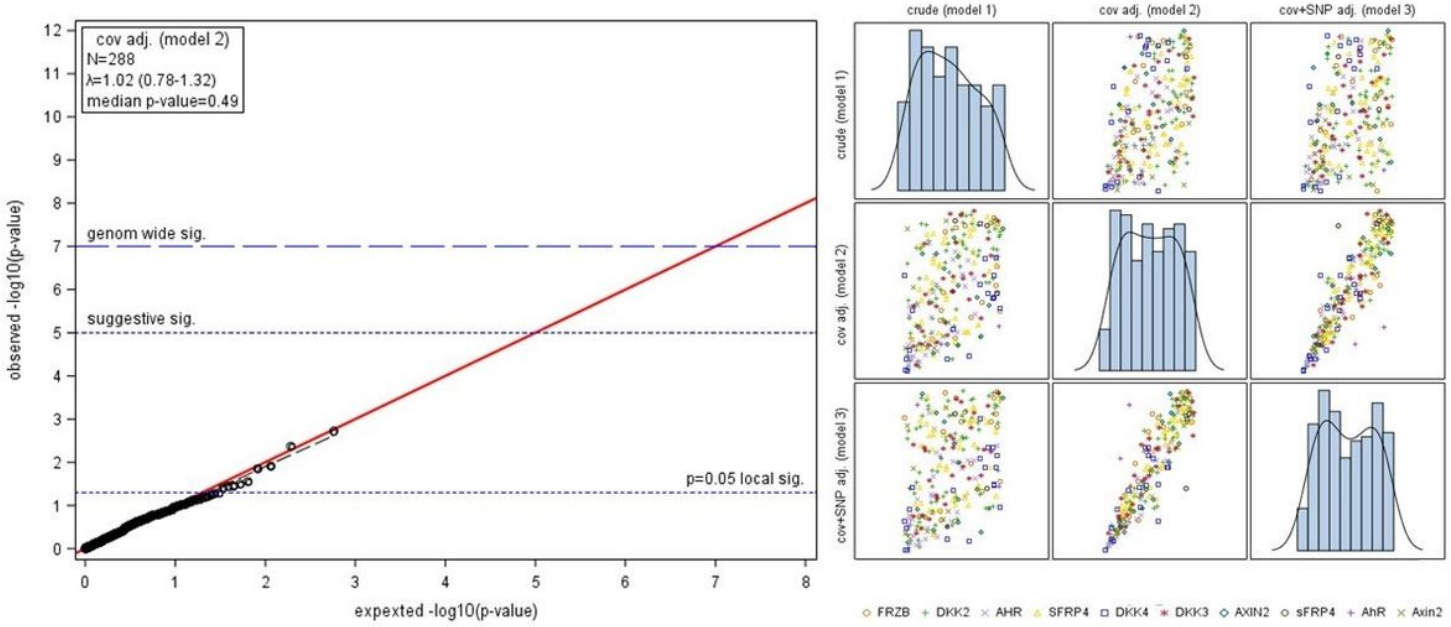

Figure 1

Association of AhR/Wnt-marker. Left panel: QQ-Plot for model 2 (adjusted for sex, age and smoking status and the first three principal components); right panel: matrix of p-values generated by model 1 (crude), model 2 (adjusted for sex, age and smoking status and the first three principal components) and model 3 (additionally adjusted for 6 selected LC-markers)

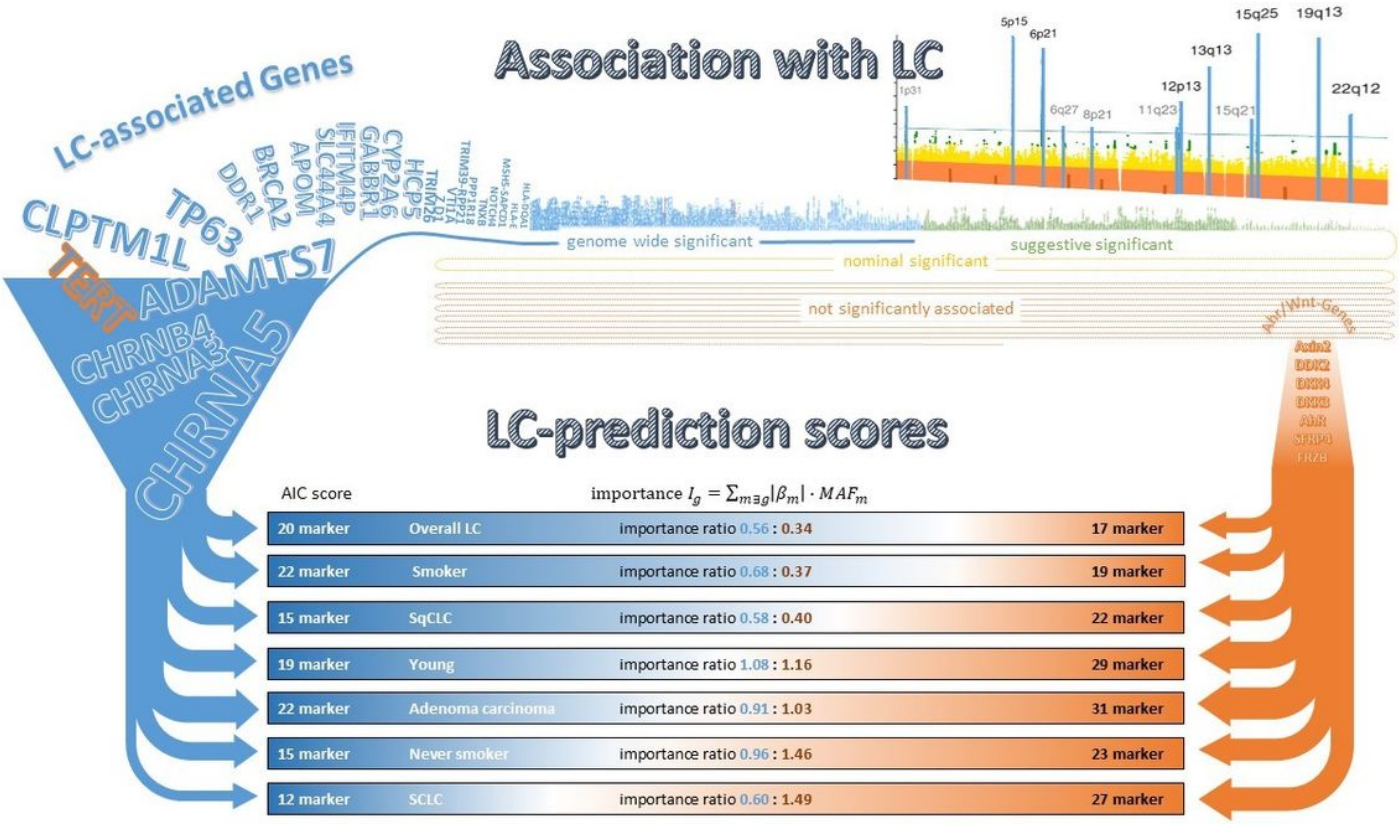

Figure 2

LC-risk score. LC: lung cancer; AIC score: score of a logistic regression model with variant selection according to the Akaike information criterion (AIC); MAFm: minor allele frequency of variant (marker) m; βm regression parameter of variant m; LC-associated genes: previously reported as associated to LC or one of its histological subtypes; Ahr/Wnt genes: selected genes assigned to Wnt-signalling, including AhR; Smoker: ever, former and current smoker; SCLC: small cell lung cancer, SqCLC : squamous cell lung cancer, young: 21 to 55 years of age

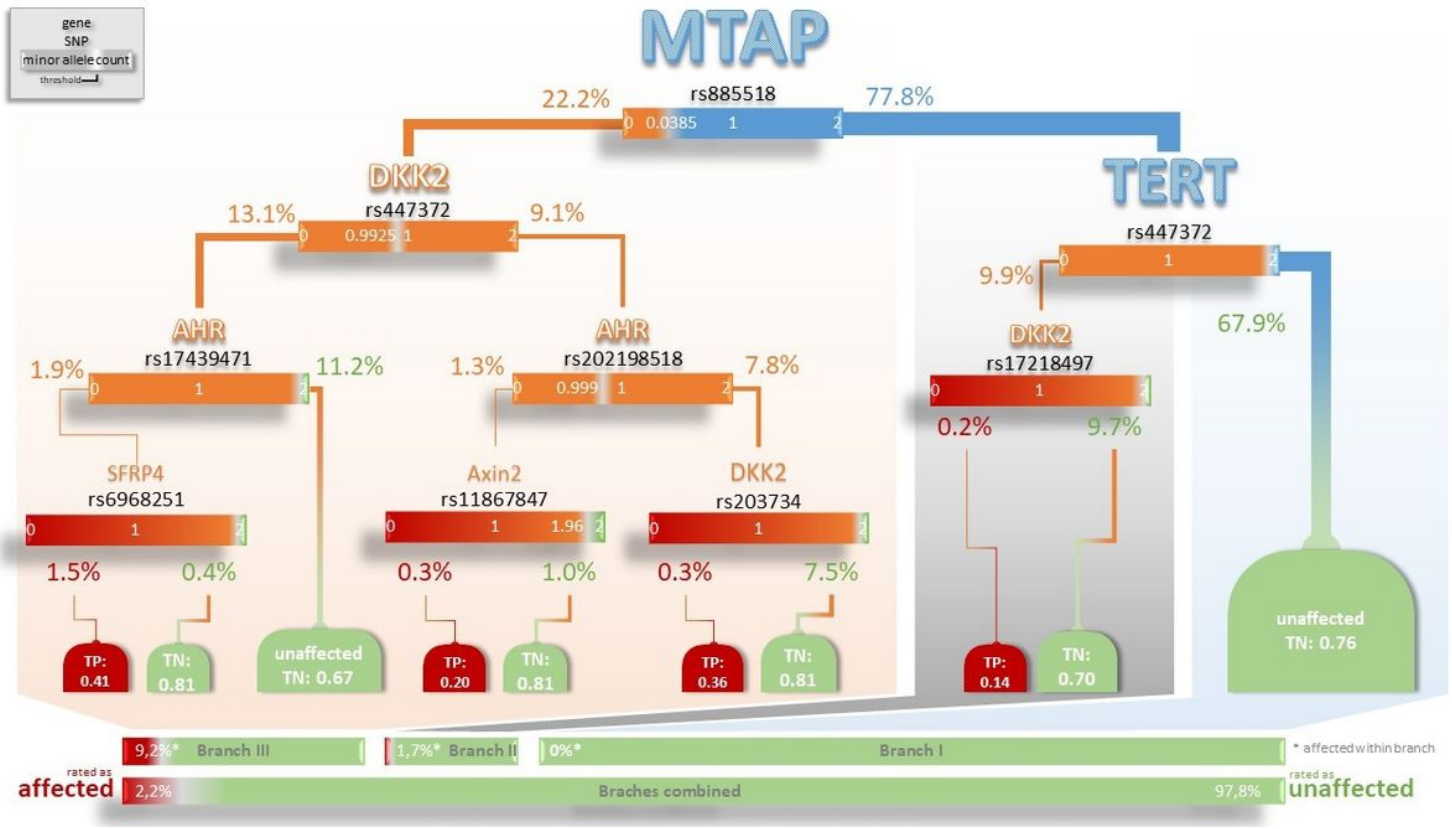

Figure 3

Decision tree for never smoker. Node information: gene name, marker; split information below the node: threshold for minor allele count; blue split nodes: LC-genes, orange split nodes: Ahr/Wnt-genes, decision nodes and bars: green for unaffected; red for affected, TN true negative rate, TP true positive red; the size of gene names, lines and decision notes is proportional to the size of the respective (sub)sample.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf