ORIGINAL ARTICLE

Systematic analyses of regulatory variants in DNase I hypersensitive sites identified two novel lung cancer susceptibility loci

Juncheng Dai^{1,2,†}, Zhihua Li^{1,3,†}, Christopher I. Amos⁴, Rayjean J. Hung^{5,6}, Adonina Tardon⁷, Angeline S. Andrew⁸, Chu Chen⁹, David C. Christiani^{10,11}, Demetrios Albanes^{12,•}, Erik H.F.M. van der Heijden¹³, Eric J. Duell¹⁴, Gad Rennert¹⁵, James D. McKay¹⁶, Jian-Min Yuan^{17,•}, John K. Field¹⁸, Jonas Manjer^{19,20}, Kjell Grankvist²¹, Loic Le Marchand²², M. Dawn Teare²³, Matthew B. Schabath²⁴, Melinda C. Aldrich²⁵, Ming-Sound Tsao²⁶, Philip Lazarus²⁷, Stephen Lam²⁸, Stig E. Bojesen^{29–31}, Susanne Arnold³², Xifeng Wu^{33,•}, Aage Haugen³⁴, Vladimir Janout³⁵, Mikael Johansson³⁶, Yonathan Brhane^{5,6}, Ana Fernandez-Somoano^{7,•}, Lambertus A. Kiemeny¹³, Michael P.A. Davies¹⁸, Shanbeh Zienolddiny³⁴, Zhibin Hu^{1,2} and Hongbing Shen^{1,2,*}

¹Department of Epidemiology, Center for Global Health, International Joint Research Center, School of Public Health and ²Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Jiangsu Collaborative Innovation Center of Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, China, ³Department of Thoracic Surgery, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China, ⁴Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA, ⁵Lunenfeld-Tanenbaum Research Institute, Sinai Health System and ⁶Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada, ⁷Faculty of Medicine, IUOPA, University of Oviedo and CIBERESP, Oviedo, Spain, ⁸Norris Cotton Cancer Center, Geisel School of Medicine, Hanover, NH, USA, ⁹Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, ¹⁰Department of Environmental Health and ¹¹Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA, ¹²Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA, ¹³Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands, ¹⁴Unit of Nutrition and Cancer, Catalan Institute of Oncology (ICO), Barcelona, Spain, ¹⁵Clalit National Cancer Control Center, Carmel Medical Center, Haifa, Israel, ¹⁶International Agency for Research on Cancer (IARC), World Health Organization, Lyon, France, ¹⁷University of Pittsburgh Cancer Institute, Pittsburgh, PA, USA, ¹⁸Roy Castle Lung Cancer Research Programme, Department of Molecular and Clinical Cancer Medicine, University of Liverpool, The William Duncan Building, Liverpool, L7 8TX, UK, ¹⁹Unit for Breast Surgery, Department of Surgery, Lund University, Malmö, Sweden, ²⁰Department of Surgery, Skåne University Hospital, Malmö, Sweden, ²¹Department of Medical Biosciences, Umeå University, Umeå, Sweden, ²²Epidemiology Program, University of Hawai'i Cancer Center, Honolulu, HI, USA, ²³School of Health and Related Research, University of Sheffield, Sheffield, South Yorkshire, UK, ²⁴Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA, ²⁵Department of Thoracic Surgery, Division of Epidemiology, Vanderbilt University Medical Center, Nashville, TN, USA, ²⁶Princess Margaret Cancer Centre, Toronto, Ontario, Canada, ²⁷Department of Pharmaceutical Sciences, College of Pharmacy, Washington State University, Spokane, WA, USA, ²⁸British Columbia Cancer Agency, Vancouver, British Columbia, Canada, ²⁹Copenhagen General Population Study and ³⁰Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Copenhagen, Denmark, ³¹Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, ³²Division of Medical Oncology Markey Cancer Center, Lexington, KY, USA, ³³Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, TX, USA, ³⁴Department of

Received: August 21, 2018; Revised: November 26, 2018; Accepted: December 22, 2018

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com.

Chemical and Biological Work Environment, National Institute of Occupational Health (STAMI), Oslo, Norway,

³⁵Department of Epidemiology and Public Health, Faculty of Health Sciences, Palacky University, Olomouc, Czech Republic

and ³⁶Department of Radiation Sciences, Umeå University, Umeå, Sweden.

*To whom correspondence should be addressed. Tel/Fax: +86 25 868 68437; Email: hbshen@njmu.edu.cn

†These authors contributed equally to this work.

Abstract

DNase I hypersensitive sites (DHS) are abundant in regulatory elements, such as promoter, enhancer and transcription factor binding sites. Many studies have revealed that disease-associated variants were concentrated in DHS-related regions. However, limited studies are available on the roles of DHS-related variants in lung cancer. In this study, we performed a large-scale case-control study with 20 871 lung cancer cases and 15 971 controls to evaluate the associations between regulatory genetic variants in DHS and lung cancer susceptibility. The expression quantitative trait loci (eQTL) analysis and pathway-enrichment analysis were performed to identify the possible target genes and pathways. In addition, we performed motif-based analysis to explore the lung-cancer-related motifs using sequence kernel association test. Two novel variants, rs186332 in 20q13.3 (C>T, odds ratio [OR] = 1.17, 95% confidence interval [95% CI]: 1.10–1.24, $P = 8.45 \times 10^{-7}$) and rs4839323 in 1p13.2 (T>C, OR = 0.92, 95% CI: 0.89–0.95, $P = 1.02 \times 10^{-6}$) showed significant association with lung cancer risk. The eQTL analysis suggested that these two SNPs might regulate the expression of MRGBP and SLC16A1, respectively. What's more, the expression of both MRGBP and SLC16A1 was aberrantly elevated in lung tumor tissues. The motif-based analysis identified 10 motifs related to the risk of lung cancer ($P < 1.71 \times 10^{-4}$). Our findings suggested that variants in DHS might modify lung cancer susceptibility through regulating the expression of surrounding genes. This study provided us a deeper insight into the roles of DHS-related genetic variants for lung cancer.

Abbreviations

AD	adenocarcinoma
CI	confidence interval
DHS	DNase I hypersensitive sites
eQTL	expression quantitative trait loci
GTex	Genotype-Tissue Expression project
LD	linkage disequilibrium
OR	odds ratio
SCC	squamous cell carcinoma
SKAT	sequence kernel association test
SNP	single nucleotide polymorphism
QC	quality control

Introduction

Lung cancer is the most frequently diagnosed cancer and the leading cause of cancer-related death worldwide. It was estimated that there were about 1.8 million new cases (12.9% of the total cancer cases) and 1.6 million deaths (19.4% of the total cancer deaths) worldwide in 2012 (1). In European and North American regions, lung cancer was the third diagnosed cancer but the leading cause of cancer-related death based on GLOBOCAN 2012. Numerous studies have shown that both environmental and genetic risk factors contributed to the process of carcinogenesis (2).

As the first genome-wide association study of lung cancer was performed about 10 years ago, 45 lung cancer susceptibility loci have been identified (3–5). Most of these loci are located in non-coding regions, suggesting that genetic variants in non-coding regions may play vital roles in the development and progress of lung cancer. DNase I hypersensitive sites (DHS), which are important components in the non-coding regions, are abundant in regulatory elements, including promoter, enhancer and transcription factor binding sites (6–8). Previous studies have highlighted that common variants associated with diverse human

diseases and phenotypic traits were concentrated in regulatory DNA marked by DHS (9–11). What's more, portion of the genetic variants in DHS could modify the affinity of DNA transcription factor binding and regulate the expression of target genes (11). Moreover, several studies indicated that trait-associated single nucleotide polymorphisms (SNPs) were more likely to be expression quantitative trait loci (eQTL) (12–14). All these findings suggested that variants (especially the regulatory variants) in DHS played vital roles in disease susceptibility. However, limited studies are available that evaluated the associations between variants of DHS and lung cancer susceptibility.

Recently, Maurano et al. (15) identified 483 415 single nucleotide variants that were likely to affect transcription occupancy by using DNase-seq data from 166 individuals and 114 cell types. These findings provided us an unprecedented opportunity to systematically estimate the associations between regulatory variants in DHS and lung cancer risk. Highlighting these bases, we hypothesized that regulatory variants in DHS could contribute to lung cancer risk by regulating gene expression through disturbing the binding of specific transcription factors.

A large-scale case-control study was performed with a total of 20 871 lung cancer cases and 15 971 controls. Regulatory variants in DHS were systematically screened and the associations between genetic variants and lung cancer risk were further evaluated using logistic regression analysis. Furthermore, we conducted eQTL analysis to explore the potential target genes and underlying biological mechanisms beyond the associations between our identified variants and lung cancer risk. Our study will provide a deeper insight into the roles of DHS-related genetic variants in lung cancer.

Materials and methods

Study subjects

This study was performed based on two European-descent datasets: OncoArray dataset (16) (for screening stage) and Division of Cancer

Epidemiology and Genetics (DCEG) Lung Cancer Study (17) (for validation stage). The OncoArray dataset was applied from the dbGap (phs001273.v1.p1); 43 398 subjects with imputed genotype information were included (16). Samples satisfied any of the following criteria were removed: (i) call rate < 95%; (ii) shared Identity By Descent (IBD) > 0.45; (iii) gender inconsistency. As a result, 18 444 cases and 14 027 controls were retained in the screening stage. For the samples of the validation stage, the data of DCEG Lung Cancer Study were downloaded from dbGap (phs000336.v1.p1), which initially included 3782 cases and 3840 controls, consisting of three cohort studies: the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study; the Prostate, Lung, Colon, Ovary Screening Trial; and the Cancer Prevention Study II Nutrition Cohort. We further eliminated the overlapping individuals between the OncoArray dataset and DCEG Lung Cancer Study (IBD $\geq$ 0.45). As shown in [Supplementary Figure 1](#), available at *Carcinogenesis* Online, the same quality control (QC) process was applied for the imputation data of the DCEG Lung Cancer Study, resulting in 2427 cases and 1944 controls. In total, 20 871 cases and 15 971 controls were available for joint analyses. All participants in this study signed an informed consent form, which was approved by the local internal review boards or ethics committees.

Genotype QC and screening for the regulatory variants

According to the database released by Maurano *et al.* (15), 483 415 single nucleotide variants in DHS could influence the binding of transcript factors. A stringent QC process was applied to these variants. Imputed SNPs that satisfied any of the following criteria were excluded: (i) poor quality: INFO < 0.8, genotype call rate < 95%, MAF (minor allele frequencies) in controls < 0.005, Hardy–Weinberg Equilibrium in cases < 1×10^{-12} or Hardy–Weinberg Equilibrium in controls < 1×10^{-7} ; (ii) SNPs located at major histocompatibility complex region from 29 to 34 Mb on chromosome 6 (NCBI Build 37); (iii) non-regulatory variants: SNPs with CATO (contextual analysis of transcription factor occupancy) value < 0.1; (iv) non-eSNPs: variants with no significant eQTL signals ($P \leq 0.05$, based on Genotype-Tissue Expression project (GTEx) v6p database in 278 lung tissues). As a result, 44 619 regulatory SNPs in DHS were remained for the association analysis. The flowchart and QC process of this study were described in [Supplementary Figure 1](#), available at *Carcinogenesis* Online.

eQTL and pathway-enrichment analysis

For revealing the regulatory relationships between our identified SNPs and the expression of surrounding genes, we performed eQTL analysis based on GTEx V6p database (including 278 normal lung tissues). Furthermore, for these regulated eQTL genes and host genes of our identified SNPs, we analyzed their expression levels in 107 paired lung tumor-normal tissues using the The Cancer Genome Atlas database. To explore the functions and pathways that eQTL genes and host genes enriched, we conducted pathway-enrichment analysis based on GTEx database and DAVID Bioinformatics Resources 6.8 website (18).

Motif-based analysis

In this study, we not only examined the separate effects of single genetic variant, but also aimed to identify lung cancer-related motifs through evaluating the overall effects of all the SNPs in each respective motif. The sequence kernel association test (SKAT) package was used to calculate the joint effects of genetic variants located in the same motif (19,20). SKAT is a powerful approach to identify phenotype-related regions (e.g. genes and other moving windows across the genome) by combining the effects of both common and rare variants in the same region. This method has been widely applied in previous studies and multiple novel trait-related genes were successfully identified, including novel oncogenes in lung cancer (21–23). In our study, all of the 44 619 SNPs were annotated into 308 motifs (only motifs containing more than four SNPs were reserved). Age, gender and principal components were adjusted. We used Bonferroni correction to correct for multiple comparisons.

Statistical analysis

The associations between genetic variants and lung cancer risk were estimated by odds ratios (ORs) and 95% confidence intervals (95% CIs)

using logistic regression analysis. Age, gender and principal components were taking as adjustments. In OncoArray dataset, the principal component analysis was performed using FlashPCA and the first three eigenvalues were adjusted as covariates (16). In DCEG dataset, Plink 1.9 was used to calculate the principal components and the first principal component was included into the regression model. For continuous variables, Student's t-test (equal-variance assumed) was adopted to compare the difference between cases and controls, whereas Fisher's exact test was used for the categorical variables. Gene differential expression analysis was performed using paired Student's t-test based on The Cancer Genome Atlas database. The Pearson's correlation model was applied to carry out gene co-expression analysis. We used meta-analysis to combine the overall effects between the OncoArray dataset and the DCEG Lung Cancer Study. For variants with P value of heterogeneity test ≥ 0.05 , the fixed-effects model was adopted; otherwise, the random-effects model was applied.

Association analyses were conducted using the SNPTTEST (v2.5.4) package and the other analyses were performed using R software (version 3.3.3). The Bonferroni correction was used to account for multiple comparisons.

Results

Basic information of study participants

The demographic characteristics and clinical information of participants in each dataset were summarized in [Supplementary Table 1](#), available at *Carcinogenesis* Online. Briefly, a total of 20 871 cases and 15 971 controls were included in this study. Among these cases, there were 4490 lung squamous cell carcinoma (SCC) cases and 6819 lung adenocarcinoma (AD) cases. The mean age for these cases was 63.79 years old and 61.77 years for controls.

Association between regulatory genetic variants and lung cancer risk

In total, 44 619 regulatory genetic variants in DHS were analyzed using a logistic regression model. Among them, 3069 SNPs were significantly ($P \leq 0.05$) associated with lung cancer risk in the screening stage. Further, 157 of 3069 SNPs were validated in the DCEG Lung Cancer Study. These 157 SNPs that showed significant association with the susceptibility of lung cancer in two stages were remained for meta-analysis. As a consequence, eight SNPs showed statistically significant association with lung cancer susceptibility after taking multiple comparison tests (Bonferroni correction, the cutoff is $0.05/44\ 619 = 1.12 \times 10^{-6}$). The Manhattan plot (44 619 variants, result from meta-analysis) was shown in [Figure 1](#). Table 1 showed the detailed association between our identified eight significant SNPs and lung cancer susceptibility.

Two novel independent lung cancer susceptibility variants were identified

Five of eight significant SNPs were located in 15q25.1 and showed low or medium linkage disequilibrium (LD) with rs55781567 and rs4886591 ([Supplementary Table 2](#), available at *Carcinogenesis* Online), which have been reported in previous genome-wide association study (16). rs12459249 in 19q13.2 was in medium LD with previously reported rs56113850 ($r^2 = 0.54$) (24). Conditional analysis was further performed to evaluate the independence of our identified SNPs. As shown in Table II, five SNPs in 15q25.1 were not significantly associated with lung cancer risk after conditioning on rs55781567 and rs4886591 in OncoArray dataset. Similarly, when conditioning on rs56113850, association between rs12459249 and lung cancer susceptibility was not significant any more. The same result was observed

Figure 1. The Manhattan plot of associations between regulatory variants in DHS and lung cancer risk. The black solid line represented the P value of 1.12×10^{-6} ; five SNPs were located in 15q25.1, the other three SNPs were located in 1p13.2 (rs4839323), 19q13.2 (rs12459249) and 20q13.33 (rs186332).

Table 1. Associations between our identified eight significant SNPs in DHS and lung cancer susceptibility

Characteristics				OncoArray			DCEG GWAS		Meta	
Cytoband	SNP	Gene	CATO ^a	E/R ^b	OR (95% CI) ^c	P	OR (95% CI) ^d	P	OR (95% CI) ^e	P _{fixed}
15q25.1	rs4886592	CHRNA5-CHRNA3-CHRNA4	0.383	C/T	1.13 (1.09–1.16)	2.92×10^{-13}	1.15 (1.05–1.25)	2.54×10^{-3}	1.13 (1.09–1.16)	3.04×10^{-15}
15q25.1	rs4886982	CHRNA5-CHRNA3-CHRNA4	0.106	G/A	1.11 (1.07–1.14)	9.12×10^{-9}	1.10 (1.00–1.21)	3.97×10^{-2}	1.11 (1.07–1.14)	1.04×10^{-9}
15q25.1	rs7182694	CHRNA5-CHRNA3-CHRNA4	0.186	T/C	1.11 (1.07–1.15)	2.13×10^{-8}	1.10 (1.00–1.21)	4.30×10^{-2}	1.11 (1.07–1.15)	2.63×10^{-9}
15q25.1	rs76412132	CHRNA5-CHRNA3-CHRNA4	0.183	T/C	1.11 (1.07–1.15)	8.12×10^{-8}	1.11 (1.01–1.22)	2.45×10^{-2}	1.11 (1.07–1.14)	6.03×10^{-9}
15q25.1	rs76681511	CHRNA5-CHRNA3-CHRNA4	0.166	T/C	1.11 (1.07–1.15)	8.12×10^{-8}	1.11 (1.01–1.22)	2.45×10^{-2}	1.11 (1.07–1.14)	6.03×10^{-9}
19q13.2	rs12459249	CYP2A6	0.168	C/T	1.10 (1.06–1.13)	1.14×10^{-7}	1.17 (1.07–1.28)	9.58×10^{-4}	1.11 (1.07–1.14)	9.44×10^{-10}
20q13.33	rs186332	PTK6	0.149	T/C	1.15 (1.08–1.22)	7.96×10^{-6}	1.24 (1.02–1.52)	2.96×10^{-2}	1.17 (1.10–1.24)	8.45×10^{-7}
1p13.2	rs4839323	MAGI3	0.164	C/T	0.92 (0.89–0.96)	1.23×10^{-5}	0.90 (0.81–0.99)	2.48×10^{-2}	0.92 (0.89–0.95)	1.02×10^{-6}

^aThe CATO (contextual analysis of transcription factor occupancy) score, SNPs with CATO ≥ 0.1 have higher possibility to influence the bindings of transcription factors.

^bE/R: Effect allele and reference allele.

^cAge, gender, principal component analysis (PCA)1, PCA2 and PCA3 were adjusted in OncoArray data.

^dAge, gender and PCA were adjusted in DCEG genome-wide association studies.

^eMeta-analysis (the fixed-effects model) was used to combine the results between these two datasets.

in DCEG Lung Cancer Study. These results suggested that five SNPs in 15q25.1 and rs12459249 in 19q13.2 were not novel lung cancer susceptibility loci. rs186332 in 20q13.33 showed no LD with rs41309931 ($r^2 = 0.015$, European population, 1000 genomes database), which has been reported by McKay *et al.* (16), suggesting that rs186332 was a novel lung cancer risk variant in this region. The 1p13.2 locus has not been reported in previous study. Therefore, rs186332 in 20q13.33 (C>T, OR = 1.17, 95% CI:

1.10–1.24, $P = 8.45 \times 10^{-7}$) and rs4839323 in 1p13.2 (T>C, OR = 0.92, 95% CI: 0.89–0.95, $P = 1.02 \times 10^{-6}$) were novel lung cancer susceptibility variants.

Subgroup analysis

We performed subgroup analysis based on the histologic types of lung cancer: AD and SCC. The same QC process as described earlier was applied in the subgroup analysis. In

the SCC subgroup population, we did not observe any novel significant loci that satisfied multiple comparison. Notably, in the AD subgroup (6819 cases and 14 027 controls), four variants showed significant association with the susceptibility of lung AD (rs77468143 in 15q21.1, T>G, OR = 0.86, 95% CI: 0.82–0.91, $P = 3.07 \times 10^{-9}$; rs55779747 in 3q28, A>C, OR = 0.87, 95% CI: 0.84–0.91, $P = 4.18 \times 10^{-9}$; rs2282245 in 9p21.3, C>T, OR = 1.24, 95% CI: 1.16–1.32, $P = 6.33 \times 10^{-9}$; rs4886592 in 15q25.1, T>C, OR = 1.11, 95% CI: 1.07–1.16, $P = 1.41 \times 10^{-7}$, [Supplementary Table 3](#), available at *Carcinogenesis* Online). However, all of them have been reported (or showed high LD with previously reported variants) in previous studies (16,25). The detailed association results in subgroup populations were shown in [Supplementary Table 3](#), available at *Carcinogenesis* Online.

eQTL analysis

In the interests of exploring the associations between identified significant variants and the expression of their surrounding genes (± 1 Mb), we conducted eQTL analysis based on GTEx database. As shown in [Supplementary Figure 2](#), available at *Carcinogenesis* Online and [Figure 2](#), the T allele of rs186332 showed a significant association with increased expression of MRGBP ($\beta = 0.279$, $P = 0.006$). Notably, the expression of MRGBP in 107 lung tumor tissues was significantly higher than that in paired adjacent tissues ([Supplementary Table 4](#), available at *Carcinogenesis* Online, upregulated in all samples, $P = 2.64 \times 10^{-42}$). Besides, the T allele of rs186332 was significantly associated with the downregulated expression of SRMS ($\beta = -0.478$, $P = 0.006$) and C20orf195 ($\beta = -0.212$, $P = 0.035$). SRMS was significantly elevated in lung tumor tissues ($P = 1.45 \times 10^{-10}$), whereas the expression of C20orf195 in tumor tissues was downregulated ($P = 2.93 \times 10^{-5}$, [Supplementary Figure 3](#), available at *Carcinogenesis* Online). For rs4839323, we found that C allele was associated with increased expression of SLC16A1 ($\beta = 0.09$, $P = 0.033$, [Figure 2](#) and [Supplementary Figure 2](#), available at *Carcinogenesis* Online), a significant upregulated gene in lung tumor tissues (upregulated in 76.64% of samples, $P = 1.25 \times 10^{-13}$, [Figure 2](#)).

Pathway-enrichment analysis

To explore the possible biological functions and potential pathways that eQTL genes and host genes enriched, we conducted pathway-enrichment analysis based on GTEx database and DAVID Bioinformatics Resources 6.8. As shown in [Supplementary Table 5](#), available at *Carcinogenesis* Online, and [Figure 3](#), MAGI3- and PTK6-related genes were abundant in metabolic-related pathways, such as metabolic pathways, propanoate metabolism, and carbon metabolism. Co-expression genes with MRGBP were mainly enriched in signaling transduction-related pathways, such as Ras signaling pathway (Fold enrichment = 2.0, $P = 3.50 \times 10^{-6}$), mitogen-activated protein kinase signaling pathway (Fold enrichment = 1.7, $P = 1.10 \times 10^{-3}$), and Notch signaling pathways (Fold enrichment = 2.6, $P = 4.40 \times 10^{-3}$). All of these pathways played important roles in carcinogenesis process (26,27). Co-expression genes with SLC16A1 were enriched in protein biogenesis or protein-processing-related pathways, for example, protein processing in endoplasmic reticulum (Fold enrichment = 2.0, $P = 5.50 \times 10^{-9}$), ribosome biogenesis in eukaryotes (Fold enrichment = 1.9, $P = 1.80 \times 10^{-6}$), valine, leucine and isoleucine degradation (Fold enrichment = 2.4, $P = 5.10 \times 10^{-3}$). All these pathways participated in the biogenesis and processing of proteins, aberrant changes in these pathways might influence the normal functions of specific proteins.

Motif-based analysis

To investigate the overall effects of motifs on lung cancer risk, we mapped genetic variants to the corresponding motifs and performed motif-based SKAT analysis. In total, there were 292 motifs with more than four genetic variants. As shown in Table III, 10 motifs showed significant association with lung cancer risk at the Bonferroni significance level of $P < 1.71 \times 10^{-4}$ (0.05/292). Among them, motif JDP2_bZIP_1 was the most significant one ($P = 2.77 \times 10^{-7}$) and 397 variants (305 common variants and 92 rare variants) were contained. This motif was the specific binding sites of transcription factor JDP2, which was the member of the activator protein-1 family. For the other motifs, all of them were the binding sites of specific transcription factors, such as runt related transcription factor 2

Table 2. Associations between six significant SNPs and lung cancer risk after conditioning on the genome-wide association studies reported variants in 15q25.1 and 19q13.2

Dataset	SNP	Cytoband	E/R ^c	Condition on rs55781567 ^d		Condition on rs4886591 ^d		Condition on rs56113850 ^e	
				OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
OncoArray ^a	rs4886592	15q25.1	C/T	1.00 (0.97–1.04)	0.856	0.93 (0.86–1.01)	0.095	—	—
	rs4886982	15q25.1	G/A	0.98 (0.94–1.02)	0.352	1.01 (0.96–1.05)	0.803	—	—
	rs7182694	15q25.1	T/C	1.05 (1.01–1.09)	0.008	1.03 (0.99–1.08)	0.124	—	—
	rs76412132	15q25.1	T/C	1.05 (1.01–1.09)	0.007	1.04 (0.99–1.08)	0.100	—	—
	rs76681511	15q25.1	T/C	1.05 (1.01–1.09)	0.007	1.04 (0.99–1.08)	0.100	—	—
	rs12459249	19q13.2	C/T	—	—	—	—	0.99 (0.93–1.05)	0.661
DCEG ^b	rs4886592	15q25.1	C/T	1.01 (0.91–1.12)	0.840	1.02 (0.85–1.21)	0.853	—	—
	rs4886982	15q25.1	G/A	0.97 (0.88–1.08)	0.633	1.01 (0.91–1.13)	0.814	—	—
	rs7182694	15q25.1	T/C	1.05 (0.95–1.15)	0.319	1.03 (0.93–1.15)	0.527	—	—
	rs76412132	15q25.1	T/C	1.06 (0.96–1.17)	0.221	1.05 (0.95–1.16)	0.367	—	—
	rs76681511	15q25.1	T/C	1.06 (0.96–1.17)	0.221	1.05 (0.95–1.16)	0.367	—	—
	rs12459249	19q13.2	C/T	—	—	—	—	1.00 (0.88–1.13)	0.968

^aAge, gender, principal component analysis (PCA)1, PCA2 and PCA3 were adjusted in OncoArray data.

^bAge, gender and PCA were adjusted in DCEG GWAS.

^cE/R: Effect alleles and reference alleles.

^drs55781567 and rs4886591 in 15q25.1 were reported in previous study (PMID: 28604730).

^ers56113850 in 19q13.2 was reported in previous study (PMID: 28604730 and 27488534).

Figure 2. Relationships between genotypes of our identified SNPs and the expression of surrounding genes and differential expression of these genes. (A) rs186332 was significantly associated with the expression of MRGBP based on GTEx v6p database (278 normal lung tissues); (B) The MRGBP was elevated in all 107 tumor tissues compared with adjacent lung tissues based on The Cancer Genome Atlas database; (C) rs4839323 showed significant association with SLC16A1 expression based on GTEx v6p database; (D) The SLC16A1 was significantly upregulated in lung tumor tissues based on The Cancer Genome Atlas database.

(RUNX2), BTB domain and CNC homolog 1 (BACH1) and CCAAT enhancer binding protein beta (CEBPB). Our findings suggested that 10 significant motifs and their corresponding transcription factors might play important roles in the carcinogenesis of lung cancer.

Discussion

DHS region contains various regulatory elements, and genetic variants in DHS may regulate the expression of surrounding genes. Therefore, genetic variants in this region could play important roles in the development of cancer. For lung cancer, many variants reported by genome-wide association study were located in DHS, such as rs2736100 (5p15.33), rs77468143 (15q21.1), and rs55781567 (15q25.1) (28). In this study, we comprehensively evaluated the relationships between regulatory variants in DHS and lung cancer risk using 20 871 lung cancer cases and 15 971 controls from Caucasian population. We successfully identified two novel lung cancer risk loci: rs186332 in 20q13.3 and rs4839323 in 1p13.2.

SNP rs186332 was located in 20q13.33 and contained in the binding site of transcription factor CEBPB. The CATO score of

rs18332 was 0.149, suggesting that this variant could disturb the binding of CEBPB. The same result was observed in JASPAR 2018 database (Supplementary Figure 4, available at Carcinogenesis Online) (29). Genotypes of this variant showed significant association with the expression of MRGBP, SRMS and C20orf195. The SRMS gene encodes a non-receptor tyrosine kinase and participates in the cell differentiation (30). To date, studies about the functions and roles of SRMS and C20orf195 in the tumorigenesis were limited. The MRGBP (also known as C20orf20) was the component of NuA4 histone acetyltransferase complex and participated in the transcriptional activation of target genes (31). Several studies indicated that MRGBP was significantly upregulated in colorectal cancer tissues and knockdown MRGBP could inhibit proliferation of colorectal cancer cells (32,33). Watt et al. (34) found that MRGBP was overexpressed in cutaneous SCC cells and knockdown of MRGBP resulted in reduced tumor cell viability. Furthermore, they revealed that MRGBP small interfering RNA knockdown could reduce tumor growth in vivo, suggesting that MRGBP could be the potential therapeutic targets for cutaneous SCC. In addition, upregulated expression of MRGBP was also observed in pancreatic ductal AD, prostate cancer and cervical cancer cells (35–37). All these studies suggested that

Figure 3. Results of Kyoto Encyclopedia of Genes and Genomes pathway-enrichment analysis. (A) Co-expression genes with PTK6 were mainly enriched in metabolic-related pathways (the number in the right of bar chart represented the fold enrichment in each pathway); (B) Co-expression genes with MRGBP were mainly enriched in signaling transduction-related pathways; (C) Co-expression genes with MAGI3 were mainly enriched in metabolic and biosynthesis of protein-related pathways; (D) Co-expression genes with SLC16A1 were mainly enriched in pathways of biosynthesis and process of proteins; Co-expression genes with SRMS were enriched in metabolic pathways. HTLV-1, human T-cell leukemia virus 1.

Table 3. Associations between 10 significant motifs and lung cancer risk using motif-based SKAT analysis

Motif	N_All ^a	N_Test ^b	N_Rare ^c	N_Common ^d	P	P_Bonferroni ^e
JDP2_bZIP_1	397	397	92	305	2.55×10^{-7}	7.44×10^{-5}
RUNX2_RUNX_3	359	359	67	292	4.88×10^{-6}	1.42×10^{-3}
V_BACH1_01	475	475	78	397	3.00×10^{-5}	8.77×10^{-3}
V_AML1_Q4	507	507	98	409	3.20×10^{-5}	9.33×10^{-3}
CEBPB_bZIP_1	1362	1362	243	1119	3.75×10^{-5}	1.10×10^{-2}
NFIX_NFI_4	737	737	151	586	5.42×10^{-5}	1.58×10^{-2}
V_NRF2_Q4	264	264	52	212	8.06×10^{-5}	2.35×10^{-2}
V_E2A_Q6	170	170	31	139	1.32×10^{-4}	3.86×10^{-2}
CREB5_bZIP_1	994	994	184	810	1.41×10^{-4}	4.13×10^{-2}
MA0136.1-ELF5	376	376	70	306	1.52×10^{-4}	4.45×10^{-2}

^aTotal number of SNPs located in each motif.

^bTotal number of SNPs included in the test.

^cTotal number of rare genetic variants (MAF < 0.05) contained in the test.

^dTotal number of common SNPs contained in the test.

^eP value after taking Bonferroni correction (292 motifs).

MRGBP functioned as an oncogene that could promote the tumorigenesis. Consistent with previous studies, we found that MRGBP was elevated in all 107 lung tumor tissues, suggesting that MRGBP could play similar role in the progression of lung tumorigenesis. Pathway-enrichment analysis indicated that co-expression genes with MRGBP were significantly enriched in signaling transduction-related pathways, such as Ras signaling pathway, mitogen-activated protein kinase signaling pathway and Notch signaling pathway. The risk T allele of rs186332 was significantly associated with upregulated expression of MRGBP, suggesting that rs186332 might regulate the expression of MRGBP. Highlighting these evidences, we speculated that rs186332 could modify lung cancer risk through regulating the expression of MRGBP by influencing the binding of transcription factor CEBPB.

SNP rs4839323 was located in 1p13.2 and in the first intron of MAGI3, which encodes the membrane-associated guanylate kinase (38). The C allele of rs4839323 was not associated with the expression of MAGI3, but showed significant association with increased expression of SLC16A1. The SLC16A1 was a member of solute carrier family 16 and encoded a monocarboxylate transporter that mediated the transport of lactate and pyruvate (39). In addition, SLC16A1 participated in a variety of biological processes, including energy metabolism, activation of T lymphocytes, spermatogenesis, drug metabolism and so on (39). Many studies have revealed that monocarboxylate transporter could influence the sensitivity of tumor cells for antitumor drugs, suggesting that the expression of monocarboxylate transporter could influence the prognosis of patients with cancers (40,41). However, limited studies were available about

the roles that SLC16A1 played in the development of malignant tumors. Fang et al. (42) found that SLC16A1 was upregulated in neuroblastoma and could be a potential therapeutic target for high-risk neuroblastoma. Li et al. (43) revealed that SLC16A1 was significantly elevated in medulloblastoma and knockdown of SLC16A1 by small-interfering-RNA-induced cell death in medulloblastoma cells. In lung cancer, upregulated SLC16A1 expression was observed and increased expression of SLC16A1 was associated with poorer prognosis of patients with lung AD (44). In this study, we further found that SLC16A1 was significantly elevated in lung tumor tissues. On the basis of these findings, rs4839323 might influence lung cancer susceptibility through regulating the expression of SLC16A1.

Apart from this, our motif-based SKAT analysis discovered 10 motifs significantly associated with lung cancer risk. Of these motifs, JDP2 belonged to the component of activator protein-1 transcription factor and involved in a variety of transcriptional responses, such as ultraviolet-induced apoptosis, cell differentiation and tumorigenesis. Previous studies have suggested that JDP2 could act as a tumor suppressor (45–47). However, the function and role of JDP2 in lung cancer was still unclear and more studies are warranted. For the other motifs, their corresponding transcription factors, such as CEBPB, RUNX2 and BACH1, played important roles in the development of tumorigenesis (48–50). More functional studies are needed to reveal their roles in lung cancer.

Besides our findings, there are several limitations that need to be considered. First, this analysis was performed based on the annotation results from multiple tissues and cells because of the limited number of regulatory variants in DHS in A549 cell. Second, for our identified novel variants, we speculated their potential target genes and possible mechanisms just relying on the bioinformatic analysis. Functional assays are warranted to validate our findings.

In conclusion, we systematically screened the regulatory variants in DHS, and two variants were identified associated with lung cancer risk. For the mechanism underlying these observations, we speculated that these SNPs could modify lung cancer susceptibility by regulating related genes' expression. These findings contribute to a better understanding of the importance of genetic variants in DHS and their potential mechanisms.

Supplementary material

Supplementary data are available at *Carcinogenesis* online.

Funding

The Key international (regional) cooperative research project (81820108028); National Natural Science of China (81521004); the Priority Academic Program for the Development of Jiangsu Higher Education Institutions (Public Health and Preventive Medicine); Top-notch Academic Programs Project of Jiangsu Higher Education Institutions (PPZY2015A067).

Acknowledgements

The authors would like to thank the patients and the supporting staff in this study.

Conflicts of Interest Statement: None declared.

Reference

1. Torre, L.A. et al. (2015) Global cancer statistics, 2012. *CA. Cancer J. Clin.*, 65, 87–108.
2. Doll, R. et al. (1981) The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *J. Natl. Cancer Inst.*, 66, 1191–1308.
3. Bossé, Y. et al. (2018) A decade of GWAS results in lung cancer. *Cancer Epidemiol. Biomarkers Prev.*, 27, 363–379.
4. Spinola, M. et al. (2006) Association of the PDCD5 locus with lung cancer risk and prognosis in smokers. *J. Clin. Oncol.*, 24, 1672–1678.
5. Spinola, M. et al. (2007) Genome-wide single nucleotide polymorphism analysis of lung cancer risk detects the KLF6 gene. *Cancer Lett.*, 251, 311–316.
6. Boyle, A.P. et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132, 311–322.
7. ENCODE Project Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447, 799–816.
8. Felsenfeld, G. et al. (2003) Controlling the double helix. *Nature*, 421, 448–453.
9. Michailidou, K. et al.; NBCS Collaborators; ABCTB Investigators; ConFab/AOCS Investigators. (2017) Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551, 92–94.
10. Maurano, M.T. et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337, 1190–1195.
11. Degner, J.F. et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482, 390–394.
12. Nicolae, D.L. et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, 6, e1000888.
13. Dubois, P.C. et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.*, 42, 295–302.
14. Westra, H.J. et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, 45, 1238–1243.
15. Maurano, M.T. et al. (2015) Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.*, 47, 1393–1401.
16. McKay, J.D. et al.; SpiroMeta Consortium. (2017) Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.*, 49, 1126–1132.
17. Landi, M.T. et al. (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, 679–691.
18. Huang da, W. et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, 4, 44–57.
19. Ionita-Laza, I. et al. (2013) Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.*, 92, 841–853.
20. Wu, M.C. et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, 89, 82–93.
21. Wang, Y. et al. (2018) Integrating expression-related SNPs into genome-wide gene- and pathway-based analyses identified novel lung cancer susceptibility genes. *Int. J. Cancer*, 142, 1602–1610.
22. Timbers, T.A. et al. (2016) Accelerating gene discovery by phenotyping whole-genome sequenced multi-mutation strains and using the Sequence Kernel Association Test (SKAT). *PLoS Genet.*, 12, e1006235.
23. Grant, R.C. et al. (2018) Exome-wide association study of pancreatic cancer risk. *Gastroenterology*, 154, 719–722 e3.
24. Patel, Y.M. et al. (2016) Novel association of genetic markers affecting CYP2A6 activity and lung cancer risk. *Cancer Res.*, 76, 5768–5776.
25. Miki, D. et al. (2010) Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. *Nat. Genet.*, 42, 893–896.
26. Kumar, M.S. et al. (2012) The GATA2 transcriptional network is requisite for RAS oncogene-driven non-small cell lung cancer. *Cell*, 149, 642–655.
27. Lim, J.S. et al. (2017) Intratumoural heterogeneity generated by Notch signalling promotes small-cell lung cancer. *Nature*, 545, 360–364.

28. Bossé, Y. et al. (2018) A decade of GWAS results in lung cancer. *Cancer Epidemiol. Biomarkers Prev.*, 27, 363–379.
29. Khan, A. et al. (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, 46(D1), D260–D266.
30. Goel, R.K. et al. (2018) Phosphoproteomics analysis identifies novel candidate substrates of the nonreceptor tyrosine kinase, src-related kinase lacking C-terminal regulatory tyrosine and N-terminal myristoylation sites (SRMS). *Mol. Cell. Proteomics*, 17, 925–947.
31. Cai, Y. et al. (2003) Identification of new subunits of the multiprotein mammalian TRRAP/TIP60-containing histone acetyltransferase complex. *J. Biol. Chem.*, 278, 42733–42736.
32. Carvalho, B. et al. (2009) Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut*, 58, 79–89.
33. Yamaguchi, K. et al. (2010) C20orf20 (MRG-binding protein) as a potential therapeutic target for colorectal cancer. *Br. J. Cancer*, 102, 325–331.
34. Watt, S.A. et al. (2011) Integrative mRNA profiling comparing cultured primary cells with clinical samples reveals PLK1 and C20orf20 as therapeutic targets in cutaneous squamous cell carcinoma. *Oncogene*, 30, 4666–4677.
35. Ding, F. et al. (2017) MRGBP as a potential biomarker for the malignancy of pancreatic ductal adenocarcinoma. *Oncotarget*, 8, 64224–64236.
36. Ito, S. et al. (2014) A genetic screen in *Drosophila* for regulators of human prostate cancer progression. *Biochem. Biophys. Res. Commun.*, 451, 548–555.
37. Scotto, L. et al. (2008) Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. *Genes Chromosomes Cancer*, 47, 755–765.
38. Ma, Q. et al. (2015) MAGI3 suppresses glioma cell proliferation via upregulation of PTEN expression. *Biomed. Environ. Sci.*, 28, 502–509.
39. Halestrap, A.P. (2013) The SLC16 gene family—structure, role and regulation in health and disease. *Mol. Aspects Med.*, 34, 337–349.
40. Afonso, J. et al. (2015) CD147 and MCT1-potential partners in bladder cancer aggressiveness and cisplatin resistance. *Mol. Carcinog.*, 54, 1451–1466.
41. Birsoy, K. et al. (2013) MCT1-mediated transport of a toxic molecule is an effective strategy for targeting glycolytic tumors. *Nat. Genet.*, 45, 104–108.
42. Fang, J. et al. (2006) The H⁺-linked monocarboxylate transporter (MCT1/SLC16A1): a potential therapeutic target for high-risk neuroblastoma. *Mol. Pharmacol.*, 70, 2108–2115.
43. Li, K.K. et al. (2009) miR-124 is frequently down-regulated in medulloblastoma and is a negative regulator of SLC16A1. *Hum. Pathol.*, 40, 1234–1243.
44. Stewart, P.A. et al. (2015) A pilot proteogenomic study with data integration identifies MCT1 and GLUT1 as prognostic markers in lung adenocarcinoma. *PLoS One*, 10, e0142162.
45. Chen, Y.L. et al. (2017) The expression of a tumor suppressor gene JDP2 and its prognostic value in hepatocellular carcinoma patients. *Hum. Pathol.*, 63, 212–216.
46. Heinrich, R. et al. (2004) The c-Jun dimerization protein 2 inhibits cell transformation and acts as a tumor suppressor gene. *J. Biol. Chem.*, 279, 5708–5715.
47. Luo, W.M. et al. (2018) Identification of four differentially methylated genes as prognostic signatures for stage I lung adenocarcinoma. *Cancer Cell Int.*, 18, 60.
48. Marigo, I. et al. (2010) Tumor-induced tolerance and immune suppression depend on the C/EBPβ transcription factor. *Immunity*, 32, 790–802.
49. Zhang, H. et al. (2011) FOXO1 inhibits Runx2 transcriptional activity and prostate cancer cell migration and invasion. *Cancer Res.*, 71, 3257–3267.
50. Cantor, S.B. et al. (2001) BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function. *Cell*, 105, 149–160.