SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
RISK FACTORS
CANCER GENETICS

Received
15 August 2014

Accepted
12 December 2014

Published
16 January 2015

Correspondence and requests for materials should be addressed to H.S. (hbshen@njmu.edu.cn)

* These authors contributed equally to this work.

Systematical analyses of variants in CTCF-binding sites identified a novel lung cancer susceptibility locus among Chinese population

Juncheng Dai[1,2]*, Meng Zhu[1]*, Cheng Wang[1], Wei Shen[1], Wen Zhou[1], Jie Sun[1], Jia Liu[1], Guangfu Jin[1,2], Hongxia Ma[1,2], Zhibin Hu[1,2,3], Dongxin Lin[4] & Hongbing Shen[1,2,3]

[1]Department of Epidemiology and Biostatistics and Ministry of Education (MOE), Key Laboratory for Modern Toxicology, School of Public Health, Nanjing Medical University, Nanjing 211166, China, [2]Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Jiangsu Collaborative Innovation Center For Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, China, [3]State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing 211166, China, [4]State Key Laboratory of Molecular Oncology and Department of Etiology and Carcinogenesis, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China.

Genome-wide association studies identified genetic susceptibility variants mostly lie outside of protein-coding regions. It suggested variants located at transcriptional regulatory region should play an important role in cancer carcinogenesis including lung cancer. In the present study, we systematically investigated the associations between the variants in the binding sites of an extensive transcription factor CTCF and lung cancer risk in Chinese population. A two-stage case-control design was conducted to evaluate the variants located at the uniform CTCF ChIP-seq peaks in a Chinese population (2,331 *vs* 3,077; 1,115 *vs* 1,346). The ChIP-seq data for CTCF, specified on lung cancer cell line A549, were downloaded from ENCODE database. Imputation was performed to increase the genome coverage in the CTCF binding regions. Three variants in CTCF binding sites were found to associate with lung cancer risk in the first stage. Further replication revealed a novel single nucleotide polymorphism rs60507107 was significantly associated with increased risk of lung cancer in two stages (Additive model: OR = 1.19, 95%CI = 1.11–1.27, $P = 6.98 \times 10^{-7}$). Our results indicate that rs60507107 in the binding site of CTCF is associated with an increased risk of lung cancer. This may further advance our understanding of regulatory DNA sequences in cancer development.

Lung cancer has been the most common cancer and the leading cause of death worldwide for several decades. There are estimated to be 1.8 million new cases (12.9% of the total) and 1.59 million deaths (19.4% of the total) in 2012. In China, lung cancer incidence accounts for 21.3% and mortality accounts for 27.1% of all cancer in 2012[1,2]. Although tobacco smoking has been confirmed to be the most common risk factor of lung cancer, just about one-tenth smokers develops lung cancer in their lifetime[3]. It indicates that many unclear factors like genetic risk factors may also play an important role in lung carcinogenesis[4].

Over the past few years, genome-wide association studies (GWAS) of lung cancer have identified more than 40 single nucleotide polymorphisms (SNPs) in 20 genome loci associated with lung cancer risk (Supplementary Table 1)[5]. The loci at 3p28, 5p15, 15q25 and 6p21 have been validated to have contribution to the susceptibility of lung cancer in multiple studies[6–11]. These findings have provided new clues for understanding lung cancer carcinogenesis.

Marc A, *et al*[12] systematically investigated the association of multiple types of ENCODE data with all GWAS identified SNPs and showed that about 20% of the SNPs lay in chromatin immunoprecipitation sequencing (ChIP-seq) peaks. When accounting for SNPs in strong linkage disequilibrium ($r^2 \geq 0.8$) with reported SNPs in CEU, the proportion reached 61%. Moreover, Rory J, *et al*[13] proved that polymorphic binding sites owned different biochemical affinities for transcription factor (TF), resulting in altered TF recruitment and differential reporter gene repression in vivo. Through comprehensive analysis GWAS identified SNPs and TF p53 ChIP-seq data, Jorge Z, *et al*[14] found a SNP rs4590952 (G/A) which influenced cancer risk through changing p53 binding

Table 1 | Associations of the 15 SNPs in discovery stage

SNP[a]	Chr.	BP.	Related gene	Info[b]	Allele	MAF 1000 Genome[c]	MAF Cases	MAF Controls	$P_{Nanjing}$[d]	$P_{Beijing}$[d]	$P_{Combined}$[d]
rs75772117	4	6744829	*BLOC1S4, KIAA0232*	0.81	G > A	0.07	0.08	0.10	3.37×10^{-2}	1.26×10^{-5}	1.09×10^{-6}
rs37010	**5**	**1349535**	***CLPTM1L, SLC6A3***	**0.98**	**A > G**	**0.17**	**0.13**	**0.16**	$\mathbf{4.16 \times 10^{-3}}$	$\mathbf{1.73 \times 10^{-4}}$	$\mathbf{2.67 \times 10^{-6}}$
rs10072980	5	33181214	*LOC340113, TARS*	0.90	T > C	0.14	0.15	0.18	2.88×10^{-3}	9.80×10^{-5}	2.77×10^{-7}
rs1632447	6	29688886	*ZFP57, HLA-F*	0.99	C > A	0.02	0.03	0.02	3.57×10^{-2}	3.79×10^{-7}	2.05×10^{-6}
rs114731497	6	29653824	*ZFP57, HLA-F*	0.98	C > G	0.02	0.04	0.02	8.65×10^{-4}	3.53×10^{-7}	3.22×10^{-8}
rs115786093	6	30451515	*TRIM39-RPP21, HLA-E*	0.92	G > A	0.05	0.04	0.03	1.45×10^{-5}	1.57×10^{-3}	1.31×10^{-6}
rs115364068	6	30325531	*TRIM39-RPP21, HLA-E*	0.98	T > C	0.07	0.07	0.05	1.62×10^{-4}	2.15×10^{-8}	1.88×10^{-9}
rs115299438	6	31164828	*HCG27*	1.00	C > A	0.09	0.10	0.13	4.88×10^{-4}	8.94×10^{-5}	7.32×10^{-6}
rs2252937	6	31461613	*HCG26, MICB*	0.89	T > C	0.08	0.14	0.11	3.38×10^{-3}	7.89×10^{-5}	3.86×10^{-6}
rs2002059	**10**	**101130855**	***CNNM1***	**0.83**	**A > G**	**0.09**	**0.07**	**0.09**	$\mathbf{1.84 \times 10^{-2}}$	$\mathbf{1.39 \times 10^{-8}}$	$\mathbf{1.07 \times 10^{-8}}$
rs3124203	10	30799716	*MAP3K8, LYZL2*	0.92	C > G	0.10	0.07	0.10	3.04×10^{-2}	3.67×10^{-4}	9.54×10^{-6}
rs60507107	**11**	**61463769**	***DAGLA***	**0.99**	**C > T**	**0.36**	**0.39**	**0.35**	$\mathbf{4.99 \times 10^{-3}}$	$\mathbf{8.48 \times 10^{-4}}$	$\mathbf{4.93 \times 10^{-6}}$
rs11617518	13	24290267	*TNFRSF19, MIPEP*	1.00	C > T	0.32	0.32	0.28	6.97×10^{-4}	9.03×10^{-4}	4.07×10^{-6}
rs12100587	14	106004323	*TMEM121, ELK2AP*	0.88	T > G	0.44	0.41	0.47	1.65×10^{-3}	1.16×10^{-9}	1.76×10^{-10}
rs2836333	21	39724700	*KCNJ15, ERG*	0.80	A > G	0.24	0.24	0.27	1.63×10^{-2}	2.71×10^{-4}	5.42×10^{-6}

a: all selected SNPs were identified by Imputation;
b: imputed quality info;
c: minor allele frequency of ASN in 1000 Genome;
d: adjusted by age, gender, pack-year of smoking and pca1.

and had undergone natural selection. All these results raise a new hypothesis that the expression levels of genes can be altered by non-coding SNPs lie within regulatory DNA sequences through altering their affinity for TFs.

CTCF is an extensive transcription factor with 11-zinc finger (ZF) protein domains and involved in many cellular processes, including transcription regulation, insulator activity and regulation of chromatin architecture. It has been reported that CTCF was one of the 127 significantly mutated genes across 12 tumor types[15]. Furthermore, the Chip-seq experiments of CTCF has been carried out and replicated in different laboratories to provide highly credible binding peaks in ENCODE project. However, there was little attention paid to the genetic variants in binding sites of CTCF. In this study, by using ENCODE and our previous Lung Cancer GWAS data, we tried to systematically evaluate the associations between genetic variants located at regulatory DNA sequences of CTCF and lung cancer risk. We first imputed and screened our GWAS data (2,331 cases *vs.* 3,077 controls) in the binding regions of CTCF, and then replicated the associations in another independent population (1,115 cases *vs.* 1,346 controls).

Results

A total of 2,331 cases and 3,077 controls were included in the discovery stage, 1,115 cases and 1,346 controls in the validation stage. The demographic and clinical information is summarized in Supplementary Table 2. Before imputation, only 3,569 genotyped SNPs in the CTCF binding peaks exited in our GWAS data set. After imputation, the data coverage increased more than eight fold, a total of 32,453 qualified SNPs in the binding sites were analyzed in the discovery stage and three SNPs (rs37010 in 5p15.33, rs2002059 in 10q24.2, rs60507107 in 11q12.2) met all the criteria above (Table 1, Figure 1 and Supplementary figure 2–3). However, rs37010 was in strong LD with rs465498 ($R^2 = 0.94$) which had been confirmed related to lung cancer risk in our previous study[16], so it was no longer validated in this study. As a result, two SNPs (rs2002059 and rs60507107) were further evaluated in the validation stage.

The replication results are shown in Table 2. Rs60507107 which located in the intron of DAGLA remained to be significantly associated with risk of lung cancer (OR = 1.13, 95%CI = 1.01–1.27, P = 0.037), consistent with the results of the discovery stage (OR = 1.22, 95%CI = 1.12–1.33, $P = 4.93 \times 10^{-6}$). After combining results from two stages, the results demonstrated that rs60507107 was significantly associated with lung cancer (OR = 1.19, 95%CI = 1.11–1.27) at a P-value of 6.98×10^{-7}, reaching the significance level after multiple comparison (Bonferroni: 1.54×10^{-6} from 0.05/32,453). The combined ORs for the heterozygote (CT) and minor homozygote (TT) are 1.14 (95%CI = 1.03–1.26) and 1.45 (95%CI = 1.25–1.67), respectively, as compared with major homozygote (CC). The regional plot of rs60507107 was shown in Figure 2.

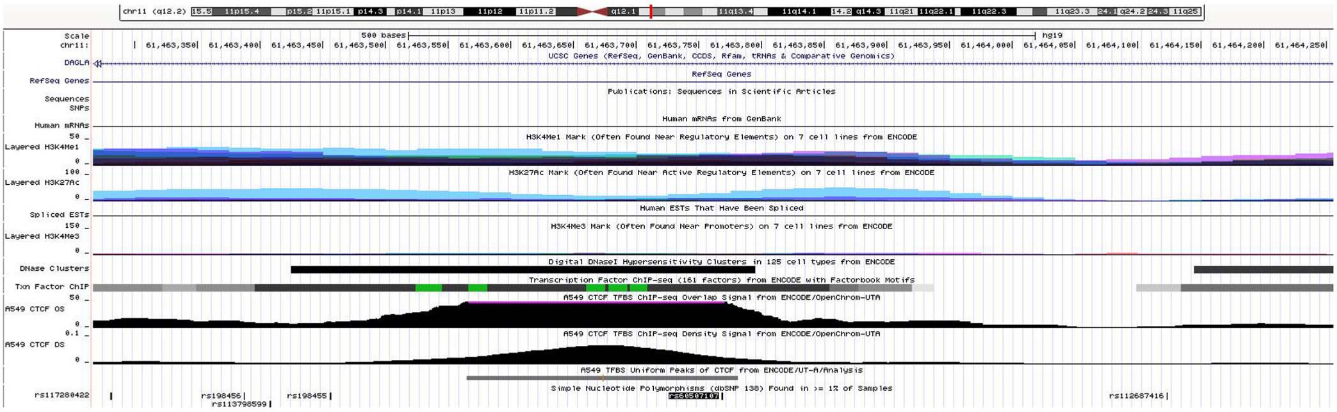
chr11 (q12.2) 15.5 11p15.4 p15.2 11p15.1 p14.3 p14.1 11p13 11p12 11p11.2 q12.1 11q13.4 11q14.1 14.2 q14.3 11q21 11q22.1 11q22.3 11q23.3 24.1 q24.2 24.3 11q25
Scale
chr11:
500 bases
hg19
61,463,350 61,463,400 61,463,450 61,463,500 61,463,550 61,463,600 61,463,650 61,463,700 61,463,750 61,463,800 61,463,850 61,463,900 61,463,950 61,464,000 61,464,050 61,464,100 61,464,150 61,464,200 61,464,250
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
DAGLA
RefSeq Genes
Publications: Sequences in Scientific Articles
Sequences SNPs
Human mRNAs
Human mRNAs from GenBank
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me1
50
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
100
0
Spliced ESTs
Human ESTs That Have Been Spliced
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K4Me3
150
0
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
A549 CTCF TFBS ChIP-seq Overlap Signal from ENCODE/OpenChrom-UTA
A549 CTCF OS
50
0
A549 CTCF TFBS ChIP-seq Density Signal from ENCODE/OpenChrom-UTA
A549 CTCF DS
0.1
0
A549 TFBS Uniform Peaks of CTCF from ENCODE/UT-A/Analysis
Simple Nucleotide Polymorphisms (dbSNP 138) Found in >= 1% of Samples
rs117280422
rs198456
rs113798599
rs198455
rs60507107
rs112687416

Figure 1 | rs60507107 in CTCF TFBS Chip-seq peaks.

Table 2 | Associations of the 2 replicated SNPs in GWAS and replicated stage

SNP		GWAS			Validation			Combined	
(cytoband, BP)	Genotype	Cases/Controls	OR (95%CI)^a	P^a	Cases/Controls	OR (95%CI)^b	P^b	OR (95%CI)^a	P^a
rs2002059	AA	2020/2587	1.00		929/1141	1.00		1.00	
10q24.2, 101130855	AG	306/469	0.73(0.63–0.85)	6.75×10^{-5}	160/185	1.06(0.85–1.34)	5.90×10^{-1}	0.91(0.80–1.04)	1.62×10^{-1}
	GG	5/21	0.23(0.10–0.55)	9.27×10^{-4}	10/7	1.77(0.67–4.68)	2.51×10^{-1}	0.73(0.38–1.39)	3.36×10^{-1}
	Additive		0.62(0.53–0.73)	1.07×10^{-8}		1.11(0.90–1.37)	3.39×10^{-1}	0.90(0.80–1.02)	1.01×10^{-1}
rs60507107	**CC**	**871/1300**	**1.00**		**428/556**	**1.00**		**1.00**	
11q12.2, 61463769	**CT**	**1118/1429**	**1.20(1.05–1.36)**	$\mathbf{5.90 \times 10^{-3}}$	**514/621**	**1.08(0.91–1.28)**	$\mathbf{4.05 \times 10^{-1}}$	**1.14(1.03–1.26)**	$\mathbf{9.87 \times 10^{-3}}$
	TT	**342/348**	**1.53(1.27–1.85)**	$\mathbf{7.87 \times 10^{-6}}$	**171/168**	**1.32(1.03–1.70)**	$\mathbf{2.60 \times 10^{-2}}$	**1.45(1.25–1.68)**	$\mathbf{4.92 \times 10^{-7}}$
	Additive		**1.22(1.12–1.33)**	$\mathbf{4.93 \times 10^{-6}}$		**1.13(1.01–1.27)**	$\mathbf{3.70 \times 10^{-2}}$	**1.19(1.11–1.27)**	$\mathbf{6.98 \times 10^{-7}}$

BP: base position;
a: Adjusted by age, gender, pack-year of smoking and pca1.
b: Adjusted by age, gender and pack-year of smoking.

However, the association between the other SNP rs2002059 observed in discovery stage was not replicated in the validation stage.

Furthermore, subgroup analyses by age, gender, smoking status and histology were conducted for the association of rs60507107 with lung cancer risk (Table 3). No significant heterogeneity between the subgroups were observed at discovery stage and validation stage. The association remained significant in females (OR = 1.29, 95%CI = 1.09–1.53, $P = 0.003$ in discovery stage; OR = 1.27, 95%CI = 1.04–1.56, $P = 0.017$ in validation stage), and subjects whose smoking level less than 25 pack-years (OR = 1.19, 95%CI = 1.07–1.33, $P = 0.002$ in discovery stage; OR = 1.17, 95%CI = 1.02–1.35, $P = 0.025$ in validation stage) in both stages.

Discussion

In our previous GWAS studies, several loci like 5p15.33, 22q12.2 or 12q23.1 were found associated with lung cancer or lung squamous cell carcinoma risk in Chinese[9,16,17]. However, these findings could only explain a small fraction of the heritability of the lung cancer because GWAS mainly focus on the peak associations and usually base on little prior information. In the current study, we systematically evaluated the association of genetic variants lay within the binding cites of CTCF, which had been proved implicated in cancer formation in recent studies, based on ENCODE database and existing GWAS data set, and further replicated the promising associations in an independent case-control study in Chinese population. We finally found a novel SNP rs60507107 located in 11q12.2 was significantly associated with the lung cancer risk, and also showed a similar signal: rs37010 which was at the same LD region with reported loci: chr5p15.33, rs465498 ($R^2 = 0.94$)[18]. The SNP rs37010 located upstream of *CLPTM1* which was associated with cisplatin-induced apoptosis, and mapped to a region of LD upstream of *TERT* which had been demonstrated implicated in carcinogenesis (Figure 2)[19–22].

The novel identified SNP rs60507107 located in the binding sites of CTCF in the first intron of *DAGLA*. CTCF is a transcription factor with 11-zinc finger (ZF) protein domains and involved in many cellular processes, including transcription regulation, insulator activity and regulation of chromatin architecture. It has been found that CTCF can mediate some important biological processes of lung cancer cells such as enhancer-promoter interactions of *TERT* and Rb2/p130 transcription[23,24]. The nuclear protein of CTCF can bind a wide variety of DNA target sequences with different ZF domains and play an important role in epigenetic regulation. It can function as transcriptional activator by binding a histone acetyltransferase (HAT)-containing complex or transcriptional repressor by binding histone deacetylase (HDAC)-containing complex[25]. The CTCF binding region identified in our study, is modified by histone H3K4Me1 and H3K27AC which usually considered as active enhancer. To measure the influence of rs60507107 to CTCF affinity, we searched the JASPA (http://jaspar.genereg.net/) and found a motif with the wild-type allele of rs60507107 "TCCATGGGAATC***G***CT" could bind CTCF with a relative score of 0.70, while the score decreased to 0.61 with the other allele "TCCATGGGAATC***A***CT". Furthermore, rs60507107 had a moderate linkage disequilibrium ($R^2 = 0.586$) with an identified *DAGLA* eQTL SNP rs198464 among CHB + JPT population in 1000 Genomes project[26]. Additionally, the rs198464 also showed a significant association with lung cancer in our GWAS data set (OR = 1.18, 95%CI = 1.09–1.28, $P = 7.44 \times 10^{-5}$). This indicates rs60507107 may be involved in lung carcinogenesis through regulating CTCF binding thus influencing the expression of *DAGLA*.

DAGLA located at region 11q12.2 which had been identified associated with colorectal cancer (CRC) in a large-scale genetic study recently[27]. Previous study also indicated the region involved in hereditary prostate cancer families with primary kindey cancer[28]. These findings suggest the region may participate in some common mechanism of carcinogenesis. *DAGLA* is a diacylglycerol lipase that cata-

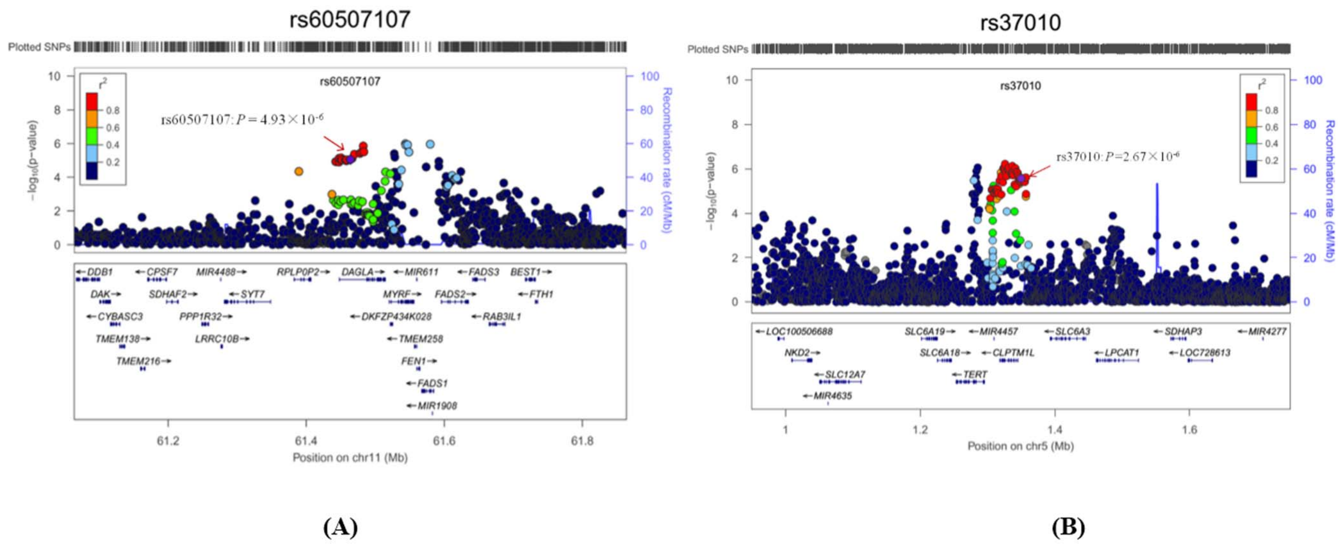

Figure 2 | regional plot of rs60507107 and rs37010.

lyzes the hydrolysis of DAG to 2-arachidonoylglycerol (2-AG), the most abundant endocannabinoid in tissues. The 2-AG is a physiological ligands for the cannabinoid receptors CB1 and CB2, two G protein-coupled receptors which are located in the central and peripheral nervous systems[29]. The cannabinoid ligand-receptor system plays an important role in a variety of physiological processes including appetite, pain-sensation, mood, memory and antitumorigenic properties[30]. Munson AE, *et al*[31] had found that the growth of Lewis lung adenocarcinoma in a mouse model was inhibited after 20 days treatment with cannabinol and Δ8-THC. As a compound of cannabinoid, 2-AG also shows anticancer properties. In glioma cell C6, 2-AG demonstrates antiproliferative effects with IC50 values of 1.8 μM[32]. In breast cancer line MCF-7, 2-AG can decrease cellular proliferation. According to MI *et al*[33], 2-AG can also decrease migration, or markers of migration in a wide range of cell lines.

In conclusion, the present study systematically investigated the association of genetic variants in the binding sites of CTCF, and identified a new loci 11q12.2 increased lung cancer risk. Considering the moderate sample size in the validation stage, further larger well-designed population-based studies are warranted to elucidate the impact of rs60507107 on lung cancer risk.

Methods

Study populations. A two-stage case-control study was designed to evaluate the associations between genetic variants in the binding sites of CTCF and the risk of lung cancer. Study subjects for discovery stage is exactly the same with our previous GWAS study on lung cancer[9,16]. Briefly, the discovery stage of 2,331 lung cancer cases and 3,077 controls included two studies (Nanjing GWAS study: 1,473 cases and 1,962 controls from Nanjing and Shanghai; and Beijing GWAS study: 858 cases and 1,115 controls from Beijing and Wuhan). The histology for each case was histopathologically or cytologically confirmed by at least two local pathologists. Cancer-free subjects were recruited in local hospitals for individuals receiving routine physical examinations or in the communities for those participating screening of chronic diseases. Demographic information was collected using standard questionnaire through interviews. Smokers were defined as individuals who had smoked at an average of one cigarette or more per day and for at least one year in their lifetime; otherwise, subjects were considered as nonsmokers. Former smokers were defined as quitting for at least one year before recruitment. Both smoke year and the number of cigarettes per day were collected to calculate pack-year. The controls were frequency-matched to lung cancer cases for age, gender and geographic regions. As a result, 2,331 cases and 3,077 controls were included in the discovery stage, and

Table 3 | Stratification analysis on rs60507107

	GWAS				Validation			
Variables	Cases/Controls	OR(95% CI)[a]	P[a]	P_het[b]	Cases/Controls	OR(95% CI)[c]	P[c]	P_het[b]
Age				1.000				0.893
≤60	1142/1521	1.22(1.08–1.38)	2.00E-03		515/689	1.14(0.97–1.35)	1.19E-01	
>60	1189/1556	1.22(1.08–1.38)	1.00E-03		600/657	1.12(0.95–1.32)	1.64E-01	
Gender				0.368				0.182
Male	1711/2086	1.20(1.08–1.33)	1.00E-03		731/875	1.07(0.92–1.24)	3.66E-01	
Female	620/991	1.29(1.09–1.53)	3.00E-03		384/471	1.27(1.04–1.56)	1.70E-02	
Smoking status				0.143				0.773
Current	1252/1083	1.21(1.07–1.37)	2.00E-03		437/515	1.14(0.94–1.38)	1.75E-01	
Former	254/226	0.95(0.73–1.24)	7.30E-01		102/114	0.98(0.65–1.48)	9.15E-01	
Never	825/1768	1.28(1.12–1.46)	3.59E-04		576/717	1.15(0.98–1.35)	8.80E-02	
Smoking level				0.888				0.405
≤25	1245/2327	1.19(1.07–1.33)	2.00E-03		747/974	1.17(1.02–1.35)	2.50E-02	
>25	1086/750	1.22(1.06–1.41)	6.00E-03		368/372	1.05(0.85–1.30)	6.34E-01	
Histology				0.275				0.819
SC	822	1.14(1.00–1.30)	4.20E-02		332	1.15(0.96–1.39)	1.29E-01	
AC	1304	1.29(1.17–1.43)	6.63E-07		783	1.12(0.98–1.27)	8.60E-02	
other	205	1.14(0.91–1.41)	2.47E-01		NA	NA	NA	

a: Adjusted by age, gender, pack-year of smoking and pca1 where is appropriate b: *P* value for Cochran's chi-square-based heterogeneity test; c: Adjusted by age, gender and pack-year of smoking where is appropriate; NA: not available.

following new recruit 1,115 cases and 1,346 controls were replicated the associations as a validation stage. All study subjects provided informed consent and both the institutional review boards of Nanjing Medical University and Chinese Academy of Medical Sciences and Peking Union Medical College approved all procedures and all experiments were conducted in accordance with the approved guidelines.

Database preparation and bioinformatics analysis. *ENCODE dataset preparation.* A total of 690 datasets of TF ChIP-seq peaks were released based on the data from five ENCODE transcription factor binding sites (TFBS) ChIP-seq production groups. In consideration of tissue specificity, only peaks appeared in lung cancer cell line A549 were analyzed in this study. The uniform peaks of CTCF were downloaded from UCSC genome browser in BED format (release at 13th-May 2013, Uniform Peaks Transcription Factor ChIP-seq from ENCODE). As the Chip-seq experiments were performed in various treatment conditions, we defined the region as CTCF binding sites if peaks were called in any condition. As a result, 101,083 peaks were identified at GWAS scale in the final data set.

SNP screening based on GWAS data. A total of 3,569 SNPs located at CTCF related regions past quality control in our GWAS study[16] were included in this study. To further increase the genome coverage of our data, we performed two-steps imputation analyses (pre-phasing and impute) in the ChIP-seq peak regions. After imputation, a total of 32,453 unique SNPs in the TFBS regions with high imputation quality (info >0.8) were included for further analysis. SNPs with $P \leq 1.0 \times 10^{-5}$ for all GWAS samples and consistent between the Nanjing and Beijing study at $P \leq 0.05$ were selected for replication, except for: (i) SNPs with MAF <0.05; (ii) SNPs located 5kb off the nearest gene; (iii) SNPs located at HLA region; (iv) SNPs in strong linkage disequilibrium (LD) ($R^2 \geq 0.5$)[17]. The detailed work-flow was shown in Supplementary figure 1. As a result, 3 SNPs, satisfied all the above criteria, were included for further validation.

Statistical analysis. The quality control in our lung cancer GWAS was described previously[16]. Ungenotyped SNPs were imputed in the GWAS discovery samples using SHAPEIT 1.0 (haplotype estimation step) and IMPUTE2 (genotype imputation step), taking haplotype information from the 1000 Genomes Project (Phase I integrated variant set across all 1,092 individuals, V3, released at May 2012) as reference[34–36]. Score test based on dosage files was used to analysis the single SNP association of discovery stage by SNPTEST (V2) under additive model adjusting for age, gender and pack years of smoking[37]. The associations between SNPs and susceptibility of lung cancer were demonstrated by calculating the odds ratios (ORs) and their 95% confidence intervals (CIs). The chi-square-based Cochran's *Q* statistic was calculated to test for heterogeneity between groups in a stratified analysis[38]. Deviations of the characteristics for lung cancer patients and control subjects were examined by the Student-*t* test (for continuous variables) or the $\chi 2$ test (for categorical variables) with R software (version 2.15.3). All tests were two-sided and the significance level was set at $P \leq 0.05$.

Genotyping. SNPscan™ kit (Genesky Biotechnologies Inc., Shanghai, China) was used to determine genotypes of the SNPs selected in the validation stage. The detailed technical description for the kit was presented elsewhere, thirty duplicated samples were genotyped to ensure the reliability (the sensitivity was 97.03% and specificity was 93.33%)[39]. A series of methods were used to control the quality of genotyping: (i) case and control samples were mixed and genotyped without knowing the case or control status; (ii) forty-two known genotypes samples were genotyped as positive control, and the concordance rates were above 99%; (iii) five percent of the samples were randomly selected to repeat the genotyping, as blind duplicates, and the reproducibility was 100%.

1. Ferlay, J. *et al.* GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No.11 [Internet].(International Agency for Research on Cancer; 2013), Available from http://globocan.iarc.fr, accessed on14 July 2014).
2. Bray, F., Ren, J. S., Masuyer, E. & Ferlay, J. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int J Cancer* **132**, 1133–1145 (2013).
3. Doll, R. & Peto, R. The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *J Natl Cancer Inst* **66**, 1191–1308 (1981).
4. Li, C. *et al.* Epidermal growth factor receptor (EGFR) pathway genes and interstitial lung disease: an association study. *Sci Rep* **4**, 4893 (2014).
5. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001–1006 (2014).
6. Amos, C. I. *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* **40**, 616–622 (2008).
7. Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* **40**, 1407–1409 (2008).
8. McKay, J. D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat Genet* **40**, 1404–1406 (2008).
9. Dong, J. *et al.* Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nat Genet* **44**, 895–899 (2012).
10. Hosgood, H. D. 3rd *et al.* Genetic variant in TP63 on locus 3q28 is associated with risk of lung adenocarcinoma among never-smoking females in Asia. *Hum Genet* **131**, 1197–1203 (2012).
11. Landi, M. T. *et al.* 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 (2009).
12. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res* **22**, 1748–1759 (2012).
13. Johnson, R. *et al.* A genome-wide screen for genetic variants that modify the recruitment of REST to its target genes. *PLoS Genet* **8**, e1002624 (2012).
14. Zeron-Medina, J. *et al.* A polymorphic p53 response element in KIT ligand influences cancer risk and has undergone natural selection. *Cell* **155**, 410–422 (2013).
15. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
16. Hu, Z. *et al.* A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat Genet* **43**, 792–796 (2011).
17. Dong, J. *et al.* Genome-wide association study identifies a novel susceptibility locus at 12q23.1 for lung squamous cell carcinoma in han chinese. *PLoS Genet* **9**, e1003190 (2013).
18. Yamamoto, K., Okamoto, A., Isonishi, S., Ochiai, K. & Ohtake, Y. A novel gene, CRR9, which was up-regulated in CDDP-resistant ovarian tumor cell line, was associated with apoptosis. *Biochem Biophys Res Commun* **280**, 1148–1154 (2001).
19. Jin, G. *et al.* Common genetic variants on 5p15.33 contribute to risk of lung adenocarcinoma in a Chinese population. *Carcinogenesis* **30**, 987–990 (2009).
20. Zienolddiny, S. *et al.* The TERT-CLPTM1L lung cancer susceptibility variant associates with higher DNA adduct formation in the lung. *Carcinogenesis* **30**, 1368–1371 (2009).
21. Kohno, T. *et al.* Individuals susceptible to lung adenocarcinoma defined by combined HLA-DQA1 and TERT genotypes. *Carcinogenesis* **31**, 834–841 (2010).
22. Pande, M. *et al.* Novel genetic variants in the chromosome 5p15.33 region associate with lung cancer risk. *Carcinogenesis* **32**, 1493–1499 (2011).
23. Eldholm, V., Haugen, A. & Zienolddiny, S. CTCF mediates the TERT enhancer-promoter interactions in lung cancer cells: identification of a novel enhancer region involved in the regulation of TERT gene. *Int J Cancer* **134**, 2305–2313 (2014).
24. Fiorentino, F. P. *et al.* CTCF and BORIS regulate Rb2/p130 gene transcription: a novel mechanism and a new paradigm for understanding the biology of lung cancer. *Mol Cancer Res* **9**, 225–233 (2011).
25. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**, 24–32 (2009).
26. Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
27. Zhang, B. *et al.* Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat Genet* **46**, 533–542 (2014).
28. Johanneson, B. *et al.* Suggestive genetic linkage to chromosome 11p11.2–q12.2 in hereditary prostate cancer families with primary kidney cancer. *Prostate* **67**, 732–742 (2007).
29. Pertwee, R. G. The diverse CB1 and CB2 receptor pharmacology of three plant cannabinoids: delta9-tetrahydrocannabinol, cannabidiol and delta9-tetrahydrocannabivarin. *Br J Pharmacol* **153**, 199–215 (2008).
30. Cridge, B. J. & Rosengren, R. J. Critical appraisal of the potential use of cannabinoids in cancer management. *Cancer Manag Res* **5**, 301–313 (2013).
31. Munson, A. E., Harris, L. S., Friedman, M. A., Dewey, W. L. & Carchman, R. A. Antineoplastic activity of cannabinoids. *J Natl Cancer Inst* **55**, 597–602 (1975).
32. Jacobsson, S. O., Wallin, T. & Fowler, C. J. Inhibition of rat C6 glioma cell proliferation by endogenous and synthetic cannabinoids. Relative involvement of cannabinoid and vanilloid receptors. *J Pharmacol Exp Ther* **299**, 951–959 (2001).
33. Rudolph, M. I. *et al.* The influence of mast cell mediators on migration of SW756 cervical carcinoma cells. *J Pharmacol Sci* **106**, 208–218 (2008).
34. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
35. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
36. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179–181 (2012).
37. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499–511 (2010).
38. Higgins, J. P. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**, 1539–1558 (2002).
39. Chen, X. *et al.* Genome-wide association study validation identifies novel loci for atherosclerotic cardiovascular disease. *J Thromb Haemost* **10**, 1508–1514 (2012).

Acknowledgments

This work was funded by the State Key Program of National Natural Science of China(81230067), the National Key Basic Research Program Grant (2011CB503805), the National Science and Technology Support Program (2011BAI09B02), the Major Program of the National Natural Science Foundation of China(81390543), National Outstanding Youth Science Foundation of China (81225020), Science Foundation for Distinguished

Young Scholars of Jiangsu (BK2012042), National Natural Science of China(81270044), Natural Science Foundation of Jiangsu Province (BK2012042), Key Grant of Natural Science Foundation of Jiangsu Higher Education Institutions (11KJA330001), Jiangsu Province Clinical Science and Technology Projects (BL2012008), and the Priority Academic Program for the Development of Jiangsu Higher Education Institutions [Public Health and Preventive Medicine]. The authors wish to thank all the study participants, research staff and students who participated in this work.

Author contributions

During the development of this project, we benefited from suggestions and critical insights provided by H.S. Valuable comments on a first draft were received from G.J., H.M. and Z.H. The samples of beijing used in this project were provided by D.L. The samples of nanjing were prepared by W.S., J.S. and J.L. C.W. and W.Z. gave valuable advices in statistical analysis. The first draft was written by M.Z. and valuable comments on the first draft were received from J.D. All analysis was conducted by M.Z. and J.D. All authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Dai, J. *et al*. Systematical analyses of variants in CTCF-binding sites identified a novel lung cancer susceptibility locus among Chinese population. *Sci. Rep.* **5**, 7833; DOI:10.1038/srep07833 (2015).