www.oncotarget.com Oncotarget, 2018, Vol. 9, (No. 32), pp: 22340-22352

Research Paper

LSCC SNP variant regulates SOX2 modulation of VDAC3

Jacqueline Chyr[1,3], Dongmin Guo[2] and Xiaobo Zhou[2,3]

[1]Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA

[2]Center for Bioinformatics and Systems Biology, Department of Radiology, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA

[3]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Correspondence to: *Xiaobo Zhou,* ***email:*** *xizhou@wakehealth.edu*

Keywords: *lung cancer; SOX2; eQTL; SNP; topologically associating domain*

Received: *June 21, 2017* ***Accepted:*** *February 28, 2018* ***Published:*** *April 27, 2018*

ABSTRACT

Lung squamous cell carcinoma (LSCC) is a genomically complex malignancy with no effective treatments. Recent studies have found a large number of DNA alterations such as SOX2 amplification in LSCC patients. As a stem cell transcription factor, SOX2 is important for the maintenance of pluripotent cells and may play a role in cancer. To study the downstream mechanisms of SOX2, we employed expression quantitative trait loci (eQTLs) technology to investigate how the presence of SOX2 affects the expression of target genes. We discovered unique eQTLs, such as rs798827-VDAC3 (FDR *p*-value = 0.0034), that are only found in SOX2-active patients but not in SOX2-inactive patients. SNP rs798827 is within strong linkage disequilibrium (r^2 = 1) to rs58163073, where rs58163073 [T] allele increases the binding affinity of SOX2 and allele [TA] decreases it. In our analysis, SOX2 silencing downregulates VDAC3 in two LSCC cell lines. Chromatin conformation capturing data indicates that this SNP is located within the same Topologically Associating Domain (TAD) of VDAC3, further suggesting SOX2's role in the regulation of VDAC3 through the binding of rs58163073. By first subgrouping patients based on SOX2 activity, we made more relevant eQTL discoveries and our analysis can be applied to other diseases.

INTRODUCTION

Lung cancer is the leading cause of cancer death in the United States with approximately 158,000 deaths in 2016 [1, 2]. Lung squamous cell carcinoma (LSCC) represents a large portion of lung cancers with over 60,000 new cases diagnosed each year [3, 4]. The five year survival rate for LSCC is only 12% [1, 4–6]. Unlike adenocarcinoma lung cancer (LUAD), the other major subtype of lung cancer, there are no targeted treatments for LSCC [7–9]. The higher molecular complexity of LSCC has hampered our understanding and the discovery of druggable targets for LSCC.

Recent studies, including the comprehensive analysis by The Cancer Genome Atlas (TCGA) network, have found a large number of DNA alterations [4]. The most notable alteration is amplification of chromosome 3q, which contains the SOX2 locus [10]. SRY (sex determining region Y)-box 2 (SOX2) is a stem cell transcription factor that plays a role in cell self-renewal and differentiation [11–14]. It has found to be amplified and/or overexpressed in 63% of LSCC at the transcript level and 80–90% at the protein level [10, 15, 16]. Other studies have shown that SOX2 overexpression is crucial in promoting LSCC formation *in vivo* upon loss of Pten and Cdkn2ab [17].

There are different molecular mechanisms at play in SOX2-active LSCC patients compared to SOX2-inactive LSCC patients. SOX2 contains a DNA-binding region and regulates the expression of many downstream genes by binding to enhancer regions and facilitating in the remodeling of chromatin and subsequent initiation and transcription of target genes [18–20]. Other studies have found that single nucleotide polymorphisms (SNPs), located within the binding region of transcription factors, that can change the binding affinity of the factors and consequently the expression of their target genes [21–25]. Therefore, some SNPs, especially those located in SOX2

binding sequences, can alter SOX2's binding affinity to DNA. The increased or reduced binding of SOX2 to DNA can affect the transcription and expression of nearby target genes. Matrix eQTL allows us to identify local and distal eQTLs that are associated with the expression of target genes [26]. We can understand the downstream mechanism of SOX2 by identifying eQTLs in SOX2-active patients that are not in SOX2-inactive patients [26–28]. Analysis of flanking sequences of the eQTL or the SNPs in strong linkage equilibrium (LD) ($r^2 < 0.8$) can elucidate which motifs the SNPs may alter. Additional analysis of ChIP-seq and high-throughput chromatin conformation capture (Hi-C) data can further confirm the interaction of SOX2 to specific regions of the genome and thereby its modulation of target genes [29, 30]. The general workflow is summarized in Figure 1. Our goal is to understand the complex mechanisms of LSCC in order to better develop targeted treatments that work for a subset of patients with SOX2 activation.

RESULTS

SOX2 in LSCC patients

Unlike LUAD, SOX2 is often seen amplified in LSCC patients [4]. Model-based analysis of regulation of gene expression (MARGE) confirms high regulatory potential of SOX2 in LSCC patients and suggests that it is a master regulator [31] (Supplementary Table 1). The increased presence of this transcription factor may lead to differential regulation of downstream genes. To understand SOX2's roles in LSCC, multiple types of data from TCGA was analyzed. Patients were first divided into two groups: SOX2-active and SOX2-inactive based on their gene expression, copy number variation, methylation of SOX2. For methylation, only the 14 SOX2 CpGs located within the promoter region of SOX2 are considered. These CpGs have high variance and are highly correlated with SOX2 gene expression (Supplementary Figure 1). Patients with activated SOX2 met all three of the following criteria (Figure 2A–2C):

1) SOX2 expression values greater than the mean (log2expression values > 10.87797)

2) SOX2 copy number variation is duplicated compared to normal (segment means > 0.5)

3) SOX2 promoter region is hypomethylated (methylation β-values ≤0.4)

Patients that did not meet all three criteria are considered as SOX2 inactive. A total of 366 LSCC patients had all three data types available and were included in our study. Out of the 366 patients, 219 had high SOX2 expression, 212 had SOX2 copy number amplification, and 292 patients had SOX2 hypomethylation. Although many patients met more than one criterion, only 159 patients which had all three criteria are grouped as SOX2-active and the other 196 patients are grouped as SOX2-inactive (Figure 2D).

Matrix eQTL identifies SNP to VDAC3 expression association

Expression quantative trait loci (eQTL) analysis correlates SNP genotypes gene expression variations [26]. There are different sets of eQTLs in SOX2-active patients that are not in SOX2-inactive patients. Using a highly efficient and accurate eQTL analysis tool called Matrix eQTL [26, 28], we identified 686,251 cis SNP-gene pairs in the SOX2-active patients and 688,591 pairs in SOX2-

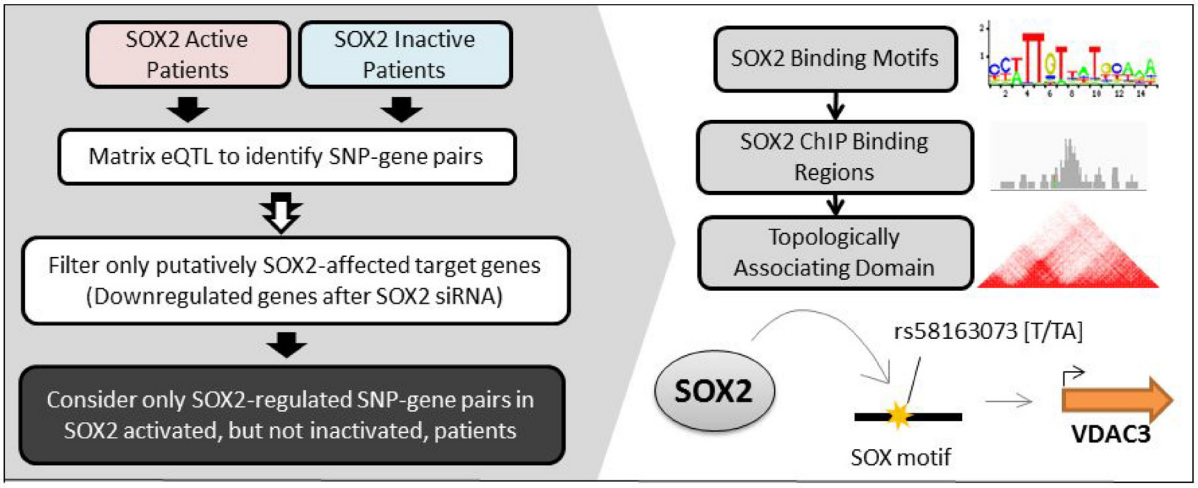

Figure 1: Brief overview of workflow. Left: LSCC patients are first clustered into two groups: SOX2-active and SOX2-inactive based on their SOX2 gene expression, copy number variation, and methylation. Matrix eQTL analyses are performed on each group of patients to identify group-specific SNP-gene pairs. Only the eQTL pairs with genes that are downregulated after SOX2 silencing are included in our analysis. Only the SNP-gene pairs that are unique in SOX2 active patients are considered. Right: Further analyses on SOX2 motifs, ChIP-seq binding peaks, and topologically associating domains are conducted to validate a SOX2→ SNP→ Gene relationship.

inactive patients ($p < 0.05$) within 1 Mbps of each other. Since we are interested in the mechanism of SOX2, we focused on SNP-gene pairs that are putatively affected by SOX2 expression. Fang WT, *et al*. significantly knocked down SOX2 expression using siRNAs in two LSCC cell lines with high expression levels of SOX2 (LK2 and NCI-H20) [32]. The gene expression of the SOX2-knocked down cells was profiled using an array. After normalization, we conducted differential expression analysis and found 266 unique genes downregulated in SOX2 siRNA cells compared to control cells as shown in Figure 3A. Of the large number of SNP-gene pairs from our Matrix eQTL analysis, 8,546 SOX2-active and 8,956 SOX2-inactive pairs contained a gene that is downregulated upon silencing SOX2. Looking at the top results with FDR $p < 0.01$, only 16 and 38 cis SNP-gene pairs remain (Table 1 and Supplementary Table 2). For each pair, the expression of the gene is correlated with the genotype of the SNP. For example, patients with active SOX2, the genotypes for SNP rs798827 is significantly correlated with the expression of the gene VDAC3 (FDR p-value = 0.0034), where allele G is correlated with lower expression of VDAC3, and allele T is correlated with higher expression of VDAC3. Not only that, but VDAC3 is downregulated upon SOX2 silencing in two LSCC cell lines. Further analysis reveals one gene-pair overlap between SOX2 active and inactive cis-Matrix eQTL results. The correlation of this SNP-gene pair may be independent of SOX2 activity.

SOX2 regulates VDAC3 expression

Among the 16 SNP-gene pairs in SOX2 active patients that were not in SOX2 inactive patients, SNP rs798827 to gene VDAC3 pair caught our interest (Table 1). Voltage-dependent anion-selective channel 3 (VDAC3) are small integral membrane channels important for controlling the flux of metabolites between mitochondria and cytoplasm [33, 34]. It has previously been shown to play a role in apoptotic and oxidative stress signaling [35–37]. It was also found to be downregulated upon SOX2 silencing in two LSCC cell lines (Figure 3B).

Not only is the expression of VDAC3 significantly higher in SOX2 active patients when compared to SOX2 inactive patients ($p = 0.0031$), the expression of VDAC3 is also correlated with the genotype of SNP rs798827 in SOX2 active patients, but not in SOX2 inactive patients (ANOVA t-test $p < 0.05$) (Figure 4A and 4B). The T genotype of rs798827 correlated with a significantly higher expression of VDAC3 than the G genotype. This association is only seen in SOX2 active patients, but not in SOX2 inactive patients (Figure 4B). SOX2 ChIP-seq data from Watanabe H, *et al*.

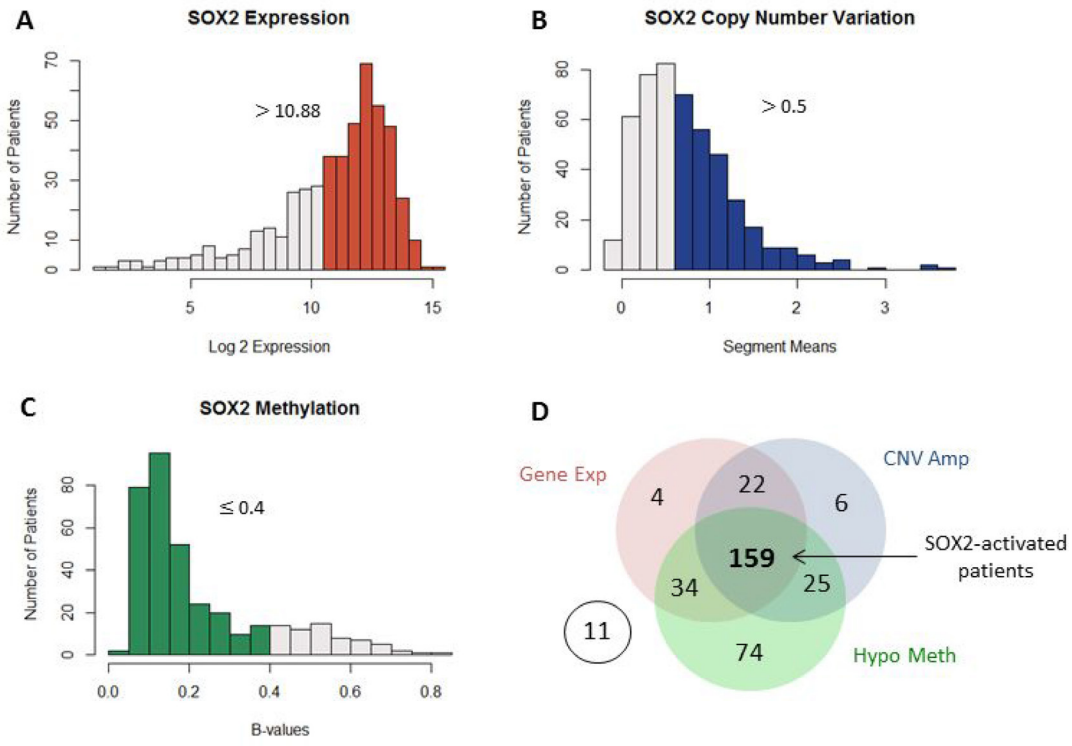

Figure 2: Patient genomic data. Multiple genomic datatypes are used to group SOX2 patients into two groups. The highlighted bars in the histograms indicate patients with high SOX2 gene expression (**A**), SOX2 copy number amplification (**B**), and SOX2 promoter-region hypomethylation (**C**). Patients with SOX2 expression > 10.88, copy number segment means > 0.5, and methylation β-values ≤ 0.4 are considered as SOX2 active patients. A total number of 366 patents has all three datatypes available and were included in our analysis. (**D**) Of the 366 patients, 159 patients were considered as SOX2-active, all other patients are considered as SOX2-inactive.

Table 1: eQTLs (FDR *p* value < 0.01) in SOX2-active patients

SNP	Allele A	Allele B	Gene	*P* value	FDR
rs4654947	C	T	NBPF3	4.60E-11	3.45E-07
rs9984519	C	T	IFNAR1	1.38E-07	5.59E-04
rs17420195	C	T	NBPF3	4.18E-07	1.45E-03
rs2465941	C	T	ZCCHC12	8.23E-07	2.60E-03
rs2290163	C	T	LMCD1	8.48E-07	2.67E-03
rs798827	**G**	**T**	**VDAC3**	**1.14E-06**	**3.43E-03**
rs4747471	A	G	MSRB2	1.22E-06	3.63E-03
rs16913776	C	G	RAB38	1.31E-06	3.83E-03
rs7624916	C	G	ARL6IP5	1.48E-06	4.27E-03
rs10968209	A	G	MOBKL2B	1.49E-06	4.27E-03
rs16850158	A	G	ALCAM	1.87E-06	5.14E-03
rs12057041	C	T	MOBKL2B	3.20E-06	7.98E-03
rs10968456	C	T	MOBKL2B	3.20E-06	7.98E-03
rs4790508	A	C	CRK	3.37E-06	8.33E-03
rs308819	A	C	RAB38	4.00E-06	9.57E-03
rs308814	C	T	RAB38	4.00E-06	9.57E-03

identified 5371 regions with high SOX2 interactions in a LSCC cell line (HCC95) [30]. SOX2 peaks were detected using MACS and normalized for copy number variation. Figure 4C shows the SOX2 ChIP-seq peak at the promoter region of VDAC3. ChIP-seq peaks for H3K27ac, an active enhancer mark, is also shown for the same LSCC cell line [38]. Collectively, our analysis indicates that SOX2 binds to the promoter region of VDAC3 and subsequently, regulates the expression of VDAC3.

SNPs or LD SNPs are located in SOX2 binding motifs

The HaploReg v4.1 web interface by Broad Institute is a well annotated database for SNP and their linkage disequilibrium (LD) SNPs [39–41]. When a SNP is in strong LD to another SNP, then those SNPs are highly associated with one another. In other words, the alleles of a few SNPs can suggest the alleles of their LD SNPs. Analyzing the 16 target SNPs from our cis-Matrix eQTL analysis of SOX2 active patients, we identified over 350 SNPs within strong LD ($r^2 > 0.80$) to our 16 SNPs. SOX2 is a member of the SOX HMG box family of transcription factors [42]. It can bind to specific regions of the genome at consensus binding sequences. Looking at the flanking sequences of those LD SNPs, seven unique SNPs are located in and may alter a SOX family binding motif (Table 2). The SNP from rs798827-VDAC3 pair is within a strong LD to SNP rs58163073. This LD SNP is actually an insertion variation where an [A] nucleotide is inserted. The flanking sequences of LD SNP rs58163073 make up the SOX2 binding motif and the SNP genotype modified the position weight matrix score. The [T] allele has a stronger binding affinity for SOX2 (11.5) than the [TA] allele (10.6) (Figure 5). The [T] allele of rs58163073 is linked to the [T] allele of rs798827 which is correlated with a higher expression of VDAC3.

VDAC3 and rs58163073 are located within the same TAD

The spatial organization of the genome plays a role in the transcriptional regulation of genes [43]. Chromatin conformation capturing methods such as Hi-C can reveal regions of the genome that have high interaction [44–46]. These regions are referred to as Topologically Associating Domains (TADs) [43, 47]. Hi-C data can be visualized in Hi-C heat maps [47]. Rao, S. S. P., Schmitt, A., and the Dekker Laboratory generated multiple Hi-C data for lung tissues and cell lines [29, 46, 48–50]. Figure 6 shows Hi-C heat map at chr8:41000000-44000000 for lung cell line IMR90. Figure 7 shows the Hi-C heap maps for lung cancer cell line A549, and two lung tissue samples. Figures were generated using 3D Genome Browser [51]. The location of VDAC3 and SNP rs58163073 are indicated. From analyzing Hi-C maps of lung cell lines (normal and cancer) and lung tissues, we have confirmed that VDAC3 and rs58163073 are located within the same TAD.

Significant association between SNP and clinical features

Minor allele frequencies (MAF) of rs798827 and rs58163073 varied by race, with Black or African Americans more likely to carry the minor alleles than Whites and Asians. The MAF of rs798827 were 0.32 for Black or African American, 0.03 for White, and 0.15 for Asian according to information from the 1000 Genomes Project. These frequencies were found to be similar in LSCC patients, 0.48, 0.07, and 0.22 for each race respectively, again, with a higher frequency of the minor alleles in Black or African American patients (Figure 8A and 8B). Our SNP is also associated with location of tumor. The malignant location of patients with the major allele were 47% left and 53% right lungs. Patients with the minor allele had tumor more predominantly located on the right lungs (24 % left and 76% right, chi-square test 0.007461) Figure 8C. There were no significant association between SNP genotype, and stage and age at diagnosis (Table 3).

DISCUSSION

LSCC remains a complex disease with many molecular alterations. SOX2 is often seen amplified in LSCC patients and plays an important role in the progression and development of LSCC, however its mechanisms are not very well understood. Utilizing multiple layers of data from genomic to epigenomic, we are able to depict a potential mechanism of SOX2 in LSCC. First, LSCC patients were separated into two groups: SOX2-active and SOX2-inactive based on

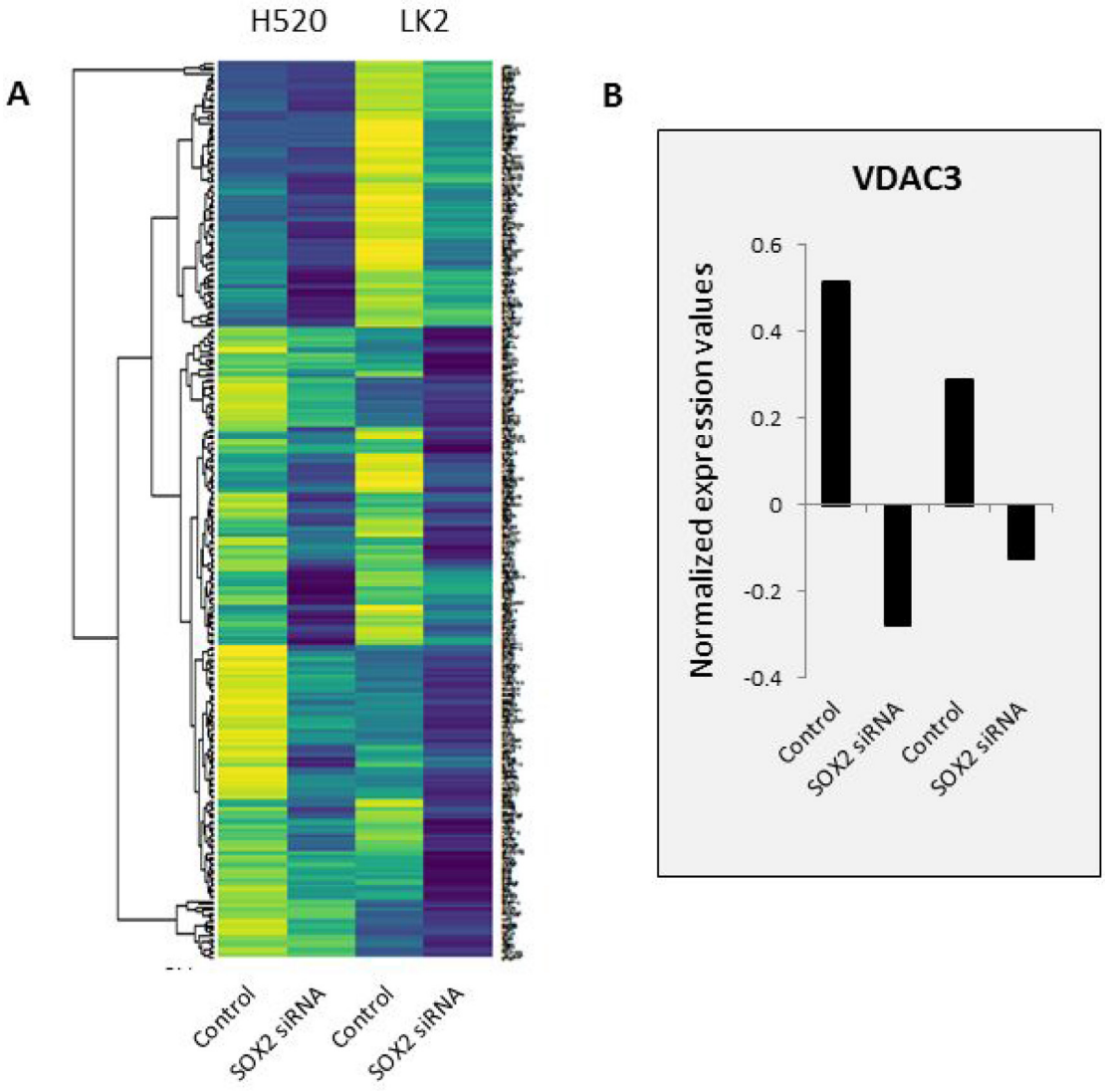

Figure 3: Gene expression profile of two SOX2-silenced LSCC cell lines. SOX2 was knocked down in two LSCC cell lines: H520 and LK2 and gene expression was profiled using an array. (**A**) Top 266 downregulated genes are shown in a heat map. Downregulated genes are defined as genes with combined value differences between SOX2 siRNA and control > 1.0. Yellow color represents higher expression and dark blue color represents lower expression. (**B**) VDAC3's normalized expression values are shown as an example. All samples were normalized to 0.

their SOX2 gene expression, copy number variation, and methylation. By first distinguishing patients based on their SOX2 activity, the mechanisms of SOX2 can be better elucidated. Using Matrix eQTL analysis, we linked variations in gene expression to SNP genotypes. Differences in SNP genotypes, especially within regulatory elements or binding motifs, can affect the binding affinity of transcription factors such as SOX2, and alter downstream transcription of target genes.

In our study, we focused on eQTLs that are found in SOX2-active patients but not in SOX2-inactive patients. We identified a SNP-gene pair (rs798827-VDAC3) that is unique in SOX2-active patients. The genotype of rs798827 is significantly correlated to the expression of VDAC3 in SOX2-active patients but not in SOX2-inactive patients. Since SOX2 is a transcription factor that binds to specific regions of the genome, variations in the binding sequences may alter the binding affinity of SOX2 to that region. Using HaploReg 4.1, we identified multiple SNPs within a strong LD to our eQTL SNPs. We found that rs798827 is within strong LD to SNP rs58163073 and motif analysis indicates that SNP rs58163073 is located within the binding motif of SOX2. The [T] allele of SNP rs58163073 increased the position weight matrix score for SOX2 and the [TA] allele decreased it. This SNP directly alter the binding sequence of SOX2 and the [T] allele increased its binding affinity.

Our cell line and ChIP-seq analysis further suggests that SOX2 regulates VDAC3. When SOX2 was silenced in two LSCC cell lines, the expression of VDAC3 was also downregulated. In addition, SOX2 ChIP-seq peaks show SOX2 binding in the promoter region of VDAC3. SNP rs58163073 and VDAC3 are also located within the same TAD in lung tissues, normal lung, and lung cancer cell lines. VDAC3 is the least investigated isoform of voltage-dependent anion channels. They are localized in the mitochondrial outer members and are pore-forming structures that control the exchange of metabolites between the mitochondria and cytoplasm [33]. They also play crucial roles in oxidative stress, maintaining redox status, and mediating cytochrome c apoptosis [34, 52]. Other studies have reported that VDAC3-deficient cancer cells have reduced permeability for ADP/ATP and decreased mitochondrial membrane potential [53–55]. The pathways VDAC3 were most involved in were cancer and reproductive system disease, according to Ingenuity Pathway Analysis (IPA) [56]. Deletion of VDAC3 significantly increases cell resistance to anti-tumor agent Erastin, which targets VDAC2 and VDAC3 [57, 58].

Further analysis shows that Black or African American patients have a higher MAF than those of White or Asian patients and patients with MAF had tumors located more predominantly on the right lungs.

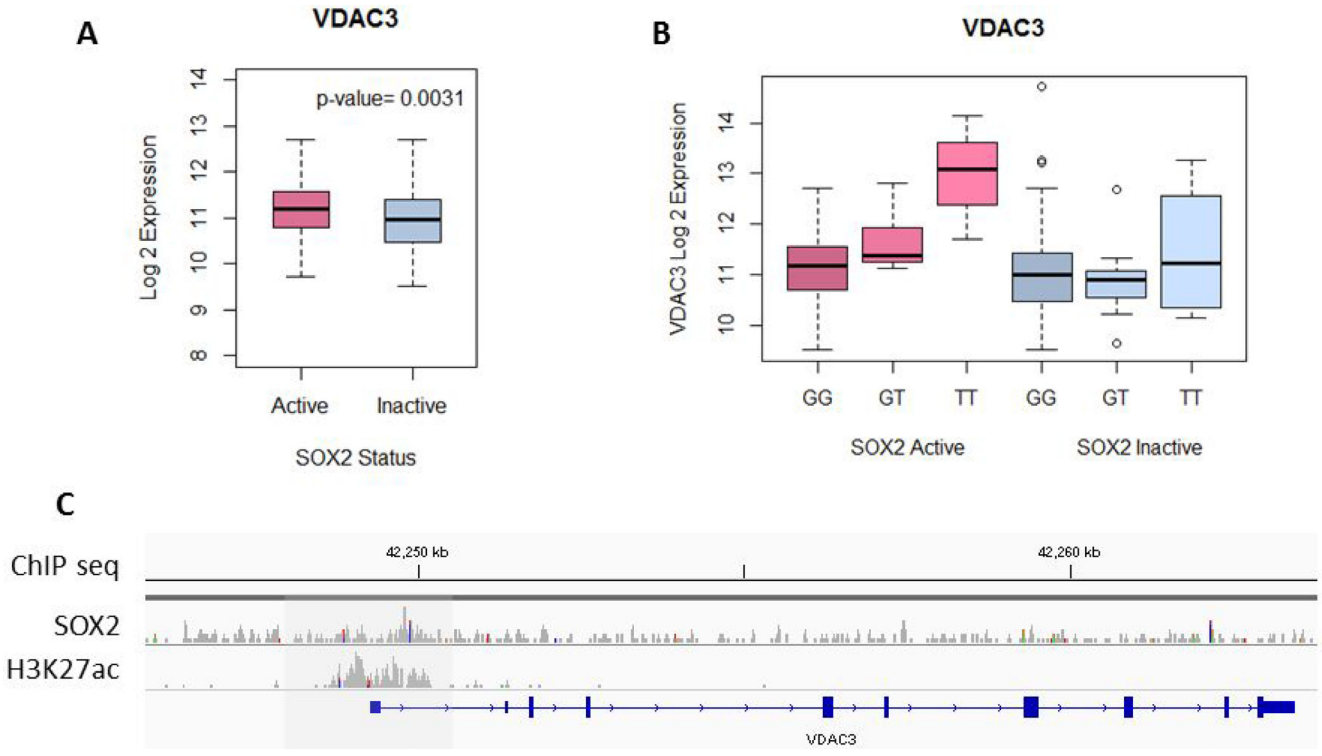

Figure 4: SOX2 regulates VDAC3 expression. (**A**) VDAC3 expression is significantly higher in SOX2-active patients compared to SOX2-inactive patients, *t*-test $p = 0.0031$. (**B**) In SOX2-active patients, the SNP genotype is associated with a significant difference in VDAC3 expression, ANOVA *t*-test $p = 6.91E\text{-}08$. In SOX2-inactive patients, the difference is not present, ANOVA *t*-test $p = 0.459$. (**C**) SOX2 and H3K27ac ChIP-seq peaks for cell line HCC95 are shown for the promoter region of VDAC3.

Table 2: Seven LD SNPs are located within and alters a SOX2 binding motif

SNP	LD SNP	Ref	Alt	D'	r^2	AFR	AMR	ASN	EUR	Motifs
rs798827	rs58163073	T	TA	–1	1	0.7	0.83	0.85	0.97	Cart1, Dbx1, Foxa2, Foxp1, HDAC2, Ncx2, **Sox2**, Sox5, Zfp105, p300
rs4747471	rs199772546	TA	T	1	1	0.04	0.02	0.21	0	Arid3a2, Dbx1, Dbx2, FAC1, Foxa2, Foxa4, Foxj2, Foxk1, Foxo2, Foxp1, HNF1, Hlx1, Hoxa10, Hoxa5, Hoxc6, Hoxd8, Lhx3, Mef2, Msx-1, Nanog, Ncx2, Nkx6-1, PLZF, Pax-6, Pou2f2, Pou3f2, Pou3f4, Pou4f3, Prrx1, Sox13, Sox18, Sox19, **Sox2**, Sox5, Sox6, Sox7, Zfp105, p300
rs4747471	rs200774383	AAT	A	1	1	0.04	0.02	0.21	0	CDP7, Dbx1, Dbx2, Evi-1, FAC1, Foxa2, Foxa4, Foxj2, Foxk1, Foxo2, Foxp1, HNF1, Hlx1, Hoxa10, Hoxa5, Hoxd8, Lhx3, Lhx3, Mef2, Nanog, Ncx_2, Nkx6-1, Nkx6-2, PLZF, Pax-6, Pou2f2, Pou3f4, Pou6f1, Prrx1, Sox13, Sox18, Sox19, **Sox2**, Sox5, Sox6, Sox7, Zfp105, p300
rs4747471	rs7087230	C	T	1	1	0.3	0.06	0.28	0	Fox, Foxk1, Foxp1, Hoxa10, Hoxd8, Lhx3, Pou2f2, Sox18, **Sox2**, Sox3, Sox7, TATA
rs4747471	rs12217320	A	C	1	1	0.03	0.02	0.25	0	FAC1, Foxa4, Foxd3, Foxk1, Foxo1, Foxo2, Foxp1, HDAC2, Irf, Mef2, Nanog, RREB-1, Sox13, **Sox2**, Sox6, Sox7, Zfp105
rs4747471	rs1398027	G	C	1	1	0.002	0.02	0.25	0	**Sox2**
rs12057041	rs10968463	C	T	1	1	0	0.03	0.2	0.01	Foxj2, **Sox2**
rs10968456	rs10968463	C	T	1	1	0	0.03	0.2	0.01	Foxj2, **Sox2**

The eQTL SNP, LD SNP, and reference and alternative alleles of the LD SNPs are shown. The coefficient of linkage disequilibrium (D`) and the square of correlation coefficient (r^2) are shown along with the frequency of the alternative allele in African (AFR), American (AMR), Asian (ASN), and European (EUR) populations. Motifs whose position weight matrix scores are affected by the LD SNPs are listed.

	Flanking Sequence	PWM
Ref [T]	AGATTATTTTTTATTATTAACATTATTATTTTTTTTTTGAGATGGAGTCTCATCCTGTC	11.5
Alt [TA]	AGATTATTTTTTATTATTAACATTATTATTATTTTTTTTTGAGATGGAGTCTCATCCTGTC	10.6
SOX2 Motif	DNDWNNDYHATTGTTHHDHVDD	

Figure 5: SNP rs58163073 alters the PWM of SOX2. The flanking sequences of rs58163073 from chromosome 8 position 42904940 +/– 29 nucleotides are shown for the reference and alternative alleles. The position weight matrix (PWM) score are obtained from HaploReg4.1. The reference [T] allele has a stronger binding affinity for SOX2 than the alternative [TA].

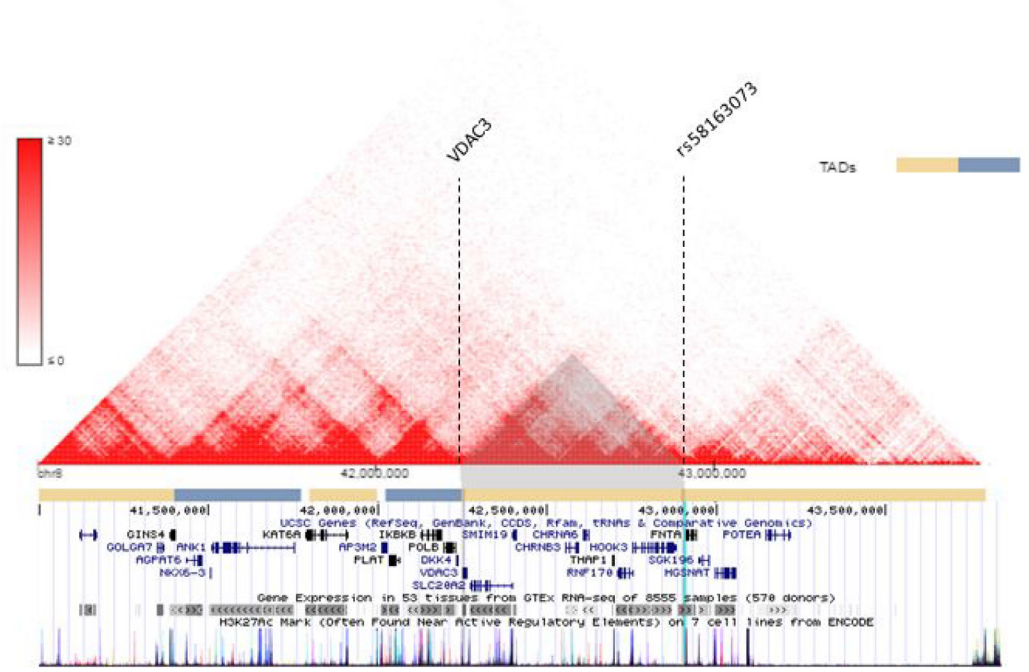

Figure 6: VDAC3 and rs58163073 are located within the same TAD in lung cell line. The Hi-C heat map of lung cell line IMR90 is shown for chr8:41000000–44000000, resolution 10 kb. The location of VDAC3 and rs58163073 are indicated with dotted lines. The different TADs are marked with pale yellow and blue bars below the heat map. The locations of other genes are shown.

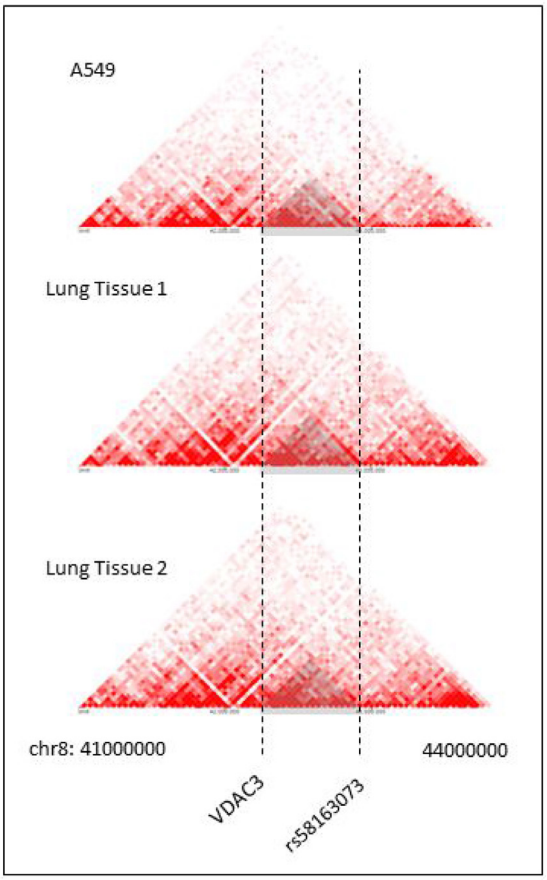

Figure 7: Hi-C heat maps for lung cancer cell line and lung tissues. Hi-C heat maps of lung cancer cell line A549 and two lung tissue samples are show for chr8:41000000–44000000, resolution 40 kb. VDAC3 gene and SNP rs58163073 are marked with a dotted line.

Collectively, these analyses support the notion that the SNP variant rs58163073 affects the binding of SOX2, which in turn affects SOX2's modulation of target genes such as VDAC3 (Figure 9). This discovery is found only when SOX2 activity is first considered. Our analysis can be used to understand the downstream mechanisms of transcription factors in other diseases and cancers.

METHODS

LSCC patient data

Patient gene expression, copy number variation, methylation, and SNP data were collected from The Cancer Genome Atlas (TCGA) database. Specifically, Illumina HiSeq RNAseqv2 was used for gene expression data, Genome Wide SNP 6.0 was used for both copy number variation and SNP genotype data, and Human Methylation 450 bioassay data set was used for methylation data. A total number of 366 LSCC patients had all three datatypes available.

Matrix eQTL computation

Expression quantitative trait loci (eQTL) analysis was conducted with Matrix eQTL: Ultra-fast eQTL analysis via large matrix operations by Andrey Shabalin [26]. Software and resources were obtained from http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/. Output threshold p-values were set at 0.05. Local eQTLs (or cis eQTLs) were set at the default distance of 1 million base pairs of each other.

SOX2 siRNA lung cancer cell line

SOX2 silencing experimental data was obtained from NCBI GSE48871 [32]. Two LSCC cell lines (H520 and LK2) were treated with either pooled siRNA sequences of SOX2 or scrambled control siRNAs) and their gene expression was profiled using Illumina's BeadChip Human HT-12v3 array. The gene expression data distribution was normalized to 0.

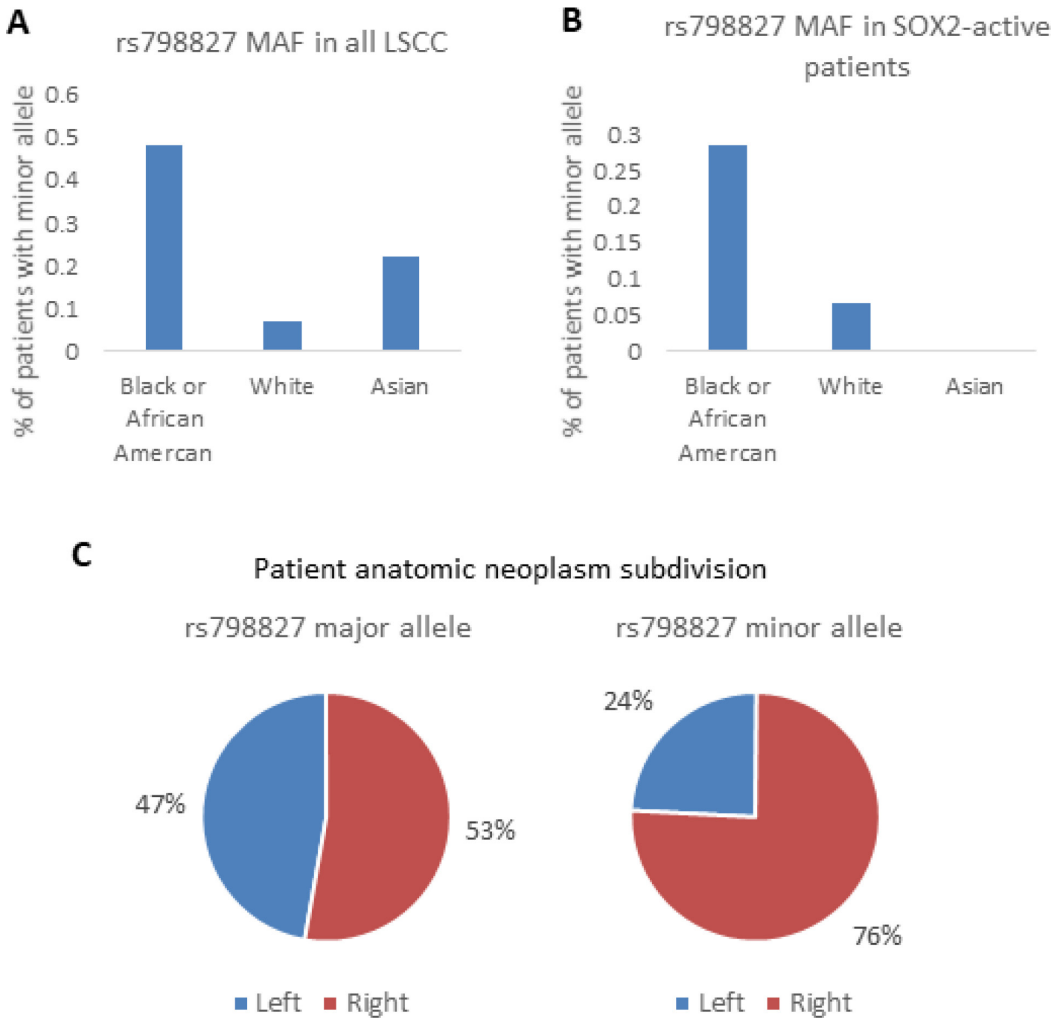

Figure 8: LSCC clinical analysis. SNP rs798827 minor allele frequency (MAF) is shown for different races in LSCC patients (**A**) and in SOX2-active patients (**B**). Frequency of tumor location (left or right) is shown for patients with the major allele and minor allele (**C**).

Table 3: Association of SNP rs798827 and clinical features in LSCC patients

	Major allele	Minor allele
Race		
Black or African American	5%	36%
White	93%	60%
Asian	2%	5%
Location		
Left-Lower	18%	14%
Left-Upper	30%	11%
Right-Lower	24%	32%
Right-Upper	29%	43%
Age		
Min	39	40
1st Q	62	63
Median	68	69
Mean	67.22	67.72
3rd Q	73	74.75
Max	90	84
Stage		
i	48%	57%
ii	33%	25%
iii	17%	16%
iv	1%	2%

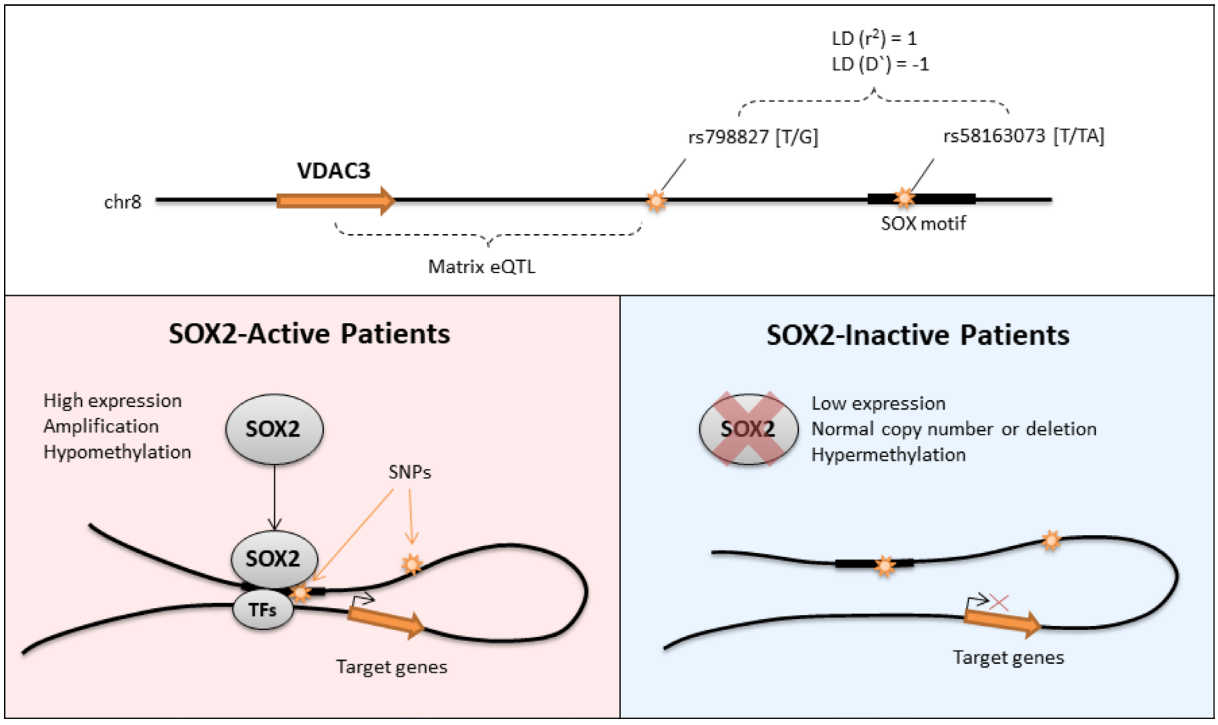

Figure 9: Graphical summary of the regulation of VDAC3 by SOX2. In SOX2-active patients, the genotype of SNP rs798827 is associated to the expression of VDAC3. This SNP is within strong LD to SNP rs58163073 which overlaps with the binding motif of SOX2. This mechanism is only seen in patients with SOX2 activity. Positions of genes and SNPs are not drawn to scale.

Linkage disequilibrium SNPs and binding motifs

LD SNPs were explored using HaploRegv4.1 hosted by Broad Institute [39]. The database table is curated from information from the 1000 Genomes Project. Pairwise LD was calculated for all pairs of SNPs within 250 kb. A LD threshold of $r^2 > 0.8$ was used to query variants. The 16 eQTLs were within strong LD to 263 SNPs. HaploRegv4.1 also contained information on regulatory motif changes. Only the SNP variants that overlapped the binding motif of SOX2 and changed the position weight matrix (PWM) score were considered. Under these conditions, only seven unique LD SNPs are identified.

SOX2 and H3K27ac ChIP-seq in HCC95 cell line

SOX2 ChIP-seq data was obtained from NCBI- GSE46837 [30]. A LSCC cell line with SOX2 amplification was used in this study. SOX2 binding sites were detected using Model-based Analysis of ChIP-Seq (MACS) and normalized for copy number variation. There were 5371 peaks with high SOX2 interaction. H3K27ac ChIP-seq data was obtained from NCBI-GSE66992 [38]. LSCC cell line, HCC95 was also used and H3K27ac binding sites were called by MACs.

Lung tissue and cell line Hi-C data

Hi-C data was obtained and visualized on 3D Genome Browser [59]. IMR90 Hi-C data was from Rao S, *et al.* (2014) [48], A549 cell line Hi-C data was from ENCODE Encyclopedia version 3 (2010) [29, 60], and two normal lung tissue Hi-C data were from Shmitt A, *et al.* (2016) [49]. Hi-C heat maps resolutions are at 10 kb, 40 kb, 40 kb, and 40 kb respectively, and between 41000000 and 44000000 position on chromosome 8, which is where VDAC3 gene and associated SNPs are located.

ACKNOWLEDGMENTS

We would like to acknowledge the members of Center for Bioinformatics and Systems Biology at Wake Forest School of Medicine for valuable discussions and advices. The authors acknowledge the Texas Advanced Computing Center (TACC) at the University of Texas at Austin (http://www.tacc.utexas.edu) and the DEMON high performance computing (HPC) cluster at Wake Forest University School of Medicine for providing HPC resources that have contributed to the research results reported within this paper.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

FUNDING

This work was supported by National Institutes of Health 1U01CA166886, 1U01AR069395 and 1R01GM123037 (to X. Zhou), and partially supported by NSFC No.61373105 and No.61672422.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015; 65:5–29.
2. Shopland DR, Eyre HJ, Pechacek TF. Smoking-attributable cancer mortality in 1991: is lung cancer now the leading cause of death among smokers in the United States? J Natl Cancer Inst. 1991; 83:1142–48.
3. Field RW, Smith BJ, Platz CE, Robinson RA, Neuberger JS, Brus CP, Lynch CF. Lung cancer histologic type in the surveillance, epidemiology, and end results registry versus independent review. J Natl Cancer Inst. 2004; 96:1105–07.
4. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–25.
5. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011; 61:69–90.
6. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ. Cancer statistics, 2008. CA Cancer J Clin. 2008; 58:71–96.
7. Ladanyi M, Pao W. Lung adenocarcinoma: guiding EGFR-targeted therapy and beyond. Mod Pathol. 2008; 21:S16–22.
8. Oka S, Uramoto H, Shimokawa H, Iwanami T, Tanaka F. The expression of Ki-67, but not proliferating cell nuclear antigen, predicts poor disease free survival in patients with adenocarcinoma of the lung. Anticancer Res. 2011; 31:4277–82.
9. Paik PK, Arcila ME, Fara M, Sima CS, Miller VA, Kris MG, Ladanyi M, Riely GJ. Clinical characteristics of patients with lung adenocarcinomas harboring BRAF mutations. J Clin Oncol. 2011; 29:2046–51.
10. Hussenet T, Dali S, Exinger J, Monga B, Jost B, Dembelé D, Martinet N, Thibault C, Huelsken J, Brambilla E, du Manoir S. SOX2 is an oncogene activated by recurrent 3q26.3 amplifications in human lung squamous cell carcinomas. PLoS One. 2010; 5:e8960.
11. Basu-Roy U, Seo E, Ramanathapuram L, Rapp TB, Perry JA, Orkin SH, Mansukhani A, Basilico C. Sox2 maintains self renewal of tumor-initiating cells in osteosarcomas. Oncogene. 2012; 31:2270–82.
12. Chen Y, Shi L, Zhang L, Li R, Liang J, Yu W, Sun L, Yang X, Wang Y, Zhang Y, Shang Y. The molecular mechanism governing the oncogenic potential of SOX2 in breast cancer. J Biol Chem. 2008; 283:17969–78.
13. Jia X, Li X, Xu Y, Zhang S, Mou W, Liu Y, Liu Y, Lv D, Liu CH, Tan X, Xiang R, Li N. SOX2 promotes tumorigenesis and increases the anti-apoptotic property of human prostate cancer cell. J Mol Cell Biol. 2011; 3:230–38.

14. Wang Q, He W, Lu C, Wang Z, Wang J, Giercksky KE, Nesland JM, Suo Z. Oct3/4 and Sox2 are significantly associated with an unfavorable clinical outcome in human esophageal squamous cell carcinoma. Anticancer Res. 2009; 29:1233–41.

15. Sholl LM, Barletta JA, Yeap BY, Chirieac LR, Hornick JL. Sox2 protein expression is an independent poor prognostic indicator in stage I lung adenocarcinoma. Am J Surg Pathol. 2010; 34:1193–98.

16. Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, Verhaak RG, Kim SY, Wardwell L, Tamayo P, Gat-Viks I, Ramos AH, Woo MS, Weir BA, et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009; 41:1238–42.

17. Ferone G, Song JY, Sutherland KD, Bhaskaran R, Monkhorst K, Lambooij JP, Proost N, Gargiulo G, Berns A. SOX2 is the determining oncogenic switch in promoting lung squamous cell carcinoma from different cells of origin. Cancer Cell. 2016; 30:519–32.

18. Scaffidi P, Bianchi ME. Spatially precise DNA bending is an essential activity of the sox2 transcription factor. J Biol Chem. 2001; 276:47296–302.

19. Xiang R, Liao D, Cheng T, Zhou H, Shi Q, Chuang TS, Markowitz D, Reisfeld RA, Luo Y. Downregulation of transcription factor SOX2 in cancer stem cells suppresses growth and metastasis of lung cancer. Br J Cancer. 2011; 104:1410–17.

20. Harley VR, Lovell-Badge R, Goodfellow PN. Definition of a consensus DNA binding site for SRY. Nucleic Acids Res. 1994; 22:1500–01.

21. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Björklund M, Wei G, Yan J, Niittymäki I, Mecklin JP, Järvinen H, Ristimäki A, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet. 2009; 41:885–90.

22. Zhang X, Zhou L, Fu G, Sun F, Shi J, Wei J, Lu C, Zhou C, Yuan Q, Yang M. The identification of an ESCC susceptibility SNP rs920778 that regulates the expression of lncRNA HOTAIR via a novel intronic enhancer. Carcinogenesis. 2014; 35:2062–67.

23. Tokuhiro S, Yamada R, Chang X, Suzuki A, Kochi Y, Sawada T, Suzuki M, Nagasaki M, Ohtsuki M, Ono M, Furukawa H, Nagashima M, Yoshino S, et al. An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. Nat Genet. 2003; 35:341–48.

24. Ng MT, Van't Hof R, Crockett JC, Hope ME, Berry S, Thomson J, McLean MH, McColl KE, El-Omar EM, Hold GL. Increase in NF-kappaB binding affinity of the variant C allele of the toll-like receptor 9 -1237T/C polymorphism is associated with Helicobacter pylori-induced gastric disease. Infect Immun. 2010; 78:1345–52.

25. Guo H, Ahmed M, Zhang F, Yao CQ, Li S, Liang Y, Hua J, Soares F, Sun Y, Langstein J, Li Y, Poon C, Bailey SD, et al. Modulation of long noncoding RNAs by risk SNPs underlying genetic predispositions to prostate cancer. Nat Genet. 2016; 48:1142–50.

26. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012; 28:1353–58.

27. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999; 401:788–91.

28. Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet. 2011; 27:72–79.

29. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293.

30. Watanabe H, Ma Q, Peng S, Adelmant G, Swain D, Song W, Fox C, Francis JM, Pedamallu CS, DeLuca DS, Brooks AN, Wang S, Que J, et al. SOX2 and p63 colocalize at genetic loci in squamous cell carcinomas. J Clin Invest. 2014; 124:1636–45.

31. Wang S, Zang C, Xiao T, Fan J, Mei S, Qin Q, Wu Q, Li X, Xu K, He HH, Brown M, Meyer CA, Liu XS. Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. Genome Res. 2016; 26:1417–29.

32. Fang WT, Fan CC, Li SM, Jang TH, Lin HP, Shih NY, Chen CH, Wang TY, Huang SF, Lee AY, Liu YL, Tsai FY, Huang CT, et al. Downregulation of a putative tumor suppressor BMP4 by SOX2 promotes growth of lung squamous cell carcinoma. Int J Cancer. 2014; 135:809–19.

33. Baines CP, Kaiser RA, Sheiko T, Craigen WJ, Molkentin JD. Voltage-dependent anion channels are dispensable for mitochondrial-dependent cell death. Nat Cell Biol. 2007; 9:550–55.

34. De Pinto V, Guarino F, Guarnera A, Messina A, Reina S, Tomasello FM, Palermo V, Mazzoni C. Characterization of human VDAC isoforms: a peculiar function for VDAC3? Biochim Biophys Acta. 2010; 1797:1268–75.

35. Cheng EH, Sheiko TV, Fisher JK, Craigen WJ, Korsmeyer SJ. VDAC2 inhibits BAK activation and mitochondrial apoptosis. Science. 2003; 301:513–17.

36. Reina S, Checchetto V, Saletti R, Gupta A, Chaturvedi D, Guardiani C, Guarino F, Scorciapino MA, Magrì A, Foti S, Ceccarelli M, Messina AA, Mahalakshmi R, et al. VDAC3 as a sensor of oxidative state of the intermembrane space of mitochondria: the putative role of cysteine residue modifications. Oncotarget. 2016; 7:2249–68. https://doi.org/10.18632/oncotarget.6850.

37. De Stefani D, Bononi A, Romagnoli A, Messina A, De Pinto V, Pinton P, Rizzuto R. VDAC1 selectively transfers apoptotic Ca2+ signals to mitochondria. Cell Death Differ. 2012; 19:267–73.

38. Zhang X, Choi PS, Francis JM, Imielinski M, Watanabe H, Cherniack AD, Meyerson M. Identification of focally

amplified lineage-specific super-enhancers in human epithelial cancers. Nat Genet. 2016; 48:176–82.

39. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res. 2016; 44:D877–81.
40. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. Linkage disequilibrium in the human genome. Nature. 2001; 411:199–204.
41. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012; 40:D930–34.
42. Wakamatsu Y, Endo Y, Osumi N, Weston JA. Multiple roles of Sox2, an HMG-box transcription factor in avian neural crest development. Dev Dyn. 2004; 229:74–86.
43. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485:376–80.
44. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013; 14:390–403.
45. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet. 2006; 38:1348–54.
46. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012; 58:268–76.
47. Jost D, Carrivain P, Cavalli G, Vaillant C. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Res. 2014; 42:9553–61.
48. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159:1665–80.
49. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, Ren B. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Reports. 2016; 17:2042–59.
50. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012; 9:999–1003.
51. Online tool available at: http://promoter.bx.psu.edu/hi-c/.
52. Shimizu S, Narita M, Tsujimoto Y. Bcl-2 family proteins regulate the release of apoptogenic cytochrome c by the mitochondrial channel VDAC. Nature. 1999; 399:483–87.
53. Maldonado EN, Lemasters JJ. Warburg revisited: regulation of mitochondrial metabolism by voltage-dependent anion channels in cancer cells. J Pharmacol Exp Ther. 2012; 342:637–41.
54. Okazaki M, Kurabayashi K, Asanuma M, Saito Y, Dodo K, Sodeoka M. VDAC3 gating is activated by suppression of disulfide-bond formation between the N-terminal region and the bottom of the pore. Biochim Biophys Acta. 2015; 1848:3188–96.
55. Maldonado EN, Sheldon KL, DeHart DN, Patnaik J, Manevich Y, Townsend DM, Bezrukov SM, Rostovtseva TK, Lemasters JJ. Voltage-dependent anion channels modulate mitochondrial metabolism in cancer cells: regulation by free tubulin and erastin. J Biol Chem. 2013; 288:11920–29.
56. Messina A, Reina S, Guarino F, Magrì A, Tomasello F, Clark RE, Ramsay RR, De Pinto V. Live cell interactome of the human voltage dependent anion channel 3 (VDAC3) revealed in HeLa cells by affinity purification tag technique. Mol Biosyst. 2014; 10:2134–45.
57. Yagoda N, von Rechenberg M, Zaganjor E, Bauer AJ, Yang WS, Fridman DJ, Wolpaw AJ, Smukste I, Peltier JM, Boniface JJ, Smith R, Lessnick SL, Sahasrabudhe S, Stockwell BR. RAS-RAF-MEK-dependent oxidative cell death involving voltage-dependent anion channels. Nature. 2007; 447:864–68.
58. Simamura E, Shimada H, Hatta T, Hirai K. Mitochondrial voltage-dependent anion channels (VDACs) as novel pharmacological targets for anti-cancer agents. J Bioenerg Biomembr. 2008; 40:213–17.
59. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, et al. The UCSC Genome Browser database: update 2010. Nucleic Acids Res. 2010; 38:D613–19.
60. Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, Raney BJ, Wang T, Hinrichs AS, Zweig AS, Fujita PA, Learned K, Rhead B, et al. ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res. 2010; 38:D620–25.