Research Article

For reprint orders, please contact: reprints@futuremedicine.com

Epigenomics

Positional integration of lung adenocarcinoma susceptibility loci with primary human alveolar epithelial cell epigenomes

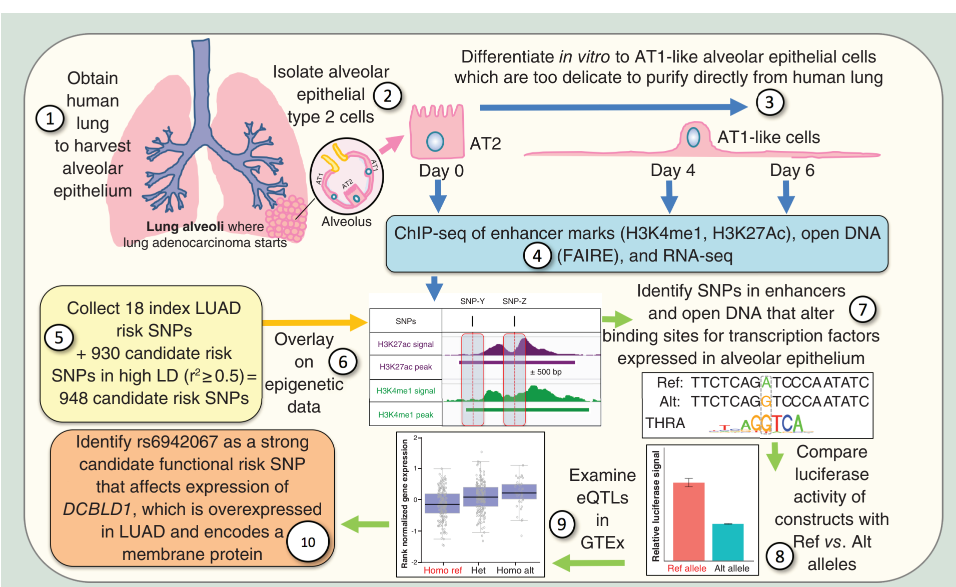
Chenchen Yang[‡,1,2,3], Theresa Ryan Stueve[‡,1,2,3,4], Chunli Yan[1,2,3], Suhn K Rhie[1,2,3], Daniel J Mullen[1,2,3], Jiao Luo[2,5], Beiyun Zhou[3,5,6], Zea Borok[2,3,5,6], Crystal N Marconett[1,2,3] & Ite A Offringa*[,1,2,3]

[1]Department of Surgery, University of Southern California, CA 90089, USA
[2]Department of Biochemistry & Molecular Medicine, University of Southern California, CA 90089, USA
[3]Norris Comprehensive Cancer Center, University of Southern California, CA 90089, USA
[4]Department of Preventive Medicine, University of Southern California, CA 90089, USA
[5]Department of Medicine, Division of Pulmonary & Critical Care & Sleep Medicine, University of Southern California, CA 90089, USA
[6]Hastings Center for Pulmonary Research, Keck School of Medicine, University of Southern California, CA 90089, USA
*Author for correspondence: ilaird@usc.edu
‡Authors contributed equally

Aim: To identify functional lung adenocarcinoma (LUAD) risk SNPs. **Materials & methods:** Eighteen validated LUAD risk SNPs ($p \leq 5 \times 10^{-8}$) and 930 SNPs in high linkage disequilibrium ($r^2 > 0.5$) were integrated with epigenomic information from primary human alveolar epithelial cells. Enhancer-associated SNPs likely affecting transcription factor-binding sites were predicted. Three SNPs were functionally investigated using luciferase assays, expression quantitative trait loci and cancer-specific expression. **Results:** Forty-seven SNPs mapped to putative enhancers; 11 located to open chromatin. Of these, seven altered predicted transcription factor-binding motifs. Rs6942067 showed allele-specific luciferase expression and expression quantitative trait loci analysis indicates that it influences expression of *DCBLD1*, a gene that encodes an unknown membrane protein and is overexpressed in LUAD. **Conclusion:** Integration of candidate LUAD risk SNPS with epigenomic marks from normal alveolar epithelium identified numerous candidate functional LUAD risk SNPs including rs6942067, which appears to affect *DCBLD1* expression.

Data deposition: Data are provided in GEO record GSE84273.

Graphical abstract:



Future Medicine

First draft submitted: 2 January 2018; Accepted for publication: 10 May 2018; Published online: 13 September 2018

Keywords: alveolar epithelial cells • *DCBLD1* • enhancer • epigenomics • eQTL • FAIRE • lung adenocarcinoma • rs6942067 • SNP

Lung cancer is the leading cause of cancer death worldwide, and in the USA, it is responsible for more deaths than breast, colorectal and prostate cancer combined [1]. The most common histological subtype of lung cancer is lung adenocarcinoma (LUAD), accounting for over 40% of new lung cancer cases. Never smokers with lung cancer are more likely to have LUAD and to be women and persons of Asian descent [2,3]. Though smoking is strongly implicated in lung cancer risk, it is estimated that half of all new cases arise in never smokers or smokers who quit many years ago [4]. Epidemiological studies indicate that in addition to environmental factors, LUAD susceptibility also has a significant genetic component [5–7].

To date, genome-wide association studies (GWAS) have uncovered 18 validated SNPs associated with LUAD risk ($p \leq 5 \times 10^{-8}$), distributed over 12 chromosomal regions (Table 1). Causal links between these SNPs and lung health remain largely unknown. The principal challenges to understanding how SNPs confer susceptibility to LUAD are the same for all post-GWAS studies [8–11]. First, because GWAS SNPs are landmark or 'index' SNPs that serve as positional references for phenotypic associations, hundreds of SNPs in linkage disequilibrium (LD), could be the actual functional SNP. Second, the majority of all GWAS SNPs, including 17 of the 18 index LUAD SNPs (Table 1), reside outside of protein-coding regions of genes and are thus hypothesized to influence gene expression by disrupting regulatory elements that could be hundreds of kilobases away from their target genes. Consistent with this hypothesis, several groups have determined that the majority of causal or 'functional' noncoding SNPs are in fact concentrated in epigenomic features characteristic of enhancers of gene transcription [8–16]. Epigenomic features make up a layer of information that is superimposed onto the genome; this does not alter the genetic sequence but affects cell type-specific gene expression [17]. Epigenomic features include chemical modifications such as DNA methylation, regulatory RNAs, transcription factors (TFs), modified histones and genome-organizing factors [17]. Epigenomic features characteristics of enhancers include regions of open chromatin (such as DNase I hypersensitive sites) that are accessible to DNA-binding proteins, H3K4me1 (a mark for poised or active enhancers [18]) and H3K27ac (a mark for active enhancers [19]). Identification of epigenetic marks disrupted by SNPs has allowed exciting progress to be made in post-GWAS investigations of genetic susceptibility to autoimmune disorders, Alzheimer's disease, diabetes and cancers of the breast, prostate and colon [10,12–16,20–22]. Importantly, in order to be relevant, the epigenomic features used for positional integration with candidate functional SNPs must be obtained from cell types pertinent to the disease at hand. In the case of LUAD, alveolar epithelial cells (AECs) are highly relevant because LUAD arises in the peripheral lung.

Alveolar lung epithelium consists of two epithelial cell types: squamous type 1 (AT1) cells and cuboidal type 2 (AT2) cells. AT1 cells are considered terminally differentiated and mediate gas exchange. They are large delicate cells that form the interface between the air in the alveoli and the microcapillaries that surround the alveolar sacs, thereby mediating gas exchange [31]. AT1 cells make up 95% of the surface area of the alveoli, and could, therefore, be disproportionately impacted by inhaled toxicants such as those present in tobacco smoke or pollution. AT2 cells produce the surfactants required to prevent the collapse of alveoli upon expiration. In adult lung, they are the progenitors of AT1 cells in response to lung injury; AT2 cells can both self-renew and differentiate into AT1 cells, and this process can be replicated *in vitro* [32–35]. Given their proliferative capacity, AT2 cells are commonly implicated as the likely progenitors of LUAD [36–39]. However, because definitive data on the roles and interactions of AT1 and AT2 cells in LUAD initiation remain to be provided, we rationalized that both cell types (here collectively referred to as alveolar epithelial cells or AECs) should be investigated in LUAD risk studies. Here, we use epigenomic data from purified primary human AT2 cells, *in vitro*-differentiated AT1-like cells, and LUAD cell lines for positional integration with LUAD risk SNPs, and apply a combination of bioinformatics and molecular approaches to identify candidate functional-risk SNPs (outlined in flow chart of Supplementary Figure 1).

Materials & methods

Ethics statement

Remnant human transplant lung was obtained in compliance with Institutional Review Board-approved protocols for the use of human source material in research (HS-07-00660) and processed within 3 days of death.

Table 1. Lung adenocarcinoma-associated risk index SNPs, listed by chromosome.

Chr	Position (hg19)	Genes	SNPs	Population	Annotation	p-values in GWAS [Ref.]
3	189356261	*TP63*	rs4488809	Asian	Intron	4.2×10^{-25} [23]
3	189357602	*TP63*	rs13314271	European	Intron	7.22×10^{-10} [24]
3	189383183	*TP63*	rs10937405	Asian	Intron	7×10^{-17} [25] 7×10^{-12} [26]
5	1286516	*TERT*	rs2736100	Asian, European	Intron	2.50×10^{-32} [25] 4×10^{-27} [6] 2.60×10^{-20} [27] 3×10^{-11} [26] 3.74×10^{-14} [28]
5	1287194	*TERT*	rs2853677	Asian	Intron	3×10^{-40} [25]
5	1325803	*CLPTM1L*	rs465498	Asian	Intron	1.20×10^{-13} [23]
5	146644115	*STK32A*	rs2895680	Asian	Intron	3.22×10^{-11} [29]
6	32368087	*BTNL2*	rs3817963	Asian	Intron	3×10^{-10} [25]
6	32433167	*HLA-DRA*	rs2395185	Asian	Intergenic	9.47×10^{-10} [6]
6	41493412	*FOXP4*	rs7741164	Asian	Intron	1.22×10^{-12} [30]
6	117786180	*ROS1, DCBLD1*	rs9387478	Asian	Intergenic	1.55×10^{-9} [6]
9	22160087	*CDKN2B-AS1*	rs72658409	Asian	Intergenic	1.94×10^{-9} [30]
10	114498476	*VTI1A*	rs7086803	Asian	Intron	1.19×10^{-11} [6]
12	52349071	*ACVR1B*	rs11610143	Asian	Intron	2.25×10^{-9} [30]
13	24293859	*MIPEP*	rs753955	Asian	Intergenic	3.90×10^{-10} [23]
15	78806023	*HYKK (AGPHD1)*	rs8034191	European	Intron	1.46×10^{-15} [28]
15	78894339	*CHRNA3*	rs1051730	European	Exon: Synonymous	7.10×10^{-19} [28]
17	65898809	*BPTF*	rs7216064	Asian	Intron	7×10^{-11} [25]

To be included in our study, LUAD risk must have been specifically mentioned as affected by the SNP and the reported p-value must be $p \leq 5 \times 10^{-8}$. Chr position (hg19): Based on human genome 19 numbering; Genes: Gene or nearest gene (if the SNP is not in a gene body) using Human Genome Organization name.
Chr: Chromosome; GWAS: Genome-wide association study; LUAD: Lung adenocarcinoma.

Cell isolation, culture & quality control

Human AT2 cells were isolated from a deidentified nonsmoker remnant transplant lung as described [40]. The subject in this study was a nonsmoking male 62-year-old male who died of cardiovascular disease. AT2 cells were plated and differentiated into AT1-like cells over the course of 6 days as described [40]. A549, H1648 and H522 cells were cultured in RPMI 1640 w/l-glutamine (Lonza #12-702F, MD, USA) supplemented with 10% fetal bovine serum and 100 units/ml of penicillin/streptomycin. All cell lines maintained by our laboratory are routinely tested and are negative for mycoplasma, and cell line identity was confirmed by DNA fingerprinting.

AEC RNA isolation, RNA-seq, protein isolation & western blot analysis

RNA was obtained from D0 (AT2), D2, D4 and D6 (AT1-like) cells and sequenced. Briefly, total cell RNA was DNase I digested and then subjected to ribosomal RNA depletion with the Ribominus™ Eukaryote v2 kit (Life Technologies/Thermo Fisher Scientific #A15020, MA, USA). Libraries were constructed with the TruSeq RNA Sample Prep Kit v2 (Illumina Inc. # RS-122-2001, CA, USA) and underwent Illumina HiSeq 2000 paired-end sequencing (2×50 bp) according to the manufacturer's instructions. Sequence reads were aligned to hg19 with TopHat2 v2.0.7 using default settings [41]. After mapping, data were further analyzed using Cufflinks [42] transcript assembly and quantification software (version 2.2.1) with default parameters and sequence bias detection and correction. Protein lysates obtained on days 0, 2, 4 and 6 were analyzed by western blot to confirm proper *in vitro* differentiation, using antibodies specific for AT1 cells (anti-aquaporin 5 [AQP-005], Alomone Labs, Jerusalem, Israel) and anti-podoplanin (8.1.1, Santa Cruz Biotechnology, CA, USA) and AT2 cells (anti-prosurfactant protein C, AB3786, MilliporeSigma, MA, USA), with loading control anti-actin (AC-15 NB600-501, Novus Biologicals, CO, USA).

LUAD risk-associated SNPs collection

LUAD risk index SNPs were collected from published GWAS papers. A p-value cutoff of $\leq 5 \times 10^{-8}$ was applied for genome-wide significance. SNPs in high LD with LUAD index SNPs ($r^2 > 0.5$) were retrieved using the online

SNP annotation tool HaploReg v3, which calculates r^2 using data from the 1000 Genomes Project [43]. Because LD varies by ethnicity, this analysis was carried out taking into account the ethnicity of the population in which each index SNP was identified. Functional predictions for coding SNPs and miRNA targets were performed using ANNOVAR [44], which integrates data from PolyPhen-2 [45], SIFT [46] and TargetScan [47].

ChIP-seq & FAIRE-seq

Cross-linking and sonication for ChIP-seq and FAIRE-seq were performed using cells from D0 (AT2), and cells from D4 and D6 (AT1-like) as described [40]. For ChIP-seq, chromatin was incubated with antibodies against H3K4me1 (catalog #pAb-037-050, Diagenode, NJ, USA) and H3K27ac (catalog #39133, Active Motif, CA, USA) after sonication, and enrichment of AEC ChIP targets was confirmed by qPCR for each cell type. Libraries were created at the University of Southern California Epigenome Center and underwent Illumina GAII single-end sequencing as previously described. We used cutadapt-1.5 to filter low-quality reads. Remaining high-quality reads were aligned to reference human genome hg19 using bwa-0.7.7 with two mismatches allowed and mapping quality thresholds set to 20. Duplicate reads were removed with picard-tools-1.107. Peaks were called using SICER [48] (for FAIRE-seq, H3K27ac ChIP-seq and H3K4me1 ChIP-seq, respectively, window sizes of: 50, 100, 200 bp; gap sizes of: 50, 200 and 200 bp were used with a false discovery rate (FDR) cutoff = 1×10^{-4}). Saturation plots (Supplementary Figure 3) were generated using the same peak-calling methods on a proportion of reads. AEC ChIP-seq density plots (Supplementary Figure 5) were generated by running annotatePeaks.pl in Hypergeometric Optimization of Motif EnRichment (HOMER) [49].

For featured LUAD cell lines, raw FASTQ data for H3K27ac and H3K4me1 were downloaded from DataBase of Transcriptional Start Sites file transfer protocol (http://dbtss.hgc.jp) [50]. Quality control, alignment, duplicate removal and peak calling were performed as described for AECs.

Identification of enhancer-associated SNPs

Candidate enhancer-associated SNPs were identified based on their position in both H3K27ac and H3K4me1 peaks in epigenomic data from D0 (AT2), and/or D4/D6 (AT1-like). Candidate enhancer-associated SNPs that were significantly enriched ($p < 1 \times 10^{-3}$) for both H3K27ac and H3K4me1 signals were designated 'enhancer-associated SNPs' (AT1- or AT2-specific or general AEC enhancers). Enrichment analysis was performed in a 0.5-kb window flanking each SNP using the HOMER script getDifferentialPeaks and the R/Bioconductor packages rtracklayer and GenomicRanges (Supplementary Figure 6 & https://www.bioconductor.org).

TF-binding prediction

Sequences (±25 bp) around each AEC enhancer-associated SNPs containing Ref/Alt alleles were extracted using the R package 'BSgenome.Hsapiens.UCSC.hg19'. Motif positional weight matrixes compiled from Encyclopedia of DNA Elements (ENCODE)-motif [51], Factorbook [52] and HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) [53] were downloaded and transformed into meme format. Find Individual Motif Occurrences (FIMO) [54] was used to predict TF-binding sites. We employed two approaches in TF motif identification. First, we used FIMO to predict TF motifs that overlap each SNP, with a significance threshold of $p = 1 \times 10^{-4}$ based on Factorbook and HOCOMOCO motif databases. Second, we identified TFs with the potential to bind specific SNPs via integrating ENCODE/ Gene Expression Omnibus (GEO) TF ChIP-seq peaks over the locations of SNPs of interest. We then used FIMO to validate whether there are corresponding TF motifs formed/disrupted by the SNP, and set a less stringent threshold of $p = 1 \times 10^{-3}$ for this step because binding predictions are supported by publicly available ChIP-seq data for each factor at each SNP described and depicted in this manuscript. In addition, we checked the genotype of the SNP in available ChIP-seq data to determine whether the expected SNP allele forms the corresponding TF motif, so that matched motif-TF events were identified. Furthermore, we limited all of our TFBS analyses to those cognate TFs with an AEC RPKM ≥3, ensuring predicted TFs are expressed in AEC (Supplementary Figure 7).

Plasmid construction & luciferase enhancer assays

Putative enhancers spanning each SNP as well as the nearest DNAse HSS in A549 cells (ENCODE) were amplified by PCR from normal human male DNA (Promega #G1471, WI, USA) using the primers listed in Supplementary Table 7. Amplicons were subcloned into the pGL4.26 luciferase plasmid (Promega, #E844A) upstream of a minimal promoter. The NEB Q5® Site-Directed Mutagenesis Kit (New England Biolabs #E0554S, MA, USA) was used to change the allele under study. All constructs were verified by sequencing. All cell lines were transfected with the indicated constructs 48 h prior to being harvested. A549 and H522 cells were transfected with Fugene HD (Promega #E2311) according to the manufacturer's instructions and H1648 cells were transfected with Lipofectamine® 3000 (Thermo Fisher Scientific #L3000008, MA, USA). Luciferase assays were performed with the Dual-Luciferase Reporter Assay System (Promega #E1960) on the 96-well LUMIstar Omega Luminescence Microplate Reader (BMG LABTECH, NC, USA) according to the manufacturer's instructions. All reported allele-specific enhancer activity represents the mean ± standard error of the mean of three or more independent biological replicates assayed as technical triplicates.

eQTL tests

Tests in GTEx: eQTLs for rs452384 and rs6942067 were identified by directly searching the website of GTEx portal website [55]. Tests in TCGA: To identify eQTLs for rs6942067 in TCGA, we downloaded genomic information for 428 LUAD samples from the TCGA data portal website (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). This included level-3 RNA-seq data, genotype and methylation data. Expression data were log2 transformed ((log2(rsem+1)). Level-4 CNV data were downloaded directly from the website (http://gdac.broadinstitute.org/runs/analyses__2014_01_15/data/LUAD/20140115/). For each gene, we calculated the methylation level as the mean β-value of all the CpGs within ±1 kb of the TSS. We used a multiple linear regression model to test the association between SNP genotype and gene expression, with adjustment for gene-level CNV and DNA methylation because CNV and DNA methylation might also influence gene expression. Since rs6942067 does not appear in the Affymetrix Genome-Wide Human SNP Array 6.0 platform (SNP 6.0), we chose rs6930292 ($r^2 = 1$, $D' = 1$ with rs6942067) as a surrogate for eQTL testing. For rs6942067, we performed eQTL analysis for rs6930292 against all genes within the flanking 1Mb region in the TCGA LUAD dataset. We also ruled out genes that are lowly expressed ((log2(rsem+1) <3). FDR was calculated using the Benjamini–Hochberg method, and the FDR cutoff was set at 0.05. Plots were generated using ggplot2 in R 3.1.1.

TCGA RNA expression analysis

Gene expression FPKM data (fragments per kilobase of transcript per million reads) from 521 LUAD tumor samples and 59 normal samples were downloaded using the TCGAbiolinks package (https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html) and log2 transformed. Differential expression between tumor and normal tissue was assessed via Student's t-test. Plots were generated using ggplot2 in R 3.4.3.

AEC RNA expression analysis

RNA-seq was performed on the AEC sample for days 0/2/4/6, respectively. Data were processed as described above in 'AEC RNA isolation, RNA-seq, protein isolation and western blot analysis'. RPKMs were log2-transformed to represent the expression level.

TCGA Kaplan–Meier analysis

We used the TCGAanalyze_SurvivalKM function included in the TCGAbiolinks package. After removing duplicate tumor samples from the same subjects, we compared the survival of subjects in the highest tertile of *DCBLD1* expression (in FPKM) to subjects in the lowest tertile of expression. This resulted in a comparison of 161 subjects in the highest expression group to 161 subjects in the lowest expression group.

Data visualization

The Integrative Genomics Viewer 2.3.34 was used to visually inspect and graph all sequencing data [56]. All ChIP-seq quality control plots were generated using R 3.1.1.

Results & discussion

Classification of SNPs associated with LUAD risk

We first carried out a literature review, collecting all SNPs reported to be significantly associated with LUAD risk. We filtered for genome-wide significance ($p < 5 \times 10^{-8}$) and for those SNPS that had been validated, in other words, observed to be significant in more than one dataset. We thus collected 18 validated GWAS index SNPs that had been significantly associated with LUAD risk (Table 1 & Figure 1A). In agreement with the general observation that most GWAS SNPs are located in noncoding regions [9], 17 were found to be either intronic or intergenic, while

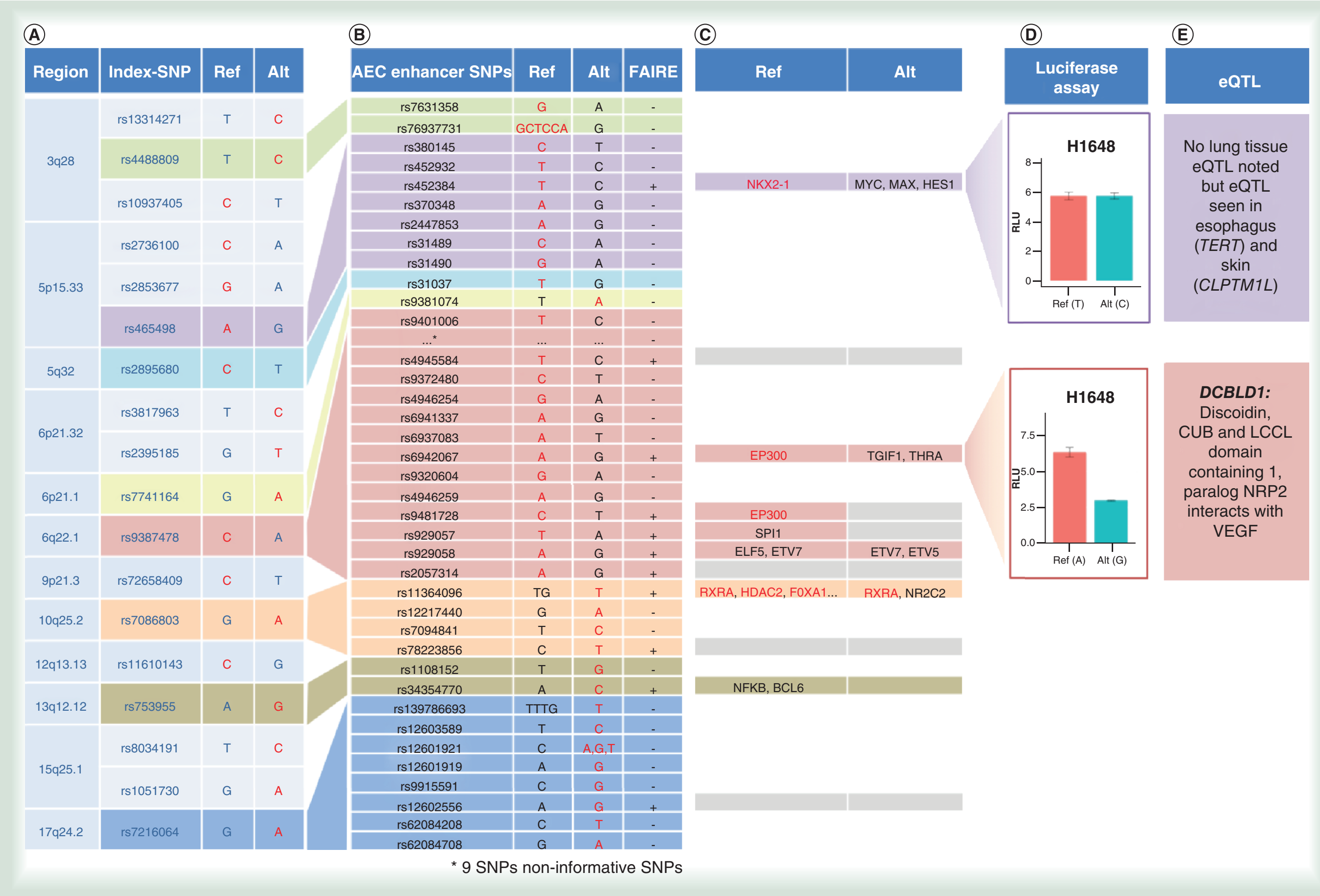


Figure 1. Data generation and analysis flow chart. (A) Eighteen validated LUAD index SNPs with alleles indicated (the risk allele is colored red). **(B)** Forty-seven AEC enhancer SNPs identified by integrating AEC ChIP-seq data, with alleles indicated (the inferred risk allele is colored red); presence in a FAIRE peak is marked as +. **(C)** SNP-affected TF-binding motifs predicted in AEC enhancer SNPs in FAIRE peaks; those TFs whose binding is also supported by ChIP-seq evidence from publicly available datasets are colored red. SNPs for whom no TF-binding motifs were predicted to be affected and no TF binding by ChIP has been reported are indicated by a gray box. **(D)** Luciferase enhancer assays for both alleles of regions containing rs452384 and rs6942067, tested in LUAD cell line H1648. **(E)** Potential eQTLs in lung tissues of the two investigated candidate AEC enhancer SNPs.
AEC: Alveolar epithelial cell; ChIP: Chromatin immunoprecipitation; eQTL: Expression quantitative trait loci; FAIRE: Formaldehyde-assisted isolation of regulatory elements; LUAD: Lung adenocarcinoma; TF: Transcription factor.
For color figures, please see online at https://www.futuremedicine.com/doi/full/10.2217/epi-2018-0003

one, rs1051730, falls within an exon resulting in a synonymous mutation. We next obtained 930 SNPs in high LD ($r^2 \geq 0.5$) with the 18 index SNPs using 1000 Genomes LD data [43] matched for the ethnicity of the population in which the index risk SNP was identified (Supplementary Table 1), and grouped these into functional classes for subsequent analysis (Supplementary Table 2). Of the 948 total index and LD SNPs, 583 were intronic, while 335 were intergenic. Only seven SNPs mapped to coding regions, including three missense mutations and four synonymous mutations. Seven SNPs were located in the 5′- or 3′-untranslated regions (UTRs), where miRNAs are most often targeted, and 16 in putative promoter regions (within 1000 bp to either side of the transcription start site or TSS).

Very little published data on the potential functionality of the 948 SNPs are available. Several SNPs in the 15q25.1 locus have been implicated as potential functional SNPs affecting expression of the cholinergic receptor, CHRNA5, a protein involved in the nicotine response [57,58]. Rs16969968 resides in the fifth exon of *CHRNA5* and changes an aspartic acid codon (GAT) into an asparagine codon (AAT) at amino acid 398 in the second intracellular loop. This makes the receptor less responsive to nicotine, thereby increasing nicotine dependence, heavy smoking and potentially, lung cancer risk [59]. Rs55853698 and rs55781567, located in the 5′UTR of *CHRNA5*, have been reported to alter *CHRNA5* promoter activity [60].

To preliminarily evaluate whether the remaining two missense-causing SNPs (rs2076530 in *BTNL2* and rs9891146 in *C17orf58*) might alter protein function, we performed structure–activity predictions using PolyPhen2 and SIFT software. Neither of the missense mutations were predicted to be deleterious by either program (Supplementary Table 3), decreasing the likelihood that these SNPs might alter the function or structure of the proteins. To investigate whether the seven SNPs in UTRs might affect miRNA targeting, we performed sequence-activity-binding predictions using the miRNA target database TargetScanHuman. No allele for any UTR SNP was highlighted as a known miRNA target (Supplementary Table 4). As noted above, of the SNPs located in promoter regions, only the two CHRNA5 SNPs have been implicated in affecting promoter function (Supplementary Table 5).

In sum, except for several SNPs in the 15q25.1 locus, all loci appear to lack a biological mechanism explaining their association with increased LUAD risk. Because enhancers have been commonly implicated in risk SNP function [10,12–16,20–22], we explored whether one or more of the SNPs might influence risk by affecting AEC enhancers. To be thorough, we included all 948 SNPs in the investigation.

Identification of AEC enhancers

As a first step in generating AEC enhancer profiles, we purified AT2 cells from remnant human transplant lung. We used cells from a nonsmoker to ensure optimally healthy cells, thereby limiting disease confounders. AT1 cells are too delicate to allow purification in sufficient numbers, necessitating derivation of AT1-like cells through *in vitro* differentiation over the course of 6 days as described [40] (Supplementary Figure 2A). At day 2 in culture, AT2 cells undergo dramatic changes in gene expression and epigenomic marks, and by days 4 and 6, they have transdifferentiated into cells that exhibit characteristics of native AT1 cells. We confirmed proper differentiation of the cells by western blot analysis (Supplementary Figure 2B). We performed chromatin immunoprecipitation and subsequent DNA sequencing (ChIP-seq) using material from day 0 (D0; AT2 cells) and days 4 and 6 (D4, D6; AT1-like cells) using antibodies against enhancer marks H3K27ac and H3K4me1. In D0, D4 and D6 cells, respectively, we identified 39,210, 44,976 and 43,743 H3K27ac peaks and 87,443, 76,292 and 75,490 H3K4me1 peaks (Supplementary Table 6; saturation plots are shown in Supplementary Figure 3A & B). Peak analysis revealed substantial overlap of peaks from different days (Supplementary Figure 4A & B).

In addition to ChIP-seq for enhancer marks, we also performed Formaldehyde-Assisted Isolation of Regulatory Elements-sequencing (FAIRE-seq [61]) on AECs from days 0, 4 and 6 in culture to identify nucleosome-depleted regions of the genome that are accessible to DNA-binding proteins (i.e., open chromatin, Supplementary Figure 3C). The D0 peaks were largely distinct from the D4/D6 peaks, while the latter showed substantial overlap (Supplementary Figure 4C). Plots of the tag density versus distance to the center of the FAIRE-seq peaks showed that FAIRE-seq signals were enriched in the center of FAIRE-seq peaks (nucleosome depletion), while as expected, signals for the histone modifications H3K27ac and H3K4me1 (on nucleosomes) were enriched in regions flanking FAIRE-seq peaks (Supplementary Figure 5).

Identification of SNPs present in enhancers

We next investigated which among the 948 candidate SNPs were located in AEC enhancer regions (i.e., were located in H3K27ac *and* H3K4me1 ChIP-seq peaks) in D0, D4 and/or D6 cells. We filtered out SNPs located on the extreme flanks of peaks by requiring that SNPs be significantly enriched in enhancer marks on both of the 0.5 kb flanking sides (Supplementary Figure 6). This yielded 47 AEC enhancer-associated SNPs that we classified as AT2 cell-specific (present only in D0 enhancers), AT1 cell-specific (present only in D4/D6 enhancers) and general AEC enhancer SNPs (present in both AT1 and AT2 enhancers). Seven AT2-specific SNPs, 23 AT1-specific SNPs and 17 AEC-specific SNPs were identified, respectively (Table 2 & Figure 1B). Notably, none of the index SNPs themselves were located in AEC, AT1 or AT2 cell enhancer elements. For ten of the 18 index SNPs, including the two near the CHRNA5 region, we detected no AEC enhancer-associated LD SNPs. This could be because these SNPs do not function through enhancers (e.g., the previously identified missense and promoter *CHRNA5* SNPs), because they affect LUAD risk through effects on other cell types (such as immune cells, lung fibroblasts, etc.) or because their function is influenced by other factors, such as smoking, which are not investigated here.

The 47 candidate AEC enhancer SNPs correspond to eight of the 18 index GWAS LUAD risk SNPs (Table 2 & Figure 1B). The number of candidate SNPs per region varied from one or two to 22 SNPs on 6q22.1, in LD

Table 2. Enhancer-associated SNPs in lung adenocarcinoma, listed by chromosome.

rsID	Chr	Pos (hg19)	Ref	Alt	r^2	D′	Index SNP_population	AEC type	Nearest gene	Annotation	FAIRE-seq peak
rs7631358	chr3	189348411	**G**	A	0.75	-0.94	rs4488809_ASN	AT1-specific	*TP63*	Intergenic	
rs76937731	chr3	189348968	**GCTCCA**	G	0.61	-0.92	rs4488809_ASN	AT1-specific	*TP63*	Intergenic	
rs380145	chr5	1328897	**C**	T	0.62	0.95	rs465498_ASN	AT1-specific	*CLPTM1L*	Intron	
rs452932	chr5	1330253	**T**	C	0.92	0.99	rs465498_ASN	AT1-specific	*CLPTM1L*	Intron	
rs452384	chr5	1330840	**T**	C	0.92	0.99	rs465498_ASN	AEC-specific	*CLPTM1L*	Intron	+
rs370348	chr5	1331219	**A**	G	0.89	0.95	rs465498_ASN	AT2-specific	*CLPTM1L*	Intron	
rs2447853	chr5	1333077	**A**	G	0.66	0.88	rs465498_ASN	AT1-specific	*CLPTM1L*	Intron	
rs31489	chr5	1342714	**C**	A	0.9	0.97	rs465498_ASN	AT1-specific	*CLPTM1L*	Intron	
rs31490	chr5	1344458	**G**	A	0.9	0.97	rs465498_ASN	AT2-specific	*CLPTM1L*	Intron	
rs31037	chr5	146615785	**T**	G	0.66	0.84	rs2895680_ASN	AT1-specific	*STK32A*	Intron	
rs9381074	chr6	41505196	T	**A**	0.7	0.86	rs7741164_ASN	AEC-specific	*FOXP4*	Intron	
rs9401006	chr6	117732924	**T**	C	0.63	0.8	rs9387478_ASN	AT1-specific	*ROS1*	Intron	
rs9385012	chr6	117733191	**T**	C	0.64	0.81	rs9387478_ASN	AT1-specific	*ROS1*	Intron	
rs9401007	chr6	117733275	**G**	A	0.64	0.81	rs9387478_ASN	AT1-specific	*ROS1*	Intron	
rs9385013	chr6	117733452	**C**	T	0.64	0.81	rs9387478_ASN	AT1-specific	*ROS1*	Intron	
rs1407185	chr6	117733650	**A**	T	0.61	0.82	rs9387478_ASN	AEC-specific	*ROS1*	Intron	
rs1407184	chr6	117733717	**G**	A	0.71	0.93	rs9387478_ASN	AEC-specific	*ROS1*	Intron	
rs1321816	chr6	117734533	**T**	C	0.66	0.84	rs9387478_ASN	AT1-specific	*ROS1*	Intron	
rs3777981	chr6	117735255	**A**	C	0.67	0.84	rs9387478_ASN	AEC-specific	*ROS1*	Intron	
rs2243	chr6	117737390	**G**	A	0.72	0.94	rs9387478_ASN	AEC-specific	*ROS1*	Intron	
rs9374658	chr6	117741495	**T**	G	0.73	0.94	rs9387478_ASN	AEC-specific	*ROS1*	Intron	
rs4945584	chr6	117750980	**T**	C	0.83	0.98	rs9387478_ASN	AT1-specific	*ROS1*	Intergenic	+
rs9372480	chr6	117760579	**C**	T	0.86	0.98	rs9387478_ASN	AT2-specific	*ROS1*	Intergenic	
rs4946254	chr6	117765313	**G**	A	0.87	0.97	rs9387478_ASN	AT2-specific	*ROS1*	Intergenic	
rs6941337	chr6	117768448	**A**	G	0.89	0.98	rs9387478_ASN	AEC-specific	*ROS1*	Intergenic	
rs6937083	chr6	117785308	**A**	T	1	1	rs9387478_ASN	AT1-specific	*DCBLD1*	Intergenic	
rs6942067	chr6	117785696	**A**	G	0.54	1	rs9387478_ASN	AT1-specific	*DCBLD1*	Intergenic	+
rs9320604	chr6	117816045	**G**	A	0.76	0.93	rs9387478_ASN	AT1-specific	*DCBLD1*	Intron	
rs4946259	chr6	117816093	**A**	G	0.56	0.95	rs9387478_ASN	AT1-specific	*DCBLD1*	Intron	
rs9481728	chr6	117817165	**C**	T	0.76	0.93	rs9387478_ASN	AEC-specific	*DCBLD1*	Intron	+
rs929057	chr6	117818911	**T**	A	0.76	0.93	rs9387478_ASN	AEC-specific	*DCBLD1*	Intron	+
rs929058	chr6	117819198	**A**	G	0.67	0.93	rs9387478_ASN	AEC-specific	*DCBLD1*	Intron	+
rs2057314	chr6	117819357	**A**	G	0.76	0.93	rs9387478_ASN	AEC-specific	*DCBLD1*	Intron	+
rs11364096	chr10	114502022	TG	**T**	1	1	rs7086803_ASN	AEC-specific	*VTI1A*	Intron	+
rs12217440	chr10	114502218	G	**A**	1	1	rs7086803_ASN	AT2-specific	*VTI1A*	Intron	
rs7094841	chr10	114502411	T	**C**	1	1	rs7086803_ASN	AEC-specific	*VTI1A*	Intron	
rs78223856	chr10	114527225	C	**T**	0.63	0.88	rs7086803_ASN	AT1-specific	*VTI1A*	Intron	+
rs1108152	chr13	24312446	T	**G**	0.65	0.92	rs753955_ASN	AT1-specific	*MIPEP*	Intron	
rs34354770	chr13	24327618	A	**C**	0.59	0.88	rs753955_ASN	AEC-specific	*MIPEP*	Intron	+
rs139786693	chr17	65825228	TTTG	**T**	0.75	0.9	rs7216064_ASN	AT1-specific	*BPTF*	Intron	
rs12603589	chr17	65825248	T	**C**	0.8	0.91	rs7216064_ASN	AT1-specific	*BPTF*	Intron	+
rs12601921	chr17	65825354	C	**A,G,T**	0.72	0.92	rs7216064_ASN	AT1-specific	*BPTF*	Intron	
rs12601919	chr17	65825374	A	**G**	0.73	0.92	rs7216064_ASN	AT2-specific	*BPTF*	Intron	
rs9915591	chr17	65826090	C	**G**	0.81	0.92	rs7216064_ASN	AEC-specific	*BPTF*	Intron	
rs12602556	chr17	65826861	A	**G**	0.81	0.92	rs7216064_ASN	AEC-specific	*BPTF*	Intron	
rs62084208	chr17	65827443	C	**T**	0.81	0.92	rs7216064_ASN	AT1-specific	*BPTF*	Intron	
rs62084708	chr17	66049707	G	**A**	0.51	0.73	rs7216064_ASN	AT2-specific	*KPNA2*	Intergenic	

AEC: Alveolar epithelial cell; Alt: Alternate allele (risk alleles indicated in bold); ASN: Asian; AT1: Alveolar-type 1 cell; AT2: Alveolar-type 2 cell; Chr: Chromosome; Chr position (hg19): Based on human genome 19 numbering; FAIRE-seq: Formaldehyde-assisted identification of regulatory elements-sequencing; Nearest gene: Gene or nearest gene (if the SNP is not in a gene body) using Human Genome Organization name; r2 and D′: Linkage disequilibrium information based on the population in which risk was identified for each SNP; Ref: Reference allele (risk alleles indicated in bold).

Table 3. Transcription factor-binding motif prediction for alveolar epithelial cell enhancer SNPs in formaldehyde-assisted isolation of regulatory elements-sequencing peaks.

rsID	Ref	Alt	ChIP-seq TF binding		
			Hom Ref	Het	Hom Alt
rs452384	**NKX2-1**	MYC, MAX, HES1	**NKX2-1**, MYC, FOS		
rs6942067	**EP300**	TGIF1, THRA	POLR2A, **EP300**	POLR2A, FOSL2, GATA2, JUN, FOS	JUND
rs9481728	**EP300**		POLR2A, **EP300**	GATA2, FOS, MAFK	
rs929057	SPI1				
rs929058	ELF5, ETV7	ETV7, ETV5			
rs2057314			PolR2A	PolR2A, TCF7L2, TCF12, SPI1, FOS	NFIC, SPI1
rs11364096	**RXRA, HDAC2, SP1, FOXA**, NR1H2, PPARG	RXRA, NR2C2	**RXRA, HDAC2, SP1, FOXA1**, FOXA2, EP300, CEBPB, GATA2, TCF7L2, NFIC, ZNF217, ESR1, MYBL2, POLR2A, ARID3A, STAT1, STAT3, FOS, HNF4G, NKX2-1		
rs34354770	NFKB, BCL6		FOXA2		
rs12602556				RBBP5, EGR1	

Two of the 11 SNPs in FAIRE peaks (rs4945584 and rs78223856) showed no predicted TF motifs or evidence of ChIP-seq TF binding and are not listed. Ref: TFs indicated in **bold** were both predicted *and* detected by ChIP-seq; Alt: TFs indicated in **bold** were both predicted and detected by ChIP-seq.
AEC: Alveolar epithelial cell; Alt: Alternate allele; ChIP-seq: Chromatin immunoprecipitation and subsequent DNA sequencing; EP300: E1A-binding protein P300; FAIRE: Formaldehyde-assisted identification of regulatory elements; Het: ChIP-seq experiment carried out with cells containing heterozygous alleles; Hom Alt: ChIP-seq experiment carried out with cells containing a homozygous alternate allele; Hom Ref: ChIP-seq experiment carried out with cells containing a homozygous reference allele; Ref: Reference allele; TF: Transcription factor; THRA: Thyroid hormone receptor alpha.

with index SNP rs9387478. The latter SNPs are located between *DCBLD1* and *ROS1* and contains numerous active chromatin marks. Two SNPs in high LD with rs4488809 on chromosome 3q28 were found in general AEC enhancer marks located in the promoter region (-1000 bp to +100 bp from TSS) of TP63 (Supplementary Table 5), suggesting an enhancer close to a TSS. Such an enhancer may or may not affect the nearest gene [62]. In studies of risk enhancers, SNPs located in nucleosome-free, TF-accessible regions of the regulatory element are the most promising candidates for functional follow-up [9]. We, thus, filtered the 47 candidate SNPs by whether they were located in FAIRE peaks, which resulted in 11 top candidates (Supplementary Figure 1B). To further evaluate these 11 SNPs, we assessed the likelihood that they would disrupt or create TF-binding sites.

Identification of SNPs that affect TF-binding sites

We compiled motif positional weight matrixes from ENCODE-motif [51], Factorbook [52] and HOCOMOCO [53] (Supplementary Figure 7). We removed TFs that are not expressed in AT1 and/or AT2 cells by excluding sites for TFs with low reads per kilobase per million (RPKM; set at RPKM ≤ 3, from RNA-seq data [63]); binding sites for TFs that are not expressed in alveolar epithelium would not be expected to have any functional effects in these cells. Our analysis identified predicted TF-binding sites overlapping with seven of the 11 candidate risk enhancer SNPs (Figure 1C & Table 3). Next, we searched publicly available ChIP-seq data from ENCODE (Encyclopedia of DNA Elements) [64] and GEO (Gene Expression Omnibus) [65] for experimental evidence supporting binding of predicted TFs to the SNP regions. This approach is limited by the fact that publicly available ChIP-seq datasets such as those generated by the ENCODE consortium are restricted to the factors that have been examined to date. In this latter analysis, we included all ENCODE cell types, including but not limited to lung-relevant cell types such as A549 (a lung cancer cell line), small airway epithelial cells (SAECs, normal primary cells from the more distal airways), IMR90 (a cell line derived from embryonic lung fibroblast), normal human lung fibroblasts and human pulmonary fibroblasts (obtained from a different source than normal human lung fibroblasts). All of these cell types are distinct from AEC, and it is well recognized that enhancers vary considerably among cell types. However, if a predicted TF was reported to bind its target by ChIP-seq in any cell type, we took this as confirmation that the factor can actually bind to the predicted target in the context of a cellular environment. For four of the seven SNPs, we found ChIP-seq data-supporting binding of the TFs we had predicted. In addition, ChIP-seq data-based TF binding was noted for two SNP locations for which we had predicted no TF-binding sites (rs2057314 and

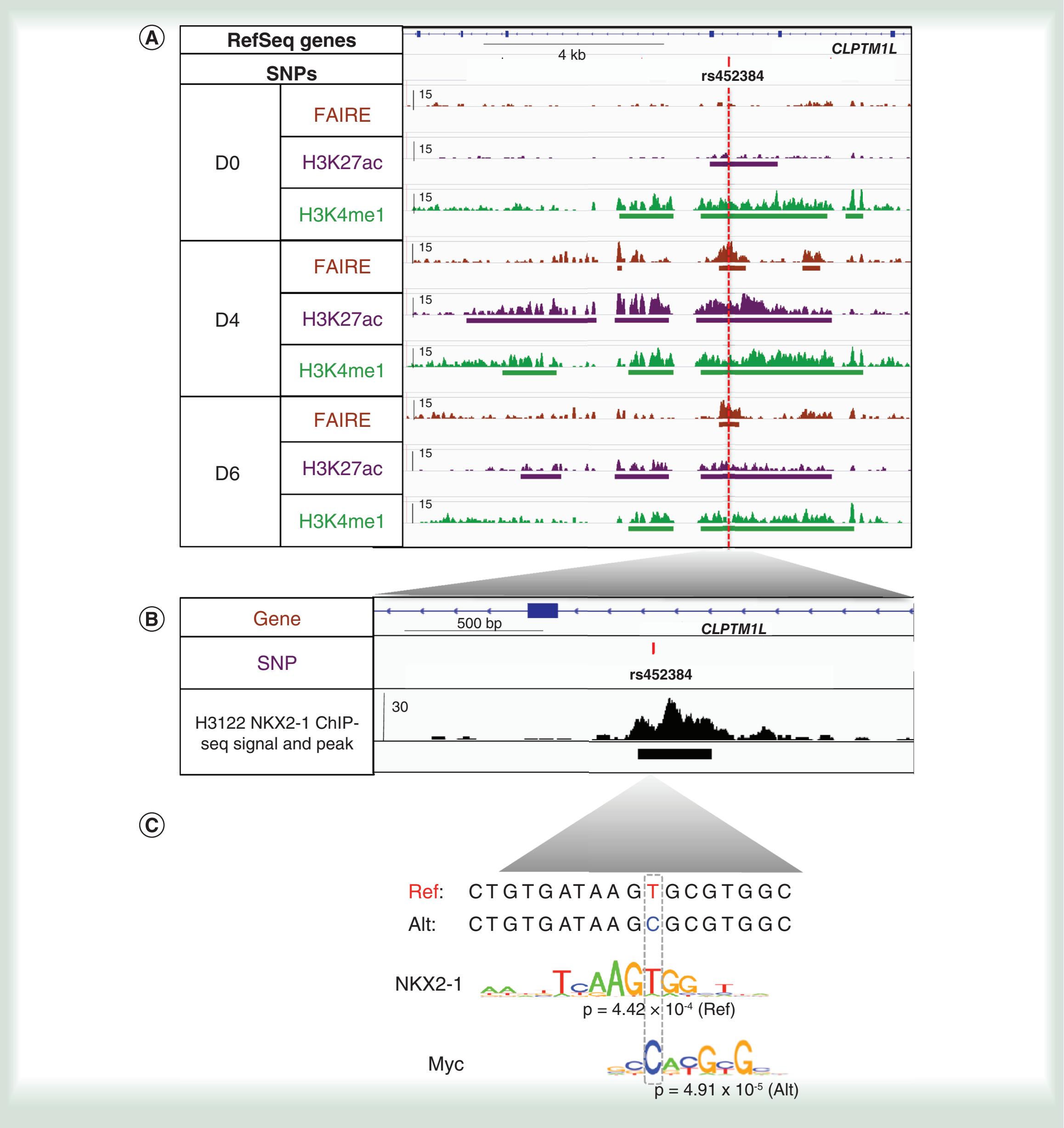


Figure 2. Epigenetic analysis of the rs452384 region. (A) Rs452384 is located in an intron of *CLPTM1L*. Histone enhancer marks and FAIRE peaks in the rs452384 region on days 0, 4 and 6 of AEC culture are indicated, and underneath each track, peaks called using SICER are marked. **(B)** ChIP-seq data for TF NKX2-1 from H3122 LUAD cells [63], which are homozygous for the reference (risk) T allele. **(C)** Predicted TF motifs for the sequences containing the reference (risk) and alternate alleles.
AEC: Alveolar epithelial cell; ChIP-seq: Chromatin immunoprecipitation and subsequent DNA sequencing; FAIRE: Formaldehyde-assisted identification of regulatory elements; LUAD: Lung adenocarcinoma; SICER: Spatial Clustering for Identification of ChIP-Enriched Regions – a peak-calling method; TF: Transcription factor.

rs12602556, Table 3). This could happen when TFs bind in multiprotein complexes or bind indirectly to DNA. In the case of indirect binding, we assumed that an effect of the SNP would be less likely. We next focused on three SNPs for which predicted TF-binding sites were supported by ChIP-based evidence (Figure 1C). These were: rs452384, corresponding to index SNP rs465498; and rs6942067 and rs9481728, corresponding to index SNP rs9387478.

Rs452384 lies in a putative general AEC enhancer in a central intron of the *CLPTM1L* gene (Figure 2A). *CLPTM1L* was recently implicated in lung tumorigenesis [66] and encodes a membrane protein that leads to apoptosis when overexpressed in cisplatin-sensitive cells [67]. This SNP also resides in a strong DNase-hypersensitive site in SAECs (ENCODE data [51,52]). Polymorphisms in *CLPTM1L* and the neighboring *TERT* gene have been reported to increase susceptibility to numerous cancers, including lung, pancreatic and breast cancers [23,68–71]. The reference (risk) allele of rs452384 forms a predicted site for NKX2-1 that is disrupted by the alternate allele (Table 3). Importantly, the ability of this TF to bind to this site is supported by ChIP-seq data in LUAD cell line H3122 (homozygous for the risk allele) [63] (Figure 2B & C). NKX2-1 plays a key role in driving lung epithelial tissue differentiation from endoderm [72,73] and is one of the most significantly amplified genes in LUAD [65]. Interestingly, a binding site for the *MYC* proto-oncogene is predicted for the alternate rs452384 allele (Figure 2C). *MYC* is implicated in numerous cancers including LUAD [74–76] and has been ChIPed at rs452384 by the ENCODE consortium in the MCF10A-Er-Src cell line (homozygous for the risk allele) [64].

Rs6942067 is located in the intergenic region between the *DCBLD1* and *ROS1* genes (Figure 3). Rs6942067 is positioned in the center of a DNase hypersensitive region between two nucleosomes carrying H3K27ac and H3K4me1 marks and the reference (risk) A allele lies in a predicted binding site for E1A-binding protein P300 (EP300) that is disrupted by the alternate allele (Table 3). EP300 binding to this site has been observed by ChIP-seq in retinoic acid-treated neuroblastoma cell line SK-N-SH_RA (A/A genotype; ENCODE data [64], Figure 3B & C). EP300 is a histone acetyltransferase that activates transcription via chromatin remodeling, and it has been implicated in a variety of cancers [77]. It is mutated and/or overexpressed in a low percentage of LUADs [78]. The alternate allele of rs6942067 forms a TF-binding site for the thyroid hormone receptor-α (Figure 3C), which is one of the several receptors for thyroid hormone known to be involved in lung development and alveolar cell function [79–81]. Notably, low levels of thyrotropin, a hormone that regulates thyroid hormone release, have been linked to an increased risk for prostate and lung cancer [82]. Like rs6942067, rs9481728 (located in the first intron of *DCBLD1*) carries enhancer histone marks and an EP300 site on the reference (risk) allele that is disrupted by the alternate allele (Figure 4 & Table 3).

The epigenomic environment of the three SNPs, coupled with the presence of overlapping TF-binding sites predicted to be affected by the SNP alleles, provided a strong rationale for functional analyses. This was further supported by the observation of active chromatin marks at the location of these three SNPs in LUAD cell lines [50] (Supplementary Figures 8–10). The first functional assay we carried out for the three candidate functional SNPs was a luciferase assay, in which we cloned the PCR-amplified putative enhancer region and inserted it upstream of a minimal promoter in the luciferase gene reporter vector pGL4.26. The assay allows both alleles to be tested in the same genetic and cellular environment. We then tested whether the genomic segment containing either allele of the SNP would exhibit enhancer activity, and whether the alleles differed in the extent to which they could enhance expression. Because immortalized human AEC are not available, we transfected each reporter construct into a LUAD cell line showing strong genomic H3K4me1 and H3K27ac peaks in the SNP region (based on publicly available H3K4me1 and H3K27ac ChIP-seq datasets for 26 lung cancer cell lines in DataBase of Transcriptional Start Sites, http://dbtss.hgc.jp) [50], reasoning that these cell lines would be enriched for TFs and coactivators required for enhancer activity at the locus of interest (Supplementary Figure 11).

Functional analyses of rs452384 on chromosome 5p15.33

To functionally study rs452384, we used a 459-bp region from chr5:1330187–1331866 (hg19) containing either the reference (risk) allele (T) or the alternate allele (C) upstream of the minimal promoter in pGL4.26 and transfected these constructs into H1648 cells, which bear strong enhancer marks in this region (http://dbtss.hgc.jp) [50] (Supplementary Figure 11). Both constructs elicited five- to sixfold higher luciferase activity over background (empty pGL4.26; Figure 5A), indicating an enhancer element is present. However, we did not observe significant differences in enhancer activity between the two alleles in H1648 cells. To ensure that we were not missing any key factors binding to adjacent sequences, we also studied a genomic fragment that had been expanded to the 5′- and 3′-end (as indicated in Supplementary Figure 12), but it did not show allele-specific activity either. We cannot exclude that differences in the TF profiles of AEC and H1648 might mask an allele-specific effect, or that specific environmental conditions (such as tobacco smoke exposure) might be required to reveal allele specificity. Because this SNP had been highly ranked in a fine mapping study [83], we also examined the online expression quantitative trait loci (eQTL) database, the Genotype-Tissue Expression (GTEx) project [55] for evidence of lung-specific eQTLs, but detected no significant lung eQTL. We also looked for eQTLs in LUAD tissues (n = 428) from The Cancer Genome Atlas (TCGA), correcting for copy number variation (CNV) and DNA methylation (the latter might affect gene expression in cancer), but detected no significant eQTLs either. The lack of an allele-specific difference

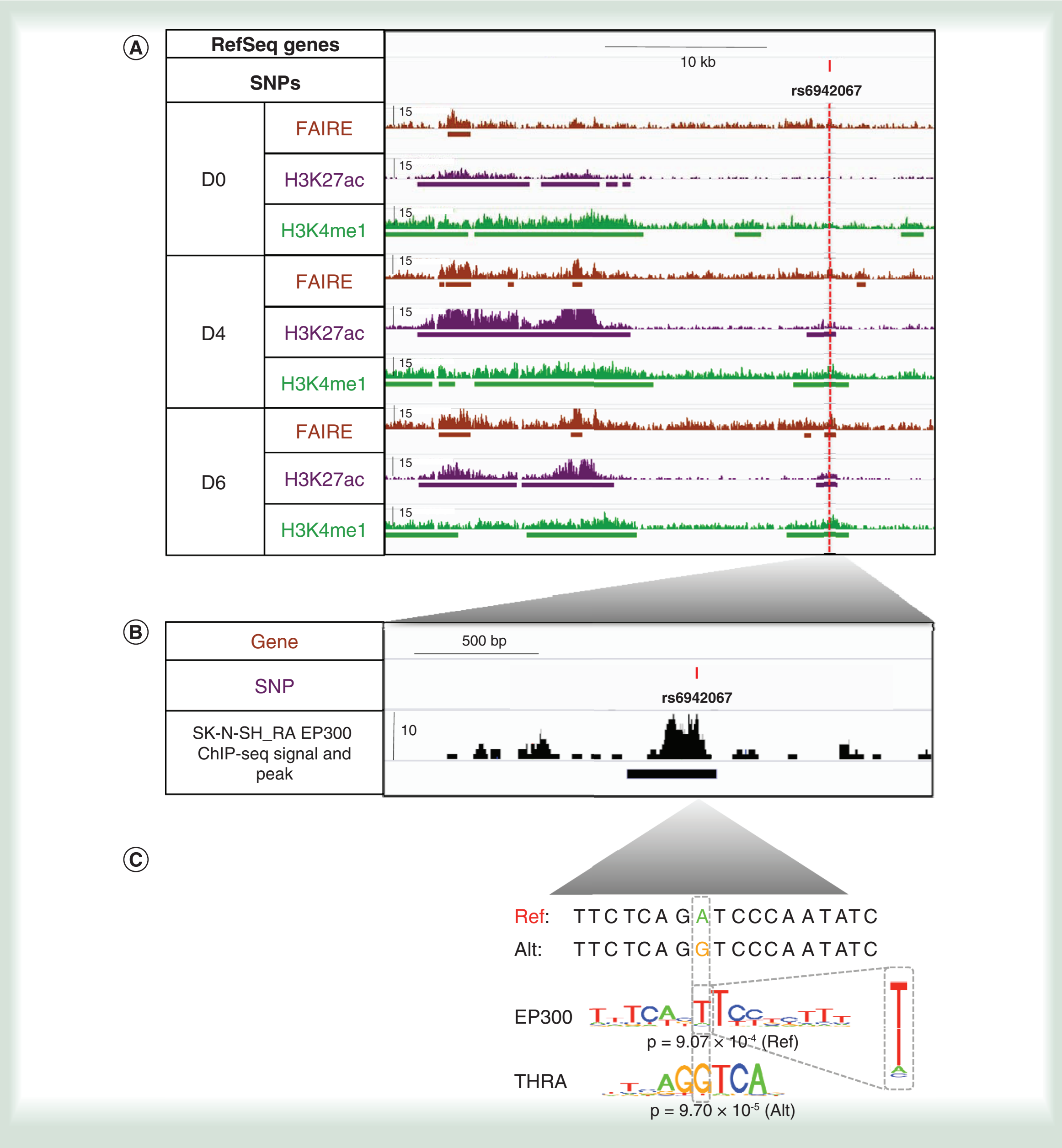


Figure 3. Epigenetic analysis of the rs6942067 region. (A) Rs6942067 is located in the intergenic region between gene *DCBLD1* and *ROS1*. Histone enhancer marks and FAIRE peaks in the rs6942067 region on days 0, 4 and 6 of AEC culture are indicated, and underneath each track, peaks called using SICER are marked. **(B)** ChIP-seq data for TF EP300 from SK-N-SH neuroblastoma cells treated with retinoic acid [64], which are homozygous for the reference (risk) A allele. **(C)** Predicted TF motifs for the sequences containing the reference (risk) and alternate alleles.
AEC: Alveolar epithelial cell; ChIP-seq: Chromatin immunoprecipitation and subsequent DNA sequencing; EP300: E1A-binding protein P300; FAIRE: Formaldehyde-assisted identification of regulatory elements; LUAD: Lung adenocarcinoma; SICER: Spatial Clustering for Identification of ChIP-Enriched Regions; TF: Transcription factor.

in the luciferase assay as well as the lack of a detectable eQTL in lung tissue suggests that, despite its presence in a TF-binding site in an AEC enhancer, this SNP may not be the functional SNP for index SNP rs465498. However, a limitation of GTEx and other available eQTL databases is that data are typically generated from whole tissue samples and rarely from purified cell populations. Thus, any true-cell-type-specific allelic expression will be diluted by the presence of other cell types, such as lung fibroblasts, macrophages or endothelial cells when 'lung eQTLs' are queried. In addition, we cannot exclude that a functional effect of this SNP may only be observable under certain

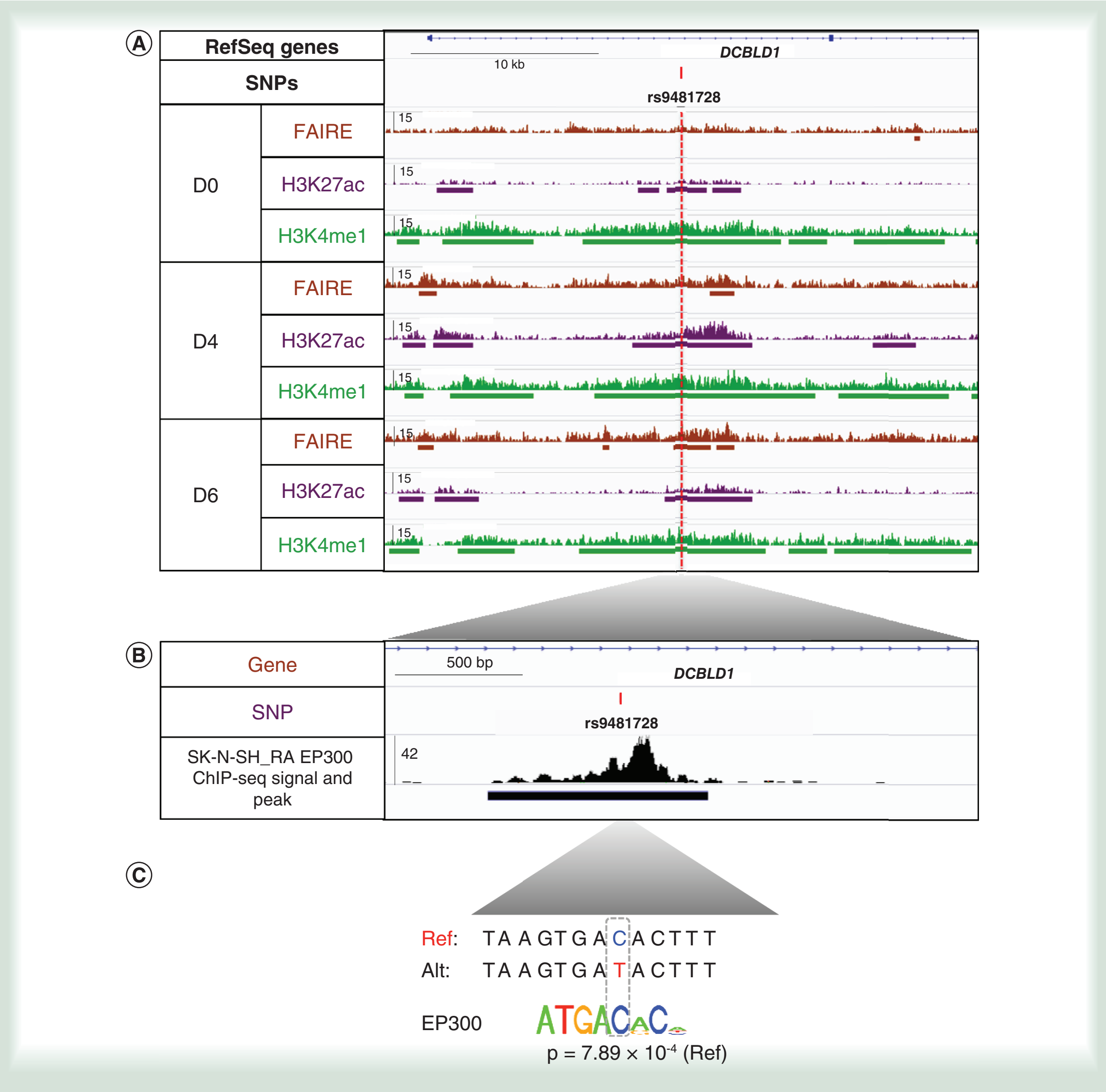


Figure 4. Epigenetic analysis of the rs9481728 region. (A) Rs9481728 is located in an intron of *DCBLD1*. Histone enhancer marks and FAIRE peaks in the rs9481728 region on days 0, 4 and 6 of AEC culture are indicated, and underneath each track, peaks called using SICER are marked. **(B)** ChIP-seq data for TF EP300 from SK-N-SH neuroblastoma cells treated with retinoic acid [64], which are homozygous for the reference (risk) A allele. **(C)** Predicted TF motifs for the sequence containing the reference allele (no sites were predicted for the alternate allele).
AEC: Alveolar epithelial cell; ChIP-seq: Chromatin immunoprecipitation and subsequent DNA sequencing; EP300: E1A-binding protein P300; FAIRE: Formaldehyde-assisted identification of regulatory elements; LUAD: Lung adenocarcinoma; SICER: Spatial Clustering for Identification of ChIP-Enriched Regions; TF: Transcription factor.

developmental or environmental conditions. We do note that in GTEx, the risk allele was associated with higher expression of the adjacent *TERT* gene in esophageal tissue ($p = 3.9 \times 10^{-9}$, Figure 5B & D) and the *CLPTM1L* gene in sun-exposed skin ($p = 2.1 \times 10^{-7}$, Figure 5D & E). *TERT* encodes telomerase, a ribonucleoprotein polymerase that is part of the complex that maintains telomere ends by addition of the telomere repeat TTAGGG, and its overexpression is a key component of the transformation process in many malignant cancer types including lung cancer [23,27,28,69,71,84,85]. *TERT* has been repeatedly implicated in lung cancer, both through mutations and polymorphisms that lie in *TERT* and *CLPTM1L*. Indeed, *TERT* is significantly overexpressed in LUAD compared with nontumor lung tissues in TCGA data, in which expression is very low ($p = 1.2 \times 10^{-38}$, Figure 5C). *CLPTM1L* is also elevated in LUAD ($p = 5.5 \times 10^{-62}$, Figure 5F) and was recently functionally implicated in lung cancer

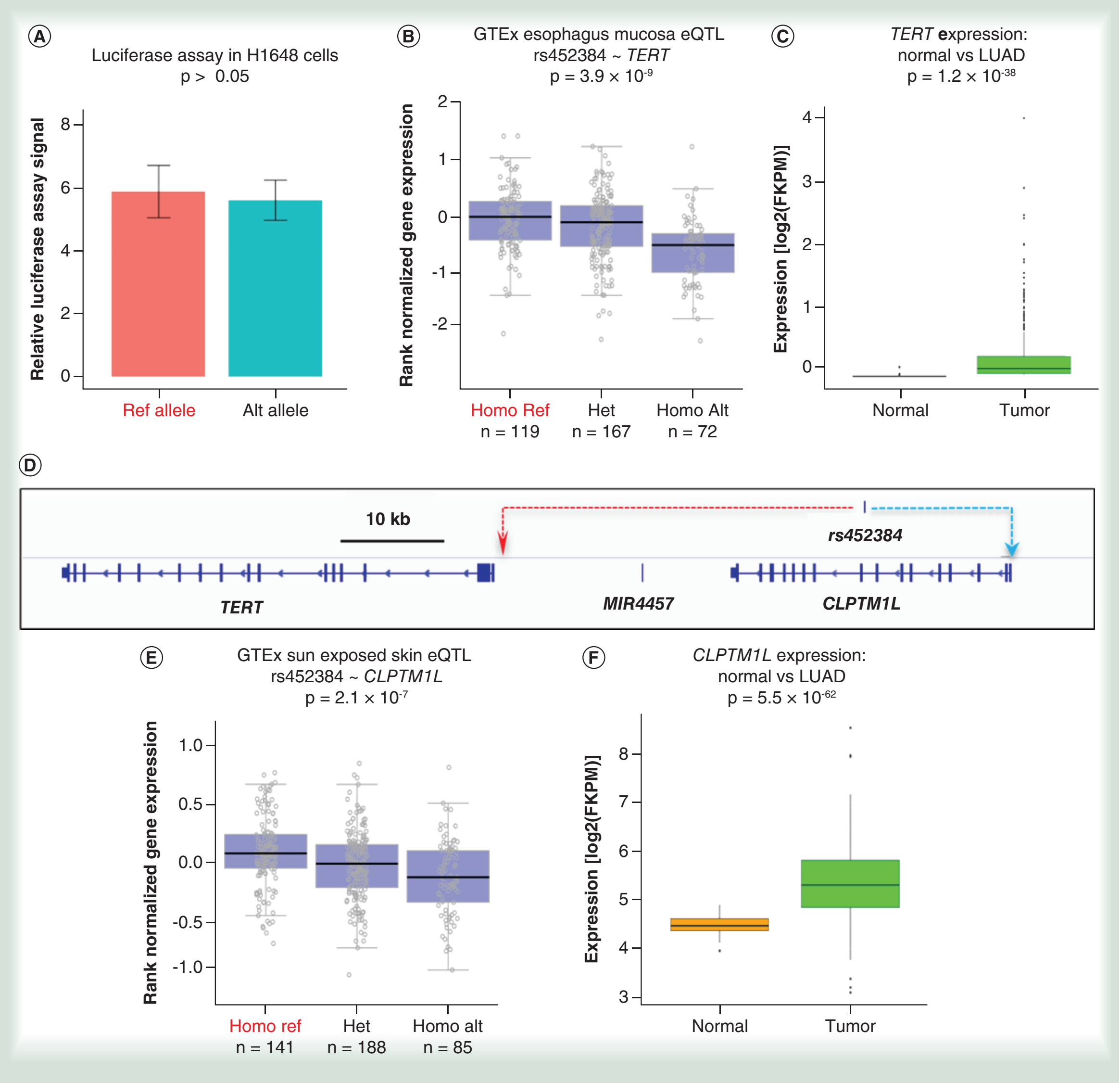


Figure 5. Functional analyses for rs452384. (A) Enhancer luciferase assay performed in H1648 cells: relative luciferase signals were compared between cells transfected with the reference (T) or alternate (C) allele plasmids. **(B)** Association between rs452384 (T/C) and *TERT* expression in esophageal mucosa from the GTEx database. **(C)** Boxplot showing statistically significantly elevated expression of *TERT* in TCGA LUAD versus nontumor lung tissue. **(D)** Relative genomic positions of *TERT, CLPTM1L* and SNP rs452384. **(E)** Association between rs452384 (T/C) and *CLPTM1L* expression in sun-exposed skin from the GTEx database. **(F)** Boxplot showing statistically significantly elevated expression of *CLPTM1L* in TCGA LUAD versus nontumor lung tissue.
GTEx: Genotype-tissue expression; LUAD: Lung adenocarcinoma; TCGA: The Cancer Genome Atlas.

susceptibility and resistance to chemotherapy [86]. One possible problem in the ability to detect a *TERT* eQTL in lung tissue may be the very low expression levels of *TERT* in alveolar epithelium (RPKM of around 1 in AT2 cells [D = 0] and absent in cells at D4 and D6), and the possibility that the SNP may only be functional under conditions that would promote cell proliferation, such as lung damage. Of note, rs452384 is located in DNase hypersensitive sites in SAECs and numerous other cell types in the publicly available SNP annotation database HaploReg v4 [87].

Functional analyses of rs6942067 & rs9481728 on chromosome 6q22.1

To test for a functional role of rs6942067, we cloned an 836-bp region (chr6: 117785006–117785841, hg19) containing either the reference (risk) allele (A) or the alternate allele (G) into pGL4.26. Based on the presence of enhancer histone marks in the rs6942067 region [50], we performed reporter assays in LUAD cell line H1648 (Supplementary Figure 11B). In H1648 cells, luciferase activity over background was 6.8-fold and 3.5-fold over background for the A and G alleles, respectively, indicating a 94% elevated enhancer activity for the reference allele ($p = 2.7 \times 10^{-2}$; Figure 6A). We replicated these findings in H522 cells and observed that the reference and alternate constructs, respectively, elicited 6.9-fold and 4.8-fold elevated activity over background, indicating a 45% stronger enhancer activity attributable to the reference (A) allele ($p = 8 \times 10^{-3}$; Supplementary Figure 13).

Rs9481728 is located 30-kb downstream of rs6942067, in the first intron of *DCBLD1* (Figure 6D). We cloned a 912-bp region (chr6: 117817044–117817955, hg19) in pGL4.26 and transfected the DNA into H1648 LUAD cells, which also showed an enhancer peak on this SNP (Supplementary Figure 11C). However, we observed no activity over background (Supplementary Figure 14). While we cannot exclude that genomic interactions with more distant factors may be required for enhancer activity of this segment, given the activity observed using the region encompassing rs6942067, we pursued only rs6942067 further.

Examination of GTEx data for rs6942067 showed a strong genome-wide significant eQTL for *DCBLD1* in lung (Figure 6B). No other genes were found to be eQTLs associated with this SNP in lung or other tissues. To investigate more thoroughly, we also examined the flanking 1Mb window in the SNP region using TCGA LUAD samples. This was done using rs6930292, a SNP in high LD ($r^2 = 1$) as a surrogate, because SNP rs6942067 is not annotated in TCGA. We detected significant allele-specific expression of *DCBLD1* ($p = 0.041$, corrected for the number of genes in the 1Mb window, Figure 6C). Surprisingly, in contrast to the luciferase-based assay, we detected lower expression of the reference (risk) allele, suggesting that in the genomic context, the reference allele might reduce, not enhance expression. However, we note that in GTEx, the risk allele is associated with higher *DCBLD1* expression in blood cells but reduced expression in thyroid tissue (Figure 6E & F), supporting a complex role of the SNP in regulating gene expression. It is of interest that we predicted binding of thyroid hormone receptor-α on the alternate allele (Figure 3C) and that this receptor can repress transcription in the absence of thyroid hormone and induce it in the presence of hormone [88]. Further investigations will be required to elucidate the role of rs6942067 and its TFs in LUAD risk.

The function of *DCBLD1* is currently unclear, making it difficult to predict how up- or downregulation of this gene would affect lung cancer risk. The encoded protein is predicted to be membrane-associated. Its paralog NRP2, or VEGF165R2, is a transmembrane protein that interacts with VEGF and is implicated in metastasis [89]. *DCBLD1* is significantly overexpressed in LUAD versus nontumor lung in TCGA data ($p = 1.7 \times 10^{-23}$, Figure 6G), and higher expression in LUAD is negatively associated with survival (Supplementary Figure 15). It is also possible that the enhancer at rs6942067 targets other gene(s) that may not be detectable as an eQTL in the mix of lung cells analyzed in GTEx. rs6942067 lies between *DCBLD1* and *ROS1* (Figure 6D). Translocations involving *ROS1* are known ‘driver’ events in lung cancer; approximately 1% of lung tumors harbor ROS1 fusions [90]. It has been reported that the level of certain histone modifications influences the predisposition to chromosome translocations [91]. Thus, effects of this SNP on the epigenetic landscape and gene translocation might need to be considered.

Conclusion

In this study, we identified 47 AEC enhancer-associated SNPs from 948 LUAD candidate risk SNPs. To ensure we examined SNPs in true enhancers, we focused on SNPs located in regions carrying *both* H3K4me1 and H3K27ac marks. However, it is possible that additional SNPs of interest are located in poised enhancers that are marked only with H3K4me1 and are activated by environmental stimuli such as exposure to tobacco smoke (the AECs used in this study were derived from a nonsmoker). Of the 47 SNPs, we chose the 11 SNPs that were in FAIRE peaks for further study. In the future, it will be important to investigate SNPs in regulatory marks in AEC from smoker’s lungs.

Our analyses do not provide strong evidence for a functional role of rs452384 on chromosome 5p15.33 in LUAD, despite recent fine-mapping data in the *TERT/CLPTM1L* region, which identified rs452384 as one of the top-ranked SNPs in the region [83]. How this SNP might function may require functional analyses using primary AEC or as yet unavailable immortalized AEC or a variety of environmental conditions such as exposure to tobacco smoke.

The reference (risk) allele of rs6942067 was associated with elevated enhancer activity in LUAD cell lines, but we detected an eQTL showing lower expression of *DCBLD1* in lung tissue carrying the reference allele. This emphasizes that luciferase assays may be helpful in detecting allele-specific activity but that the cell type or genomic context

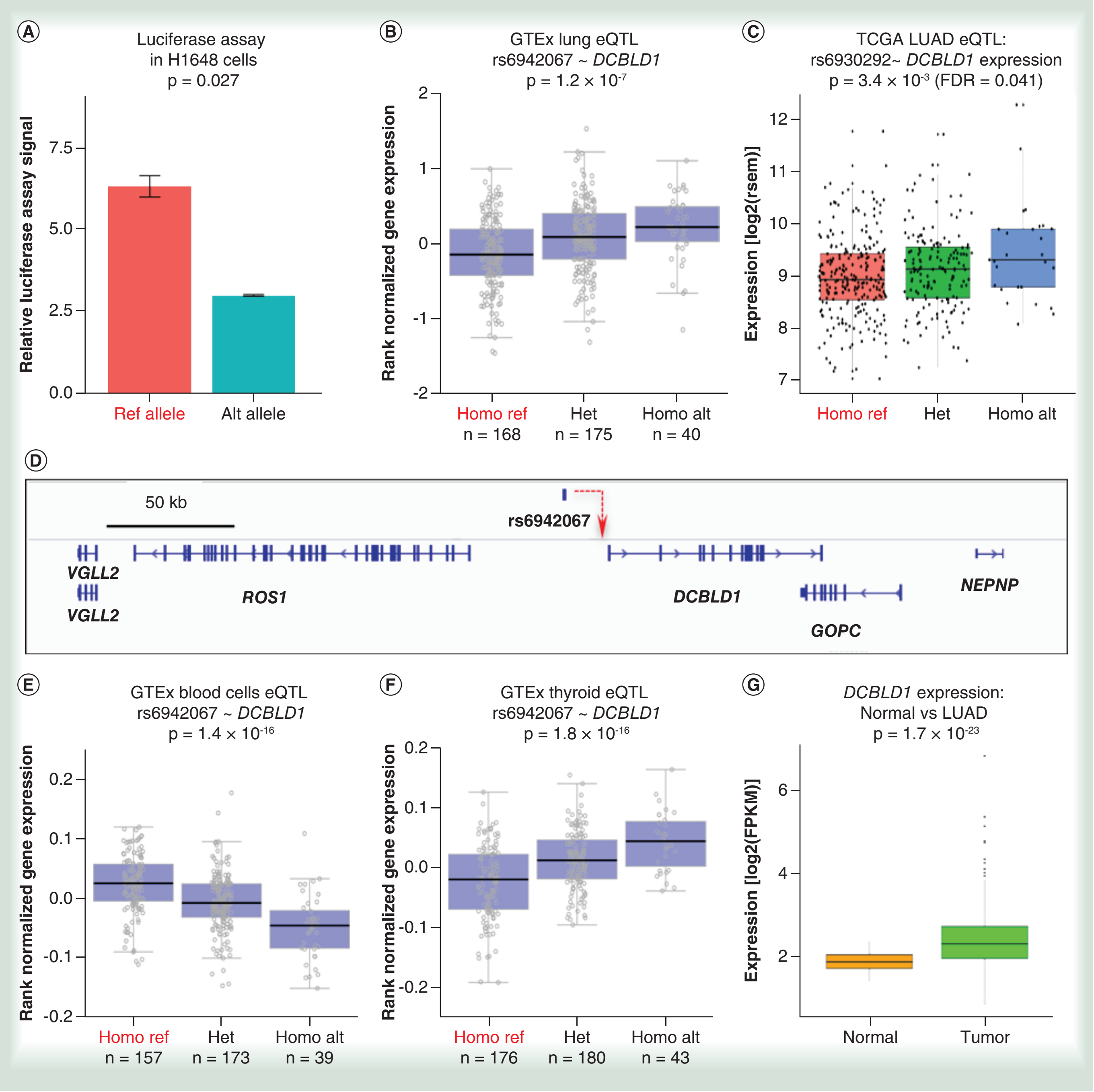


Figure 6. Functional analyses for rs6942067. (A) Enhancer luciferase assay performed in H1648 cells (836-bp fragment, Supplementary Figure 11B); relative luciferase signals were compared between cells transfected with either the reference (A) or alternate (G) allele plasmids. **(B)** Association between rs6942067 (A/G) and *DCBLD1* expression in lung from the GTEx database. **(C)** Association between rs6942067 and *DCBLD1* expression in LUAD tumors from the TCGA database. **(D)** Relative genomic positions of *DCBLD1* and SNP rs6942067. **(E & F)** The association between rs6942067 (A/G) and *DCBLD1* expression in blood cells **(E)** and thyroid **(F)** from the GTEx database. **(G)** Boxplot showing statistically significantly elevated expression of *DCBLD1* in TCGA LUAD versus nontumor lung tissue.
GTEx: Genotype-tissue expression; LUAD: Lung adenocarcinoma; TCGA: The Cancer Genome Atlas.

may affect the observed regulatory effects. The versatility of this SNP is supported by the differential *DCBLD1* eQTLs seen in different tissues (Figure 6B, E & F). EP300, which has been predicted to bind to the reference allele and was observed to do so in ChIP studies, is a histone acetyltransferase associated with gene activation. A lower activity of the reference allele might, therefore, be associated with competition of other factors with EP300 for binding. A role for *DCBLD1* in lung cancer is supported by its increased expression in LUAD versus nontumor lung and by its negative association with survival. A recent study of nonsmoking Asian women found that the association of rs9387478 (the index SNP for rs6942067) with LUAD was stronger in EGFR mutation-positive cases [92], suggesting that the environment, ethnicity and gender could all influence the manifestation of this genetic effect. Taken together, further studies into the function of *DCBLD1* in lung cancer are certainly warranted.

Implications of our findings

Our work emphasizes the importance of integrating epigenomes of purified disease-relevant cell types to elucidate the genetic basis for lung cancer risk. However, it also illustrates that focusing on disease-relevant cells may not be sufficient to identify functional SNPs for most index SNPs. In the case of lung cancer, one may require tobacco-exposed disease-relevant cells. Besides examining tobacco-smoke exposed cells and tissues, purification of other lung cell types will be required to investigate the role of SNPs in other histological subtypes of lung cancer [17]. A recent large-scale study of lung cancer susceptibility loci highlights the differences in genetic susceptibility between histological lung cancer subtypes [93]. Basal cells, which are airway epithelial cells that lie on the basement membrane and are thought to be involved in airway regeneration upon injury, are implicated as progenitors of squamous cell carcinoma [94,95]. Basal cells should be purified and epigenetically profiled in a similar fashion to the work described here. Further investigation of SNPs and their targets may ultimately yield more effective and personalized strategies for lung cancer risk assessment, prevention and treatment.

Summary points

- Integration of 948 candidate lung adenocarcinoma (LUAD) risk-associated SNPs with alveolar epithelial cell epigenomic information identified 47 SNPs in putative enhancer regions, marking them as candidate functional SNPs of increased interest.
- Focusing specifically on the 11 of the 47 SNPs that were located in open DNA as indicated by formaldehyde-assisted isolation of regulatory elements (FAIRE) analyses, seven were predicted to disrupt transcription factor-binding sites.
- Genomic fragments containing rs452384 showed increased luciferase activity over background (reporter vector lacking a genomic fragment), suggesting the region contains an enhancer, but no allele-specific activity or lung expression quantitative trait loci (eQTLs) were detected.
- A genomic fragment carrying rs6942067 showed increased expression of the risk allele when examined in a luciferase reporter assay, indicating presence of an enhancer and the potential of the SNP to affect gene expression.
- eQTL of *DCBLD1* was observed in the Genotype-Tissue Expression project for rs6942067 in lung as well as thyroid, with higher expression of the alternate allele, while blood showed eQTL of *DCBLD1* with higher expression of the reference allele, suggesting the role of this SNP is likely complex.
- TCGA LUAD data showed increased *DCBLD1* expression in cells containing homozygous alternate alleles, and *DCBLD1* is significantly overexpressed in LUAD tumor versus nontumor lung in TCGA, supporting a possible role of this gene in LUAD development or progression.

Acknowledgements

The H1648 and H522 cells were kind gifts from E Haura.

Financial & competing interests disclosure

This work was supported by NIH (https://www.nih.gov/) grants R01 HL114094 (to IA Offringa and Z Borok), R01 HL126877 and R01 HL112638 (to Z Borok), the Norris Comprehensive Cancer Center core grant (National Cancer Institute P30CA01408, supporting IA Offringa), Department of Defense (http://cdmrp.army.mil/funding/lcrp) Concept Award W81XWH-14-1-0174-1 (to IA Offringa), the Tobacco-Related Disease Research Program (award ID 500806, to IA Offringa) and the California Community Foundation (http://www.calfund.org/) (to IA Offringa, supporting C Yang and J Luo) and support from the Thomas G. Labrecque Foundation (https://tglfoundation.com/) (to IA Offringa), the Whittier Foundation (to IA Offringa), the Hastings Foundation (to Z Borok) and generous donations from Conya and Wallace Pembroke (to IA Offringa). Z Borok is the Ralph Edgington Chair in Medicine and Hastings Professor of Medicine. TR Stueve was supported by the National Institute of Health National Institute of Environmental Health Sciences (NIEHS, NIH T32ES013678) and by the USC Provost's Postdoctoral Scholar Research Grant. DJ Mullen was supported by a University of Southern California Provost Fellowship and a Roy E. Thomas Foundation graduate scholarship. CN Marconett was supported by American Cancer Society (https://www.cancer.org/research.html)/Canary Foundation (http://www.canaryfoundation.org/) postdoctoral fellowship # PFTED-10-207-01-SIED and later the Department of Surgery, Keck School of Medicine. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Author contributions

Conceptualization was done by C Yang, TR Stueve, CN Marconett and IA Offringa. Data curation was done by C Yang and CN Marconett. Formal analysis was done by C Yang, TR Stueve and DJ Mullen. Funding acquisition was done by CN Marconett, TR Stueve, B Zhou, Z Borok and IA Offringa. Investigation was done by C Yang, TRS and CL Yan. C Yang, TR Stueve, CN Marconett, SK Rhie and DJ Mullen contributed in methodology. Project administration was done by IA Offringa. Resources were provided by Z Borok, J Luo, BZ and IA Offringa. Software was provided by C Yang and SK Rhie. CN Marconett and IA Offringa were involved in supervision. Validation was done by C Yang, TR Stueve and CL Yan. Visualization was done by C Yang, TR Stueve, DJ Mullen and IA Offringa. C Yang, TR Stueve and IA Offringa wrote the original draft. All the authors were involved in writing and editing this review.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J. Clin.* 61(2), 69–90 (2011).
2. Toh CK, Lim WT. Lung cancer in never-smokers. *J. Clin. Pathol.* 60(4), 337–340 (2007).
3. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers – a different disease. *Nat. Rev. Cancer.* 7(10), 778–790 (2007).
4. American Cancer Society. Cancer facts & figures 2015 (2015). www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2015/index
5. Brennan P, Hainaut P, Boffetta P. Genetics of lung-cancer susceptibility. *Lancet Oncol.* 12(4), 399–408 (2011).
6. Lan Q, Hsiung CA, Matsuo K *et al.* Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat. Genet.* 44(12), 1330–1335 (2012).
7. Li Y, Sheu CC, Ye Y *et al.* Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *Lancet Oncol.* 11(4), 321–330 (2010).
8. Coetzee SG, Rhie SK, Berman BP, Coetzee GA, Noushmehr H. FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res.* 40(18), e139 (2012).
9. Tak YG, Farnham PJ. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin.* 8, 57 (2015).

•• **A thorough review about the utility of epigenomic information to identify functional risk SNPs.**

10. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res.* 22(9), 1748–1759 (2012).
11. Freedman ML, Monteiro AN, Gayther SA *et al.* Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.* 43(6), 513–518 (2011).
12. Farh KK, Marson A, Zhu J *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518(7539), 337–343 (2015).
13. Maurano MT, Haugen E, Sandstrom R *et al.* Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.* 47(12), 1393–1401 (2015).

• **An great example or large-scale functional analysis of SNPs to identify functional variants implicated in transcription factor binding.**

14. Hazelett DJ, Rhie SK, Gaddis M *et al.* Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet.* 10(1), e1004102 (2014).
15. Rhie SK, Coetzee SG, Noushmehr H *et al.* Comprehensive functional annotation of seventy-one breast cancer risk Loci. *PLoS ONE* 8(5), e63925 (2013).
16. Yao L, Tak YG, Berman BP, Farnham PJ. Functional annotation of colon cancer risk SNPs. *Nat. Commun.* 5, 5114 (2014).

17. Stueve TR, Marconett CN, Zhou B *et al.* The importance of detailed epigenomic profiling of different cell types within organs. *Epigenomics* 8(6), 817–829 (2016).

- **A manuscript emphasizing the importance and applications of obtaining epigenomes from many different cell types.**

18. Heintzman ND, Stuart RK, Hon G *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39(3), 311–318 (2007).

19. Heintzman ND, Hon GC, Hawkins RD *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459(7243), 108–112 (2009).

20. Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O *et al.* Epigenomic enhancer profiling defines a signature of colon cancer. *Science* 336(6082), 736–739 (2012).

21. Pasquali L, Gaulton KJ, Rodríguez-Seguí SA *et al.* Pancreatic islet enhancer clusters enriched in Type 2 diabetes risk-associated variants. *Nat. Genet.* 46(2), 136–143 (2014).

22. Gjoneska E, Pfenning AR, Mathys H *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* 518(7539), 365–369 (2015).

23. Hu Z, Wu C, Shi Y *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(8), 792–796 (2011).

24. Wang Y, McKay JD, Rafnar T *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat. Genet.* 46(7), 736–741 (2014).

25. Shiraishi K, Kunitoh H, Daigo Y *et al.* A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. *Nat. Genet.* 44(8), 900–903 (2012).

26. Miki D, Kubo M, Takahashi A *et al.* Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. *Nat. Genet.* 42(10), 893–896 (2010).

27. Hsiung CA, Lan Q, Hong YC *et al.* The 5p15.33 locus is associated with risk of lung adenocarcinoma in never-smoking females in Asia. *PLoS Genet.* 6(8), e1001051 (2010).

28. Landi MT, Chatterjee N, Yu K *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(5), 679–691 (2009).

29. Dong J, Hu Z, Wu C *et al.* Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nat. Genet.* 44(8), 895–899 (2012).

30. Wang Z, Seow WJ, Shiraishi K *et al.* Meta-analysis of genome-wide association studies identifies multiple lung cancer susceptibility loci in never-smoking Asian women. *Hum. Mol. Genet.* 25(3), 620–629 (2016).

31. Dobbs LG, Johnson MD, Vanderbilt J, Allen L, Gonzalez R. The great big alveolar TI cell: evolving concepts and paradigms. *Cell. Physiol. Biochem.* 25(1), 55–62 (2010).

32. Rackley CR, Stripp BR. Building and maintaining the epithelium of the lung. *J. Clin. Invest.* 122(8), 2724–2730 (2012).

•• **A review that provides a clear overview of cell types within the lung, information that is required when considering progenitors of lung cancer.**

33. Wang J, Edeen K, Manzer R *et al.* Differentiated human alveolar epithelial cells and reversibility of their phenotype *in vitro*. *Am. J. Respir. Cell Mol. Biol.* 36(6), 661–668 (2007).

34. Ballard PL, Lee JW, Fang X *et al.* Regulated gene expression in cultured type II cells of adult human lung. *Am. J. Physiol. Lung Cell Mol. Physiol.* 299(1), L36–L50 (2010).

35. Danto SI, Shannon JM, Borok Z, Zabski SM, Crandall ED. Reversible transdifferentiation of alveolar epithelial cells. *Am. J. Respir. Cell Mol. Biol.* 12(5), 497–502 (1995).

36. Desai TJ, Brownfield DG, Krasnow MA. Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature* 507(7491), 190–194 (2014).

- **A discussion of the alveolar compartment of the lung and its contribution to lung cancer development.**

37. Xu X, Rock JR, Lu Y *et al.* Evidence for type II cells as cells of origin of K-Ras-induced distal lung adenocarcinoma. *Proc. Natl Acad. Sci. USA* 109(13), 4910–4915 (2012).

38. Lin C, Song H, Huang C *et al.* Alveolar type II cells possess the capability of initiating lung tumor development. *PLoS ONE* 7(12), e53817 (2012).

39. Mainardi S, Mijimolle N, Francoz S, Vicente-Dueñas C, Sánchez-García I, Barbacid M. Identification of cancer initiating cells in K-Ras driven lung adenocarcinoma. *Proc. Natl Acad. Sci. USA* 111(1), 255–260 (2014).

40. Marconett CN, Zhou B, Rieger ME *et al.* Integrated transcriptomic and epigenomic analysis of primary human lung epithelial cell differentiation. *PLoS Genet.* 9(6), e1003513 (2013).

41. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14(4), R36 (2013).

42. Trapnell C, Williams BA, Pertea G *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28(5), 511–515 (2010).

43. Khurana E, Fu Y, Colonna V *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342(6154), 1235587 (2013).

44. San Lucas FA, Wang G, Scheet P, Peng B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* 28(3), 421–422 (2012).

45. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7, Unit 7.20 (2013).

46. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31(13), 3812–3814 (2003).

47. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19(1), 92–105 (2009).

48. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 25(15), 1952–1958 (2009).

49. Heinz S, Benner C, Spann N *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell.* 38(4), 576–589 (2010).

50. Suzuki A, Wakaguri H, Yamashita R *et al.* DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. *Nucleic Acids Res.* 43(Database issue), D87–D91 (2015).

- **Description of DataBase of Transcriptional Start Sites, a great resource for epigenomic information on lung cancer cell lines.**

51. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* 42(5), 2976–2987 (2014).

52. Wang J, Zhuang J, Iyer S *et al.* Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res.* 41(Database issue), D171–D176 (2013).

53. Kulakovskiy IV, Medvedeva YA, Schaefer U *et al.* HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* 41(Database issue), D195–D202 (2013).

54. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27(7), 1017–1018 (2011).

55. Consortium G. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45(6), 580–585 (2013).

•• **Description of the Genotype-Tissue Expression project, a public resource to identify expression quantitative trait loci.**

56. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14(2), 178–192 (2013).

57. Falvella FS, Galvan A, Frullanti E *et al.* Transcription deregulation at the 15q25 locus in association with lung adenocarcinoma risk. *Clin. Cancer Res.* 15(5), 1837–1842 (2009).

58. Nguyen JD, Lamontagne M, Couture C *et al.* Susceptibility loci for lung cancer are associated with mRNA levels of nearby genes in the lung. *Carcinogenesis* 35(12), 2653–2659 (2014).

59. Improgo MR, Scofield MD, Tapper AR, Gardner PD. The nicotinic acetylcholine receptor CHRNA5/A3/B4 gene cluster: dual role in nicotine addiction and lung cancer. *Prog. Neurobiol.* 92(2), 212–226 (2010).

60. Doyle GA, Wang MJ, Chou AD *et al. In vitro* and *ex vivo* analysis of CHRNA3 and CHRNA5 haplotype expression. *PLoS ONE* 6(8), e23373 (2011).

61. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* 17(6), 877–885 (2007).

62. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature* 489(7414), 109–113 (2012).

- **A manuscript providing information on long-range interactions between regulatory elements such as enhancers and their target genes.**

63. Marconett CN, Zhou B, Siegmund KD, Borok Z, Laird-Offringa IA. Transcriptomic profiling of primary alveolar epithelial cell differentiation in human and rat. *Genom. Data* 2, 105–109 (2014).

64. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414), 57–74 (2012).

•• **A solid primer on ENCODE, which provides and important resource of public epigenomic data.**

65. Watanabe H, Francis JM, Woo MS *et al.* Integrated cistromic and expression analysis of amplified NKX2-1 in lung adenocarcinoma identifies LMO3 as a functional transcriptional target. *Genes Dev.* 27(2), 197–210 (2013).

66. James MA, Vikis HG, Tate E, Rymaszewski AL, You M. CRR9/CLPTM1L regulates cell survival signaling and is required for Ras transformation and lung tumorigenesis. *Cancer Res.* 74(4), 1116–1127 (2014).

67. James MA, Wen W, Wang Y *et al.* Functional characterization of CLPTM1L as a lung cancer risk candidate gene in the 5p15.33 locus. *PLoS ONE* 7(6), e36116 (2012).

68. Wang Z, Zhu B, Zhang M *et al.* Imputation and subset-based association analysis across different cancer types identifies multiple independent risk loci in the TERT-CLPTM1L region on chromosome 5p15.33. *Hum. Mol. Genet.* 23(24), 6616–6633 (2014).

69. Chen XF, Cai S, Chen QG *et al.* Multiple variants of TERT and CLPTM1L constitute risk factors for lung adenocarcinoma. *Genet. Mol. Res.* 11(1), 370–378 (2012).

70. Timofeeva MN, Hung RJ, Rafnar T *et al.* Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum. Mol. Genet.* 21(22), 4980–4995 (2012).

71. Mocellin S, Verdi D, Pooley KA *et al.* Telomerase reverse transcriptase locus polymorphisms and cancer risk: a field synopsis and meta-analysis. *J. Natl. Cancer Inst.* 104(11), 840–854 (2012).

72. Minoo P. Transcriptional regulation of lung development: emergence of specificity. *Respir. Res.* 1(2), 109–115 (2000).

73. Herriges M, Morrisey EE. Lung development: orchestrating the generation and regeneration of a complex organ. *Development* 141(3), 502–513 (2014).

74. Dang CV. c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Mol. Cell. Biol.* 19(1), 1–11 (1999).

75. Seo AN, Yang JM, Kim H *et al.* Clinicopathologic and prognostic significance of c-MYC copy number gain in lung adenocarcinomas. *Br. J. Cancer.* 110(11), 2688–2699 (2014).

76. Iwakawa R, Kohno T, Kato M *et al.* MYC amplification as a prognostic marker of early-stage lung adenocarcinoma identified by whole genome copy number analysis. *Clin. Cancer Res.* 17(6), 1481–1489 (2011).

77. Iyer NG, Ozdag H, Caldas C. p300/CBP and cancer. *Oncogene* 23(24), 4225–4231 (2004).

78. Kan Z, Jaiswal BS, Stinson J *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* 466(7308), 869–873 (2010).

79. Rajatapiti P, Kester MH, de Krijger RR, Rottier R, Visser TJ, Tibboel D. Expression of glucocorticoid, retinoid, and thyroid hormone receptors during human lung development. *J. Clin. Endocrinol. Metab.* 90(7), 4309–4314 (2005).

80. Kolla V, Gonzales LW, Gonzales J *et al.* Thyroid transcription factor in differentiating type II cells: regulation, isoforms, and target genes. *Am. J. Respir. Cell Mol. Biol.* 36(2), 213–225 (2007).

81. Hume R, Richard K, Kaptein E, Stanley EL, Visser TJ, Coughtrie MW. Thyroid hormone metabolism and the developing human lung. *Biol. Neonate.* 80(Suppl. 1), 18–21 (2001).

82. Hellevik AI, Asvold BO, Bjøro T, Romundstad PR, Nilsen TI, Vatten LJ. Thyroid function and cancer risk: a prospective population study. *Cancer Epidemiol. Biomarkers Prev.* 18(2), 570–574 (2009).

83. Kachuri L, Amos CI, McKay JD *et al.* Fine mapping of chromosome 5p15.33 based on a targeted deep sequencing and high density genotyping identifies novel lung cancer susceptibility loci. *Carcinogenesis* 37(1), 96–105 (2016).

84. Ma X, Gong R, Wang R *et al.* Recurrent TERT promoter mutations in non-small cell lung cancers. *Lung Cancer* 86(3), 369–373 (2014).

85. Calado RT. Telomeres in lung diseases. *Prog. Mol. Biol. Transl. Sci.* 125, 173–183 (2014).

86. James MA, Wen W, Wang Y *et al.* Functional characterization of CLPTM1L as a lung cancer risk candidate gene in the 5p15.33 locus. *PLoS ONE* 7(6), e36116 (2012).

87. 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.* 44(D1), D877–D881 (2016).

•• **Description of HaploReg, a key resource to carry out data mining on candidate risk variants.**

88. Astapova I. Role of co-regulators in metabolic and transcriptional actions of thyroid hormone. *J. Mol. Endocrinol.* 56(3), 73–97 (2016).

89. Caunt M, Mak J, Liang WC *et al.* Blocking neuropilin-2 function inhibits tumor cell metastasis. *Cancer Cell.* 13(4), 331–342 (2008).

90. Takeuchi K, Soda M, Togashi Y *et al.* RET, ROS1 and ALK fusions in lung cancer. *Nat. Med.* 18(3), 378–381 (2012).

91. Burman B, Zhang ZZ, Pegoraro G, Lieb JD, Misteli T. Histone modifications predispose genome regions to breakage and translocation. *Genes Dev.* 29(13), 1393–1402 (2015).

92. Seow WJ, Matsuko K, Hsiung CA *et al.* Association between GWAS-identified lung adenocarcinoma susceptibility loci and EGFR mutations in never-smoking Asian women, and comparison with findings from Western populations. *Hum. Mol. Genet.* 26(2), 454–465 (2017).

93. McKay JD, Hung RJ, Han Y *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histologial subtypes. *Nature Genet.* 49(7), 1126–1132 (2017).

• **A recent large analysis showing that SNPs exhibit differental risk-based lung cancer histological subtype.**

94. Sutherland KD, Berns A. Cell of origin of lung cancer. *Mol. Oncol.* 4(5), 397–403 (2010).

95. Van de Laar E, Clifford M, Hasenoeder S *et al.* Cell surface marker profiling of human tracheal basal cells reveals distinct subpopulations, identifies MST1/MSP as a mitogenic signal, and identifies new biomarkers for lung squamous cell carcinomas. *Respir. Res.* 15, 160 (2014).