EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

# A Genome-Wide Association Study implicates *NR2F2* in Lymphangioleiomyomatosis Pathogenesis

Wonji Kim, Krinio Giannikou, John R. Dreier, Sanghun Lee, Magdalena E. Tyburczy, Edwin K. Silverman, Elżbieta Radzikowska, Shulin Wu, Chin-Lee Wu, Elizabeth P. Henske, Gary Hunninghake, Havi Carel, Antonio Roman, Miquel Angel Pujana, Joel Moss, Sungho Won, David J. Kwiatkowski

Please cite this article as: Kim W, Giannikou K, Dreier JR, *et al*. A Genome-Wide Association Study implicates *NR2F2* in Lymphangioleiomyomatosis Pathogenesis. *Eur Respir J* 2019; in press (https://doi.org/10.1183/13993003.00329-2019).

# A Genome-Wide Association Study implicates *NR2F2* in Lymphangioleiomyomatosis Pathogenesis

Wonji Kim[1,2*], Krinio Giannikou[3*], John R. Dreier[3], Sanghun Lee[4], Magdalena E. Tyburczy[3], Edwin K. Silverman[2,3], Elżbieta Radzikowska[5], Shulin Wu[6], Chin-Lee Wu[6], Elizabeth P. Henske[3], Gary Hunninghake[3], Havi Carel[7], Antonio Roman[8], Miquel Angel Pujana[9], Joel Moss[10], Sungho Won[11,12], David J. Kwiatkowski[3]

[1]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, Korea

[2]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115, USA

[3]Division of Pulmonary and Critical Care Medicine and of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, 02115, Massachusetts, United States of America

[4]Department of Medical Consilience, Graduate School, Dankook University, Seoul, Korea

[5]National Tuberculosis and Lung Diseases Research Institute, Warsaw, Poland

[6]Urology Research Laboratory, Massachusetts General Hospital, Boston, MA, 02114, USA

[7]Department of Philosophy, University of Bristol, UK

[8]Vall d'Hebron University Hospital, CIBERES, Barcelona, Spain

[9]ProCURE, Catalan Institute of Oncology, Oncobell, Bellvitge Institute of Biomedical Research (IDIBELL), Barcelona, Spain

[10]Cardiovascular and Pulmonary Branch, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, MDGTEx

[11]Department of Public Health Sciences, Seoul National University, Seoul, Korea

[12]Institute of Health and Environment, Seoul National University, Seoul, Korea

Corresponding authors:

Sungho Won

1 Kwanak-ro Kwanak-gu, Department of Public Health Sciences, Seoul National University, Seoul 151-742, Korea

sunghow@gmail.com

phone:+82-2-880-2714 fax: +82 82-303-0942-2714

David J. Kwiatkowski

20 Shattuck Street, Division of Pulmonary Medicine, Brigham and Women's Hospital, Boston, MA 02115

dk@rics.bwh.harvard.edu

phone: +1-857-307-0781 fax: +1-6173942769

[*]equal contribution

Author contributions:

Contributions to the conception and design of the work: SW, EKS, GH, DJK

Contributions to the acquisition, analysis, or interpretation of data: all authors

Drafting the work or revising it critically for important intellectual content: all authors

Final approval of the version to be published: all authors

Agreement to be accountable for all aspects of the work: all authors

## SUMMARY

We conducted a genome wide association study to identify variants outside of *TSC2*/*TSC1* associated with Lymphangioleiomyomatosis (LAM) pathogenesis. Two SNPs were identified near *NR2F2*. Pathology studies indicate this transcription factor is expressed highly in a LAM-related tumor.

## ABSTRACT

**Rationale:** Lymphangioleiomyomatosis occurs either associated with Tuberous Sclerosis Complex or as sporadic disease (S-LAM). Risk factors for development of S-LAM are unknown.

**Objectives:** We hypothesized that DNA sequence variants outside of *TSC2*/*TSC1* might be associated with susceptibility for S-LAM, and performed a Genome Wide Association Study (GWAS).

**Methods:** Genotyped and imputed data on 5,426,936 SNPs in 426 S-LAM subjects were compared, using conditional logistic regression, to similar data from 852 females from COPDGene in a matched case-control design. For replication studies, genotypes for 196 non-Hispanic white (NHW) female S-LAM subjects were compared with three different sets of controls. RNA-seq and immunohistochemistry analyses were also performed.

**Results:** Two non-coding genotyped SNPs met genome-wide significance; rs4544201 and rs2006950 (P-value=$4.2\times10^{-8}$, $6.1\times10^{-9}$, respectively) which are in the same 35kb linkage disequilibrium block on chr15q26.2. This association was replicated in an independent cohort. *NR2F2,* a nuclear receptor and transcription factor, was the only nearby protein-coding gene. *NR2F2* expression was higher by RNA-seq in one abdominal LAM tumor and four kidney angiomyolipomas, a LAM-related tumor, compared to all TCGA cancers. Immunohistochemistry showed strong nuclear expression in both LAM and angiomyolipoma tumors.

**Conclusions:** SNPs on chr15q26.2 are associated with S-LAM, and chromatin and expression data suggest that this association may occur through effects on *NR2F2* expression, which potentially plays an important role in S-LAM development.

## KEYWORDS

GWAS, S-LAM, LAM, NR2F2, SNP, TSC2

## INTRODUCTION

Lymphangioleiomyomatosis (LAM) is a rare aggressive low-grade neoplasm which affects almost exclusively women at reproductive age or older and causes progressive cystic lung destruction leading to fatal respiratory failure in subjects with severe disease [1-6]. LAM is characterized by an abnormal proliferation of smooth muscle-like and epithelioid cells in innumerable tiny clusters in the lungs, in association with thin-walled cysts and lung parenchymal destruction [7, 8]. Progressive cyst enlargement and inflammation contribute to decline in lung function measured as both decreased $FEV_1$ and $DL_{CO}$. The diagnosis of LAM is based on clinical features, chest computed tomography findings of thin-walled cysts, and either pathology seen on lung biopsy or elevated serum vascular endothelial growth factor D (VEGF-D) levels.

LAM occurs at high frequency (> 10%) in women with Tuberous Sclerosis Complex (TSC); and at much lower frequency in women (about 1 in 100,000) without that disorder, in which it is called sporadic (S-LAM). TSC is due to germline and/or mosaic mutations in either *TSC1* (25%) or *TSC2* (75%) [9]. Tumor development in TSC follows the classic Knudson model of a germline mutation complemented by a somatic second hit mutation in the other corresponding allele in tumors [9, 10]. Limited data are available for S-LAM, but it appears that *TSC2* mutations are seen in the vast majority of S-LAM lesions. About 50% S-LAM subjects have kidney angiomyolipoma, a tumor which is seen in 70-80% of adults with TSC. Angiomyolipoma share histologic, expression, and genetic features with LAM, though are not identical pathologic lesions.

Genome-wide association studies (GWAS) are utilized to identify genetic variants and susceptibility loci associated with complex traits and common diseases.

Although there is no precedent for genetic influence on the development of S-LAM, we hypothesized that DNA sequence variants outside of *TSC2*/*TSC1* might be associated with disease risk, and go unrecognized due to the low prevalence of this disorder.

## METHODS

### Discovery cohort

Over 600 female S-LAM patients were initially identified and collected through international solicitation from 2010 to 2014 from 14 countries (Supplemental Table 1). S-LAM patients were diagnosed using standard diagnostic criteria [1-5, 7] by their treating physicians. Genomic DNA was extracted from saliva using the QIAamp DNA mini kit (Qiagen, Germany), and 479 S-LAM DNA samples were genotyped with the Infinium OmniExpress-24 v1.2 BeadChip, which assesses 716,503 SNPs across the entire genome. 34 non-white S-LAM subjects were excluded from further analyses. There were no self-declared Hispanics in this set of subjects.

Genotype data from the same genotyping chip were available for 1261 healthy female volunteers from the COPDGene Consortium, and were obtained from dbGaP (phs000951.v2.p2.c1). These COPDGene participants had smoked at least 10 pack years and were 45 to 80 years old, and were without known COPD [11, 12].

### Quality control analyses of SNP genotype data

We evaluated the quality of SNPs and subjects in the discovery data set using PLINK [13] and ONETOOL [14]. We excluded all SNPs for which: the Hardy-Weinberg equilibrium (HWE) test [15] gave $P < 1\times10^{-5}$; minor allele frequency (MAF) was < 0.05; or genotype call rates were less than 95%. We also discarded any subjects whose missing genotype rates were > 5%, or showed identity-by-state > 80% with any other subject (Figure 1). These filtering procedures were first applied separately to cases and controls, and were repeated on the pooled dataset. In addition, any SNP showing a difference in missing data rate between cases and controls by Fisher's exact test [16], with $P < 1\times10^{-5}$ was removed (Figure 1).

### Genome-wide imputation

We performed genome-wide imputation for all autosomes to enable discovery of associations for both genotyped and imputed SNPs. Imputation was conducted using the Sanger Imputation Service (https://imputation.sanger.ac.uk). We used Haplotype Reference Consortium release v1.1 for the reference panel and considered predominantly European ancestry [17]. Pre-phasing was performed first with EAGLE2 v2.0.5 [25], and then the Positional Burrows–Wheeler Transform (PBWT) package [26] was used for imputation according to the imputation pipeline recommended by Sanger Imputation Service. Imputation accuracy was evaluated with the INFO metric [18]. Imputed SNPs were filtered out if INFOs, MAFs or P-values for the HWE test were < 0.3, 0.05, or $1\times10^{-5}$, respectively.

### Statistical analyses with genetic data

EIGENSTRAT [19] was also applied to the pooled data and principal component (PC) scores were calculated. PC scores were used to detect subjects with an outlying genetic background, and such outliers (3 subjects) were then removed (Figure 1).

To ensure matching of cases and controls for primary analysis, we used conditional logistic regression (CLR). Each case was matched with two controls using the *Matching* R package [20]. Matching quality is affected by the number of PC scores used, and we assessed how many PC scores were required for effective matching. Two PC scores gave the genomic inflation factor closest to 1 (Supplementary Figure 1). Thus CLR was conducted by conditioning on the matched cases and controls with the first 2 PC scores. Our CLR can be expressed as follows: for $i^{\text{th}}$ strata,

$$\Pr\left(Y_{i1} = 1, Y_{i2} = 0, Y_{i3} = 0 | \mathbf{X}_{i1}, \mathbf{X}_{i2}, \mathbf{X}_{i3}, \sum_{j=1}^{3} Y_{ij} = 1\right)$$

$$= \frac{\Pr(Y_{i1} = 1|\mathbf{X}_{i1}) \Pr(Y_{i2} = 0|\mathbf{X}_{i2}) \Pr(Y_{i3} = 0|\mathbf{X}_{i3})}{\sum_{y_{ij} \in (y_{i1}+y_{i2}+y_{i3}=1)} \Pr(Y_{i1} = y_{i1}|\mathbf{X}_{i1}) \Pr(Y_{i2} = y_{i2}|\mathbf{X}_{i2}) \Pr(Y_{i3} = y_{i3}|\mathbf{X}_{i3})}$$

$$= \frac{\exp(\mathbf{X}_{i1}\beta)}{\sum_{j=1}^{3} \exp\left(\mathbf{X}_{ij}\beta\right)}$$

where $Y_{ij}$ and $\mathbf{X}_{ij}$ indicate the phenotype and covariates including SNP of $j^{\text{th}}$ subject in the $i^{\text{th}}$ matched strata, respectively. For covariates, 10 PC scores were included to adjust the additional population substructure. CLR analyses were performed with the R package *survival* [21] and genome-wide significance was assessed by P-value < $5\times10^{-8}$.

We applied the PICS software to all imputed and genotyped SNPs showing association with LAM to calculate the probability of each individual SNP being the causal SNP [22].

We also conducted gene-based analyses for association with LAM for those genes near the genome-wide significant SNPs using the SKAT-O statistic [23]. We included all genotyped SNPs in this analysis with no MAF cut-off for inclusion. Age, squared age, and 10 PC scores were included as covariates.

## Replication data

Replication analysis was done on an independent set of 196 non-Hispanic white (NHW) female S-LAM subjects, seen at the NIH Clinical Center by one co-author (JM, Supplementary Table 1). Careful scrutiny was performed by a third independent party ('honest broker') to compare the names of subjects used in the primary analysis and patient candidates for the replication population to select those that were not analyzed in the primary analysis. Genotyping was performed by TaqMan SNP genotyping assays C_832391_10 and C_27296040_10 for SNPs rs2006950 and rs4544201, respectively (ThermoFisher Scientific). Nine randomly selected S-LAM subjects from the discovery study were also genotyped by this method to confirm genotyping accuracy in the replication analysis. Their discovery study genotypes matched the TaqMan analysis genotypes perfectly, and these 9 subjects were not included in the replication analyses. We used three independent datasets as controls for comparison in the replication study: 1) 409 NHW healthy females from COPDGene Consortium who were not used for discovery analyses; 2) 1,121 Hispanic white females in the Multi-Ethnic Study of Atherosclerosis (MESA) dataset obtained from dbGaP (phs000209.v13.p3) [24]; and 3) 225,731 white British females from UK Biobank dataset [25]. For each control dataset, we used genotyped or imputed data for the genome-wide significant SNPs.

## Topologically associated domains (TADs) and chromatin interactions

To identify chromatin interactions in the region of interest on chromosome 15q26.2, we used two 3D genome browsers (www.3dgenome.org and https://yunliweb.its.unc.edu/hugin/) to predict TADs [26, 27]. We checked for TADs around the genome-wide significant SNPs and protein coding genes belonging to each TAD were investigated. We analyzed TADs from four cell lines/tissues judged closest to LAM: (i) human fetal lung fibroblast (IMR90), (ii) lung-related tissues (LUNG), (iii) H1 derived mesenchymal stem cells (H1-MSC), and (iv) Human Umbilical Vein Endothelial Cells (HUVEC).

**Statistical analyses with RNA sequencing data**

Whole transcriptome RNA-Seq analysis was performed on one abdominal LAM tumor and four kidney angiomyoliopomas at the Broad Institute of Harvard and MIT. Briefly, mRNA-Seq was performed using polyA cDNA capture followed by cDNA library synthesis (Illumina Truseq RNA Library Prep Kit), and sequencing on Illumina machines, following the same methods and in the same facility in which the Gene and Tissue Expression (GTEx) RNA-seq project occurred [24]. Read data was processed into FASTQ files with standard QC methods, and aligned to the genome (hg19, NCBI37) using Tophat v2.0.10 [28]. Fastq files were also converted into RSEM format [29]. RSEM values were compared to RNA-seq data from 2,463 tumors of 27 different histologic types from the TCGA [30]. RPKM values for *NR2F2* were compared to the GTEx data set of normal human tissues with Limma statistic (11,688 RNA-Seq samples from 53 normal tissue types, v7 release) [31].

We also searched for any *cis*-expression quantitative trait loci (eQTL) for all SNPs in the LD block with association to LAM using GTEx release v7 database [33]. This resource provides results of eQTL analysis for each SNP-gene pair for all SNPs within 1 Mb upstream and downstream of the transcription start site. FastQTL is used by this resource (https://www.gtexportal.org/home) for *cis*-eQTL mapping [32] with covariate adjustment of top three PC scores, genotyping platform, sex and a set of relevant variables identified using PEER method [33].

### Immunohistochemistry analyses

Immunochistochemistry was performed as described elsewhere [34] using a primary mouse monoclonal antibody against *NR2F2* [Abcam Cat.Num # ab41859 Concentration 1:100 (10ug/ml)]. Briefly, formalin-fixed, paraffin-embedded tumor sections were deparaffinized in xylene, rehydrated~~,~~ and antigen retrieval was performed in EDTA (pH 8.0, Diagnostic BioSystems). Endogenous peroxidase activity was blocked with 3% $H_2O_2$, blocking was done with 5% goat serum, followed by incubation overnight with antibody at 4°C, washing in TBST, and incubation with anti-goat secondary antibody (Vector Labs, Burlingame, CA, dilution 1:300). The peroxidase reaction was developed using DAB substrate (DakoCytomation). Both LAM lung samples and kidney angiomyolipomas were stained by similar methods.

## RESULTS

### GWAS analysis of S-LAM identifies two intergenic SNPs on chromosome 15

After multiple filtration steps and elimination of SNPs and samples as described in the Methods and shown in Figure 1, GWAS was performed on 426 S-LAM subjects and 852 control subjects from the COPDGene project, for 5,426,936 SNPs (549,591 genotyped and 4,877,345 imputed) using CLR. Twenty non-coding SNPs on chromosome 15 met genome-wide significance, of which 2 had been directly genotyped (Table 1, rs4544201: P-value=$4.19\times10^{-8}$; rs2006950: P-value=$6.12\times10^{-9}$).

Quantile-quantile plots for CLRs and Manhattan plots demonstrated that the distribution of observed P-values met the expected distribution, with the exception of the 20 SNPs (Figure 2), indicating that the analyses were free of systematic P-value inflation (genomic inflation factor = 1.025). Scatter plots of PC scores indicated similarity between cases and controls in the discovery analyses (Supplementary Figure 2). All subjects from the COPDGene cohort were smokers, and this might have caused an association between SNPs associated with nicotine addiction. We checked p-values for SNPs associated with nicotine addiction from the GWAS catalog [35] and other SNPs correlated with those ($r^2$ >0.8) (Supplementary Table 2). None of those SNPs showed a significant difference in allele frequency in the LAM and COPDGene cohorts, indicating that our findings are not affected by nicotine addiction SNPs.

Linkage disequilibrium (LD) blocks near genome-wide significant SNPs were identified using Haploview with default options [36]. All 20 SNPs, including the two directly genotyped, rs4544201 and rs2006950, belong to the same LD block on 15q26.2; the latter two SNPs were 11,563nt apart, and were strongly correlated ($D$'=0.977, $r^2$=0.854; Supplementary Figure 3). Based on the proximity of the two SNPs to each other and their LD relationship, it is likely that there is a single disease susceptibility locus in the region. They are located in an intergenic gene desert between *MCTP2* (1.1Mb away) and *NR2F2* (700kb away), that contains many long non-coding RNAs (lncRNAs) (Figure 3). Both SNPs have minor and major alleles of A and G, and showed a lower minor allele frequency (MAF) in the S-LAM cohort than the control population. The odds ratios (ORs) of a single minor allele in the S-LAM cohort were 0.49 and 0.47 respectively, in comparison to the control population (Table 2). To adjust for the possible effect of the 'Winner's curse', we used br2 [37], and found that the bias-adjusted OR for rs4544201 and rs2006950 were 0.57 and 0.53, respectively.

We calculated the proportion of phenotypic variance explained by the genotyped SNPs, $h^2_{\mathrm{SNP}}$. Estimates of $h^2_{\mathrm{SNP}}$ vary according to disease prevalence (Supplementary Figure 4). With prevalence set at 1 in 100,000 women, $h^2_{\mathrm{SNP}}$ was 15% (0.3% on the observed 0-1 scale).

Given that *TSC2* mutations occur consistently in LAM cells, genetic variants in each of *TSC1* and *TSC2* were considered a priori candidates for association with S-LAM. Hence, we checked SNPs within or < 1 Mb away from either gene. There were 566 and 416 SNPs for *TSC1* and *TSC2*, respectively, and only rs11552431 (located at 16:1823024, 274kb away from *TSC2*) was significant in CLR after Bonferroni correction at q <0.1 (nominal P-value = 5.97 x $10^{-5}$). We included that SNP and 9 others with the lowest P-values from these genes as covariates in the CLR. The significance of rs4544201 and rs2006950 changed minimally following this adjustment (Supplementary Table 3).

Replication analysis was performed for the two genome-wide significant and genotyped SNPs, which were genotyped in 196 additional non-Hispanic white (NHW) S-LAM patients and compared with SNP allele frequencies in each of three control datasets: 1) 409 NHW healthy COPDGene females who were not used for discovery analyses; 2) 1,121 Hispanic white females from the MESA dataset [38]; and 3) 225,731 British white females in the UK Biobank dataset [25]. Similar ORs for association of the minor allele of these SNPs with S-LAM were observed in all three comparisons (Table 2). Furthermore, we compared the MAFs of the 2 SNPs in LAM patients with those available from 7 other studies (composed of NHW European or USA populations), including all UK Biobank individuals. The MAFs of the 2 SNPs in LAM patients were significantly smaller than those reported in every other cohort (Supplemental Table 4).

To attempt to identify the causal SNP(s) among the SNPs with low P-values, we performed PICS analysis for all SNPs in Table 1. rs41374846 had both significant association with LAM, and the largest PICS probability ($P_{PICS}$=0.65, Supplementary Table 5), making it the candidate causal SNP in this association [22].

We also queried the GTEx database for SNPs in this LD block that might have an eQTL relationship with expression levels of any gene. None were identified.

### Association of GWAS-significant SNPs with *NR2F2*

The majority of SNPs associated with human disease or other phenotypes are thought to cause the association through effects on enhancer regions or other regulatory elements of a coding gene within the topologically associated domain (TAD) containing the SNP [39]. To identify the TAD containing these SNPs, we used TAD information available for four tissues: IMR90 cells, a fetal lung myofibroblast cell line [40]; lung tissue [41]; H1-MSC, a mesenchymal stem cell line [42]; and HUVEC, human umbilical vein endothelial cells [40]. Supplementary Figures 5-8 display Hi-C heatmaps for the 3 Mb region containing the GWAS SNPs and NR2F2 for these cells/tissues. HUGIN showed that P-values between rs4544201 and NR2F2 were $<10^{-18}$ for IMR90, $<10^{-16}$ for H1-MSC, and $\approx 0.1$ for lung tissue (not available for HUVEC) [27]. Thus the region containing our significant SNPs interacts with the NR2F2 genomic region in IMR90 and H1-MSC cells.

NR2F2 is the only protein-coding gene within the TAD containing the associated SNPs. This suggests that this SNP region may influence expression of *NR2F2* as its mechanism of association with S-LAM.

To examine this possibility in further detail, we conducted gene-based analyses of association of SNPs within each of the three protein-coding genes in the

2 MB region of chromosome 15 surrounding the GWAS-SNPs using SKAT-O. *NR2F2* was the only one of the three genes located in this chromosomal region that showed a significant association (P-value=0.03, Table 3).

*NR2F2*, also known as COUP-transcription factor II, encodes a member of the steroid/thyroid hormone superfamily of nuclear receptors [43], and plays important roles in many developmental processes, including the neural crest [44], which is considered a potential candidate cell of origin of LAM [45], as well as in lymphangiogenesis and in angiogenesis [46]. Hence, we considered it a potential target of regulation by one of the SNPs showing a strong association with LAM (Table 2), and performed further studies.

## Analysis of *NR2F2* in kidney angiomyolipoma and LAM

Using RNA-seq data, we compared the gene expression of four kidney angiomyolipomas and one abdominal LAM tumor with an extensive set of human cancers (from TCGA [30]), and normal tissues (from GTEX [31]) (Figure 4). *NR2F2* expression was higher in the LAM-related tumors than any TCGA cancer (Figure 4a), and was also relatively highly expressed in LAM-related tumors in comparison to normal tissues (Figure 4b, P = $6.38\times10^{-6}$, Limma statistic)[47]. In contrast, two other genes, *SPATA8* and *MCTP2*, that were next closest to the SNP region showing association with LAM (1.1 and 1.2Mb distant, Figure 3b) had no expression in the LAM-related tumors (data not shown).

Immunohistochemistry (IHC) analysis also demonstrated strong nuclear expression of *NR2F2* in both LAM lung (n=8) and kidney angiomyolipoma sections (n=4) (Figure 5).

## DISCUSSION

LAM occurs almost exclusively in women of childbearing age. Most LAM patients who come to medical attention are sporadic cases without TSC, and the origins of LAM in S-LAM patients are completely unknown. In the present study, we conducted a GWAS in a large cohort of S-LAM subjects. Twenty intergenic SNPs were identified in a 34kb LD block on chromosome 15, that met genome-wide significance for association with LAM, including rs4544201 and rs2006950 that were directly genotyped (Table 1). The association was replicated in a validation population.

The SNPs with association to S-LAM lie in a gene desert on distal chromosome 15q26.2. The nearest protein-coding gene is *NR2F2*, 700kb away, and consideration of chromatin TADs in this region indicates that only *NR2F2* is in/on the border of the TAD region containing the SNPs showing association with S-LAM in four relevant cells/tissues, suggesting that these SNP alleles may influence *NR2F2* expression as the potential mechanism of their association with S-LAM development.

*NR2F2* is an orphan nuclear receptor with known critical functions in development and tumorigenesis [48], making it a promising candidate driver gene in LAM pathogenesis. LAM occurs nearly exclusively in women, and estrogen levels influence LAM development and progression [49, 50]. siRNA knockdown of ERα (Estrogen Receptor) in MCF-7 breast cancer cells decreased *NR2F2* expression, while treatment with estradiol increased its expression [51]. This interaction between ERα and *NR2F2* may also play a role in LAM development.

*NR2F2* is highly expressed in LAM and angiomyolipoma by RNA-Seq analysis in comparison to large cancer and normal tissue data sets, and *NR2F2* shows high expression with nuclear localization in both LAM and angiomyolipoma by IHC.

Although we did not identify an eQTL relationship for any of the 20 SNPs associated with S-LAM for any gene in any normal tissue or cancer type [31], it is possible that such an eQTL relationship exists for LAM cells. We also note that the region of these SNPs contains several non-coding long RNAs, some antisense transcripts, and miR1469 (Figure 3a). It is possible that expression of one or more of these noncoding genes are affected by these SNP alleles, and have a role in LAM development, a possibility which requires further investigation.

Lymphatic involvement in LAM is a hallmark pathologic feature with LAM cell clusters in the lung showing marked enrichment for lymphatic vessels [52, 53]. VEGF-D is a probable driver of lymphatic vessel growth in LAM, as serum VEGF-D levels are increased in the majority of LAM patients, and serves as a diagnostic biomarker of LAM [54]. In mice, *NR2F2* has been shown to be required, with *SOX18*, for the polarized expression of *PROX1* in a subset of endothelial cells within the cardinal vein at embryonic day 9.5, an event that leads to development of the lymphatic endothelium [55]. Hence there is also a potential connection between *NR2F2*, VEGF-D, lymphatic development, and LAM pathogenesis.

There are potential limitations to our study. Although our cohort of samples was large for a rare disease like S-LAM, it was of only moderate size for GWAS. In order to obtain sufficient patient samples, we employed a worldwide recruitment strategy for S-LAM patients of European origin. Although our controls were all from the USA, they were selected for European ancestry to minimize population stratification issues. In addition, we employed EIGENSTRAT to remove genetic outliers from both S-LAM patients and controls. Finally we used a CLR design, matching each case with two controls to further minimize confounding due to genetic heterogeneity. Previous studies have shown that CLR is superior to unconditional logistic regression (LR) if variables used for matching are true confounding variables, and only a moderate number of controls are excluded through matching [56-62]. We also found that CLR generated more significant results than LR (Supplemental Table 6). Functional analyses to confirm our hypothesis that *NR2F2* is the gene affected by this SNP are limited due to the absense of a reliable LAM tumor cell line, the very low abundance of LAM cells in LAM lung specimens (often <5%), and lack of a LAM animal model.

In conclusion, our GWAS has identified non-coding SNPs on chr15q26.2 whose alleles are associated with S-LAM, that are located in a TAD containing the orphan nuclear receptor *NR2F2*, suggesting a model in which these SNP alleles influence *NR2F2* expression and thereby LAM pathogenesis. *NR2F2* is relatively highly expressed in LAM and LAM-related tumors. *NR2F2* has not previously been implicated in LAM, and these novel and unexpected findings will hopefully lead to better understanding of the pathogenesis of this often progressive and lethal lung disorder.

**Acknowledgements**

We thank all the LAM patients who participated in this research project; and the following individuals who contributed to subject recruitment: Dr. Hubert Wirtz, Universitätsklinikum Leipzig, Germany; Iris Bassi, LAM Italia; Corine Durand, FLAM Association (France Lymphangioléiomyomatose); Lesley Bogoslavski, Israeli LAM Organization; and Dr. Mordechai Kramer, Beilinson Hospital, Israel. This research was supported by the LAM Foundation, John Adler, The Engles Fund for TSC and LAM research, and the Bio-Synergy Research Project (NRF-2017M3A9C4065964) of the Ministry of Science, ICT and Future Planning through the Korean National Research Foundation.

## Data and Code Availability

The primary GWAS and replication data will be made available on publication of this work through dbGaP.

## REFERENCES

1. Kitaichi M, Nishimura K, Itoh H, Izumi T. Pulmonary lymphangioleiomyomatosis: a report of 46 patients including a clinicopathologic study of prognostic factors. *Am J Resp Crit Care* 1995: 151(2): 527-533.

2. Chu SC, Horiba K, Usuki J, Avila NA, Chen CC, Travis WD, Ferrans VJ, Moss J. Comprehensive evaluation of 35 patients with lymphangioleiomyomatosis. *CHEST Journal* 1999: 115(4): 1041-1052.

3. Urban T, Lazor R, Lacronique J, Murris M, Labrune S, Valeyre D, Cordier J-F. Pulmonary lymphangioleiomyomatosis: a study of 69 patients. *MEDICINE-BALTIMORE-* 1999: 78: 321-337.

4. Cunha B, Conceição DM, Cabo C, Jesus N, Santos L, de Carvalho A. Pulmonary Lymphangioleiomyomatosis on a Post-Menopausal Woman with Chronic Lymphocytic Leukaemia. *Case Reports in Clinical Medicine* 2016: 5(03): 101.

5. Youssef AL, Alami B, Sahnoun F, Boubbou M, Kamaoui I, Maâroufi M, Houssaini NS, Amara B, Tizniti S. Lymphangioleiomyomatosis: An unusual age of diagnosis with literature review. *International Journal of Diagnostic Imaging* 2014: 1(1): 17.

6. Soler-Ferrer C, Gómez-Lozano A, Clemente-Andrés C, De Cendra-Morera E, Custal-Teixidor M, Colomer-Pairés J. Lymphangioleiomyomatosis in a post-menopausal women. *Archivos de Bronconeumología ((English Edition))* 2010: 46(3): 148-150.

7. Taylor JR, Ryu J, Colby TV, Raffin TA. Lymphangioleiomyomatosis. *New England Journal of Medicine* 1990: 323(18): 1254-1260.

8. Kalassian KG, Doyle R, Kao P, Ruoss S, Raffin TA. Lymphangioleiomyomatosis: new insights. *Am J Resp Crit Care* 1997: 155(4): 1183-1186.

9. Giannikou K, Malinowska IA, Pugh TJ, Yan R, Tseng Y-Y, Oh C, Kim J, Tyburczy ME, Chekaluk Y, Liu Y. Whole exome sequencing identifies TSC1/TSC2 biallelic loss as the primary and sufficient driver event for renal angiomyolipoma development. *PLoS genetics* 2016: 12(8): e1006242.

10. Carsillo T, Astrinidis A, Henske EP. Mutations in the tuberous sclerosis complex gene TSC2 are a cause of sporadic pulmonary lymphangioleiomyomatosis. *Proceedings of the National Academy of Sciences* 2000: 97(11): 6085-6090.

11. Moss J, Avila NA, Barnes PM, Litzenberger RA, Bechtle J, Brooks PG, Hedin CJ, Hunsberger S, Kristof AS. Prevalence and clinical characteristics of lymphangioleiomyomatosis (LAM) in patients with tuberous sclerosis complex. *Am J Resp Crit Care* 2001: 164(4): 669-671.

12. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD. Genetic epidemiology of COPD (COPDGene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease* 2011: 7(1): 32-43.

13. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 2007: 81(3): 559-575.

14. Song YE, Lee S, Park K, Elston RC, Yang H-J, Won S. ONETOOL for the analysis of family-based big data. *Bioinformatics* 2018: 1: 3.

15. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *The American Journal of Human Genetics* 2005: 76(5): 887-

893.

16. Raymond M, Rousset F. An exact test for population differentiation. *Evolution* 1995: 49(6): 1280-1283.

17. Consortium HR. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* 2016: 48(10): 1279-1283.

18. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nature reviews Genetics* 2010: 11(7): 499.

19. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* 2006: 38(8): 904.

20. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. 2011.

21. Therneau TM, Lumley T. Package ‘survival’. *R package version* 2017: 2.41-43.

22. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 2015: 518(7539): 337-343.

23. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Team ELP, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics* 2012: 91(2): 224-237.

24. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, JacobsJr DR, Kronmal R, Liu K. Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology* 2002: 156(9): 871-881.

25. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* 2015: 12(3): e1001779.

26. 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(7398): 376-380.

27. Martin JS, Xu Z, Reiner AP, Mohlke KL, Sullivan P, Ren B, Hu M, Li Y. HUGIn: Hi-C unifying genomic interrogator. *Bioinformatics* 2017: 33(23): 3793-3795.

28. 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 biology* 2013: 14(4): R36.

29. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 2011: 12(1): 323.

30. Network CGAR. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008: 455(7216): 1061.

31. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N. The genotype-tissue expression (GTEx) project. *Nature genetics* 2013: 45(6): 580-585.

32. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 2015: 32(10): 1479-1485.

33. Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology* 2010: 6(5): e1000770.

34. Bongaarts A, Giannikou K, Reinten RJ, Anink JJ, Mills JD, Jansen FE, Spliet GW, den Dunnen WF, Coras R, Blümcke I. Subependymal giant cell astrocytomas in

Tuberous Sclerosis Complex have consistent TSC1/TSC2 biallelic inactivation, and no BRAF mutations. *Oncotarget* 2017: 8(56): 95516.

35. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research* 2016: 45(D1): D896-D901.

36. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2004: 21(2): 263-265.

37. Poirier JG, Faye LL, Dimitromanolakis A, Paterson AD, Sun L, Bull SB. Resampling to Address the Winner's Curse in Genetic Association Analysis of Time to Event. *Genetic epidemiology* 2015: 39(7): 518-528.

38. Hankinson JL, Kawut SM, Shahar E, Smith LJ, Stukovsky KH, Barr RG. Performance of American Thoracic Society-recommended spirometry reference values in a multiethnic sample of adults: the multi-ethnic study of atherosclerosis (MESA) lung study. *Chest* 2010: 137(1): 138-145.

39. Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, Greenside P, Srivas R, Phanstiel DH, Pekowska A. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* 2015: 162(5): 1051-1065.

40. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014: 159(7): 1665-1680.

41. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell reports* 2016: 17(8): 2042-2059.

42. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W. Chromatin architecture reorganization during stem cell differentiation. *Nature* 2015: 518(7539): 331.

43. Qiu Y, Krishnan V, Zeng Z, Gilbert DJ, Copeland NG, Gibson L, Yang-Feng T, Jenkins NA, Tsai MJ, Tsai SY. Isolation, characterization, and chromosomal localization of mouse and human COUP-TF I and II genes. *Genomics* 1995: 29(1): 240-246.

44. Rada-Iglesias A, Bajpai R, Prescott S, Brugmann SA, Swigut T, Wysocka J. Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell stem cell* 2012: 11(5): 633-648.

45. Julian LM, Delaney SP, Wang Y, Goldberg AA, Doré C, Yockell-Lelièvre J, Tam RY, Giannikou K, McMurray F, Shoichet MS. Human Pluripotent Stem Cell–Derived TSC2-Haploinsufficient Smooth Muscle Cells Recapitulate Features of Lymphangioleiomyomatosis. *Cancer Res* 2017: 77(20): 5491-5502.

46. Qin J, Chen XP, Xie X, Tsai MJ, Tsai SY. COUP-TFII regulates tumor growth and metastasis by modulating tumor angiogenesis. *P Natl Acad Sci USA* 2010: 107(8): 3687-3692.

47. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015: 43(7): e47.

48. Xu MF, Qin J, Tsai SY, Tsai MJ. The role of the orphan nuclear receptor COUP-TFII in tumorigenesis. *Acta Pharmacol Sin* 2015: 36(1): 32-36.

49. Juvet SC, Hwang D, Downey GP. Rare lung diseases I--Lymphangioleiomyomatosis. *Canadian respiratory journal* 2006: 13(7): 375-380.

50. McCormack FX, Gupta N, Finlay GR, Young LR, Taveira-DaSilva AM, Glasgow CG, Steagall WK, Johnson SR, Sahn SA, Ryu JH, Strange C, Seyama K, Sullivan EJ, Kotloff RM, Downey GP, Chapman JT, Han MK, D'Armiento JM, Inoue Y, Henske EP, Bissler JJ, Colby TV, Kinder BW, Wikenheiser-Brokamp KA, Brown KK, Cordier JF, Meyer C, Cottin V, Brozek JL, Smith K, Wilson KC, Moss J, Lymphangioleiomyomato AJC. Official American Thoracic Society/Japanese Respiratory Society Clinical Practice Guidelines: Lymphangioleiomyomatosis Diagnosis and Management. *Am J Resp Crit Care* 2016: 194(6): 748-761.

51. Riggs KA, Wickramasinghe NS, Cochrum RK, Watts MB, Klinge CM. Decreased chicken ovalbumin upstream promoter transcription factor II expression in tamoxifen-resistant breast cancer cells. *Cancer Res* 2006: 66(20): 10188-10198.

52. Glasgow CG, Taveira-DaSilva AM, Darling TN, Moss J. Lymphatic involvement in lymphangioleiomyomatosis. *Ann Ny Acad Sci* 2008: 1131: 206-214.

53. Seyama K, Mitani K, Kumasaka T. Lymphangioleiomyoma Cells and Lymphatic Endothelial Cells Expression of VEGFR-3 in Lymphangioleiomyoma Cell Clusters. *Am J Pathol* 2010: 176(4): 2051-2052.

54. Young LR, Lee HS, Inoue Y, Moss J, Singer LG, Strange C, Nakata K, Barker AF, Chapman JT, Brantly ML, Stocks JM, Brown KK, Lynch JP, Goldberg HJ, Downey GP, Swigris JJ, Taveira-DaSilva AM, Krischer JP, Trapnell BC, McCormack FX, Grp MT. Serum VEGF-D concentration as a biomarker of lymphangioleiomyomatosis severity and treatment response: a prospective analysis of the Multicenter International Lymphangioleiomyomatosis Efficacy of Sirolimus (MILES) trial. *Lancet Resp Med* 2013: 1(6): 445-452.

55. Srinivasan RS, Geng X, Yang Y, Wang Y, Mukatira S, Studer M, Porto MP, Lagutin O, Oliver G. The nuclear hormone receptor Coup-TFII is required for the initiation and early maintenance of Prox1 expression in lymphatic endothelial cells. *Genes & development* 2010: 24(7): 696-707.

**Table 1. Statistical analyses of imputed SNPs with CLR.** Imputation was conducted using EAGEL2 and PBWT for pre-phasing. Imputation was conducted by using the Haplotype Reference Consortium as reference panel.

| CHR | SNP | POS | Alleles[*] | MAF | Imputed vs genotyped | INFO[†] | P-value for CLR[‡] |
|---|---|---|---|---|---|---|---|
| 15 | rs41374846 | 96143559 | A/G | 0.2605 | imputed | 0.9097 | $1.322\times10^{-7}$ |
| 15 | rs59125351 | 96144157 | G/T | 0.2510 | imputed | 0.9771 | $2.741\times10^{-9}$ |
| 15 | rs17581137 | 96146414 | C/A | 0.2336 | imputed | 0.9893 | $1.250\times10^{-10}$ |
| 15 | rs6496126 | 96148439 | C/G | 0.2330 | imputed | 0.9890 | $6.982\times10^{-9}$ |
| 15 | rs2397810 | 96148765 | C/T | 0.2330 | imputed | 0.9890 | $6.691\times10^{-9}$ |
| 15 | rs10520790 | 96151040 | T/G | 0.2478 | imputed | 0.9958 | $6.691\times10^{-9}$ |
| 15 | rs55804812 | 96151256 | A/T | 0.2475 | imputed | 0.9952 | $4.008\times10^{-8}$ |
| 15 | rs16975389 | 96153782 | C/T | 0.2463 | imputed | 0.9967 | $1.173\times10^{-8}$ |
| 15 | rs16975396 | 96158705 | G/T | 0.2466 | imputed | 0.9983 | $3.547\times10^{-8}$ |
| 15 | rs4544201 | 96167827 | A/G | 0.2469 | genotyped | 1.0000 | $4.186\times10^{-8}$ |
| 15 | rs4628911 | 96167905 | T/C | 0.2472 | imputed | 1.0000 | $3.547\times10^{-8}$ |
| 15 | rs6496128 | 96168303 | G/A | 0.2472 | imputed | 1.0000 | $3.547\times10^{-8}$ |
| 15 | rs8029996 | 96168770 | A/G | 0.2472 | imputed | 0.9998 | $3.547\times10^{-8}$ |
| 15 | rs4551988 | 96169589 | C/G | 0.2472 | imputed | 0.9998 | $3.547\times10^{-8}$ |
| 15 | rs58878263 | 96171069 | A/C | 0.2493 | imputed | 0.9979 | $3.632\times10^{-8}$ |
| 15 | rs8040665 | 96175692 | G/T | 0.2487 | imputed | 0.9976 | $2.375\times10^{-8}$ |
| 15 | 15:96175733 | 96175733 | A/G | 0.2466 | imputed | 0.9975 | $2.227\times10^{-8}$ |
| 15 | rs8040168 | 96176096 | G/C | 0.2466 | imputed | 0.9981 | $2.227\times10^{-8}$ |
| 15 | rs17504029 | 96177670 | T/A | 0.2478 | imputed | 0.9876 | $2.289\times10^{-8}$ |
| 15 | rs2006950 | 96179390 | A/G | 0.2262 | genotyped | 1.0000 | $6.117\times10^{-9}$ |

Definition of abbreviations: CHR = Chromosome; POS = SNP Position according to NCBI genome build 37 (hg19); MAF = Minor allele frequency; CLR = Conditional Logistic Regression.

[*] Minor/Major alleles are listed.

[†] INFO is a metric for imputation quality determined by IMPUTE2.

[‡] CLR was applied to imputed SNP genotype data to identify SNPs with significant association ($P < 5\times10^{-8}$) with S-LAM.

**Table 2. Genome-wide significant genotyped SNPs.**

| | **rs4544201** | **rs2006950** |
|---|---|---|
| ***Chromosome*** | 15q26.2 | 15q26.2 |
| ***SNP position (hg19)*** | 96167827 | 96179390 |
| ***Minor / Major alleles*** | A / G | A / G |
| ***Minor allele frequency*** | | |
| S-LAM | 0.1655 | 0.1420 |
| Control | 0.2750 | 0.2529 |
| ***Discovery data*** | | |
| ***Genotype counts (AA / AG / GG / Missing)*** | | |
| S-LAM | 16 / 108 / 299 / 3 | 11 / 99 / 316 / 0 |
| Control | 62 / 343 / 444 / 3 | 58 / 315 / 479 / 0 |
| **Odds ratio** | | |
| Original | 0.4973 | 0.4673 |
| Bias adjusted | 0.5925 | 0.5272 |
| **P-value** | $4.19\times10^{-8}$ | $6.12\times10^{-9}$ |
| ***Replication data*** | | |
| ***Genotype counts (AA / AG / GG / Missing)*** | | |
| S-LAM | 4 / 48 / 144 / 0 | 3 / 39 / 154 / 0 |
| COPDGene | 26 / 171 / 212 / 0 | 26 / 159 / 224 |
| MESA | 69 / 417 / 635 / 0 | 64 / 385 / 672 / 0 |
| UK BioBank | 14468 / 85721 / 125542 / 0 | 12765 / 81784 / 131182 / 0 |
| S-LAM vs COPDGene | | |
| Odds ratio | 0.3288 | 0.2731 |
| P-value | $4.32\times10^{-5}$ | $1.56\times10^{-5}$ |
| S-LAM vs MESA | | |
| Odds ratio | 0.5070 | 0.4448 |
| P-value | $9.28\times10^{-6}$ | $1.04\times10^{-6}$ |
| S-LAM vs UK BioBank | | |
| Odds ratio | 0.4888 | 0.4159 |

| | | |
|---|---|---|
| P-value | $7.30\times10^{-7}$ | $3.11\times10^{-8}$ |

Definition of abbreviations: SNP = Single-Nucleotide Polymorphism; S-LAM = Sporadic Lymphangioleiomyomatosis.

**Table 3. Gene-based analyses of SNP association with LAM.** Three protein-coding genes were found on chromosome 15 from 94.2 Mb to 98.2 Mb, the 4 Mb region surrounding the GWAS-SNPs, and gene-based analysis for association with LAM was performed using SKAT-O.

| Gene | CHR | Start* | End† | Number of SNPs | P-value |
|---|---|---|---|---|---|
| *NR2F2* | 15 | 96869157 | 96883492 | 5 | 0.0307 |
| *MCTP2* | 15 | 94774767 | 95027181 | 4 | 0.3579 |
| *SPATA8* | 15 | 97326619 | 97328845 | 3 | 0.5250 |

Definition of abbreviations: SNP= Single-Nucleotide Polymorphism; LAM = Lymphangioleiomyomatosis; GWAS = Genome-Wide Association Study; CHR = Chromosome

* Start position of the corresponding gene.

† End position of the corresponding gene.

**Figure Legends**

**Figure 1. Workflow of statistical analysis and quality control for the LAM GWAS discovery data set.** Multiple standard quality controls were performed for both cases (S-LAM subjects) and controls (healthy women without COPD from COPDGene consortium) to exclude outlier SNPs and subjects. HWE, Hardy-Weinberg equilibrium test; MAF, minor allele frequency; IBS, identity-by-state.

**Figure 2. Quantile-quantile and Manhattan plots for the discovery LAM GWAS.**

a) The observed distributions of P-values for 5,426,936 SNPs including 549,591 directly genotyped are plotted relative to the expected (null) distribution for the Conditional logistic regression (CLR) analysis. b) Manhattan plot. Each dot represents the P-value of a single SNP, plotted on the genome scale at bottom. The Y-axis value is the negative logarithm of the P-value for association between each genotyped SNP and S-LAM. Two SNPs on 15q met genome-wide significance.

**Figure 3. Genomic region on chr15 containing the SNPs associated with LAM.**

a. Ideogram of chromosome 15.

b. Three Mb region containing the SNPs associated with LAM. Manhattan plot at top shows P-values for directly genotyped SNPs in this region, including the two SNPs meeting genome-wide significance (red dots). There are 3 protein-coding genes *NR2F2*, *MCTP2*, and *SPATA8* which are highlighted by yellow backbround, and many lncRNAs in this region.

c. Expanded Manhattan plot of the 250kb region showing both genotyped and imputed SNPs. SNP rs41374846, the candidate causal SNP, is indicated by purple, and other SNPs are colored according to their $r^2$ value in relation to that SNP.

**Figure 4. Comparison of *NR2F2* expression in kidney angiomyolipoma/LAM with cancer (TCGA) and normal (GTEx) tissues.**

Boxplot figures are shown to compare expression of *NR2F2* in 4 angiomyolipoma tumors and one abdominal LAM lesion with 2463 cancers of 27 types (from TCGA) in RSEM units (a); and with ~7,000 samples of 47 normal tissues (from GTEx) in RPKM units (b). The median value, interquartile range, and 95% ranges are shown, with outliers indicated by circles. Abbreviations used here for TCGA cancer types are explained in Supplemental Table 7.

**Figure 5. Immunohistochemistry for *NR2F2* in LAM and kidney angiomyolipoma.** Strong nuclear staining is seen in lung LAM cells (A) and angiomyolipoma cells (B) (brown stain). Some other cells also have nuclear staining for *NR2F2* but most do not. This is a representative field obtained from 8 LAM lung samples and 4 angiomyolipoma samples examined by IHC.

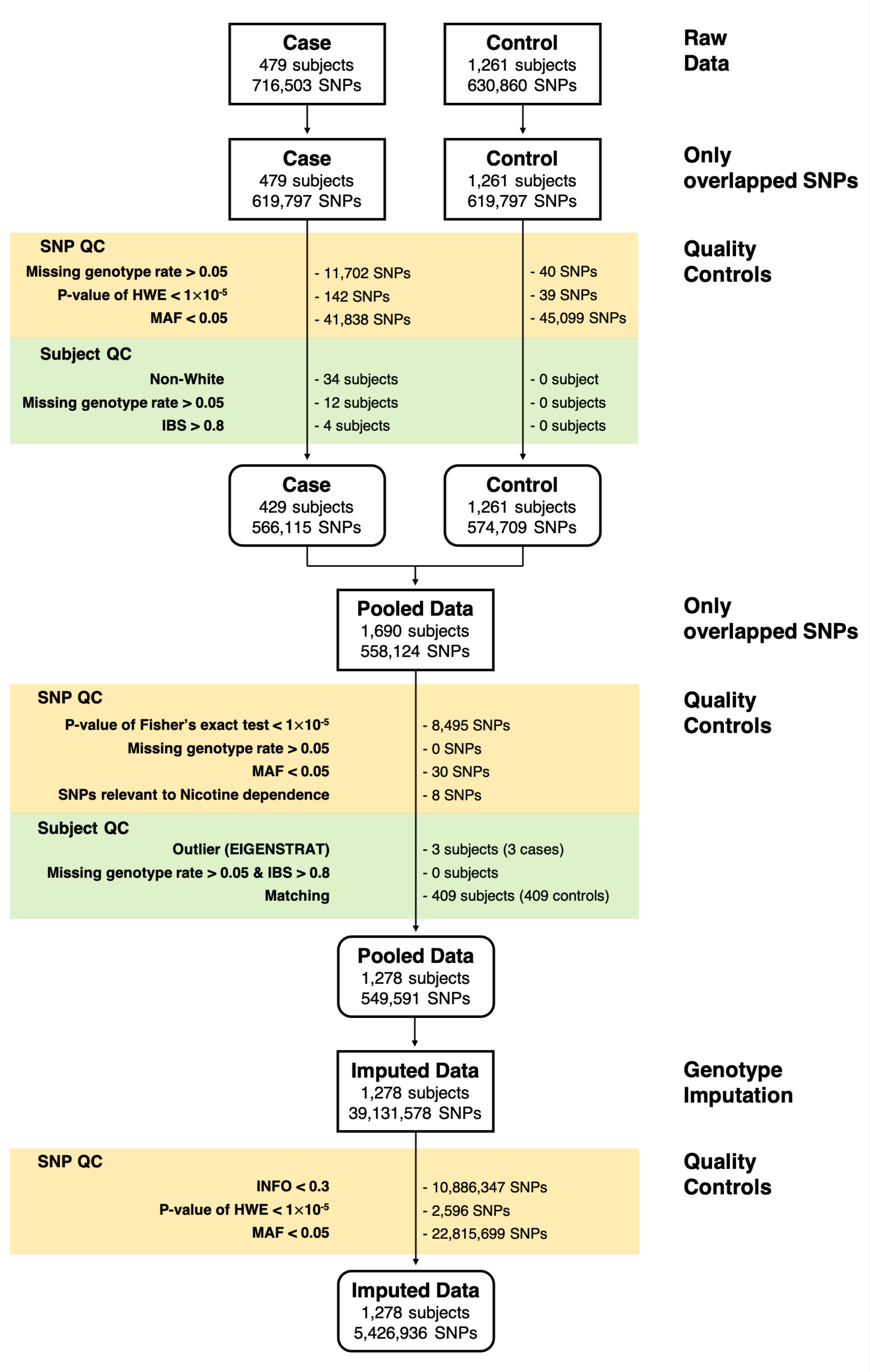

Case
479 subjects
716,503 SNPs
Control
1,261 subjects
630,860 SNPs
Raw Data
Case
479 subjects
619,797 SNPs
Control
1,261 subjects
619,797 SNPs
Only overlapped SNPs
SNP QC
Missing genotype rate > 0.05
P-value of HWE < $1\times10^{-5}$
MAF < 0.05
- 11,702 SNPs
- 142 SNPs
- 41,838 SNPs
- 40 SNPs
- 39 SNPs
- 45,099 SNPs
Quality Controls
Subject QC
Non-White
Missing genotype rate > 0.05
IBS > 0.8
- 34 subjects
- 12 subjects
- 4 subjects
- 0 subject
- 0 subjects
- 0 subjects
Case
429 subjects
566,115 SNPs
Control
1,261 subjects
574,709 SNPs
Pooled Data
1,690 subjects
558,124 SNPs
Only overlapped SNPs
SNP QC
P-value of Fisher's exact test < $1\times10^{-5}$
Missing genotype rate > 0.05
MAF < 0.05
SNPs relevant to Nicotine dependence
- 8,495 SNPs
- 0 SNPs
- 30 SNPs
- 8 SNPs
Quality Controls
Subject QC
Outlier (EIGENSTRAT)
Missing genotype rate > 0.05 & IBS > 0.8
Matching
- 3 subjects (3 cases)
- 0 subjects
- 409 subjects (409 controls)
Pooled Data
1,278 subjects
549,591 SNPs
Imputed Data
1,278 subjects
39,131,578 SNPs
Genotype Imputation
SNP QC
INFO < 0.3
P-value of HWE < $1\times10^{-5}$
MAF < 0.05
- 10,886,347 SNPs
- 2,596 SNPs
- 22,815,699 SNPs
Quality Controls
Imputed Data
1,278 subjects
5,426,936 SNPs

Figure 1.

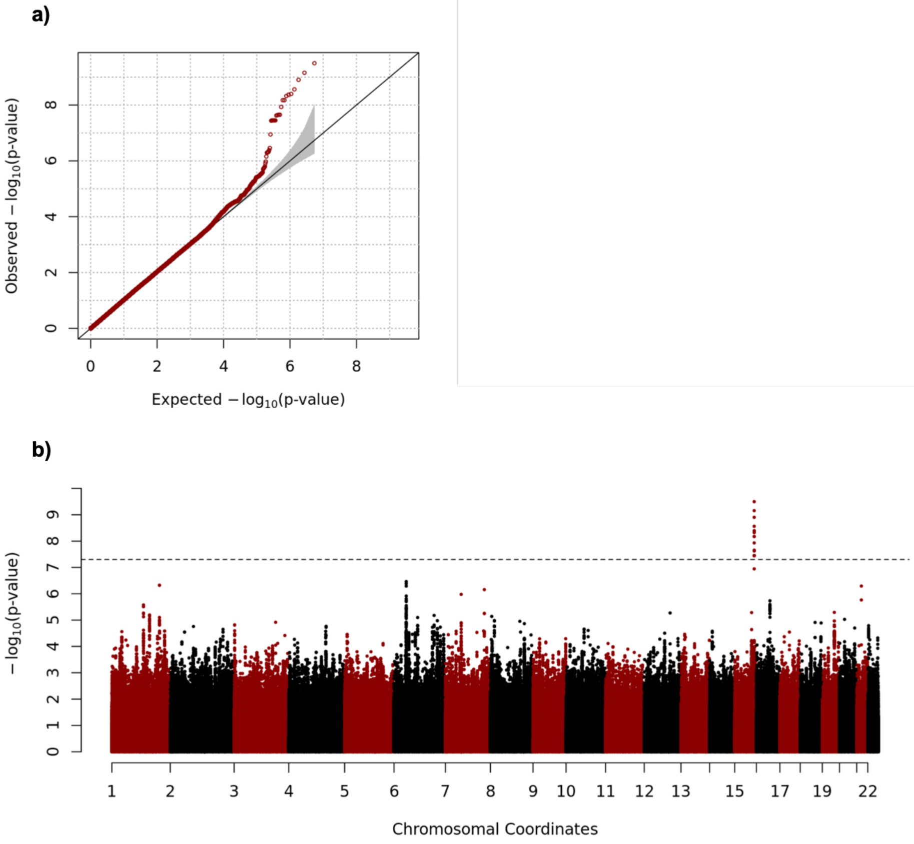

a)
Observed $-\log_{10}$(p-value)
0
2
4
6
8
Expected $-\log_{10}$(p-value)
0
2
4
6
8
b)
$-\log_{10}$(p-value)
0
1
2
3
4
5
6
7
8
9
1
2
3
4
5
6
7
8
9
10
11
12
13
15
17
19
22
Chromosomal Coordinates

Figure 2.

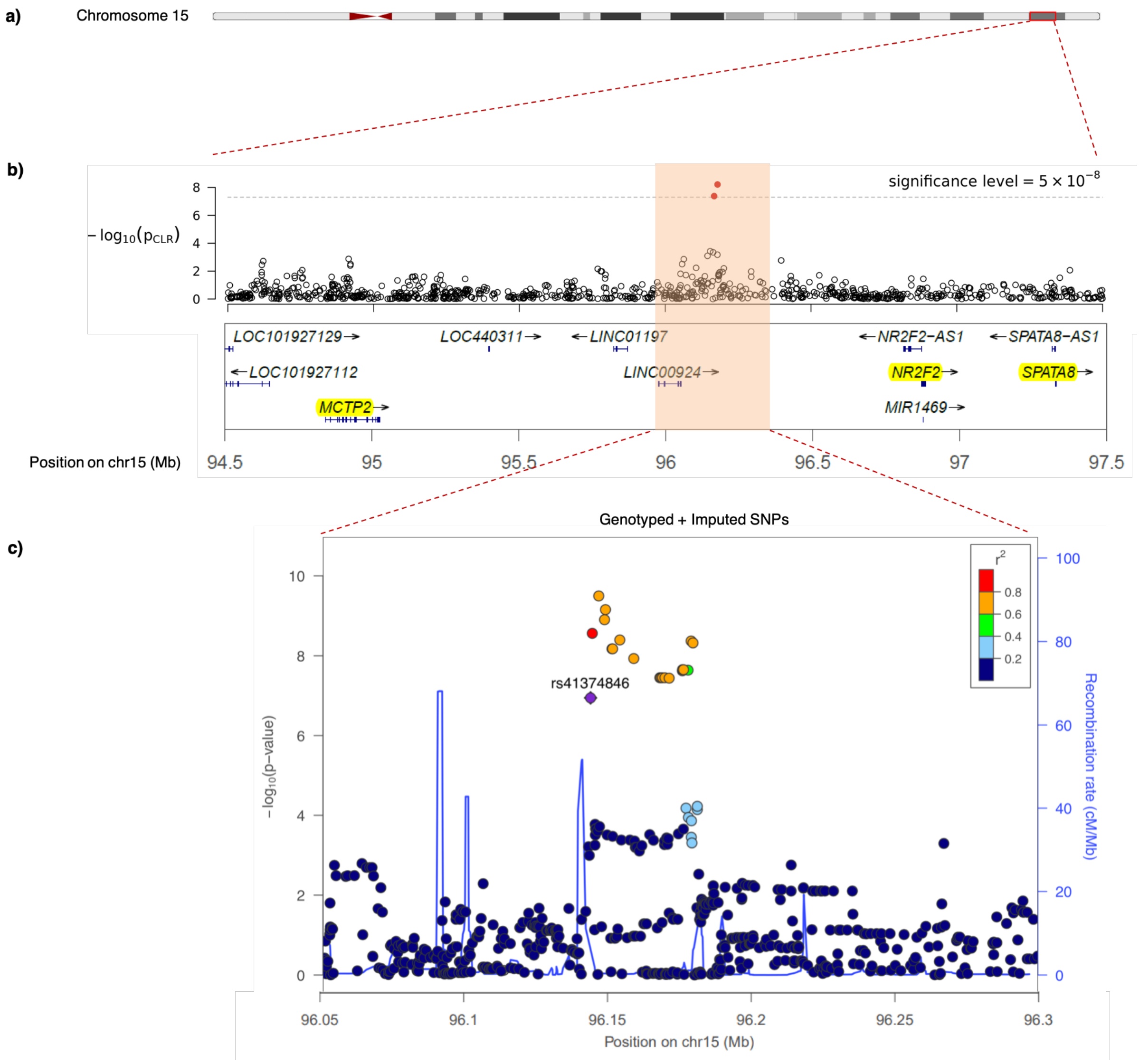

a)
Chromosome 15
b)
significance level = 5 × 10$^{-8}$
$-\log_{10}(p_{CLR})$
8
6
4
2
0
LOC101927129→
LOC440311→
←LINC01197
←NR2F2−AS1
←SPATA8−AS1
←LOC101927112
LINC00924→
NR2F2→
SPATA8→
MCTP2→
MIR1469→
Position on chr15 (Mb)
94.5
95
95.5
96
96.5
97
97.5
c)
Genotyped + Imputed SNPs
$-\log_{10}$(p-value)
10
8
6
4
2
0
rs41374846
r²
0.8
0.6
0.4
0.2
Recombination rate (cM/Mb)
100
80
60
40
20
0
96.05
96.1
96.15
96.2
96.25
96.3
Position on chr15 (Mb)

Figure 3.

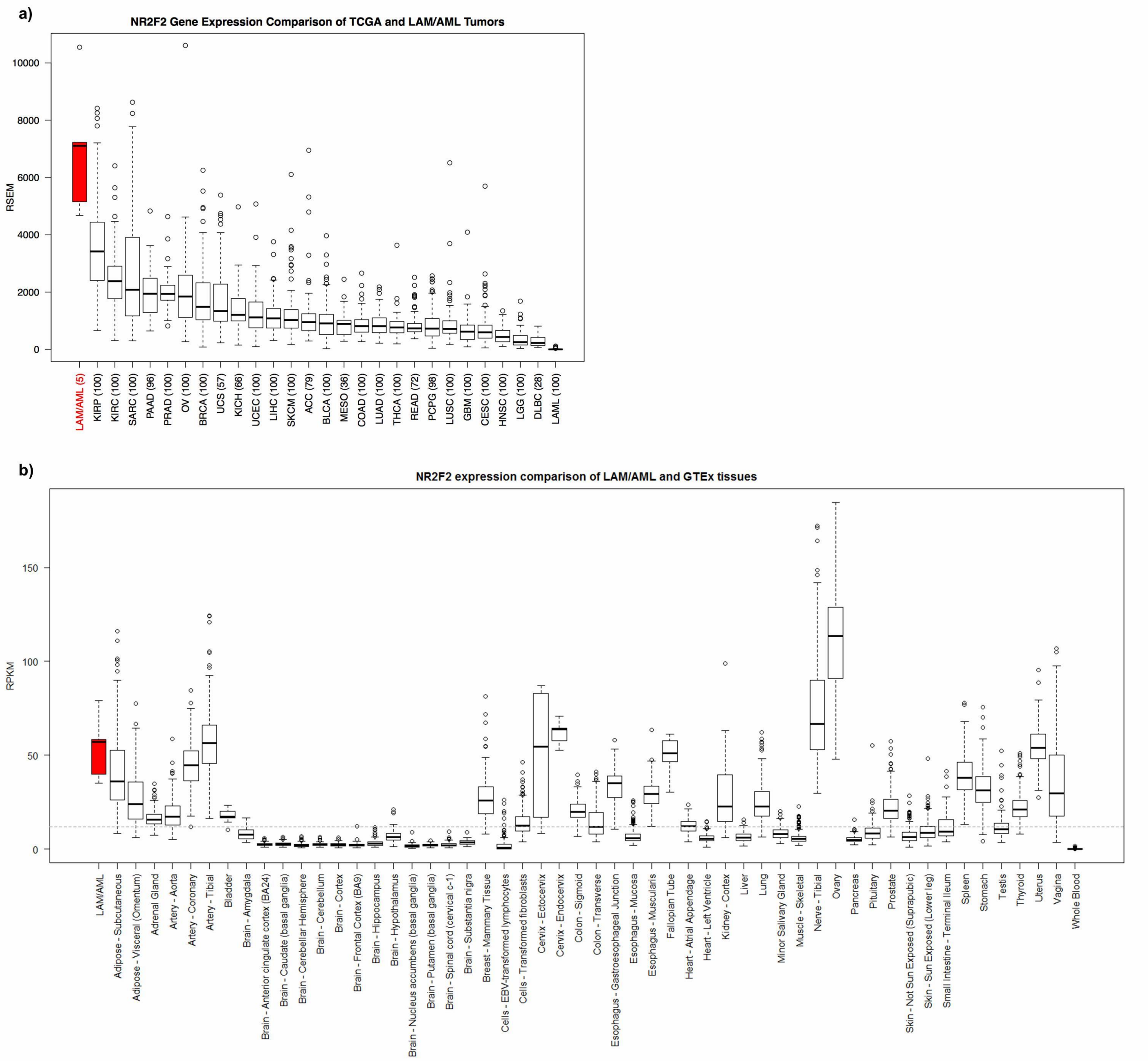

a)
NR2F2 Gene Expression Comparison of TCGA and LAM/AML Tumors
RSEM
10000
8000
6000
4000
2000
0
LAM/AML (5)
KIRP (100)
KIRC (100)
SARC (100)
PAAD (96)
PRAD (100)
OV (100)
BRCA (100)
UCS (57)
KICH (66)
UCEC (100)
LIHC (100)
SKCM (100)
ACC (79)
BLCA (100)
MESO (36)
COAD (100)
LUAD (100)
THCA (100)
READ (72)
PCPG (98)
LUSC (100)
GBM (100)
CESC (100)
HNSC (100)
LGG (100)
DLBC (28)
LAML (100)
b)
NR2F2 expression comparison of LAM/AML and GTEx tissues
RPKM
150
100
50
0
LAM/AML
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cells - Transformed fibroblasts
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood

Figure 4.

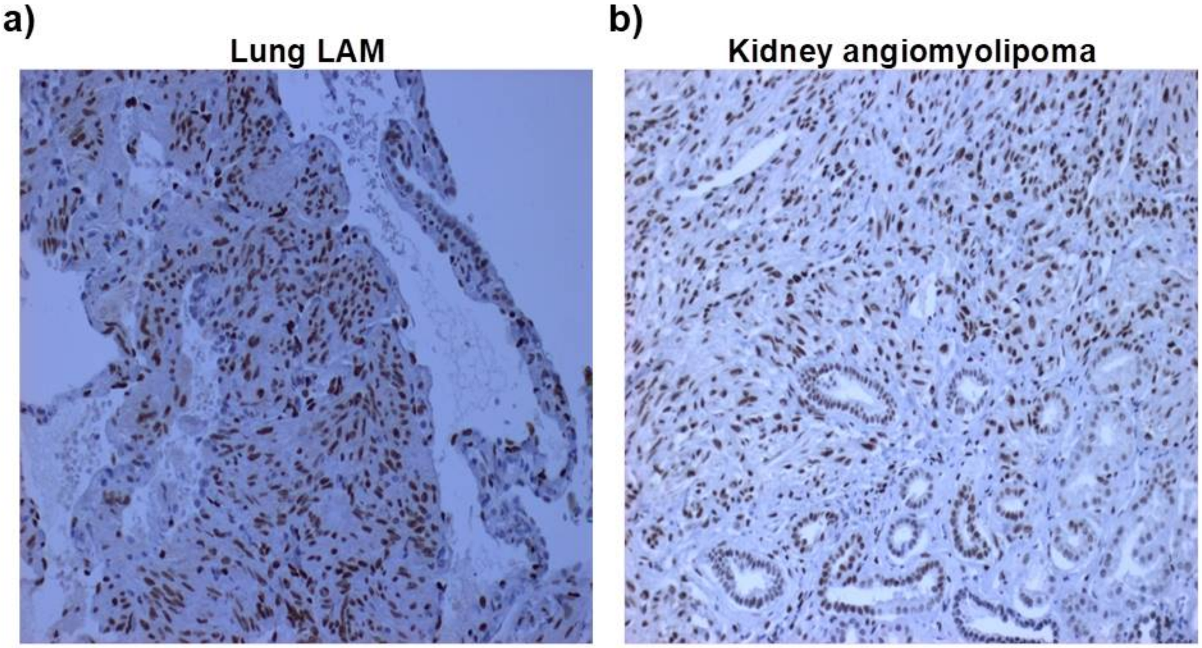


Figure 5.

# Supplementary Materials for “A Genome Wide Association Study implicates *NR2F2* in Lymphangioleiomyomatosis Pathogenesis”

**Supplementary Table 1.** Distribution of LAM patients according to their nationality

| | Discovery LAM | Replication LAM |
|---|---|---|
| USA | 190 | 196 |
| France | 54 | 0 |
| Spain | 40 | 0 |
| Italy | 35 | 0 |
| United Kingdom | 32 | 0 |
| Germany | 21 | 0 |
| Australia | 20 | 0 |
| Poland | 15 | 0 |
| Israel | 7 | 0 |
| Canada | 4 | 0 |
| Panama | 1 | 0 |
| Puerto Rico | 1 | 0 |
| Scotland | 1 | 0 |
| Unknown | 5 | 0 |
| Total | 426 | 196 |

**Supplementary Table 2. P-values for SNPs associated with nicotine addiction.**

P values are shown in comparison of allele frequencies for the S-LAM discovery cohort and the COPDGene controls.

| CHR | SNP | Mapped gene | P-value |
|---|---|---|---|
| 1 | rs1060061 | *NR5A2* | 0.4885 |
| 6 | rs9503551 | *SLC22A23* | 0.0840 |
| 7 | rs4285401 | *LINC01287* | 0.3263 |
| 8 | rs804292 | *NEIL2* | 0.8145 |
| 8 | rs6470120 | *ZHX2* | 0.1152 |
| 9 | rs10491551* | *GLIS3* | 0.7217 |
| 15 | rs1051730 | *CHRNA3* | 0.9759 |
| 21 | rs2836823 | *AF064858.3* | 0.1560 |

* rs10491551 is included due to its high correlation with rs12348139 in the GWAS catalogue ($r^2$ = 1).

**Supplementary Table 3. P-values for rs4544201 and rs2006950 adjusted by effect of *TSC1/2* genes.**

| *TSC1/2* | rs4544201 | rs2006950 |
|---|---|---|
| rs11552431 | $4.56\times10^{-8}$ | $3.98\times10^{-9}$ |
| Top 10 SNPs | $1.08\times10^{-7}$ | $1.13\times10^{-8}$ |

**Supplementary Table 4. Minor allele frequencies for SNPs rs4544201 and rs2006950 in multiple populations.**

| SNP | LAM patients | | | Normal | | |
|---|---|---|---|---|---|---|
| | Data | N | MAF<br>(95% CI) | Data | N | MAF<br>(95% CI) |
| rs4544201 | Discovery<br>(USA/NHW/females) | 190 | 0.1684<br>(0.131, 0.206) | COPDGene<br>(USA/NHW/females) | 1,258 | 0.2742<br>(0.257, 0.292) |
| | Discovery<br>(EUR/NHW/females) | 233 | 0.1631<br>(0.130, 0.197) | COPDGene<br>(USA/NHW/males) | 1,224 | 0.2774<br>(0.260, 0.295) |
| | Replication<br>(USA/NHW/females) | 186 | 0.1429<br>(0.107, 0178) | MESA-Lung[*]<br>(USA/HW/females) | 1,153 | 0.2563<br>(0.238, 0.274) |
| | | | | 1000GP[**]<br>(USA/NHW/females) | 50 | 0.2600<br>(0.174, 0.346) |
| | | | | 1000GP[**]<br>(EUR/NHW/females) | 213 | 0.2300<br>(0.190, 0.270) |
| | | | | ECLIPSE[***]<br>(EUR/NHW/females) | 792 | 0.2563<br>(0.235, 0.278) |
| | | | | UKBiobank[†]<br>(EUR/NHW/both) | 337,199 | 0.2605<br>(0.259, 0.262) |
| | | | | GnomAD[‡]<br>(EUR/NHW/both) | 7,482 | 0.2601<br>(0.253, 0.267) |
| rs2006950 | Discovery<br>(USA/NHW/females) | 190 | 0.1474<br>(0.112, 0.183) | COPDGene<br>(USA/NHW/females) | 1,261 | 0.2546<br>(0.238, 0.272) |
| | Discovery<br>(EUR/NHW/females) | 230 | 0.1377<br>(0.107, 0.169) | COPDGene<br>(EUR/NHW/males) | 1,226 | 0.2557<br>(0.238, 0.273) |
| | Replication<br>(USA/NHW/females) | 186 | 0.1148<br>(0.082, 0.147) | MESA-Lung[*]<br>(USA/HW/females) | 1,128 | 0.2283<br>(0.211, 0.246) |
| | | | | 1000GP[**]<br>(USA/NHW/females) | 50 | 0.2300<br>(0.148, 0.312) |
| | | | | 1000GP[**]<br>(EUR/NHW/females) | 213 | 0.2160<br>(0.177, 0.255) |
| | | | | ECLIPSE[***]<br>(EUR/NHW/females) | 792 | 0.2431<br>(0.222, 0.264) |
| | | | | UKBiobank[†]<br>(EUR/NHW/both) | 337,199 | 0.2432<br>(0.242, 0.244) |
| | | | | GnomAD[‡]<br>(EUR/NHW/both) | 7,496 | 0.2421<br>(0.235, 0.249) |

[*] MESA = Multi-Ethnic Study of Atherosclerosis. Hispanic whites females were chosen and MAFs were calculated.

** 1000GP = 1000 Genome Project

*** ECLIPSE = Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points

† http://pheweb.sph.umich.edu:5000/

‡ http://gnomad.broadinstitute.org

**Supplementary Table 5.** PICS analysis to identify probable causal SNPs in the chr 15q region.

| CHR | SNP[*] | POS | P-value | $D'$[†] | $r^2$[‡] | PICS probability |
|---|---|---|---|---|---|---|
| **15** | **rs41374846** | **96143559** | $1.322\times10^{-7}$ | 1.0000 | 1.0000 | 0.6485 |
| 15 | rs59125351 | 96144157 | $2.741\times10^{-9}$ | 0.9703 | 0.7941 | 0.0352 |
| 15 | rs55804812 | 96151256 | $4.008\times10^{-8}$ | 0.9557 | 0.7758 | 0.0290 |
| 15 | rs16975389 | 96153782 | $3.547\times10^{-8}$ | 0.9555 | 0.7700 | 0.0272 |
| 15 | rs10520790 | 96151040 | $6.691\times10^{-9}$ | 0.9486 | 0.7698 | 0.0271 |
| 15 | rs16975396 | 96158705 | $3.547\times10^{-8}$ | 0.9480 | 0.7581 | 0.0239 |
| 15 | rs58878263 | 96171069 | $4.008\times10^{-8}$ | 0.9328 | 0.7287 | 0.0172 |
| 15 | rs8029996 | 96168770 | $3.547\times10^{-8}$ | 0.9325 | 0.7230 | 0.0161 |
| 15 | rs6496128 | 96168303 | $6.982\times10^{-9}$ | 0.9325 | 0.7230 | 0.0161 |
| 15 | rs4628911 | 96167905 | $3.547\times10^{-8}$ | 0.9325 | 0.7230 | 0.0161 |
| 15 | rs8040665 | 96175692 | $2.227\times10^{-8}$ | 0.9254 | 0.7171 | 0.0151 |
| 15 | rs17581137 | 96146414 | $1.250\times10^{-10}$ | 0.9529 | 0.7125 | 0.0143 |
| 15 | rs4544201 | 96167827 | $3.547\times10^{-10}$ | 0.9317 | 0.7116 | 0.0142 |
| 15 | rs4551988 | 96169589 | $3.547\times10^{-8}$ | 0.9183 | 0.7113 | 0.0141 |
| 15 | rs2397810 | 96148765 | $6.691\times10^{-9}$ | 0.9451 | 0.7008 | 0.0125 |
| 15 | rs6496126 | 96148439 | $6.982\times10^{-9}$ | 0.9380 | 0.7005 | 0.0124 |
| 15 | rs8040168 | 96176096 | $2.227\times10^{-8}$ | 0.9233 | 0.6887 | 0.0108 |

Definition of abbreviations: CHR = Chromosome; POS = SNP Position according to NCBI genome build 37 (hg19); CLR = Conditional Logistic Regression.

SNP rs41374846 (shown in bold) was identified as the probable causal SNP, with the highest PICS probability. SNPs are sorted by PIC probability.

[†] $D' = D_{AB}/D_{\max}$ where $D_{AB}$: the frequency of the haplotype AB and $D_{\max}$: theoretical maximum difference between the observed and expected haplotype frequencies.

[‡] $r^2$: squared correlation coefficient

**Supplementary Table 6. Unconditional logistic regression results for genome-wide significant SNPs.** We performed unconditional logistic regression using 479 cases and 1,261 controls for rs4544201 and rs2006950. Two PC scores corresponding two greatest eigenvalues and age were included as covariates.

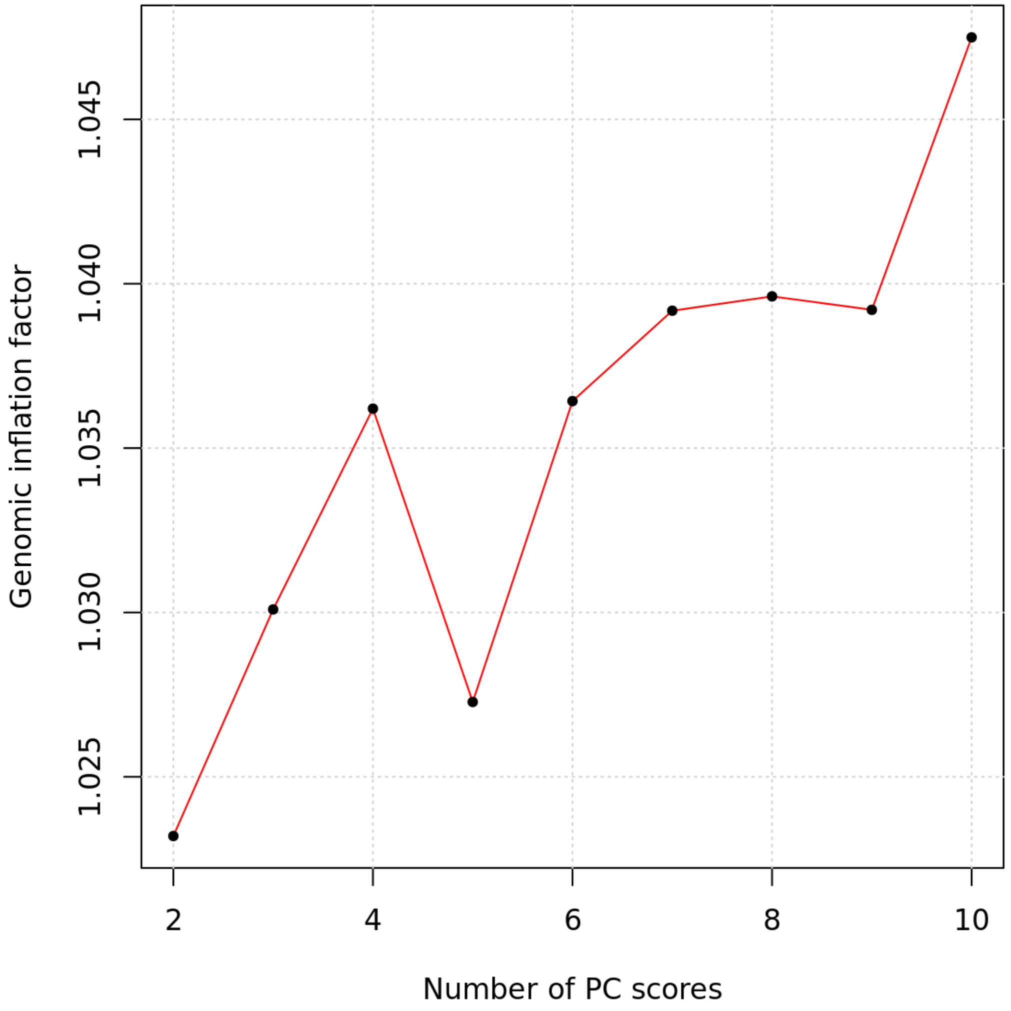
| | **rs4544201** | **rs2006950** |
|---|---|---|
| ***Chromosome*** | 15q26.2 | 15q26.2 |
| ***SNP position (hg19)*** | 96167827 | 96179390 |
| ***Minor / Major alleles*** | A / G | A / G |
| ***Minor allele frequency*** | | |
| S-LAM | 0.1655 | 0.1420 |
| Control | 0.2742 | 0.2546 |
| ***Genotype counts (AA / AG / GG / Missing)*** | | |
| S-LAM | 16 / 108 / 299 / 3 | 11 / 99 / 316 / 0 |
| Control | 88 / 514 / 656 / 3 | 84 / 474 / 703 / 0 |
| LR results | | |
| Odds ratio | 0.5728 | 0.5152 |
| P-value | $5.00\times10^{-7}$ | $1.23\times10^{-8}$ |

Definition of abbreviations: LR = Unconditional logistic regression

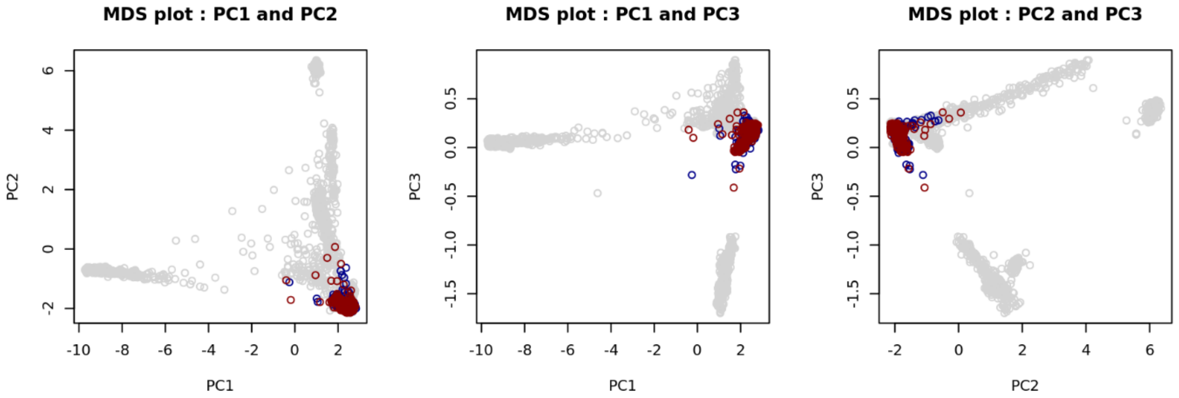
**Supplementary Table 7. TCGA tumor abbreviations**

| Abbreviation | Cancer type |
|---|---|
| **KIRP** | Kidney renal papillary cell carcinoma |
| **KIRC** | Kidney Renal Clear Cell Carcinoma |
| **SARC** | Sarcoma |
| **PAAD** | Pancreatic Adenocarcinoma |
| **OV** | Ovarian Serous Cystadenocarcinoma |
| **BRCA** | Breast Invasive Carcinoma |
| **UCS** | Uterine Carcinosarcoma |
| **KICH** | Kidney Chromophobe |
| **UCEC** | Uterine Corpus Endometrial Carcinoma |
| **LIHC** | Liver Hepatocellular Carcinoma |
| **SKCM** | Skin Cutaneous Melanoma |
| **ACC** | Adrenocortical Carcinoma |
| **BLCA** | Bladder Urothelial Carcinoma |
| **MESO** | Mesothelioma |
| **COAD** | Colon Adenocarcinoma |
| **LUAD** | Lung Adenocarcinoma |
| **THCA** | Thyroid Carcinoma |
| **READ** | Rectum Adenocarcinoma |
| **PCPG** | Pheochromocytoma and Paraganglioma |
| **LUSC** | Lung Squamous Cell Carcinoma |
| **GBM** | Glioblastoma Multiforme |
| **CESC** | Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma |
| **HNSC** | Head and Neck Squamous Cell Carcinoma |
| **LGG** | Low Grade Glioma |
| **DLBC** | Lymphoid Neoplasm Diffuse Large B-cell Lymphoma |
| **LAML** | Acute Myeloid Leukemia |

**Supplementary Figure 1. Genomic inflation factors according to the number of PC scores used for the discovery data.** Cases and controls were matched with different numbers of PC scores (2 – 10 PC scores) and age, and CLR was applied to matched cases and controls. Variance inflation factors were calculated for different numbers of PC scores, and plotted against the numbers of PC scores.



**Supplementary Figure 2. Scatter plot of PC scores.** Multi-dimensional scaling plots were generated using a pool of our Discovery S-LAM cohort, our COPDGene controls, and 1000 Genome project data. Red and blue circles indicate S-LAM and COPDGene samples used for our discovery analyses, respectively, and grey circles represent participants for 1000Genome projects.

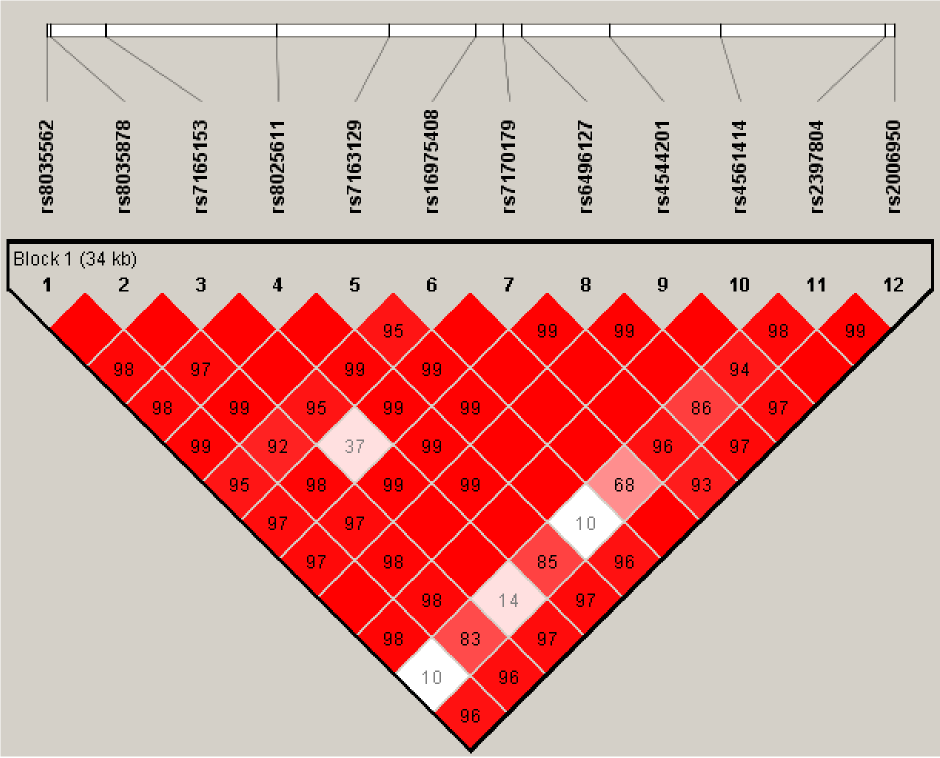


MDS plot : PC1 and PC2

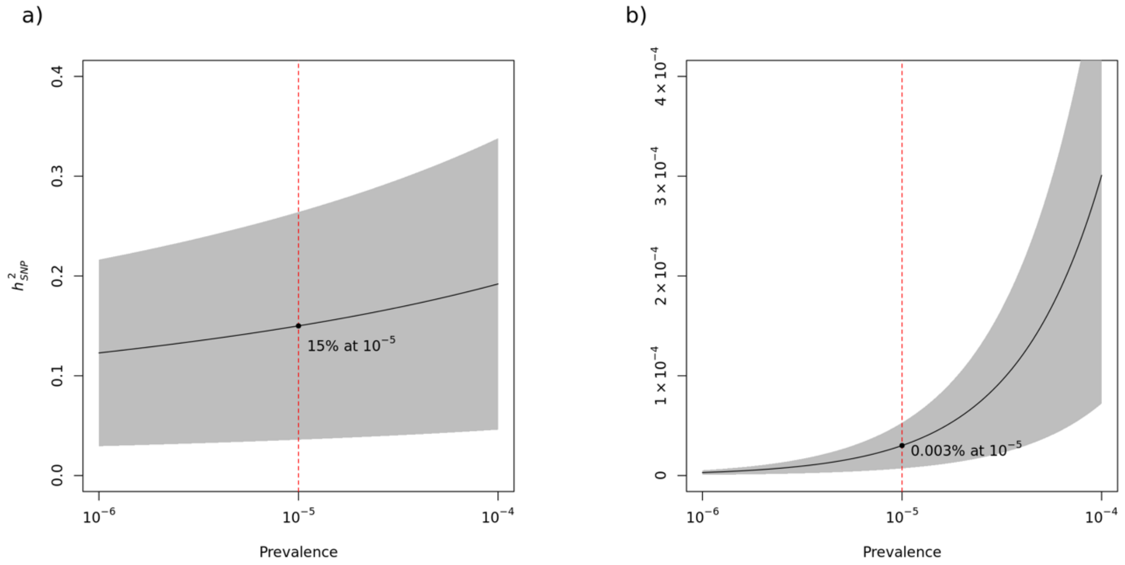
MDS plot : PC1 and PC3

MDS plot : PC2 and PC3

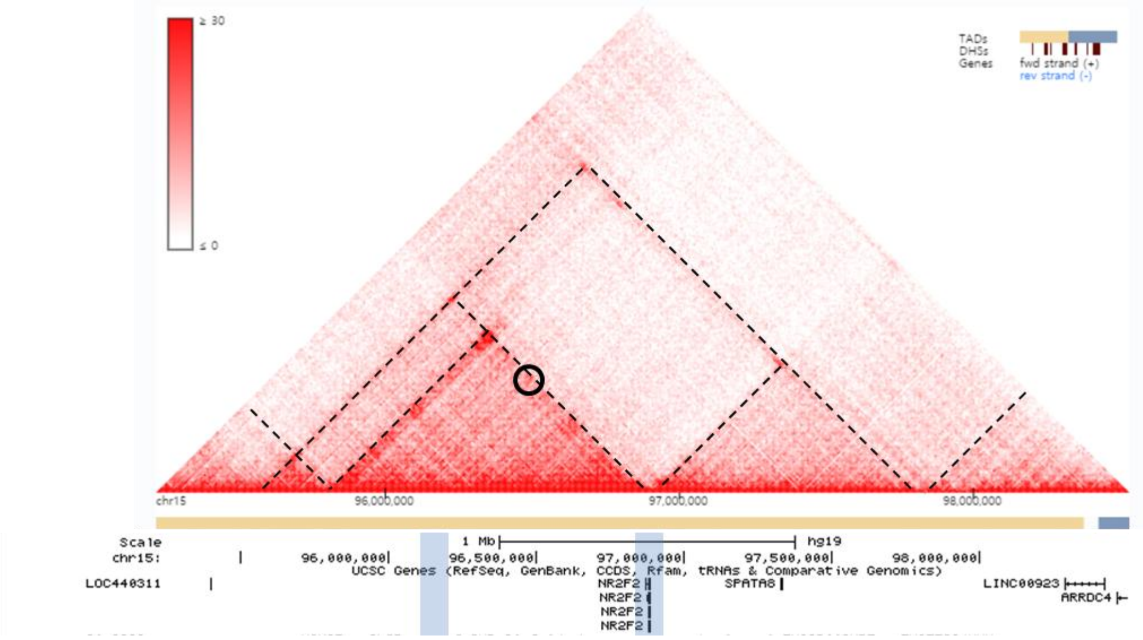
**Supplementary Figure 3. Linkage disequilibrium (LD) block around genome wide significant and genotyped SNPs, rs4544201 and rs2006950.** Graph represents all genotyped SNPs in the 34kb LD block on chromosome 15q26.2. The color of each rectangle and number within indicates the level of LD between a pair of SNPs, with complete LD (D'=100%, no number shown) indicated by red, and no LD indicated by white.



**Supplementary Figure 4. The proportion of phenotypic variance explained by the genotyed SNPs according to disease prevalences ranging from 10$^{-6}$ to 10$^{-4}$.** The proportion of phenotypic variance explained by genotyped SNPs was calculated with GCTA on a) the liability scale and b) the observed 0-1 scale. Shaded area indicates the 95% confidence interval for $h^2_{SNP}$.

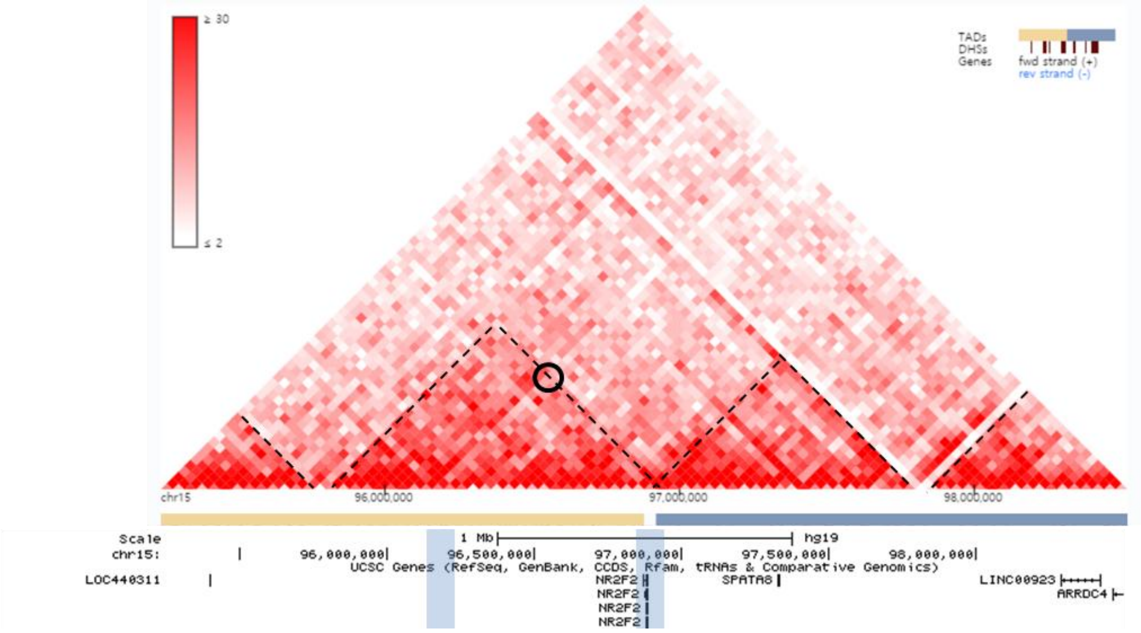


**Supplementary Figure 5. Hi-C heatmap and TADs defined in IMR90 cells.** The heatmap shows the degree of physical interaction defined by Hi-C analysis for genomic region pairs from a 3Mb region of chromosome 15q. A deeper red color at the intersection point reflects a greater degree of interaction between the two genomic regions. The dotted lines indicate probable TAD structures in this region. The two blue shaded regions at bottom indicate the genome wide significant SNP region (left) and *NR2F2* (right). The black circle reflects the interaction point between the SNP region and *NR2F2.*



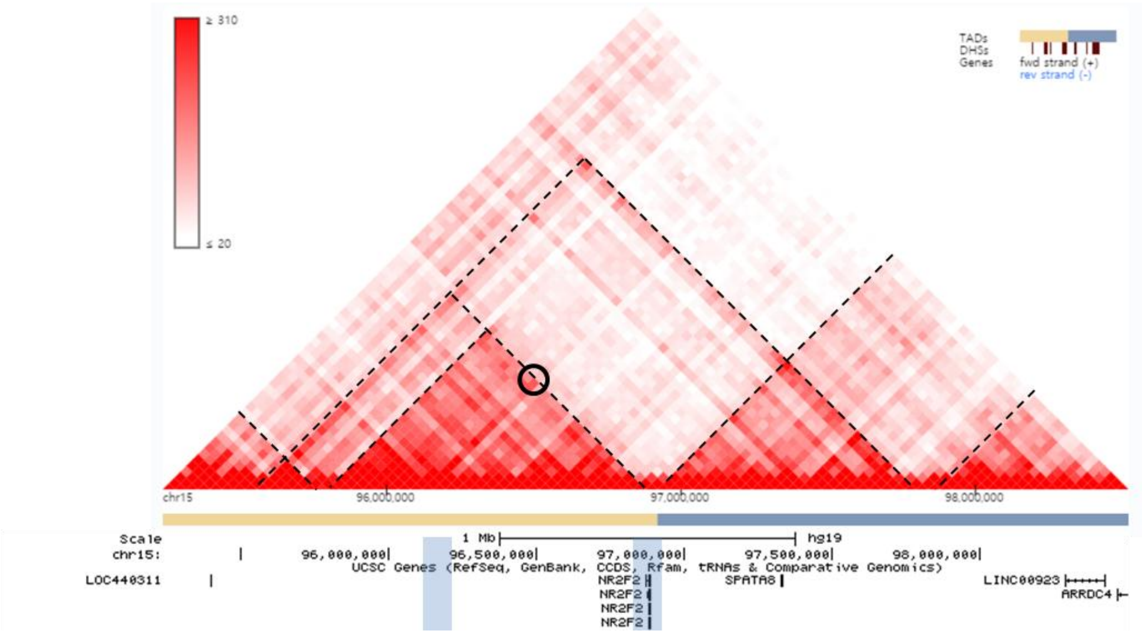
**Supplementary Figure 6. Hi-C heatmap and TADs defined in lung tissue.**

The heatmap shows the degree of physical interaction defined by Hi-C analysis for genomic region pairs from a 3Mb region of chromosome 15q. A deeper red color at the intersection point reflects a greater degree of interaction between the two genomic regions. The dotted lines indicate probable TAD structures in this region. The two blue shaded regions at bottom indicate the genome wide significant SNP region (left) and *NR2F2* (right). The black circle reflects the interaction point between the SNP region and *NR2F2.*



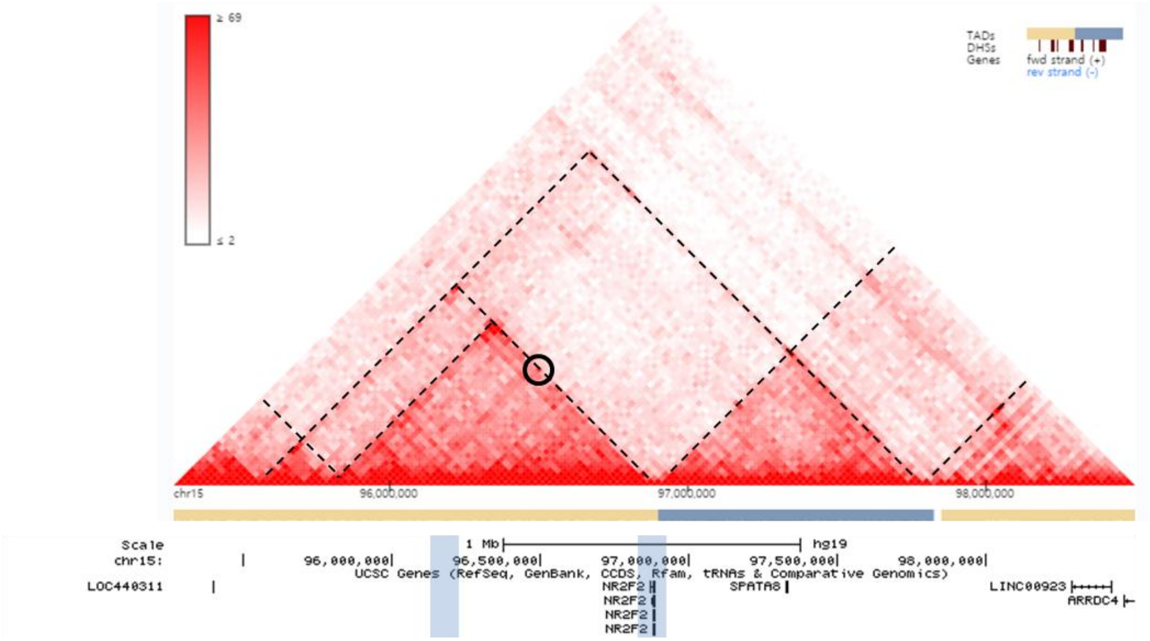
**Supplementary Figure 7. Hi-C heatmap and TADs defined in H1 derived mesenchymal stem cells (h1-MSC) cells.**

The heatmap shows the degree of physical interaction defined by Hi-C analysis for genomic region pairs from a 3Mb region of chromosome 15q. A deeper red color at the intersection point reflects a greater degree of interaction between the two genomic regions. The dotted lines indicate probable TAD structures in this region. The two blue shaded regions at bottom indicate the genome wide significant SNP region (left) and *NR2F2* (right). The black circle reflects the interaction point between the SNP region and *NR2F2.*



**Supplementary Figure 8. Hi-C heatmap and TADs defined in HUVEC cells.**

The heatmap shows the degree of physical interaction defined by Hi-C analysis for genomic region pairs from a 3Mb region of chromosome 15q. A deeper red color at the intersection point reflects a greater degree of interaction between the two genomic regions. The dotted lines indicate probable TAD structures in this region. The two blue shaded regions at bottom indicate the genome wide significant SNP region (left) and *NR2F2* (right). The black circle reflects the interaction point between the SNP region and *NR2F2.*


≥ 69
≤ 2
TADs
DHSs
Genes
fwd strand (+)
rev strand (-)
chr15
Scale
1 Mb
hg19
chr15:
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
LOC440311
NR2F2
NR2F2
NR2F2
NR2F2
SPATA8
LINC00923
ARRDC4