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

Wonji Kim1,2\*, Krinio Giannikou3\*, Sungho Won1,4,5\*, John R. Dreier3, Sanghun Lee6, Magdalena E. Tyburczy3, Edwin K. Silverman2,3, Elżbieta Radzikowska7, Shulin Wu8, Chin-Lee Wu8 , Elizabeth P. Henske3, Gary Hunninghake3, Havi Carel9, Antonio Roman10, Miquel Angel Pujana11, Joel Moss12, David J. Kwiatkowski3

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

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

3Division 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

4Department of Public Health Sciences, Seoul National University, Seoul, Korea

5Institute of Health and Environment, Seoul National University, Seoul, Korea

6Department of Medical Consilience, Graduate School, Dankook University, Seoul, Korea

7National Tuberculosis and Lung Diseases Research Institute, Warsaw, Poland

8Urology Research Laboratory, Massachusetts General Hospital, Boston, MA, 02114, USA

9Department of Philosophy, University of Bristol, UK

10Vall d'Hebron University Hospital, CIBERES, Barcelona, Spain

11ProCURE, Catalan Institute of Oncology, Oncobell, Bellvitge Institute of Biomedical Research (IDIBELL), Barcelona, Spain

12Cardiovascular and Pulmonary Branch, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, MD

**Corresponding Author:**

David J. Kwiatkowski

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

[dk@rics.bwh.harvard.edu](mailto:dk@rics.bwh.harvard.edu)

phone: 8573070781 fax: 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

Short running head: LAM GWAS implicates *NR2F2*

**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:** Saliva DNA samples from 479 S-LAM subjects were genotyped on the Infinium OmniExpress-24 v1.2 BeadChip, and compared with genotype data from 1,261 COPDGene controls using conditional logistic regression. A replication study, RNA-seq and immunohistochemistry analyses were also performed.

**Results:** Two non-coding SNPs met genome-wide significance; rs4544201 and rs2006950 (P-value=9.2×10 -10, 4.1×10 -10, 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 FEV1 and DLCO. 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 or/and somatic 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.

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 < 1×10-5; 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. 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 < 1×10-5 was removed. Last, EIGENSTRAT [17] was 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 were then removed. Each case was matched with two controls, and 409 controls were removed. These filters led to retention of 426 S-LAM cases and 852 female controls for analysis in the discovery phase with 549,591 SNP genotypes (Figure 1).

**Replication data**

Replication analysis was done on an additional independent set of 196 non-Hispanic white (NHW) female S-LAM subjects, for the two significant SNPs identified in the discovery study, provided by one co-author (JM, Supplementary Table 1). Careful scrutiny was performed by a third independent party to ensure that there was no overlap between the primary analysis population and the replication population. 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 NHW females in the Multi-Ethnic Study of Atherosclerosis (MESA) dataset obtained from dbGaP (phs000209.v13.p3) [18] and 3) 255,247 NHW females from UK Biobank dataset . For each control dataset, we used genotype data when rs4544201 and rs2006950 were available or imputed data with 1000 Genomes Project [19] when they were not. We included only NHW female and excluded variants if P-value of HWE test is less than 1×10-5, MAF is less than 0.05 or missing genotype rate is greater than 5%. With these quality control (QC) criteria, no variants were excluded for all datasets.

**Statistical analyses with genetic data**

GWAS analyses with discovery data were conducted using conditional logistic regression (CLR).

Principal Components (PC) Analysis scores were estimated with EIGENSTRAT [17], and used to adjust population substructure. CLR requires matching of cases and controls, and matching quality is affected by the number of PC scores matching. Each case was matched with two controls using the *Matching* R package [20]. Supplementary Figure 1 shows that matching with two PC scores corresponding to the two greatest eigenvalues provide the genomic inflation factor closest to 1. Thus CLR was conducted conditioning on the SNP and the number of cases in each strata consisting of cases and controls matched with two PC scores. In this study where each case was matched two controls, this can be expressed as follows:

for th strata,

where is the phenotype and is the genotype. CLR analyses were performed with the R package *survival* [21]{Team, 2016 #31;Therneau, 2017 #94} and genome-wide significance was assessed by P-value < 5×10-8.

We also conducted gene-based analyses to identify genes with significant association with S-LAM using the SKAT-O statistic [22]. All SNPs after QC except MAF criteria within each gene were included for SKAT-O analysis, and age, squared age and 10 PC scores were included as covariates.

**Genotype imputation and statistical analyses with imputed genotypes**

We performed genome-wide imputation for autosomes to find potential novel loci. 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 [23]. Before imputation, pre-phasing was conducted with EAGLE2 v2.0.5 [24], and then the Positional Burrows–Wheeler Transform (PBWT) package [25] was used for the imputation according to the imputation pipeline recommended by Sanger Imputation Service in terms of speed and accuracy. Imputation accuracy was evaluated with the INFO metric [26]. Imputed SNPs were filtered out if INFOs, MAFs or P-values for the HWE test were < 0.3, 0.05, or 1×10-5, respectively. We applied CLR to 5,427,337 SNPs which satistied standards of QC. Furthermore, we applied PICS software to significant imputed and genotyped SNPs containing the genome-wide significant SNPs to calculate the probability of each individual SNP being the causal SNP [27].

**Topologically associated domains (TADs) and chromatin interactions**

To identify chromatin interactions in the region of interest on chromosome 15q26.2, we used a 3D genome browser ([www.3dgenome.org](http://www.3dgenome.org)) to predict TADs[28]. 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 [29]. Fastq files were also converted into RSEM format [30]. RSEM values were compared to RNA-seq data from 2,463 tumors of 27 different histologic types from the TCGA [31]. 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) [32].

We also searched the single-tissue *cis*-eQTL of all SNPs in the LD block from public database provided by GTEx in release v7. They provide results of eQTL analysis for each SNP-gene pair consisting of all SNPs within 1 Mb upstream and downstream of transcription start site of each gene. FastQTL was used for *cis-*eQTL mapping [33] including top three PC scores, genotyping platform, sex and a set of relavant variables identified using PEER method [34] as covariates. The detailed workflow can be found on the GTEx website (https://www.gtexportal.org/home).

**Immunohistochemistry analyses**

Immunochistochemistry was performed as described elsewhere [35] 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% H2O2, 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 549,591 SNPs using CLR. Two non-coding SNPs rs4544201 and rs2006950 on chromosome 15 met genome-wide significance (rs4544201: P-value=8.5110-10; rs2006950: P-value=3.9210-10).

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 two SNPs (Figure 2ab), indicating that the analyses were free of systematic P-value inflation (genomic inflation factor = 1.07). Scatter plots of PC scores indicated netic similarity between cases and controls in the discovery analyses (Supplementary Figure 2). Since the control COPDGene cohort were smokers, this association analysis might have selection bias caused by SNP alleles associated with nicotine addiction. We checked p-values for SNPs associated with nicotine addiction from the GWAS catalog [36] and other SNPs correlated with those (r2 >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. Table 1 provides summaries for the two genome-wide significant SNPs.

Linkage disequilibrium (LD) blocks near genome-wide significant SNPs were chosen by using Haploview with default options [37]. rs4544201 and rs2006950 belong to the same LD block on 15q26.2, 11,563nt apart, and are strongly correlated (*D*’=0.977, *r*2=0.854; Supplementary Figure 3). Based on the location of the two SNPs and their LD relationship, these two associations might be derived from a unique signal. They are located in 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 1). To adjust for the possible effect of the ‘Winner’s curse’, we used br2 [38], 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, . Estimates of vary according to disease prevalence (Supplementary Figure 4). With prevalence set at 1 in 100,000 women, is very low at 0.76%, and the proportion of phenotypic variance explained by the LD block containing genome-wide significant SNPs is 3.04% suggesting that other genetic or environmental components contribute to disease susceptibility.

To examine the potential association of other SNPs in this region with S-LAM, we used the genotyped SNP data to impute genotype data. Eighteen imputed SNPs in the 34kb LD block had P-values for association with LAM similar to rs4544201 and rs2006950 (Table 2, Supplementary Figure 5).

Replication analysis was performed for the 2 SNPs with association with LAM using 196 additional non-Hispanic white (NHW) S-LAM patients and 409 NHW healthy females from COPDGene participants who were not used for discovery analyses. Similar ORs for association of the minor allele of these SNPs with S-LAM were seen in the replication data (Table 1, ORrs4544201=0.33, ORrs2006950 = 0.28). 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 the UK Biobank study of 337,199 individuals. The MAFs of the 2 SNPs in LAM patients were significantly smaller than those reported in every other cohort (Supplemental Table 3).

To attempt to identify the causal SNP(s) among these SNPs with low P-values, we performed PICS analysis for all SNPs in Table 2, and the original two SNPs showing association. rs41374846 had both significant association with LAM, and the largest PICS probability (PPICS=0.65, Supplementary Table 4), making it the candidate causal SNP in this association [27].

We queried the GTEx database for SNPs in the LD block to find SNP-gene associations in eQTL studies. Unfortunately, there was no significant association with an FDR threshold of 0.05.

**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 element function 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]. In Supplementary Figures 6-9, Hi-C heatmaps display about 1 Mb sub-TADs containing the GWAS SNPs and *NR2F2* near the boundary of the sub-TAD within a larger TAD in all four of these cells/tissues, and *NR2F2* was the only protein-coding gene. 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 all 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 [31]), and normal tissues (from GTEX [32]) (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 = 6.38×10-6, 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. Two intergenic SNPs, rs4544201 and rs2006950, were identified in a 34kb LD block on chromosome 15, that met genome-wide significance for association with LAM (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 [32], 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 4a). 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 overcome this issue, we used CLR rather than the classical logistic regression (LR). According to the some literatures, CLR can improve the efficiency of the study rather than LR for the same sample size if the variables used for matching are true confounding variables and moderate number of controls were excluded from the analysis as a result of matching [56-61]. In this study, we considered sex and population substructure as the potential confounders because S-LAM occurs almost exclusively for women and S-LAM patients were collected from several countries. To control confounders, we only included women for both discovery and replication studies, and included two PC scores as matching variables. Even though we dropped some controls out after matching, CLR was more efficient than LR to identify significant SNPs on chromosome 15 (Supplementary Figure 10). Second, to collect sufficient LAM subjects, 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 identify genetic outliers from both our S-LAM and control cohorts to further reduce genetic heterogeneity. Further functional analyses to confirm our hypothesis that *NR2F2* is the gene affected by this SNP is 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, and The Engles Fund for TSC and LAM research.

**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. 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.

18. 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.

19. Consortium GP. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012: 491(7422): 56.

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. 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.

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

24. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, Schoenherr S, Forer L, McCarthy S, Abecasis GR. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics* 2016: 48(11): 1443.

25. Durbin R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* 2014: 30(9): 1266-1272.

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

27. 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.

28. 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.

29. 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.

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

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

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

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

38. 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.

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.

56. Breslow N, Day N, Halvorsen K, Prentice R, Sabai C. Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology* 1978: 108(4): 299-307.

57. Kupper LL, Karon JM, Kleinbaum DG, Morgenstern H, Lewis DK. Matching in epidemiologic studies: validity and efficiency considerations. *Biometrics* 1981: 271-291.

58. McKinlay SM. Pair-matching--A reappraisal of a popular technique. *Biometrics* 1977: 725-735.

59. THOMPSON WD, KELSEY JL, WALTER SD. Cost and efficiency in the choice of matched and unmatched case-control study designs. *American journal of epidemiology* 1982: 116(5): 840-851.

60. Thomas DC, Greenland S. The relative efficiencies of matched and independent sample designs for case-control studies. *Journal of chronic diseases* 1983: 36(10): 685-697.

61. Miettinen OS. Estimation of relative risk from individually matched series. *Biometrics* 1970: 75-86.

**Table 1. Genome-wide significant 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 |
| ***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 |
| ***Discovery data*** |  |  |
| Odds ratio |  |  |
| Original | 0.4916 | 0.4732 |
| Bias adjusted | 0.5677 | 0.5315 |
| P-value | 8.5110-10 | 3.9210-10 |
| ***Replication data*** | | |
| COPDGene |  |  |
| Odds ratio | 0.3288 | 0.2731 |
| P-value | 4.32×10-5 | 1.5610-5 |
| MESA |  |  |
| Odds ratio | 0.5070 | 0.4448 |
| P-value | 9.28×10-6 | 1.04×10-6 |
| UK BioBank |  |  |
| Odds ratio | 0.4900 | 0.4229 |
| P-value | 7.77×10-7 | 5.46×10-8 |

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

**Table 2. Statistical analyses of imputed SNPs with CLR.** Imputation was conducted using EAGEL2 (pre-phasing) and PBWT (Imputation) with the Haplotype Reference Consortium as reference panel.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CHR | SNP | POS | Alleles\* | MAF | INFO† | P-value for CLR‡ |
| 15 | rs41374846 | 96143559 | A/G | 0.2605 | 0.9097 | 3.43210-9 |
| 15 | rs59125351 | 96144157 | G/T | 0.2510 | 0.9771 | 3.22910-10 |
| 15 | rs17581137 | 96146414 | C/A | 0.2336 | 0.9893 | 1.38410-10 |
| 15 | rs6496126 | 96148439 | C/G | 0.2330 | 0.9890 | 1.81410-10 |
| 15 | rs2397810 | 96148765 | C/T | 0.2330 | 0.9890 | 1.81410-10 |
| 15 | rs10520790 | 96151040 | T/G | 0.2478 | 0.9958 | 3.57110-10 |
| 15 | rs55804812 | 96151256 | A/T | 0.2475 | 0.9952 | 4.17810-10 |
| 15 | rs16975389 | 96153782 | C/T | 0.2463 | 0.9967 | 5.80110-10 |
| 15 | rs16975396 | 96158705 | G/T | 0.2466 | 0.9983 | 9.59210-10 |
| 15 | rs4628911 | 96167905 | T/C | 0.2472 | 1.0000 | 5.14710-10 |
| 15 | rs6496128 | 96168303 | G/A | 0.2472 | 1.0000 | 5.14710-10 |
| 15 | rs8029996 | 96168770 | A/G | 0.2472 | 0.9998 | 5.14710-10 |
| 15 | rs4551988 | 96169589 | C/G | 0.2472 | 0.9998 | 5.14710-10 |
| 15 | rs58878263 | 96171069 | A/C | 0.2493 | 0.9979 | 6.36110-10 |
| 15 | rs8040665 | 96175692 | G/T | 0.2487 | 0.9976 | 7.35610-10 |
| 15 | 15:96175733 | 96175733 | A/G | 0.2466 | 0.9975 | 5.22410-10 |
| 15 | rs8040168 | 96176096 | G/C | 0.2466 | 0.9981 | 5.22410-10 |
| 15 | rs17504029 | 96177670 | T/A | 0.2478 | 0.9876 | 1.90010-10 |

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 the metric about imputation quality provided by IMPUTE2.

‡ CLR was applied to imputed SNP genotype data to identify SNPs with significant association (P < 5×10-8) with S-LAM.

**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 2 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 549,591 genotyped SNPs 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 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 containing the genotyped and imputed SNPs showing association with LAM. 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 5.

**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 tumor samples and 4 angiomyolipoma samples examined by IHC.