A genome-wide association study implicates *NR2F2* in lymphangioleiomyomatosis pathogenesis

Wonji Kim[1,2,13], Krinio Giannikou[3,13], 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[11,12,14] and David J. Kwiatkowski[3,14]

@ERSpublications
GWAS identified alleles of two SNPs near *NR2F2* that were associated with sporadic lymphangioleiomyomatosis. NR2F2 is a transcription factor that is expressed highly in both LAM and a LAM-related tumour, and *NR2F2* is a new LAM gene. http://ow.ly/87xx30oiVhZ

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; 53: 1900329 [https://doi.org/10.1183/13993003.00329-2019].

ABSTRACT

Introduction: Lymphangioleiomyomatosis (LAM) occurs either associated with tuberous sclerosis complex (TSC) or as sporadic disease (S-LAM). Risk factors for development of S-LAM are unknown. We hypothesised 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 single nucleotide polymorphisms (SNPs) in 426 S-LAM subjects were compared, using conditional logistic regression, with similar data from 852 females from COPDGene in a matched case–control design. For replication studies, genotypes for 196 non-Hispanic White female S-LAM subjects were compared with three different sets of controls. RNA sequencing and immunohistochemistry analyses were also performed.

Results: Two noncoding genotyped SNPs met genome-wide significance: rs4544201 and rs2006950 ($p=4.2\times10^{-8}$ and 6.1×10^{-9}, respectively), which are in the same 35 kb linkage disequilibrium block on chromosome 15q26.2. This association was replicated in an independent cohort. *NR2F2* (nuclear receptor subfamily 2 group F member 2), a nuclear receptor and transcription factor, was the only nearby protein-coding gene. *NR2F2* expression was higher by RNA sequencing in one abdominal LAM tumour and four kidney angiomyolipomas, a LAM-related tumour, compared with all cancers from The Cancer Genome Atlas. Immunohistochemistry showed strong nuclear expression in both LAM and angiomyolipoma tumours.

Conclusions: SNPs on chromosome 15q26.2 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.

This article has supplementary material available from erj.ersjournals.com

Received: Nov 27 2018 | Accepted after revision: March 19 2019

Data and code availability: The primary GWAS and replication data will be made available on publication of this work through dbGaP (www.ncbi.nlm.nih.gov/gap).

Introduction

Lymphangioleiomyomatosis (LAM) is a rare aggressive low-grade neoplasm which affects almost exclusively females 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 characterised 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 forced expiratory volume in 1 s and diffusing capacity of the lung for carbon monoxide. 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 (VEGF)-D levels.

LAM occurs at high frequency (>10%) in females with tuberous sclerosis complex (TSC) and at much lower frequency in females (∼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]. Tumour 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 tumours [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% of S-LAM subjects have kidney angiomyolipoma, a tumour which is seen in 70–80% of adults with TSC. Angiomyolipoma shares histological, expression and genetic features with LAM, although the lesions are not pathologically identical.

Genome-wide association studies (GWASs) are utilised 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 hypothesised that DNA sequence variants outside of *TSC2*/*TSC1* might be associated with disease risk and go unrecognised 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 during 2010–2014 from 14 countries (supplementary table S1). S-LAM patients were diagnosed using standard diagnostic criteria by their treating physicians [1–5, 7]. Genomic DNA was extracted from saliva using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) and 479 S-LAM DNA samples were genotyped with the Infinium OmniExpress-24 version 1.2 BeadChip (Illumina, San Diego, CA, USA), which assesses 716 503 single nucleotide polymorphisms (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 (www.ncbi.nlm.nih.gov/gap; phs000951.v2.p2.c1). These COPDGene participants had smoked at least 10 pack-years and were 45–80 years old, and were without known chronic obstructive pulmonary disease [11].

Quality control analyses of SNP genotype data

We evaluated the quality of SNPs and subjects in the discovery dataset using PLINK [12] and ONETOOL [13]. We excluded all SNPs for which the Hardy–Weinberg equilibrium test [14] gave $p<1\times10^{-5}$, minor allele frequency (MAF) was <0.05 or genotype call rates were <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 both cases and controls, and were repeated on

Affiliations: [1]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, Korea. [2]Channing Division of Network Medicine, Dept of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [3]Division of Pulmonary and Critical Care Medicine and of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [4]Dept of Medical Consilience, Graduate School, Dankook University, Yongin-si, Korea. [5]National Tuberculosis and Lung Diseases Research Institute, Warsaw, Poland. [6]Urology Research Laboratory, Massachusetts General Hospital, Boston, MA, USA. [7]Dept of Philosophy, University of Bristol, 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]Pulmonary Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [11]Dept of Public Health Sciences, Seoul National University, Seoul, Korea. [12]Institute of Health and Environment, Seoul National University, Seoul, Korea. [13]These two authors contributed equally to this work. [14]Joint senior authors.

Correspondence: David J. Kwiatkowski, Division of Pulmonary Medicine, Brigham and Women's Hospital, 20 Shattuck Street, Boston, MA 02115, USA. E-mail: dk@rics.bwh.harvard.edu

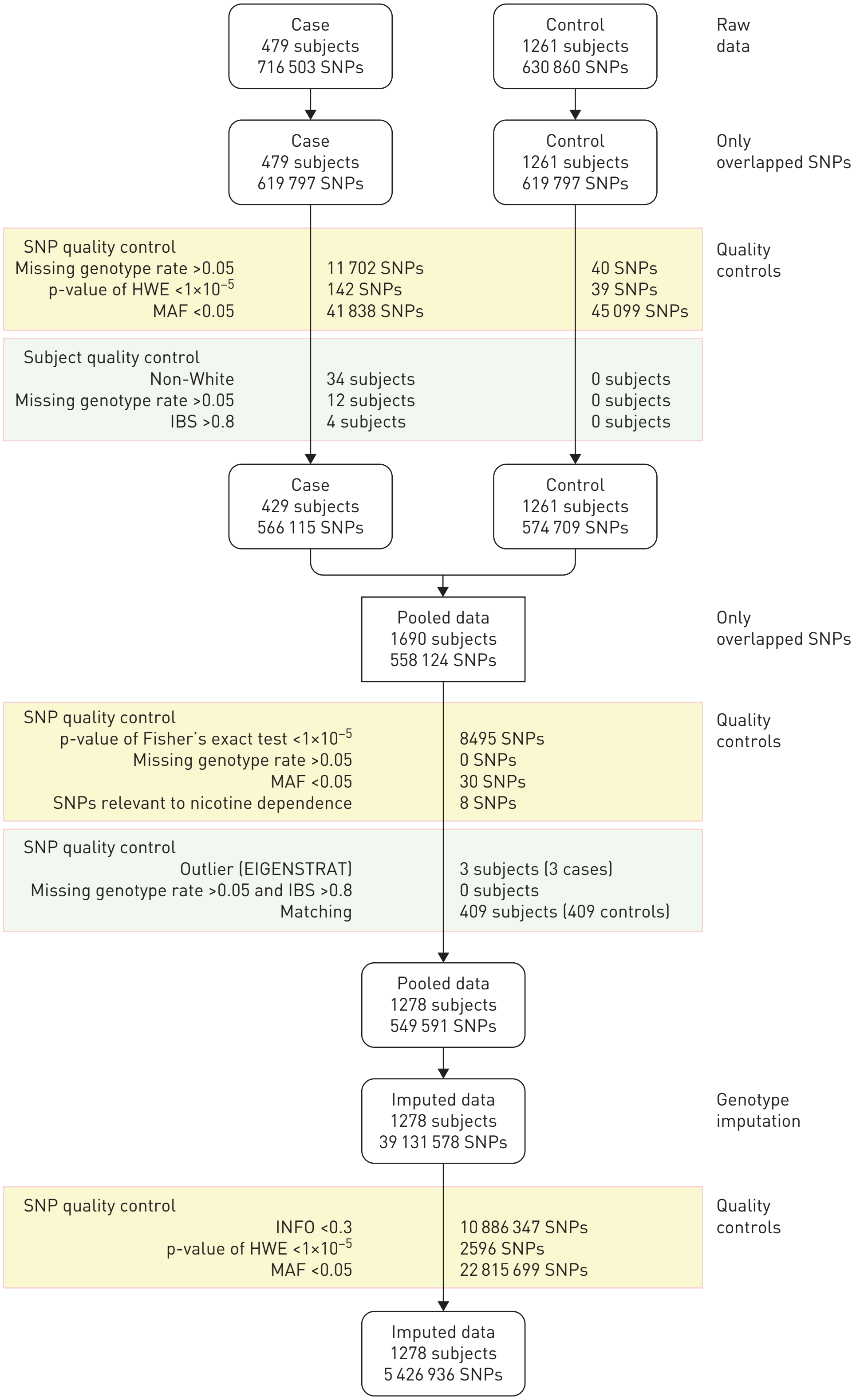

FIGURE 1 Workflow of statistical analysis and quality control for the lymphangioleiomyomatosis (LAM) genome-wide association study discovery dataset. SNP: single nucleotide polymorphism; HWE: Hardy–Weinberg equilibrium; MAF: minor allele frequency; IBS: identity-by-state. Multiple standard quality controls were performed for both cases (sporadic (S)-LAM subjects) and controls (healthy females without chronic obstructive pulmonary disease from the COPDGene consortium) to exclude outlier SNPs and subjects.

the pooled dataset. In addition, any SNP showing a difference in missing data rate between cases and controls by Fisher's exact test [15] 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 version 1.1 for the reference panel and considered predominantly European ancestry [16]. Pre-phasing was performed first with Eagle2 version 2.0.5 [17] and then the PBWT (Positional Burrows–Wheeler Transform) package [18] used for imputation according to the imputation pipeline recommended by the Sanger Imputation Service. Imputation accuracy was evaluated with the INFO metric [19]. Imputed SNPs were filtered out if INFOs, MAFs or p-values for the Hardy–Weinberg equilibrium test were <0.3, <0.05 or $<1\times10^{-5}$, respectively.

Statistical analyses with genetic data

EIGENSTRAT [20] 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 (three 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 R package Matching [21]. 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 S1). Thus, CLR was conducted by conditioning on the matched cases and controls with the first two PC scores. Our CLR can be expressed as follows.

For the *i*th strata:

$$
\begin{aligned}
&\Pr\left(Y_{i1}=1,\ Y_{i2}=0,\ Y_{i3}=0|\mathbf{X}_{i1},\mathbf{X}_{i2},\mathbf{X}_{i3},\ \sum_{j=1}^{3}Y_{ij}=1\right)\\
&=\frac{\Pr(Y_{i1}=1|\mathbf{X}_{i1})\Pr(Y_{i2}=0|\mathbf{X}_{i2})\Pr(Y_{i3}=0|\mathbf{X}_{i3})}{\sum_{y_{ij}\in(y_{i1}+y_{i2}+y_{i3}=1)}\Pr(Y_{i1}=y_{i1}|\mathbf{X}_{i1})\Pr(Y_{i2}=y_{i2}|\mathbf{X}_{i2})\Pr(Y_{i3}=y_{i3}|\mathbf{X}_{i3})}\\
&=\frac{\exp(\mathbf{X}_{i1}\beta)}{\sum_{j=1}^{3}\exp(\mathbf{X}_{ij}\beta)}
\end{aligned}
$$

where Y_{ij} and $\mathbf{X}_{ij}$ indicate the phenotype and covariates including SNPs of the *j*th subject in the *i*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 [22] and genome-wide significance was assessed by $p<5\times10^{-8}$.

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 [23].

We also conducted gene-based analyses for association with LAM for those genes near the genome-wide significant SNPs using the SKAT-O statistic [24]. We included all genotyped SNPs in this analysis with no MAF cut-off for inclusion. Age, age squared and 10 PC scores were included as covariates.

Replication data

Replication analysis was done on an independent set of 196 non-Hispanic White female S-LAM subjects, seen at the NIH Clinical Centre (Bethesda, MD, USA) by one co-author (J.M.) (supplementary table S1). 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 analysed 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 (Thermo Fisher Scientific, Waltham, MA, USA). 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 nine subjects were not included in the replication analyses. We used three independent datasets as controls for comparison in the replication study: 1) 409 non-Hispanic White healthy females from the COPDGene consortium who were not used for discovery analyses, 2) 1121 Hispanic White females in the Multi-Ethnic Study of Atherosclerosis (MESA) dataset obtained from dbGaP (phs000209.v13.p3) [25] and 3) 225 731 British

White females from the UK Biobank dataset [26]. For each control dataset, we used genotyped or imputed data for the genome-wide significant SNPs.

Topologically associated domains and chromatin interactions

To identify chromatin interactions in the region of interest on chromosome 15q26.2, we used two three-dimensional genome browsers (www.3dgenome.org and https://yunliweb.its.unc.edu/hugin) to predict topologically associated domains (TADs) [27, 28]. We checked for TADs around the genome-wide significant SNPs and protein-coding genes belonging to each TAD were investigated. We analysed TADs from four cell lines/tissues judged closest to LAM: 1) human fetal lung fibroblasts (IMR90), 2) lung-related tissues, 3) H1-derived mesenchymal stem cells (H1-MSCs) and 4) human umbilical vein endothelial cells (HUVECs).

Statistical analyses with RNA sequencing data

Whole transcriptome RNA sequencing analysis was performed on one abdominal LAM tumour and four kidney angiomyolipomas at the Broad Institute of Harvard and MIT (Boston, MA, USA). Briefly, mRNA sequencing was performed using poly(A) cDNA capture followed by cDNA library synthesis (Truseq RNA Library Prep Kit; Illumina) and sequencing on Illumina machines, following the same methods and in the same facility in which the Gene and Tissue Expression (GTEx) RNA sequencing project occurred [29]. Read data were processed into FASTQ files with standard quality control methods and aligned to the genome (Genome Reference Consortium GRCh37 (hg19)) using TopHat version 2.0.10 [30]. FASTQ files were also converted into RSEM (RNA sequencing by expectation maximisation) format [31]. RSEM values were compared with RNA sequencing data from 2463 tumours of 27 different histological types from The Cancer Genome Atlas (TCGA) [32]. RPKM (reads per kilobase per million mapped reads) values for *NR2F2* (nuclear receptor subfamily 2 group F member 2) were compared with the GTEx dataset of normal human tissues with the limma statistic (11 688 RNA sequencing samples from 53 normal tissue types, version 7 release) [29].

We also searched for any *cis*-expression quantitative trait loci (eQTL) for all SNPs in the linkage disequilibrium block with association to LAM using the GTEx release version 7 database [29]. 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 (www.gtexportal.org/home) for *cis*-eQTL mapping [33] with covariate adjustment of the top three PC scores, genotyping platform, sex and a set of relevant variables identified using the PEER method [34].

Immunohistochemistry analyses

Immunohistochemistry was performed as described elsewhere [35] using a primary mouse monoclonal antibody against *NR2F2* (ab41859; Abcam, Cambridge, MA, USA; concentration 1:100 (10 $\mu g \cdot mL^{-1}$)). Briefly, formalin-fixed, paraffin-embedded tumour sections were deparaffinised in xylene, rehydrated and antigen retrieval was performed in EDTA (pH 8.0) (Diagnostic BioSystems, Pleasanton, CA, USA). Endogenous peroxidase activity was blocked with 3% hydrogen peroxide; blocking was done with 5% goat serum, followed by incubation overnight with antibody at 4°C, washing in TBS/Tween 20 and incubation with anti-goat secondary antibody (Vector, Burlingame, CA, USA; dilution 1:300). The peroxidase reaction was developed using 3,3′-diaminobenzidine substrate (DakoCytomation, Glostrup, Denmark). 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. 20 noncoding SNPs on chromosome 15 met genome-wide significance, of which two had been directly genotyped (rs4544201: $p=4.19\times10^{-8}$; rs2006950: $p=6.12\times10^{-9}$) (table 1).

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 S2). 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 catalogue [36] and other SNPs correlated with those ($r^2>0.8$) (supplementary table S2). 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 Statistical analyses of imputed single nucleotide polymorphisms (SNPs) with conditional logistic regression (CLR)

Chromosome	SNP	Position[#]	Alleles[¶]	MAF	Imputed *versus* genotyped	INFO[+]	p-value for CLR[§]
15	rs41374846	96 143 559	A/G	0.2605	Imputed	0.9097	1.322×10^{-7}
15	rs59125351	96 144 157	G/T	0.2510	Imputed	0.9771	2.741×10^{-9}
15	rs17581137	96 146 414	C/A	0.2336	Imputed	0.9893	1.250×10^{-10}
15	rs6496126	96 148 439	C/G	0.2330	Imputed	0.9890	6.982×10^{-9}
15	rs2397810	96 148 765	C/T	0.2330	Imputed	0.9890	6.691×10^{-9}
15	rs10520790	96 151 040	T/G	0.2478	Imputed	0.9958	6.691×10^{-9}
15	rs55804812	96 151 256	A/T	0.2475	Imputed	0.9952	4.008×10^{-8}
15	rs16975389	96 153 782	C/T	0.2463	Imputed	0.9967	1.173×10^{-8}
15	rs16975396	96 158 705	G/T	0.2466	Imputed	0.9983	3.547×10^{-8}
15	rs4544201	96 167 827	A/G	0.2469	Genotyped	1.0000	4.186×10^{-8}
15	rs4628911	96 167 905	T/C	0.2472	Imputed	1.0000	3.547×10^{-8}
15	rs6496128	96 168 303	G/A	0.2472	Imputed	1.0000	3.547×10^{-8}
15	rs8029996	96 168 770	A/G	0.2472	Imputed	0.9998	3.547×10^{-8}
15	rs4551988	96 169 589	C/G	0.2472	Imputed	0.9998	3.547×10^{-8}
15	rs58878263	96 171 069	A/C	0.2493	Imputed	0.9979	3.632×10^{-8}
15	rs8040665	96 175 692	G/T	0.2487	Imputed	0.9976	2.375×10^{-8}
15	15:96 175 733	96 175 733	A/G	0.2466	Imputed	0.9975	2.227×10^{-8}
15	rs8040168	96 176 096	G/C	0.2466	Imputed	0.9981	2.227×10^{-8}
15	rs17504029	96 177 670	T/A	0.2478	Imputed	0.9876	2.289×10^{-8}
15	rs2006950	96 179 390	A/G	0.2262	Genotyped	1.0000	6.117×10^{-9}

Imputation was conducted using Eagle2 and PBWT for pre-phasing. Imputation was conducted by using the Haplotype Reference Consortium as reference panel. MAF: minor allele frequency. [#]: SNP position according to Genome Reference Consortium GRCh37 (hg19); [¶]: 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 sporadic lymphangioleiomyomatosis.

Linkage disequilibrium blocks near genome-wide significant SNPs were identified using Haploview with default options [37]. All 20 SNPs, including the two directly genotyped, rs4544201 and rs2006950, belong to the same linkage disequilibrium block on 15q26.2; the latter two SNPs were 11 563 nucleotides apart and were strongly correlated (D'=0.977, r^2=0.854) (supplementary figure S3). Based on the proximity of the two SNPs to each other and their linkage disequilibrium 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* (multiple C2 and transmembrane domain containing 2; 1.1 Mb away) and *NR2F2* (700 kb away) that contains many long noncoding RNAs (figure 3). Both SNPs have minor and major alleles of A and G, and showed a lower MAF in the S-LAM cohort than the control population. The ORs of a single minor allele in the S-LAM cohort were 0.49 and 0.47, respectively, in comparison with the control population (table 2). To adjust for the possible effect of the "winner's curse", we used br^2 [38], and found that the bias-adjusted ORs 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_{SNP}. Estimates of h^2_{SNP} vary according to disease prevalence (supplementary figure S4). With prevalence set at 1 in 100 000 females, h^2_{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 1 Mb away from either gene. There were 566 and 416 SNPs for *TSC1* and *TSC2*, respectively, and only rs11552431 (located at 16:1 823 024, 274 kb away from *TSC2*) was significant in CLR after Bonferroni correction at q<0.1 (nominal $p=5.97\times10^{-5}$). We included that SNP and nine 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 S3).

Replication analysis was performed for the two genome-wide significant and genotyped SNPs, which were genotyped in 196 additional non-Hispanic White S-LAM patients and compared with SNP allele frequencies in each of three control datasets: 1) 409 non-Hispanic White healthy COPDGene females who were not used for discovery analyses, 2) 1121 Hispanic White females from the MESA dataset [25, 39] and 3) 225 731 British White females in the UK Biobank dataset [26]. Similar ORs for association of the minor

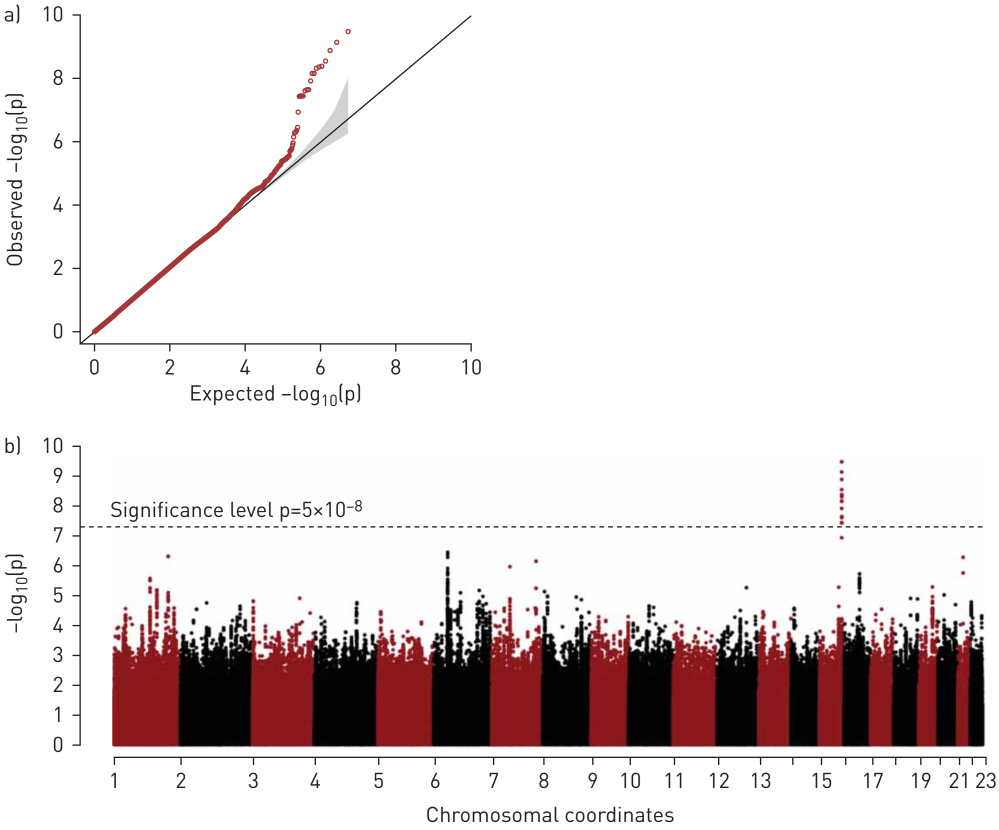

FIGURE 2 a) Quantile–quantile and b) Manhattan plots for the discovery lymphangioleiomyomatosis (LAM) genome-wide association study using the imputed data. SNP: single nucleotide polymorphism. a) The observed distributions of p-values for 5 426 936 SNPs including 549 591 directly genotyped are plotted relative to the expected (null) distribution for the conditional logistic regression analysis. The grey shading indicates the 95% confidence interval for observed p-values relative to those expected. b) Manhattan plot. Each dot represents the p-value of a single SNP, plotted on the genome scale at the bottom. The y-axis value is the negative logarithm of the p-value for association between each genotyped SNP and sporadic LAM. Twenty SNPs on chromosome 15q met genome-wide significance.

allele of these SNPs with S-LAM were observed in all three comparisons (table 2). Furthermore, we compared the MAFs of the two SNPs in LAM patients with those available from seven other studies (composed of non-Hispanic White European or USA populations), including all UK Biobank individuals. The MAFs of the two SNPs in LAM patients were significantly smaller than those reported in every other cohort (supplementary table S4).

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 S5), making it the candidate causal SNP in this association [23].

We also queried the GTEx database for SNPs in this linkage disequilibrium 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 TAD containing the SNP [40]. To identify the TAD containing these SNPs, we used TAD information available for four tissues: IMR90 cells [41], lung tissue [42], H1-MSCs [43] and HUVECs [41]. Supplementary figures S5–S8 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 cells, $<10^{-16}$ for H1-MSCs and ~0.1 for lung tissue (not available for HUVECs) [28]. Thus, the region containing our significant SNPs interacts with the *NR2F2* genomic region in IMR90 cells and H1-MSCs.

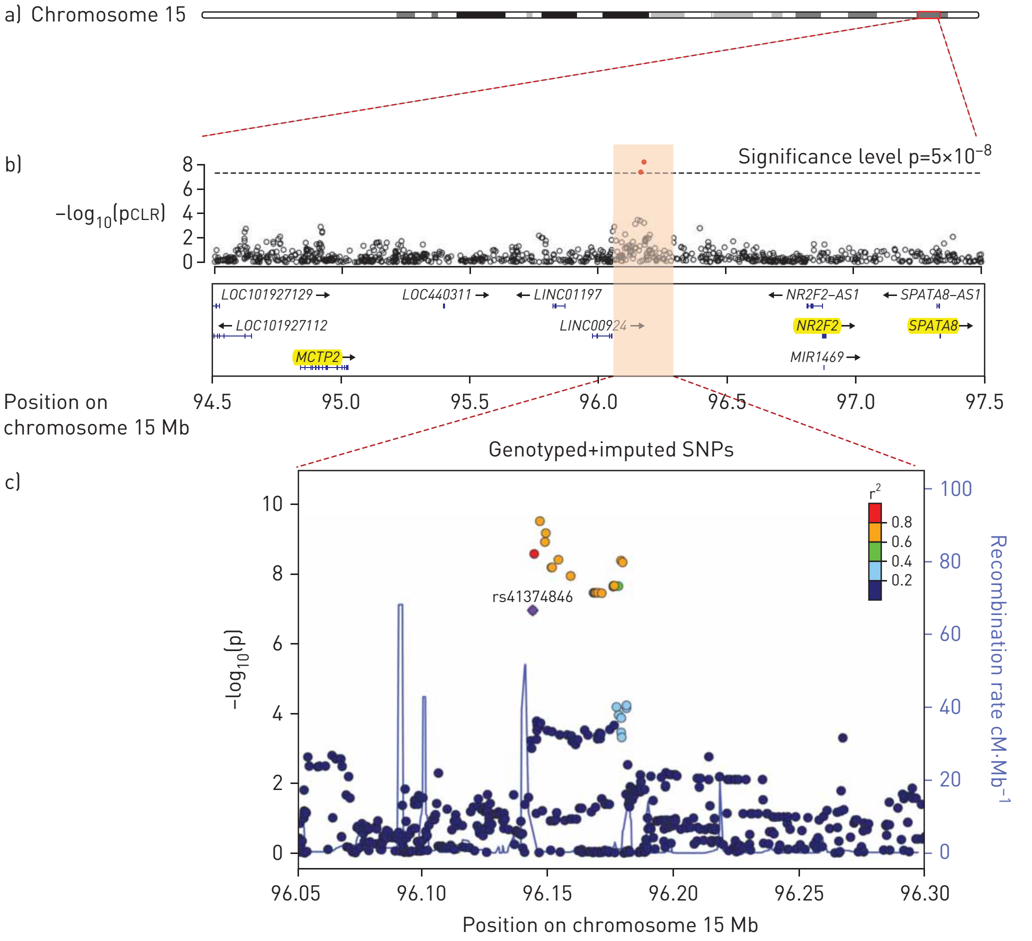

FIGURE 3 Genomic region on chromosome 15 containing the single nucleotide polymorphisms (SNPs) associated with lymphangioleiomyomatosis (LAM). CLR: conditional logistic regression. a) Ideogram of chromosome 15. b) 3 Mb region containing the SNPs associated with LAM. Manhattan plot at the top shows p-values for directly genotyped SNPs in this region, including the two SNPs meeting genome-wide significance (red dots). There are three protein-coding genes, *NR2F2*, *MCTP2* and *SPATA8*, which are highlighted in yellow, and many long noncoding RNAs in this region. c) Expanded Manhattan plot of the 250 kb region showing both genotyped and imputed SNPs. SNP rs41374846, the candidate causal SNP, is indicated in purple and other SNPs are coloured according to their r^2 value in relation to that SNP.

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=0.03) (table 3).

NR2F2, also known as COUP-transcription factor II, encodes a member of the steroid/thyroid hormone superfamily of nuclear receptors [44] and plays important roles in many developmental processes, including the neural crest [45], which is considered a potential candidate cell of origin of LAM [46], as well as in lymphangiogenesis and in angiogenesis [47]. 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 sequencing data, we compared the gene expression of four kidney angiomyolipomas and one abdominal LAM tumour with an extensive set of human cancers (from TCGA [32]) and normal tissues (from GTEx [29]) (figure 4). *NR2F2* expression was higher in the LAM-related tumours than any TCGA cancer (figure 4a) and was also relatively highly expressed in LAM-related tumours in comparison with normal tissues (p=6.38×10^{-6}, limma statistic) (figure 4b) [48]. In contrast, two other genes, *SPATA8*

TABLE 2 Genome-wide significant genotyped single nucleotide polymorphisms (SNPs)

	rs4544201	rs2006950
Chromosome	15q26.2	15q26.2
SNP position (hg19)	96 167 827	96 179 390
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	4.19×10^{-8}	6.12×10^{-9}
Replication data		
Genotype counts (AA/AG/GG/missing)		
S-LAM	4/48/144/0	3/39/154/0
COPDGene	26/171/212/0	26/159/224/0
MESA	69/417/635/0	64/385/672/0
UK Biobank	14 468/85 721/125 542/0	12 765/81 784/131 182/0
S-LAM *versus* COPDGene		
Odds ratio	0.3288	0.2731
p-value	4.32×10^{-5}	1.56×10^{-5}
S-LAM *versus* MESA		
Odds ratio	0.5070	0.4448
p-value	9.28×10^{-6}	1.04×10^{-6}
S-LAM *versus* UK Biobank		
Odds ratio	0.4888	0.4159
p-value	7.30×10^{-7}	3.11×10^{-8}

S-LAM: sporadic lymphangioleiomyomatosis; MESA: Multi-Ethnic Study of Atherosclerosis.

(spermatogenesis associated 8) and *MCTP2*, that were next closest to the SNP region showing association with LAM (1.1 and 1.2 Mb distant) (figure 3b) had no expression in the LAM-related tumours (data not shown).

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

Discussion

LAM occurs almost exclusively in females 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. 20 intergenic SNPs were identified in a 34 kb linkage disequilibrium block on chromosome 15, which met genome-wide

TABLE 3 Gene-based analyses of single nucleotide polymorphism (SNP) association with lymphangioleiomyomatosis (LAM)

Gene	Chromosome	Start[#]	End[¶]	SNPs n	p-value
NR2F2	15	96 869 157	96 883 492	5	0.0307
MCTP2	15	94 774 767	95 027 181	4	0.3579
SPATA8	15	97 326 619	97 328 845	3	0.5250

Three protein-coding genes were found on chromosome 15 from 94.2 to 98.2 Mb, the 4 Mb region surrounding the genome-wide association study SNPs, and gene-based analysis for association with LAM was performed using SKAT-O. [#]: start position of the corresponding gene; [¶]: end position of the corresponding gene.

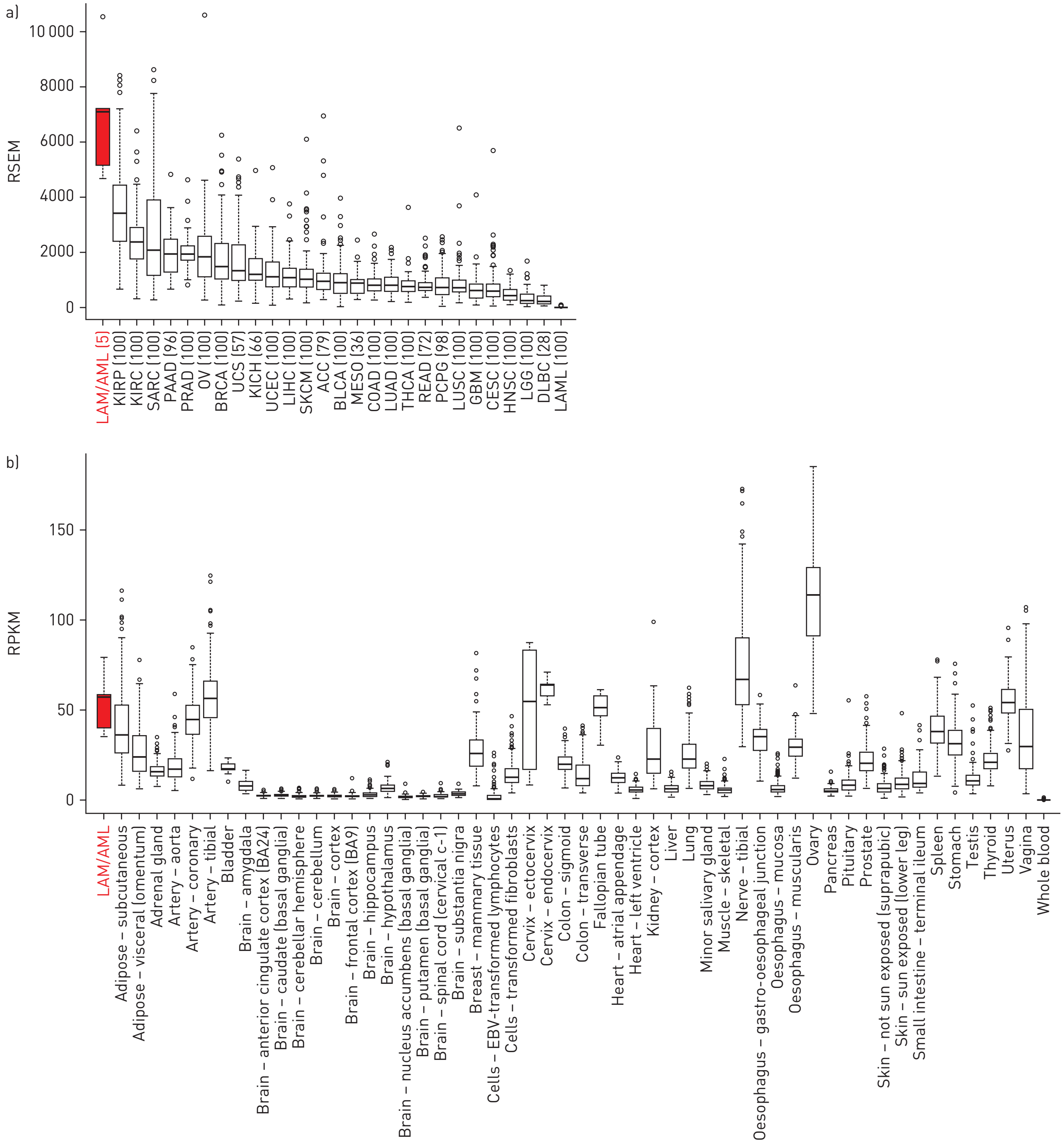

FIGURE 4 Comparison of *NR2F2* expression in kidney angiomyolipoma (ALM)/lymphangioleiomyomatosis (LAM) with a) cancer (The Cancer Genome Atlas (TCGA)) and b) normal (Gene and Tissue Expression (GTEx)) tissues. RSEM: RNA sequencing by expectation maximisation; RPKM: reads per kilobase per million mapped reads; EBV: Epstein–Barr virus. TCGA tumour abbreviations: KIRP: kidney renal papillary cell carcinoma; KIRC: kidney renal clear cell carcinoma; SARC: sarcoma; PAAD: pancreatic adenocarcinoma; PRAD: prostate 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 leukaemia. Box plot figures are shown to compare expression of *NR2F2* a) in four angiomyolipoma tumours and one abdominal LAM lesion with 2463 cancers of 27 types (from TCGA) in RSEM units (the number of samples of each type is given in brackets), and b) with ~7000 samples of 47 normal tissues (from GTEx) in RPKM units. Medians, interquartile ranges and 95% ranges are shown, with outliers indicated by circles.

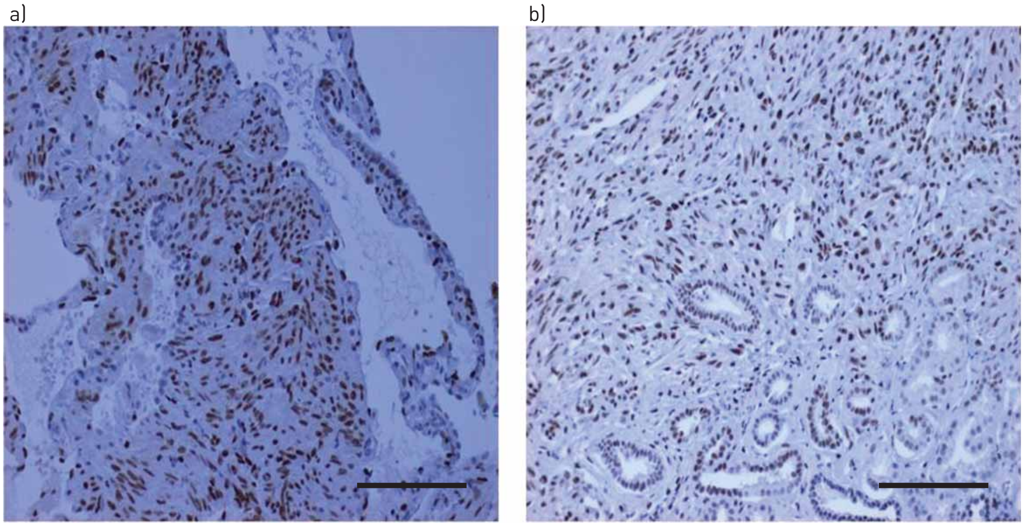

FIGURE 5 Immunohistochemistry for NR2F2 in lymphangioleiomyomatosis (LAM) and kidney angiomyolipoma. Strong nuclear staining is seen in a) lung LAM cells and b) angiomyolipoma cells (brown stain). Some other cells also have nuclear staining for *NR2F2*, but most do not. Representative fields obtained from eight LAM lung samples and four angiomyolipoma samples examined by immunohistochemistry. Scale bar: 150 μm.

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*, 700 kb 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 [49], making it a promising candidate driver gene in LAM pathogenesis. LAM occurs nearly exclusively in females, and oestrogen levels influence LAM development and progression [50, 51]. Small interfering RNA knockdown of oestrogen receptor α in MCF-7 breast cancer cells decreased *NR2F2* expression, while treatment with oestradiol increased its expression [52]. This interaction between oestrogen receptor α and *NR2F2* may also play a role in LAM development.

NR2F2 is highly expressed in LAM and angiomyolipoma by RNA sequencing analysis in comparison with large cancer and normal tissue datasets, and *NR2F2* shows high expression with nuclear localisation in both LAM and angiomyolipoma by immunohistochemistry. 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 [29], it is possible that such an eQTL relationship exists for LAM cells. We also note that the region of these SNPs contains several noncoding long RNAs, some antisense transcripts and microRNA miR1469 (figure 3b). It is possible that expression levels 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 pathological feature, with LAM cell clusters in the lung showing marked enrichment for lymphatic vessels [53, 54]. VEGF-D is a probable driver of lymphatic vessel growth in LAM, as serum VEGF-D level is increased in the majority of LAM patients, and serves as a diagnostic biomarker of LAM [55]. In mice, *NR2F2* has been shown to be required, with *SOX18* (SRY-box 18) for the polarised expression of *PROX1* (Prospero homeobox) 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 [56]. 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 a 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 minimise 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 minimise confounding due to genetic heterogeneity. Previous studies have shown that CLR is superior to unconditional logistic regression if variables used for matching are true confounding variables and only a moderate number of controls are excluded through matching [57–63]. We also found that CLR generated more significant results than unconditional logistic regression (supplementary table S6). Functional analyses to confirm our hypothesis that *NR2F2* is the gene affected by this SNP are limited due to the absence of a reliable LAM tumour 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 noncoding SNPs on chromosome 15q26.2 whose alleles are associated with S-LAM, which 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 tumours. *NR2F2* has not previously been implicated in LAM, and these novel and unexpected findings will hopefully lead to a 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: Hubert Wirtz (Universitätsklinikum, Leipzig, Germany), Iris Bassi (LAM Italia, Pavia di Udine, Italy), Corine Durand (Association FLAM, Paris, France) Lesley Bogoslavski (Israeli LAM Organization, Jerusalem, Israel) and Mordechai Kramer (Beilinson Hospital, Petah Tikva, Israel). M.A. Pujana wishes to thank the Spanish Association of LAM (AELAM) for its support.

Author contributions: Contributions to the conception and design of the work: S. Won, E.K. Silverman, G. Hunninghake and D.J. Kwiatkowski. 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.

Conflict of interest: W. Kim has nothing to disclose. K. Giannikou has nothing to disclose. J.R. Dreier has nothing to disclose. S. Lee has nothing to disclose. M.E. Tyburczy has nothing to disclose. E.K. Silverman reports grants from NIH, during the conduct of the study; honoraria for symposium participation from Novartis; and grant and travel support from GlaxoSmithKline, outside the submitted work. E Radzikowska has nothing to disclose. S. Wu has nothing to disclose. C-L. Wu has nothing to disclose. E.P. Henske has nothing to disclose. G. Hunninghake reports personal fees from Genentech, Boehringer Ingelheim, Gerson Lehrman Group and Mitsubishi Chemical, outside the submitted work. H. Carel has nothing to disclose. A. Roman has nothing to disclose. M.A. Pujana reports grants from Roche Pharma outside the submitted work. J. Moss has nothing to disclose. S Won has nothing to disclose. D.J. Kwiatkowski has nothing to disclose.

Support statement: 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 and by the Intramural Research Program, National Institutes of Health, National Heart, Lung, and Blood Institute. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Kitaichi M, Nishimura K, Itoh H, *et al.* Pulmonary lymphangioleiomyomatosis: a report of 46 patients including a clinicopathologic study of prognostic factors. *Am J Respir Crit Care Med* 1995; 151: 527–533.

2 Chu SC, Horiba K, Usuki J, *et al.* Comprehensive evaluation of 35 patients with lymphangioleiomyomatosis. *Chest J* 1999; 115: 1041–1052.

3 Urban T, Lazor R, Lacronique J, *et al.* Pulmonary lymphangioleiomyomatosis: a study of 69 patients. *Medicine* 1999; 78: 321–337.

4 Cunha B, Conceição DM, Cabo C, *et al.* Pulmonary lymphangioleiomyomatosis on a post-menopausal woman with chronic lymphocytic leukaemia. *Case Rep Clin Med* 2016; 5: 101–104.

5 Youssef AL, Alami B, Sahnoun F, *et al.* Lymphangioleiomyomatosis: an unusual age of diagnosis with literature review. *Int J Diagn Imaging* 2014; 1: 17–20.

6 Soler-Ferrer C, Gómez-Lozano A, Clemente-Andrés C, *et al.* Lymphangioleiomyomatosis in a post-menopausal women. *Arch Bronconeumol* 2010; 46: 148–150.

7 Taylor JR, Ryu J, Colby TV, *et al.* Lymphangioleiomyomatosis. *N Engl J Med* 1990; 323: 1254–1260.

8 Kalassian KG, Doyle R, Kao P, *et al.* Lymphangioleiomyomatosis: new insights. *Am J Respir Crit Care Med* 1997; 155: 1183–1186.

9 Giannikou K, Malinowska IA, Pugh TJ, *et al.* Whole exome sequencing identifies *TSC1*/*TSC2* biallelic loss as the primary and sufficient driver event for renal angiomyolipoma development. *PLoS Genet* 2016; 12: e1006242.

10 Carsillo T, Astrinidis A, Henske EP. Mutations in the tuberous sclerosis complex gene *TSC2* are a cause of sporadic pulmonary lymphangioleiomyomatosis. *Proc Natl Acad Sci USA* 2000; 97: 6085–6090.

11 Regan EA, Hokanson JE, Murphy JR, *et al.* Genetic epidemiology of COPD (COPDGene) study design. *COPD* 2011; 7: 32–43.

12 Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81: 559–575.

13 Song YE, Lee S, Park K, *et al.* ONETOOL for the analysis of family-based big data. *Bioinformatics* 2018; 34: 2851–2853.

14 Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy–Weinberg equilibrium. *Am J Hum Genet* 2005; 76: 887–893.

15 Raymond M, Rousset F. An exact test for population differentiation. *Evolution* 1995; 49: 1280–1283.

16 McCarthy S, Das S, Kretzschmar W, *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; 48: 1279–1283.

17 Loh P-R, Danecek P, Palamara PF, *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 2016; 48: 1443–1448.

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

19 Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; 11: 499–511.

20 Price AL, Patterson NJ, Plenge RM, *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; 38: 904–909.

21 Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J Stat Softw* 2011; 42: 1–52.

22 Therneau TM, Lumley T. Package "survival". 2017. https://CRAN.R-project.org/package=survival Date last accessed: April 4, 2019.

23 Farh KK-H, Marson A, Zhu J, *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 2015; 518: 337–343.

24 Lee S, Emond MJ, Bamshad MJ, *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012; 91: 224–237.

25 Bild DE, Bluemke DA, Burke GL, *et al.* Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol* 2002; 156: 871–881.

26 Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; 12: e1001779.

27 Dixon JR, Selvaraj S, Yue F, *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012; 485: 376–380.

28 Martin JS, Xu Z, Reiner AP, *et al.* HUGIn: Hi-C unifying genomic interrogator. *Bioinformatics* 2017; 33: 3793–3795.

29 Lonsdale J, Thomas J, Salvatore M, *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013; 45: 580–585.

30 Kim D, Pertea G, Trapnell C, *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013; 14: R36.

31 Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011; 12: 323.

32 Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008; 455: 1061–1068.

33 Ongen H, Buil A, Brown AA, *et al.* Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 2015; 32: 1479–1485.

34 Stegle O, Parts L, Durbin R, *et al.* A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* 2010; 6: e1000770.

35 Bongaarts A, Giannikou K, Reinten RJ, *et al.* Subependymal giant cell astrocytomas in Tuberous Sclerosis Complex have consistent *TSC1*/*TSC2* biallelic inactivation, and no *BRAF* mutations. *Oncotarget* 2017; 8: 95516–95529.

36 MacArthur J, Bowler E, Cerezo M, *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2016; 45: D896–D901.

37 Barrett JC, Fry B, Maller J, *et al.* Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2004; 21: 263–265.

38 Poirier JG, Faye LL, Dimitromanolakis A, *et al.* Resampling to address the winner's curse in genetic association analysis of time to event. *Genet Epidemiol* 2015; 39: 518–528.

39 Hankinson JL, Kawut SM, Shahar E, *et al.* 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: 138–145.

40 Grubert F, Zaugg JB, Kasowski M, *et al.* Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* 2015; 162: 1051–1065.

41 Rao SS, Huntley MH, Durand NC, *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014; 159: 1665–1680.

42 Schmitt AD, Hu M, Jung I, *et al.* A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep* 2016; 17: 2042–2059.

43 Dixon JR, Jung I, Selvaraj S, *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* 2015; 518: 331–336.

44 Qiu Y, Krishnan V, Zeng Z, *et al.* Isolation, characterization, and chromosomal localization of mouse and human COUP-TF I and II genes. *Genomics* 1995; 29: 240–246.

45 Rada-Iglesias A, Bajpai R, Prescott S, *et al.* Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell Stem Cell* 2012; 11: 633–648.

46 Julian LM, Delaney SP, Wang Y, *et al.* Human pluripotent stem cell-derived *TSC2*-haploinsufficient smooth muscle cells recapitulate features of lymphangioleiomyomatosis. *Cancer Res* 2017; 77: 5491–5502.

47 Qin J, Chen XP, Xie X, *et al.* COUP-TFII regulates tumor growth and metastasis by modulating tumor angiogenesis. *Proc Natl Acad Sci USA* 2010; 107: 3687–3692.

48 Ritchie ME, Phipson B, Wu D, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; 43: e47.

49 Xu MF, Qin J, Tsai SY, *et al.* The role of the orphan nuclear receptor COUP-TFII in tumorigenesis. *Acta Pharmacol Sin* 2015; 36: 32–36.

50 Juvet SC, Hwang D, Downey GP. Rare lung diseases I – lymphangioleiomyomatosis. *Can Respir J* 2006; 13: 375–380.

51 McCormack FX, Gupta N, Finlay GR, *et al.* Official American Thoracic Society/Japanese Respiratory Society Clinical Practice Guidelines: Lymphangioleiomyomatosis Diagnosis and Management. *Am J Respir Crit Care Med* 2016; 194: 748–761.

52 Riggs KA, Wickramasinghe NS, Cochrum RK, *et al.* Decreased chicken ovalbumin upstream promoter transcription factor II expression in tamoxifen-resistant breast cancer cells. *Cancer Res* 2006; 66: 10188–10198.

53 Glasgow CG, Taveira-DaSilva AM, Darling TN, *et al.* Lymphatic involvement in lymphangioleiomyomatosis. *Ann NY Acad Sci* 2008; 1131: 206–214.

54 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: 2051–2052.

55 Young LR, Lee HS, Inoue Y, *et al.* 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 Respir Med* 2013; 1: 445–452.

56 Srinivasan RS, Geng X, Yang Y, *et al.* The nuclear hormone receptor Coup-TFII is required for the initiation and early maintenance of *Prox1* expression in lymphatic endothelial cells. *Genes Dev* 2010; 24: 696–707.

57 Breslow N, Day N, Halvorsen K, *et al.* Estimation of multiple relative risk functions in matched case-control studies. *Am J Epidemiol* 1978; 108: 299–307.

58 Kupper LL, Karon JM, Kleinbaum DG, *et al.* Matching in epidemiologic studies: validity and efficiency considerations. *Biometrics* 1981; 37: 271–291.

59 McKinlay SM. Pair-matching – a reappraisal of a popular technique. *Biometrics* 1977; 33: 725–735.

60 Thompson WD, Kelsey JL, Walter SD. Cost and efficiency in the choice of matched and unmatched case-control study designs. *Am J Epidemiol* 1982; 116: 840–851.

61 Thomas DC, Greenland S. The relative efficiencies of matched and independent sample designs for case-control studies. *J Chronic Dis* 1983; 36: 685–697.

62 Miettinen OS. Estimation of relative risk from individually matched series. *Biometrics* 1970; 26: 75–86.

63 Luca D, Ringquist S, Klei L, *et al.* On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am J Hum Genet* 2008; 82: 453–463.