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

Wonji Kim1\*, Krinio Giannikou2\*, Sungho Won1,3,4\*, John R. Dreier2, Sanghun Lee5, Magdalena E. Tyburczy2, Edwin K. Silverman2,6, Elżbieta Radzikowska7, Shulin Wu8, Chin-Lee Wu8 , Elizabeth P. Henske2, Gary Hunninghake2, Havi Carel9, Antonio Roman10, Miquel Angel Pujana11, Joel Moss12, David J. Kwiatkowski2

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

2Division 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

3Department of Public Health Sciences, Seoul National University, Seoul, Korea

4Institute of Health and Environment, Seoul National University, Seoul, Korea

5Department of Medical Consilience, Graduate School, Dankook University, Seoul, Korea

6Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA, 02115, USA

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

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

12 Cardiovascular 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

Supported by the LAM Foundation, John Adler, and The Engles Fund for TSC and LAM research.

Short running head: LAM GWAS implicates *NR2F2*

Subject Category List: 9.30 Other Diffuse Lung Diseases/Drug Related Lung Diseases

The total word count of the body of the manuscript is 3032 words.

This article has an online data supplement, which is accessible from the issue’s table of content online at [www.atsjornals.org](http://www.atsjornals.org).

**At a Glance Commentary**

**Scientific Knowledge on the Subject**

Lymphangioleiomyomatosis (LAM) is a rare aggressive low-grade neoplasm which affects almost exclusively women, causes cystic lung destruction, and is due to *TSC2* mutations in most cases. However, risk factors and causes of the development of sporadic LAM are not known.

**What This Study Adds to the Field**

Through GWAS, we identified two SNPs on chromosome 15 that are strongly associated with LAM. *NR2F2* is the only protein-coding gene near these SNPs, and it appears that the SNP alleles associated with LAM may influence LAM development through effects on *NR2F2* expression. No previous GWAS on LAM has ever been reported.

**ABSTRACT**

**Rationale:** Lymphangioleiomyomatosis is a rare aggressive low-grade neoplasm which affects almost exclusively women and causes cystic lung destruction. Lymphangioleiomyomatosis can either be associated with Tuberous Sclerosis Complex (TSC) or be sporadic. Risk factors for development of sporadic Lymphangioleiomyomatosis (S-Lymphangioleiomyomatosis) are unknown.

**Objectives:** We hypothesized that DNA sequence variants outside of *TSC2*/*TSC1* might be associated with susceptibility/risk for Lymphangioleiomyomatosis, and performed an S-Lymphangioleiomyomatosis Genome Wide Association Study.

**Methods:** Saliva DNA samples from 479 S-Lymphangioleiomyomatosis subjects were genotyped on the Infinium OmniExpress-24 v1.2 BeadChip, and compared with genotype data for 1,261 COPDGene controls. Standard quality controls were employed, and both standard and conditional logistic regression for matched cases and controls was done. A replication study was performed. RNA-seq and immunohistochemistry analyses were also performed.

**Measurements and Main Results:** Two non-coding SNPs met genome-wide significance; rs4544201 (minor allele frequency (MAF) 0.166 in cases vs. 0.274 in controls, P-value=9.252×10 -10) and rs2006950 (MAF 0.143 in cases vs. 0.255 in controls, P-value=4.100×10 -10), which are in the same 35kb linkage disequilibrium (LD) block. 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 analysis in kidney angiomyolipoma, a Lymphangioleiomyomatosis-related tumor, than any cancer type available, and immunohistochemistry showed strong nuclear expression in both Lymphangioleiomyomatosis and angiomyolipoma.

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

Abstract word count is 249.

**KEYWORDS**

TSC2*,* GWAS, S-LAM, LAM, NR2F2

**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 (> 40%) 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 mosaic mutations in either *TSC1* (25%) or *TSC2* (75%) [9]. Tumor development in TSC follows the classic Knudson model of a germline mutation complemented by a somatic second hit mutation in the other corresponding allele in tumors [9, 10]. Limited data are available for S-LAM, but it appears that *TSC2* mutations are seen in the vast majority of S-LAM lesions. About 50% S-LAM subjects have kidney angiomyolipoma, a tumor which is very common in TSC. Angiomyolipoma share histologic, expression, and genetic features with LAM, though are not identical pathologic lesions.

Treatment options for women with LAM include rapamycin and related drugs (rapalogs), which have been shown to provide significant benefit in stabilizing lung function. Lung transplant is also an option for patients with severe and progressive disease [11-13]. Although rapalog therapy is beneficial, prolonged treatment is necessary, since cessation of therapy leads to recurrence of progressive respiratory functional decline [11-14].

Genome-wide association studies (GWAS) are utilized to identify genetic variants and susceptibility loci associated with human complex traits and common disease. 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 frequency of this disorder.

**METHODS**

**Discovery cohort**

Over 600 female S-LAM patients were identified and collected through international solicitation. S-LAM was diagnosed using standard diagnostic criteria [1-3, 7-8] by their treating physicians. Subjects were recruited from 2010 to 2014 from 14 countries: Australia, Canada, Colombia, France, Germany, Israel, Italy, Panama, Poland, Puerto Rico, Scotland, Spain, United Kingdom, and the USA. Genomic DNA was extracted from saliva using the QIAamp DNA mini kit (Qiagen), 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 at the Harvard Cancer Center Highthroughput Genotyping Core, Harvard T.H Chan School of Public Health. The S-LAM patient cohort consisted of 445 females of European ancenstry and 34 non-white females, and non-white females were excluded from our 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 [13, 15].

**Quality control analyses of SNP genotype data**

We evaluated the quality of SNPs and subjects in the discovery data set using standard methods, the PLINK [16] and ONETOOL [17] which are the analysis toolsets for genetic data. We excluded all SNPs for which: the Hardy-Weinberg equilibrium test [18] 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 then cases and controls were pooled/merged, and this procedure was 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 [19], with P < 1×10-5 was removed. Last, EIGENSTRAT [20] 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 outliers were then removed. These filters led to retention of 428 S-LAM cases and 1,261 female controls for analysis in the discovery phase with 549,599 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 SNPs identified in the discovery study. Need to add info on source of these samples; some local some from J Moss. Genotyping was performed by TaqMan SNP assay Need to add info on the specific TaqMan probe sets, as requested. (AppliedBiosystems; Life Technologies or ThermoFisher Scientific), and 9 randomly selected S-LAM subjects from the discovery study were also genotyped by this method to confirm the accuracy of genotyping in the replication analysis. Their discovery study genotypes matched the TaqMan analysis genotypes perfectly. Those 9 subjects were not included in the replication analyses. 1,214 NHW healthy males from the COPDgene consortium were used as controls for the replication study.

**Statistical analyses with genetic data**

GWAS analyses with discovery data were conducted using two different statistical approaches: logistic regression (LR) and conditional logistic regression (CLR).

PC scores were estimated with EIGENSTRAT [20], 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. Matching was conducted with *Matching* R package [21], and Supplementary Figure 1 shows that matching with age and two PC scores corresponding to the 2 greatest eigenvalues provide the variance inflation factor closest to 1. Thus CLR was conducted with cases and controls matched with age and 2 PC scores. For LR, we used all available cases and controls, and ten PC scores corresponding to 10 greatest eigenvalues and ages were included as covariates to adjust for population substructure.

LR and CLR analyses were performed using logistic regression in PLINK and conditional logistic regression in the R package *survival* [22]{Team, 2016 #31;Therneau, 2017 #94}, respectively. Genome-wide significance was assessed by P-value < 5×10-8 between cases and controls.

We also conducted gene-based analyses to identify genes with significant association with S-LAM using the optimal sequence kernel association test (SKAT-O) [23]. SNPs within each gene were used to provide a SNP set file, and age, squared age and 10 PC scores were included as covariates.

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

We imputed untyped SNPs located within 1 mega-base of the two genome-wide significant SNPs on chromosome 15 to do fine-mapping. Imputation was conducted using the Sanger Imputation Service (<https://imputation.sanger.ac.uk>). We used Haplotype Reference Consortium release v1.1 and considered predominantly European ancestry [24] which consists of 64,940 haplotypes as reference panel. Pre-phasing and imputation was conducted with SHAPEIT [25] and the PBWT package [26], respectively, and imputation accuracy was evaluated with the INFO metric [27]. Imputed SNPs were filtered out if INFOs, MAFs or P-values for the Hardy-Weinberg equilibrium test were < 0.3, 0.05, or 1×10-5, respectively. LD blocks were chosen by using Haploview with default options [28] and we applied CLR to all SNPs in the LD block with the genome-wide significant SNPs from the initial genotyping. Furthermore, we applied Probabilistic Identification of Causal SNPs (PICS) software to imputed and genotyped SNPs within the 34kb LD block containing the genome-wide significant SNPs and calculated PICS probability for each SNP. PICS assumes that there is a single causal SNP, and calculates the probability of each individual SNP being the causal SNP [29].

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

To identify chromatin interactions in the region of interest on chromosome 15q26.2, we used a publicly available web-based query tool, 3D genome browser ([www.3dgenome.org](http://www.3dgenome.org)). 3D genome browser predicts TADs using 86 published high-throughput chromosome conformation capture data sets with a pipeline provided by Dixon *et al* [30]. 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 GTEx RNA-seq project occurred [31]. Read data was processed into FASTQ files with standard QC methods, and aligned to the genome (hg19, NCBI37) using Tophat v2.0.10. **Some program** yielded RPKM. Fastq files were also converted into RSEM format by David Marron and Joel Parker (UNC)]. RSEM values were compared to RNA-seq data from 2463 tumors of 27 different histologic types from the TCGA {need ref}. RPKM values for *NR2F2* were compared to the GTEx data set of normal human tissues (~7,000 samples from 53 normal tissue types, v6p release) [31].

**Immunohistochemistry analyses**

Immunochistochemistry was performed as described elsewhere [32] using a primary mouse monoclonal antibody against *NR2F2* [Abcam Cat.Num # ab41859 Concentration 1:100 (10ug/ml) ]. Briefly, 5 µm sections of formalin-fixed, paraffin-embedded tumor specimens were deparaffinized in xylene, rehydrated with graded concentrations of alcohol to distilled water, washed with Tris-buffered saline with 0.1% Tween 20 (TBST), and antigen retrieval was performed in EDTA (pH 8.0, Diagnostic BioSystems) using a pressure cooker. Endogenous peroxidase activity was blocked with 3% H2O2, blocking was done with 5% goat serum for 30 minutes, followed by incubation overnight with antibody at 4°C, washing three times in TBST, incubation with anti-goat secondary antibody (Vector Labs, Burlingame, CA) at a dilution of 1:300. The peroxidase reaction was developed using DakoCytomation Liquid DAB plus Substrate Chromogen System (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, GWAS was performed on 428 S-LAM subjects and 1,261 control subjects for 549,599 SNPs using LR and CLR. Two non-coding SNPs rs4544201 and rs2006950 on chromosome 15 met genome-wide significance (rs4544201: P-value=9.252×10-10; rs2006950: P-value=4.100×10-10) using CLR. With LR, rs2006950 met genome-wide significance (P-value=7.782×10-9), while rs4544201 was close but did not meet that level (P-value=3.091×10-7).

Quantile-quantile plots for LRs and CLRs demonstrated that the distribution of observed P-values met the expected distribution, with the exception of the two SNPs (Figure 2), indicating that the analyses were free of systematic P-value inflation and likely robust. Manhattan plots of the genome demonstrated that only the two SNPs met genome-wide significance, with only one attaining significance by LR as mentioned above (Figure 3). Table 1 provides summaries for the two genome-wide significant SNPs.

SNPs rs4544201 and rs2006950 are located on 15q26.2, 11,563 nt apart, in an intergenic gene desert between *MCTP2* (1.1Mb away) and *NR2F2* (700kb away), that contains many lncRNAs (Figure 4B). Both SNPs have minor and major alleles of A and G, and showed lower MAFs in S-LAM cohort than the control population. The odds ratios (ORs) of a single minor allele in the S-LAM cohort were 0.56 and 0.51 respectively, in comparison to the control population (Table 1). ORs for genome-wide significant SNPs tend to be inflated by the Winner’s curse [33]. To adjust for this potential effect, bias-adjusted ORs for rs4544201 and rs2006950 were calculated using br2 [34] and were still highly significant at 0.58 and 0.52, respectively.

Replication was performed for the 2 SNPs with association with LAM using 196 additional S-LAM patients and 1,214 NHW healthy males from COPDGene. A similar OR for association of the minor allele of these SNPs with S-LAM was seen in the replication data (Table 1, ORrs4544201=0.47, ORrs2006950 = 0.40), confirming the results from the initial GWAS.

rs4544201 and rs2006950 belong to the same LD block on 15q26.2 [28], and are strongly correlated (*D*’=0.977, *r*2=0.854; Supplementary Figure 2). To examine the potential association of other SNPs in this region with S-LAM, we used the genotyped SNP data to impute genotype data for all SNPs within 1 megabase of these two SNPs. Eighteen imputed SNPs in the 34kb LD block had P-values for association with LAM similar to the two genotyped SNPs (Table 2).

To attempt to identify the causal SNP 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 1), suggesting that it is the candidate causal SNP in this association [29].

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

To assess potential mechanisms by which these SNP alleles are associated with LAM development, several analyses were performed.

First, we considered the chromatin TAD which contained these SNPs. TADs are defined using high-throughput chromosome conformation capture data to identify in a global manner regions of chromatin that have physical interaction through looping. It is postulated that the majority of SNPs associated with human disease or other phenotypes cause the association through interaction with regulatory elements of a coding gene within the TAD containing the SNP [35]. To identify the TAD containing these SNPs, we used TAD information available for four tissues: IMR90 cells, a fetal lung myofibroblast cell line; lung tissue; H1-MSC, a mesenchymal stem cell line; and HUVEC, human umbilical vein endothelial cells (Supplementary Figures 3-6). In all four of these cells/tissues, *NR2F2* was the only protein-coding gene within or near the boundary of the TAD containing the GWAS 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 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 [36], and plays important roles in many developmental processes, including the human neural crest [37], which is considered a potential candidate cell of origin of LAM [38], as well as in lymphangiogenesis and in angiogenesis [39]. 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 angiomyolipoma and LAM**

Using RNA-seq data, we compared the gene expression of 4 kidney angiomyolipoma and 1 abdominal LAM tumor with an extensive set of human cancers (from TCGA [40]), and normal tissues (from GTEX [31]) (Figure 5). *NR2F2* was more highly expressed in the LAM-related tumors than in any TCGA cancer (Figure 5A), and was also relatively highly expressed in LAM-related tumors in comparison to normal tissues (Figure 5B, P-value=6.38×10-6). 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 4B) 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 and kidney angiomyolipoma sections (Figure 6).

**DISCUSSION**

LAM occurs almost exclusively in women of childbearing age. Most LAM patients presenting with symptoms 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 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). Eighteen other SNPs in this 34kb region were also strongly associated with LAM as assessed by genotype imputation analysis (Table 2). The association was replicated in a validation population. Despite these findings, there were limitations to our study. Although our cohort of samples was large for a rare disease such as LAM, it was of only moderate size for GWAS. Second, to collect sufficient LAM subjects, we employed a successful worldwide recruitment strategy. However, our controls were all from the USA. Hence, we employed EIGENSTRAT to identify genetic outliers, and make our control and LAM subject sets more uniform from a genetic background perspective.

The SNPs with association to S-LAM lie in a gene desert on distal chromosome 15. The nearest protein-coding gene is *NR2F2*, 700kb away, and other nearby protein-coding genes are *MCTP2*, 1.2Mb away, and *SPATA8,* 1.1Mb away. 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. Furthermore *NR2F2* is highly expressed in LAM and angiomyolipoma by RNA-Seq analysis in comparison to large cancer and normal tissue data sets, and *NR2F2* shows high expression with nuclear localization in both LAM and angiomyolipoma by IHC. Although we did not identify an eQTL relationship for any of the 20 SNPs associated with S-LAM for any gene in any normal tissue or cancer type [31], it is possible that such an eQTL relationship exists for LAM cells. We also note that the region of these SNPs also contains several non-coding long RNAs, some antisense transcripts, and microRNA 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.

*NR2F2* is an orphan nuclear receptor known to play important roles in both normal tissue development and in tumorigenesis [41], making it a promising candidate driver gene in LAM pathogenesis. LAM occurs nearly exclusively in women, and estrogen levels influence LAM development and progression [42, 43]. siRNA knockdown of ERα (Estrogen Receptor) in MCF-7 breast cancer cells decreased *NR2F2* expression, while treatment with estradiol increased its expression [44]. This interaction between ERα and *NR2F2* may also play a role in LAM development.

Lymphatic involvement in LAM is a hallmark pathologic feature with LAM cell clusters in the lung showing marked enrichment for lymphatic vessels [45, 46]. 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 [47]. In mice, *NR2F2* has been shown to be required, with *SOX18*, for the polarized expression of *PROX1* in a subset of endothelial cells (ECs) within the cardinal vein at embryonic day 9.5, an event that leads to development of the lymphatic endothelium [48]. Hence there is also a potential connection between *NR2F2*, VEGF-D, lymphatic development, and LAM pathogenesis.

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 may influence *NR2F2* expression and thereby LAM pathogenesis. *NR2F2* is relatively highly expressed in LAM and LAM-related tumors. *NR2F2* has not previously been associated with LAM pathogenesis, and these findings will hopefully lead to further analysis of its role in LAM, leading to better insight into 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.

**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., et al., *Pulmonary lymphangioleiomyomatosis: a report of 46 patients including a clinicopathologic study of prognostic factors.* American journal of respiratory and critical care medicine, 1995. **151**(2): p. 527-533.

2. Chu, S.C., et al., *Comprehensive evaluation of 35 patients with lymphangioleiomyomatosis.* CHEST Journal, 1999. **115**(4): p. 1041-1052.

3. Urban, T., et al., *Pulmonary lymphangioleiomyomatosis: a study of 69 patients.* MEDICINE-BALTIMORE-, 1999. **78**: p. 321-337.

4. Cunha, B., et al., *Pulmonary Lymphangioleiomyomatosis on a Post-Menopausal Woman with Chronic Lymphocytic Leukaemia.* Case Reports in Clinical Medicine, 2016. **5**(03): p. 101.

5. Youssef, A.L., et al., *Lymphangioleiomyomatosis: An unusual age of diagnosis with literature review.* International Journal of Diagnostic Imaging, 2014. **1**(1): p. 17.

6. Soler-Ferrer, C., et al., *Lymphangioleiomyomatosis in a post-menopausal women.* Archivos de Bronconeumología ((English Edition)), 2010. **46**(3): p. 148-150.

7. Taylor, J.R., et al., *Lymphangioleiomyomatosis.* New England Journal of Medicine, 1990. **323**(18): p. 1254-1260.

8. Kalassian, K.G., et al., *Lymphangioleiomyomatosis: new insights.* American journal of respiratory and critical care medicine, 1997. **155**(4): p. 1183-1186.

9. Giannikou, K., et al., *Whole exome sequencing identifies TSC1/TSC2 biallelic loss as the primary and sufficient driver event for renal angiomyolipoma development.* PLoS genetics, 2016. **12**(8): p. e1006242.

10. Carsillo, T., A. Astrinidis, and E.P. Henske, *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): p. 6085-6090.

11. Harari, S., et al., *Sirolimus therapy for patients with lymphangioleiomyomatosis leads to loss of chylous ascites and circulating LAM cells.* Chest, 2016. **150**(2): p. e29-e32.

12. Cai, X., et al., *Sirolimus decreases circulating lymphangioleiomyomatosis cells in patients with lymphangioleiomyomatosis.* Chest, 2014. **145**(1): p. 108-112.

13. Moss, J., et al., *Prevalence and clinical characteristics of lymphangioleiomyomatosis (LAM) in patients with tuberous sclerosis complex.* American journal of respiratory and critical care medicine, 2001. **164**(4): p. 669-671.

14. Black, J., et al., *In vitro studies of lymphangioleiomyomatosis.* European Respiratory Journal, 2005. **26**(4): p. 569-576.

15. Regan, E.A., et al., *Genetic epidemiology of COPD (COPDGene) study design.* COPD: Journal of Chronic Obstructive Pulmonary Disease, 2011. **7**(1): p. 32-43.

16. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* The American Journal of Human Genetics, 2007. **81**(3): p. 559-575.

17. Song, Y.E., et al., *ONETOOL for the analysis of family-based big data.* Bioinformatics, 2018. **1**: p. 3.

18. Wigginton, J.E., D.J. Cutler, and G.R. Abecasis, *A note on exact tests of Hardy-Weinberg equilibrium.* The American Journal of Human Genetics, 2005. **76**(5): p. 887-893.

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

20. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature genetics, 2006. **38**(8): p. 904.

21. Sekhon, J.S., *Multivariate and propensity score matching software with automated balance optimization: the matching package for R.* 2011.

22. Therneau, T.M. and T. Lumley, *Package ‘survival’.* R package version, 2017: p. 2.41-3.

23. Lee, S., et al., *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): p. 224-237.

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

25. Delaneau, O., J. Marchini, and G.P. Consortium, *Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel.* Nature communications, 2014. **5**: p. 3934.

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

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

28. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2004. **21**(2): p. 263-265.

29. Farh, K.K.-H., et al., *Genetic and epigenetic fine mapping of causal autoimmune disease variants.* Nature, 2015. **518**(7539): p. 337-343.

30. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-380.

31. Lonsdale, J., et al., *The genotype-tissue expression (GTEx) project.* Nature genetics, 2013. **45**(6): p. 580-585.

32. Bongaarts, A., et al., *Subependymal giant cell astrocytomas in Tuberous Sclerosis Complex have consistent TSC1/TSC2 biallelic inactivation, and no BRAF mutations.* Oncotarget, 2017. **8**(56): p. 95516.

33. Kraft, P., *Curses—winner's and otherwise—in genetic epidemiology.* Epidemiology, 2008. **19**(5): p. 649-651.

34. Poirier, J.G., et al., *Resampling to Address the Winner's Curse in Genetic Association Analysis of Time to Event.* Genetic epidemiology, 2015. **39**(7): p. 518-528.

35. Grubert, F., et al., *Genetic control of chromatin states in humans involves local and distal chromosomal interactions.* Cell, 2015. **162**(5): p. 1051-1065.

36. Qiu, Y., et al., *Isolation, characterization, and chromosomal localization of mouse and human COUP-TF I and II genes.* Genomics, 1995. **29**(1): p. 240-6.

37. Rada-Iglesias, A., et al., *Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest.* Cell stem cell, 2012. **11**(5): p. 633-648.

38. Julian, L.M., et al., *Human Pluripotent Stem Cell–Derived TSC2-Haploinsufficient Smooth Muscle Cells Recapitulate Features of Lymphangioleiomyomatosis.* Cancer research, 2017. **77**(20): p. 5491-5502.

39. Qin, J., et al., *COUP-TFII regulates tumor growth and metastasis by modulating tumor angiogenesis.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(8): p. 3687-3692.

40. Network, C.G.A.R., *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061.

41. Xu, M.F., et al., *The role of the orphan nuclear receptor COUP-TFII in tumorigenesis.* Acta Pharmacologica Sinica, 2015. **36**(1): p. 32-36.

42. Juvet, S.C., D. Hwang, and G.P. Downey, *Rare lung diseases I--Lymphangioleiomyomatosis.* Can Respir J, 2006. **13**(7): p. 375-80.

43. McCormack, F.X., et al., *Official American Thoracic Society/Japanese Respiratory Society Clinical Practice Guidelines: Lymphangioleiomyomatosis Diagnosis and Management.* American Journal of Respiratory and Critical Care Medicine, 2016. **194**(6): p. 748-761.

44. Riggs, K.A., et al., *Decreased chicken ovalbumin upstream promoter transcription factor II expression in tamoxifen-resistant breast cancer cells.* Cancer Research, 2006. **66**(20): p. 10188-10198.

45. Glasgow, C.G., et al., *Lymphatic involvement in lymphangioleiomyomatosis.* Lymphatic Continuum Revisited, 2008. **1131**: p. 206-214.

46. Seyama, K., K. Mitani, and T. Kumasaka, *Lymphangioleiomyoma Cells and Lymphatic Endothelial Cells Expression of VEGFR-3 in Lymphangioleiomyoma Cell Clusters.* American Journal of Pathology, 2010. **176**(4): p. 2051-2052.

47. Young, L.R., 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 Respiratory Medicine, 2013. **1**(6): p. 445-452.

48. Srinivasan, R.S., et al., *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): p. 696-707.

**Figure Legends**

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

**Figure 2. Quantile-quantile plots for discovery LAM GWAS dataset.** The observed distributions of P-values for 549,599 genotyped SNPs are plotted relative to the expected (null) distribution for CLR analyses.

**Figure 3. Manhattan plot of p-values from the discovery LAM GWAS**. 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 LAM. Two SNPs on 15q26.2 met genome-wide significance by CLR.

**Figure 4. 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 were represented in yellow shaded boxs, 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 is indicated by purple, and other SNPs are colored according to their *r*2 value in relation to rs41374846.

**Figure 5. Comparison of *NR2F2* expression in kidney angiomyolipoma/LAM with cancer (TCGA) and normal tissues (GTEx).**

Boxplot figures are shown to compare expression of *NR2F2* in 4 angiomyolipoma 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.

**Figure 6. 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.

**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*** | | |
| Odds ratio | 0.3288 | 0.2731 |
| P-value | 4.32×10-5 | 1.5610-5 |

Definition of abbreviations: LR = Logistic Regression; CLR = Conditional Logistic Regression.

**Table 2.** Statistical analyses of imputed SNPs with CLR.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CHR | SNP | POS | Alleles\* | MAF | INFO† | P-value for CLR‡ |
| 15 | rs41374846 | 96143559 | A/G | 0.2605 | 0.9097 | 3.04410-9 |
| 15 | rs59125351 | 96144157 | G/T | 0.2510 | 0.9771 | 3.47610-10 |
| 15 | rs17581137 | 96146414 | C/A | 0.2336 | 0.9893 | 1.50010-10 |
| 15 | rs6496126 | 96148439 | C/G | 0.2330 | 0.9890 | 1.96310-10 |
| 15 | rs2397810 | 96148765 | C/T | 0.2330 | 0.9890 | 1.96310-10 |
| 15 | rs10520790 | 96151040 | T/G | 0.2478 | 0.9958 | 3.84810-10 |
| 15 | rs55804812 | 96151256 | A/T | 0.2475 | 0.9952 | 4.50110-10 |
| 15 | rs16975389 | 96153782 | C/T | 0.2463 | 0.9967 | 6.23910-10 |
| 15 | rs16975396 | 96158705 | G/T | 0.2466 | 0.9983 | 1.03010-9 |
| 15 | rs4628911 | 96167905 | T/C | 0.2472 | 1.0000 | 5.54010-10 |
| 15 | rs6496128 | 96168303 | G/A | 0.2472 | 1.0000 | 5.54010-10 |
| 15 | rs8029996 | 96168770 | A/G | 0.2472 | 0.9998 | 5.54010-10 |
| 15 | rs4551988 | 96169589 | C/G | 0.2472 | 0.9998 | 5.54010-10 |
| 15 | rs58878263 | 96171069 | A/C | 0.2493 | 0.9979 | 6.83510-10 |
| 15 | rs8040665 | 96175692 | G/T | 0.2487 | 0.9976 | 7.90410-10 |
| 15 | 15:96175733 | 96175733 | A/G | 0.2466 | 0.9975 | 5.61610-10 |
| 15 | rs8040168 | 96176096 | G/C | 0.2466 | 0.9981 | 5.61610-10 |
| 15 | rs17504029 | 96177670 | T/A | 0.2478 | 0.9876 | 2.03510-10 |

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 LAM.

**Table 3.** Gene-based analyses of SNP association with LAM.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |

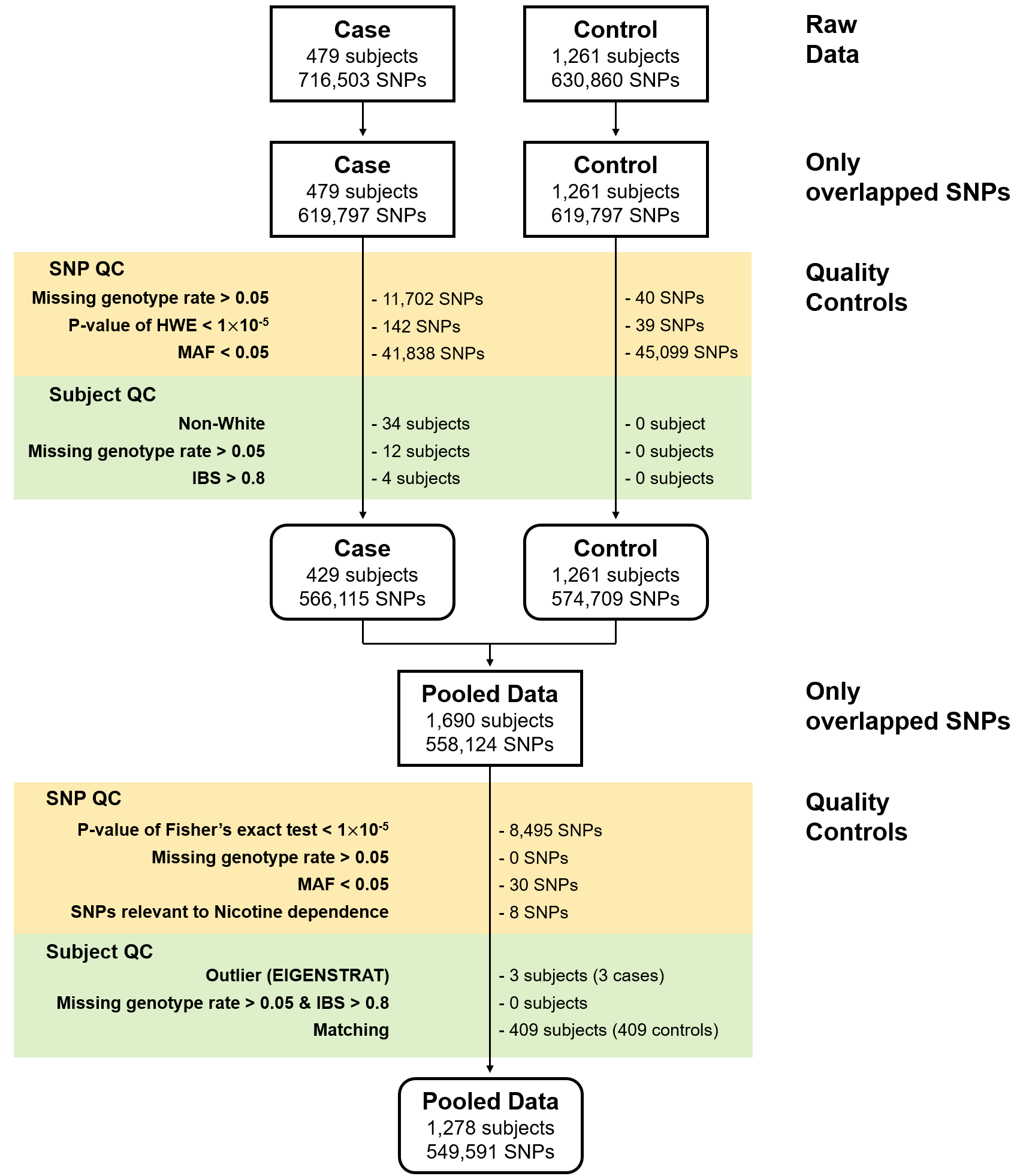
Definition of abbreviations: CHR = Chromosome

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.

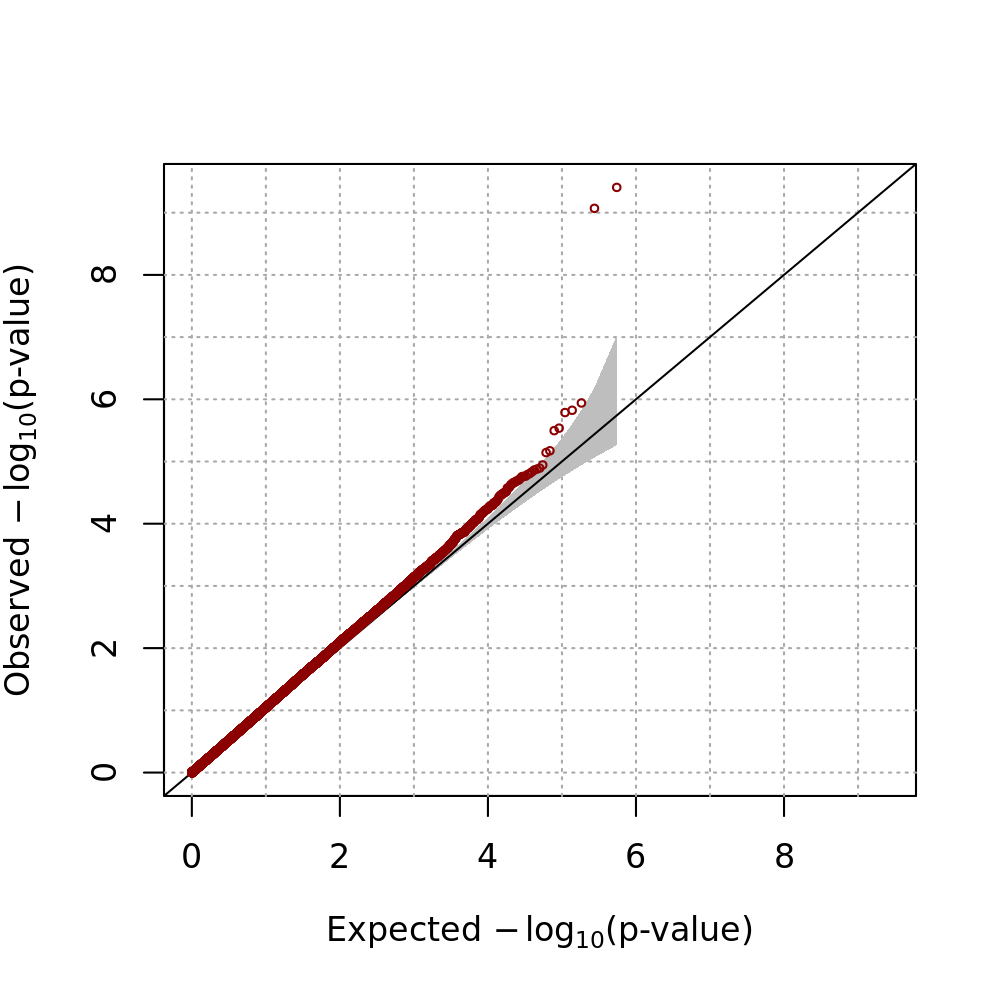
\* Start position of the corresponding gene.

† End position of the corresponding gene.

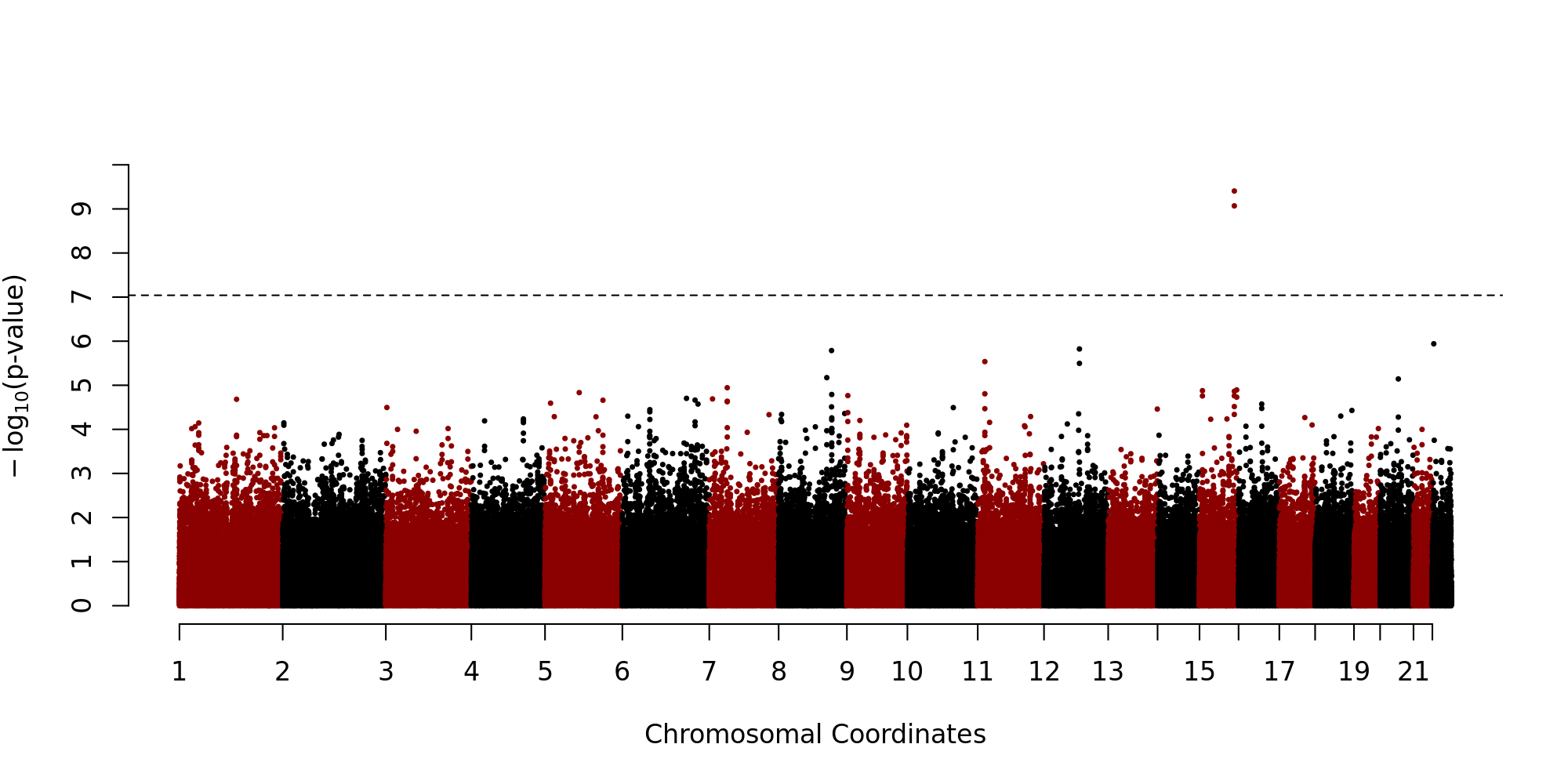
**Figure 1. Workflow of quality control for the LAM GWAS discovery data set.** Multiple standard quality controls were performed for both cases (female S-LAM subjects) and controls (healthy women without COPD from COPD consortium) to exclude outlier SNPs and subjects. Here, HWE, MAF and IBS are abbreviations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



**Figure 2. Quantile-quantile plots for discovery LAM GWAS dataset.** The observed distributions of P-values for 549,599 genotyped SNPs are plotted relative to the expected (null) distribution for CLR analyses.

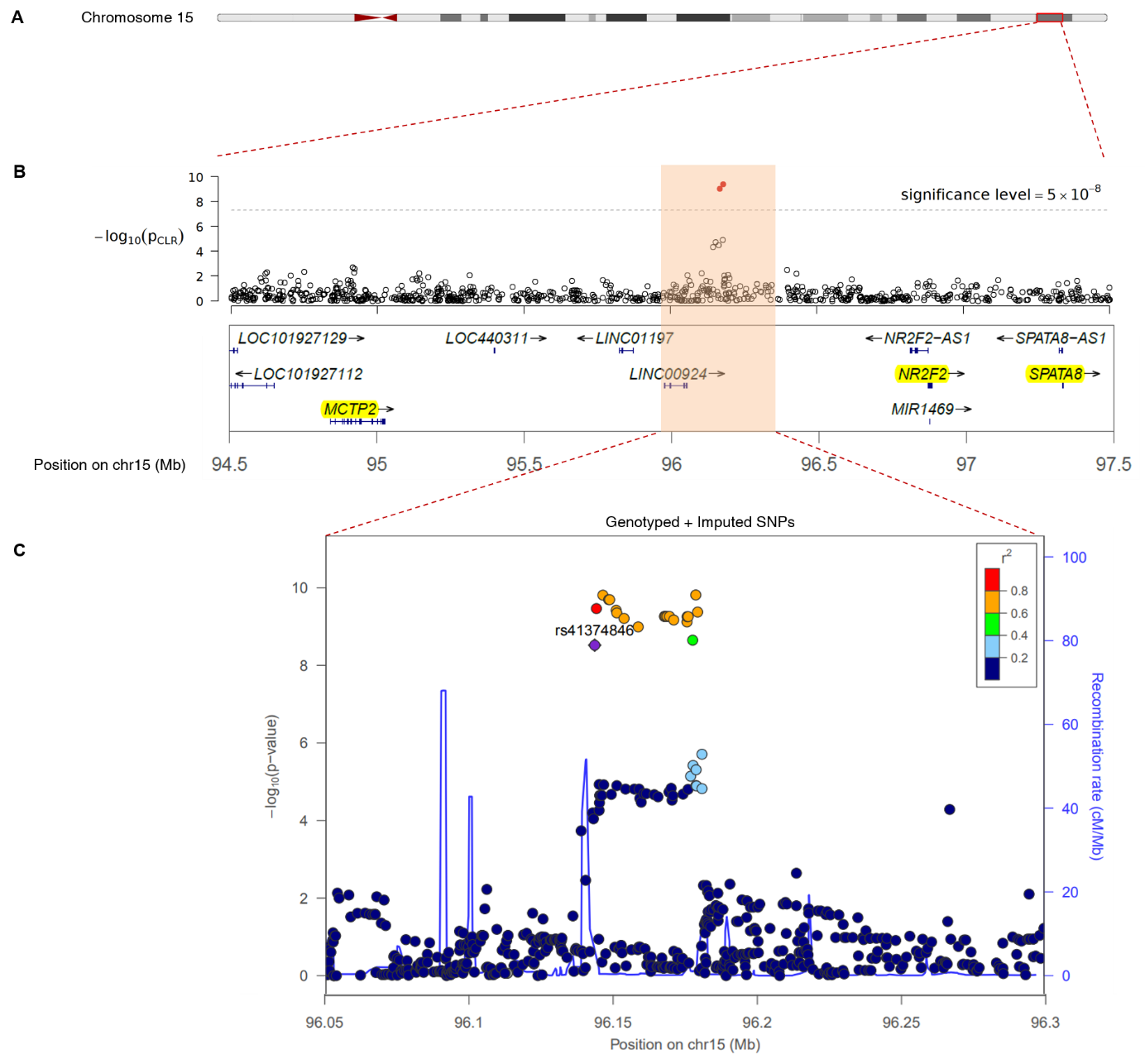
****

**Figure 3. Manhattan plot of p-values from the discovery LAM GWAS**. 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 LAM. Two SNPs on 15q26.2 met genome-wide significance by CLR.

****

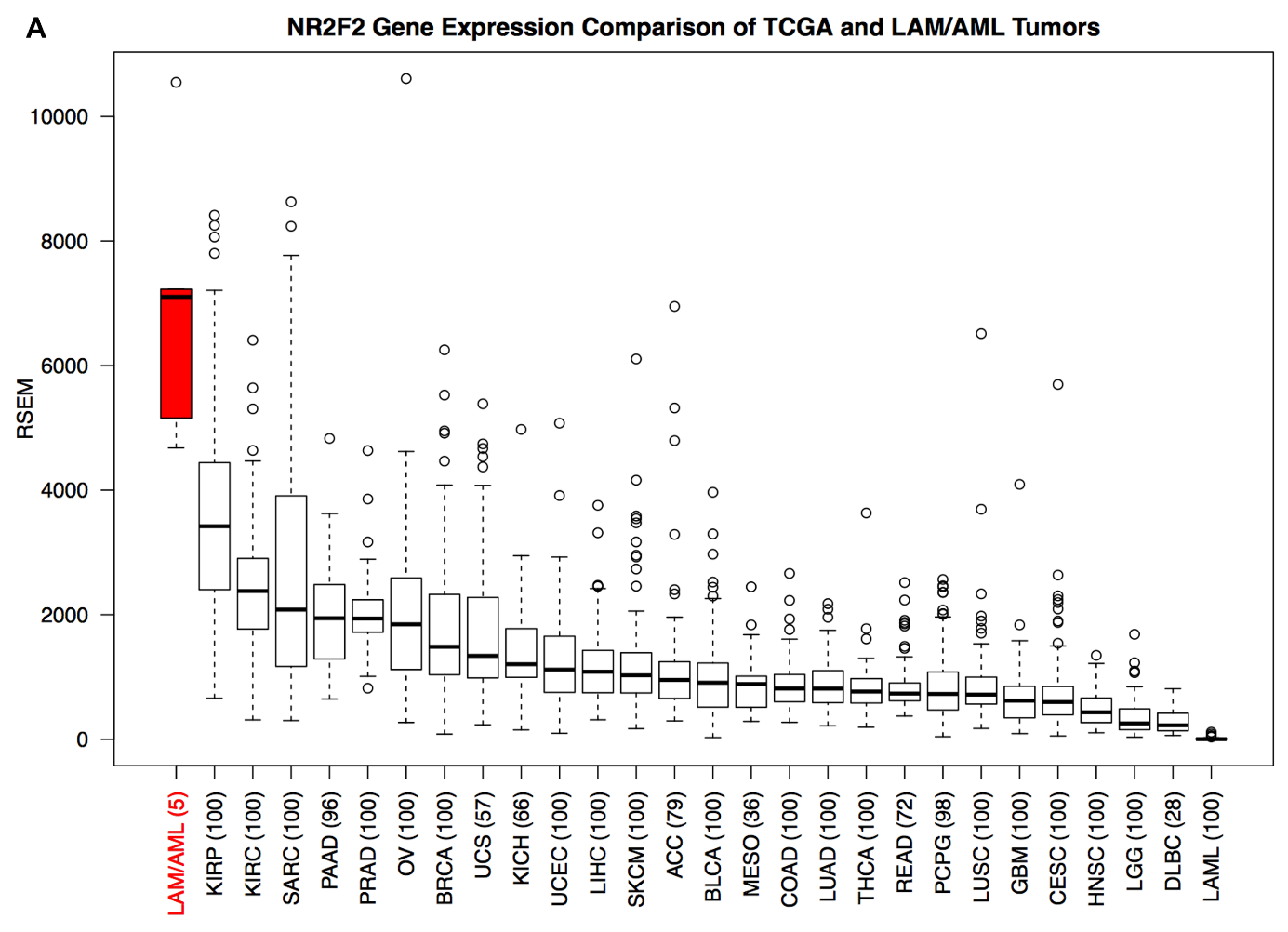
**Figure 4. Genomic region on chr15 containing the SNPs associated with LAM.**

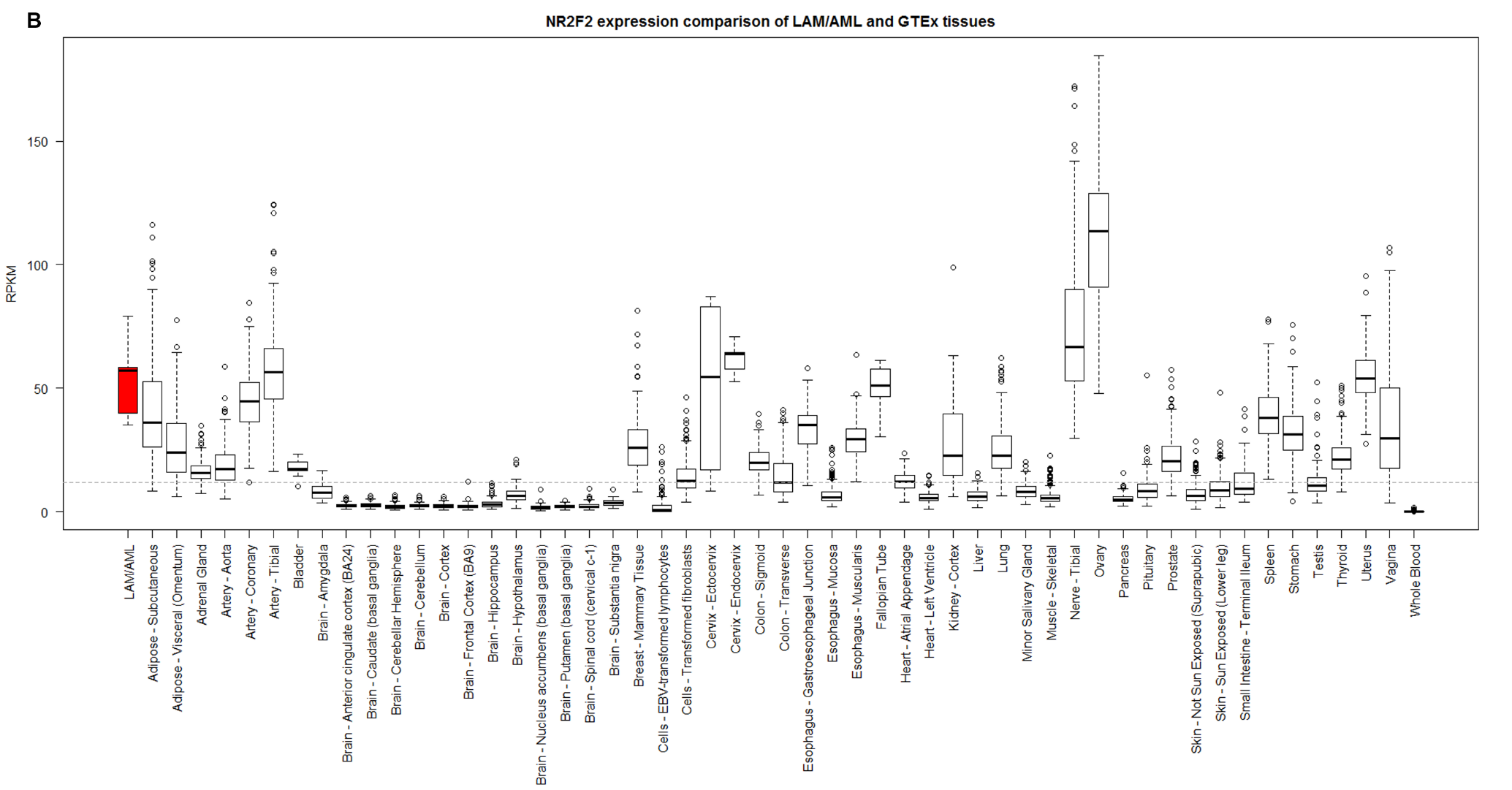
(A) Ideogram of chromosome 15. (B) Three Mb region containing the SNPs associated with LAM. Manhattan plot at top shows P-values for SNPs in this region, including the two SNPs meeting genome-wide significance (red dots). There are 3 protein-coding genes *NR2F2*, *MCTP2*, and *SPATA8* which were represented in yellow shaded boxs, 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 is indicated by purple, and other SNPs are colored according to their *r*2 value in relation to rs41374846.

****

**Figure 5. Comparison of *NR2F2* expression in kidney angiomyolipoma/LAM with cancer (TCGA) and normal tissues (GTEx).**

Boxplot figures are shown to compare expression of *NR2F2* in 4 angiomyolipoma and one abdominal LAM lesion with 2463 cancers of 27 types (from TCGA, brackets on x-axis include the number of samples analyzed per tumor type; abbreviations are explained in Supplementary Table 2) in RSEM units (A); and with ~7,000 samples of 47 normal tissues (from GTEx) in RPKM units (B). Remarkably, NR2F2 gene expression is the highest compared to all TCGA tumors and higher compared to most GTEx normal tissues; similar to cervix, fallopian tubes, uterus and ovaries. The median value, interquartile range, and 95% ranges are shown, with outliers indicated by circles. In the X axis, the each number in brackets is the number of samples corresponding each tissue.

****

****

**Figure 6. 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. 