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

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

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

**Supplementary Table 2. P-values for SNPs associated with nicotine addiction.** P values are shown in comparison of allele frequencies for the S-LAM discovery cohort and the COPDGene controls.

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

**\*** rs10491551 is included due to high correlation with rs12348139 on chromosome 9 (r2 = 1) although it is not found in the GWAS catalogue.

**Supplementary Table 3. Minor allele frequencies for SNPs rs4544201 and rs2006950 in multiple populations.**

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

\* MESA = Multi-Ethnic Study of Atherosclerosis. Nonhispanic whites females were chosen and MAFs were calculated.

\*\* 1000GP = 1000 Genome Project

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

† http://pheweb.sph.umich.edu:5000/

‡ http://gnomad.broadinstitute.org

**Supplementary Table 4.** PICS analysis to identify probable causal SNPs in the chr 15q region.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CHR | SNP\* | POS | P-value | † | ‡ | PICS probability |
| **15** | **rs41374846** | **96143559** | **3.43210-9** | 1.0000 | 1.0000 | 0.6485 |
| 15 | rs59125351 | 96144157 | 3.22910-10 | 0.9703 | 0.7941 | 0.0352 |
| 15 | rs55804812 | 96151256 | 4.17810-10 | 0.9557 | 0.7758 | 0.0290 |
| 15 | rs16975389 | 96153782 | 5.80110-10 | 0.9555 | 0.7700 | 0.0272 |
| 15 | rs10520790 | 96151040 | 3.57110-10 | 0.9486 | 0.7698 | 0.0271 |
| 15 | rs16975396 | 96158705 | 9.59210-10 | 0.9480 | 0.7581 | 0.0239 |
| 15 | rs58878263 | 96171069 | 6.36110-10 | 0.9328 | 0.7287 | 0.0172 |
| 15 | rs8029996 | 96168770 | 5.14710-10 | 0.9325 | 0.7230 | 0.0161 |
| 15 | rs6496128 | 96168303 | 5.14710-10 | 0.9325 | 0.7230 | 0.0161 |
| 15 | rs4628911 | 96167905 | 5.14710-10 | 0.9325 | 0.7230 | 0.0161 |
| 15 | rs8040665 | 96175692 | 7.35610-10 | 0.9254 | 0.7171 | 0.0151 |
| 15 | rs17581137 | 96146414 | 1.38410-10 | 0.9529 | 0.7125 | 0.0143 |
| 15 | rs4544201 | 96167827 | 5.14710-10 | 0.9317 | 0.7116 | 0.0142 |
| 15 | rs4551988 | 96169589 | 5.14710-10 | 0.9183 | 0.7113 | 0.0141 |
| 15 | rs2397810 | 96148765 | 1.81410-10 | 0.9451 | 0.7008 | 0.0125 |
| 15 | rs6496126 | 96148439 | 1.81410-10 | 0.9380 | 0.7005 | 0.0124 |
| 15 | rs8040168 | 96176096 | 5.22410-10 | 0.9233 | 0.6887 | 0.0108 |

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

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

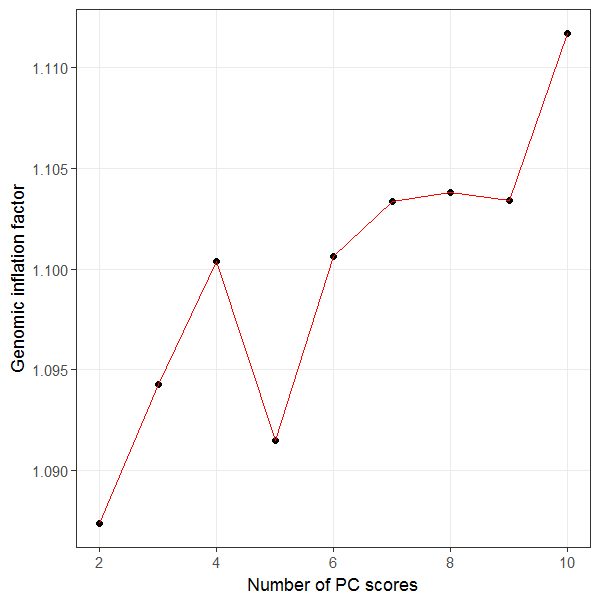
†  where : the frequency of the haplotype AB and : theoretical maximum difference between the observed and expected haplotype frequencies.

‡ : squared correlation coefficient

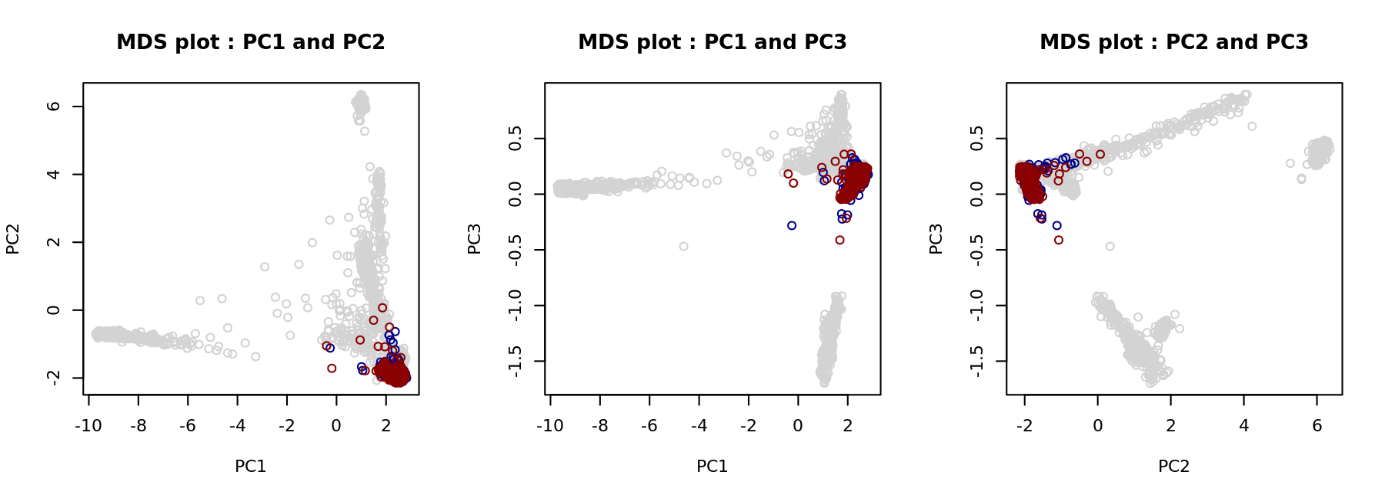
**Supplementary Table 5. TCGA tumor abbreviations**

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

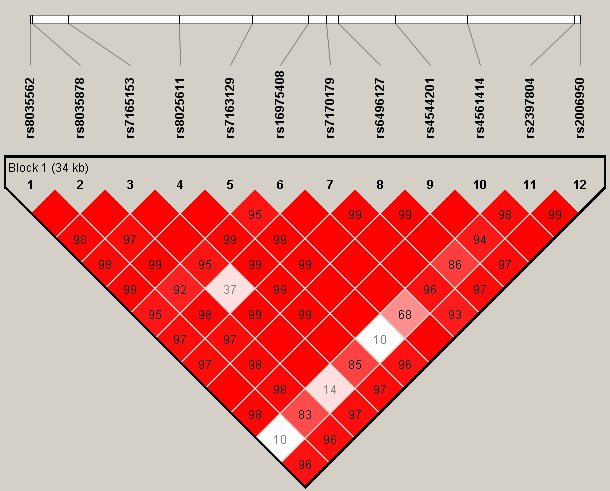
**Supplementary Figure 1. Variance inflation factors according to the number of PC scores used for the discovery data.** Cases and controls were matched with different numbers of PC scores (2 – 10 PC scores) and age, and CLR was applied to matched cases and controls. Variance inflation factors were calculated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

****

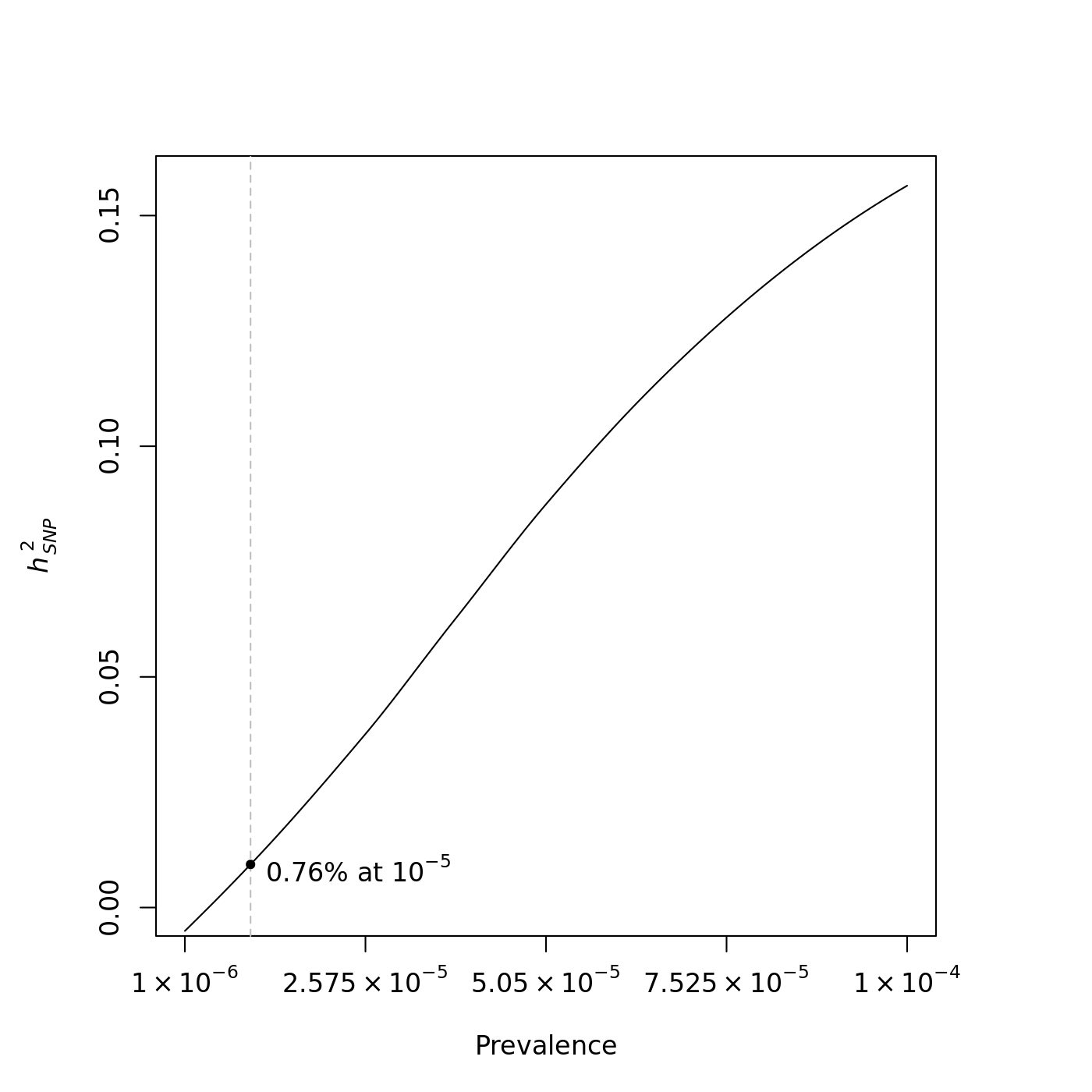
**Supplementary Figure 2. Scatter plot of PC scores.** Multi-dimensional scaling plots were generated using a pool of our Discovery S-LAM cohort, our COPDGene controls, and 1000 Genome project data. Red and blue circles indicate S-LAM and COPDGene samples used for our discovery analyses, respectively,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

****

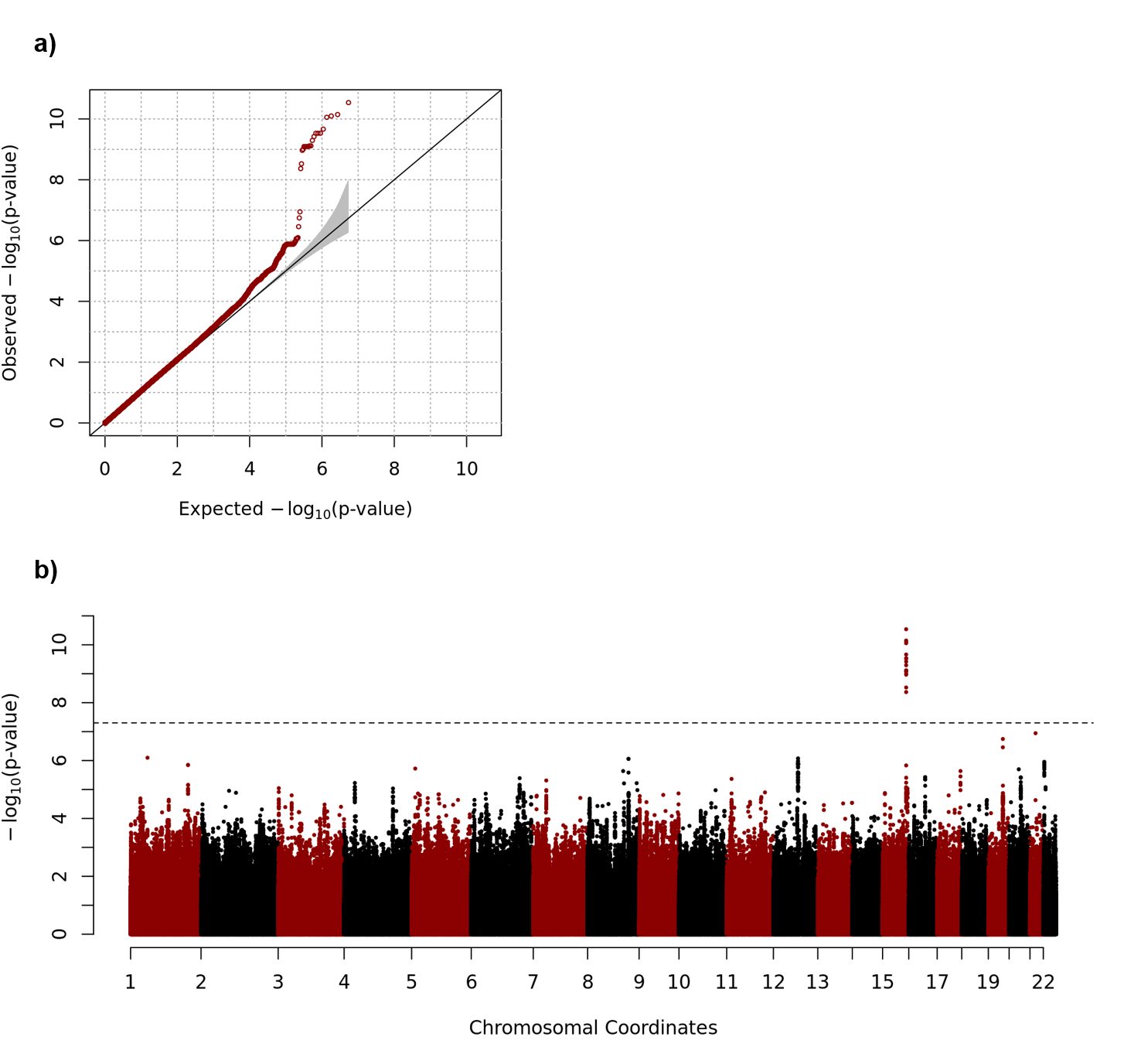
**Supplementary Figure 3. Linkage disequilibrium (LD) block around the two genome wide significant SNPs, rs4544201 and rs2006950.** Graph represents all genotyped SNPs in the 34kb LD block on chromosome 15q26.2. The color of each rectangle and number within indicates the level of LD between a pair of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

****

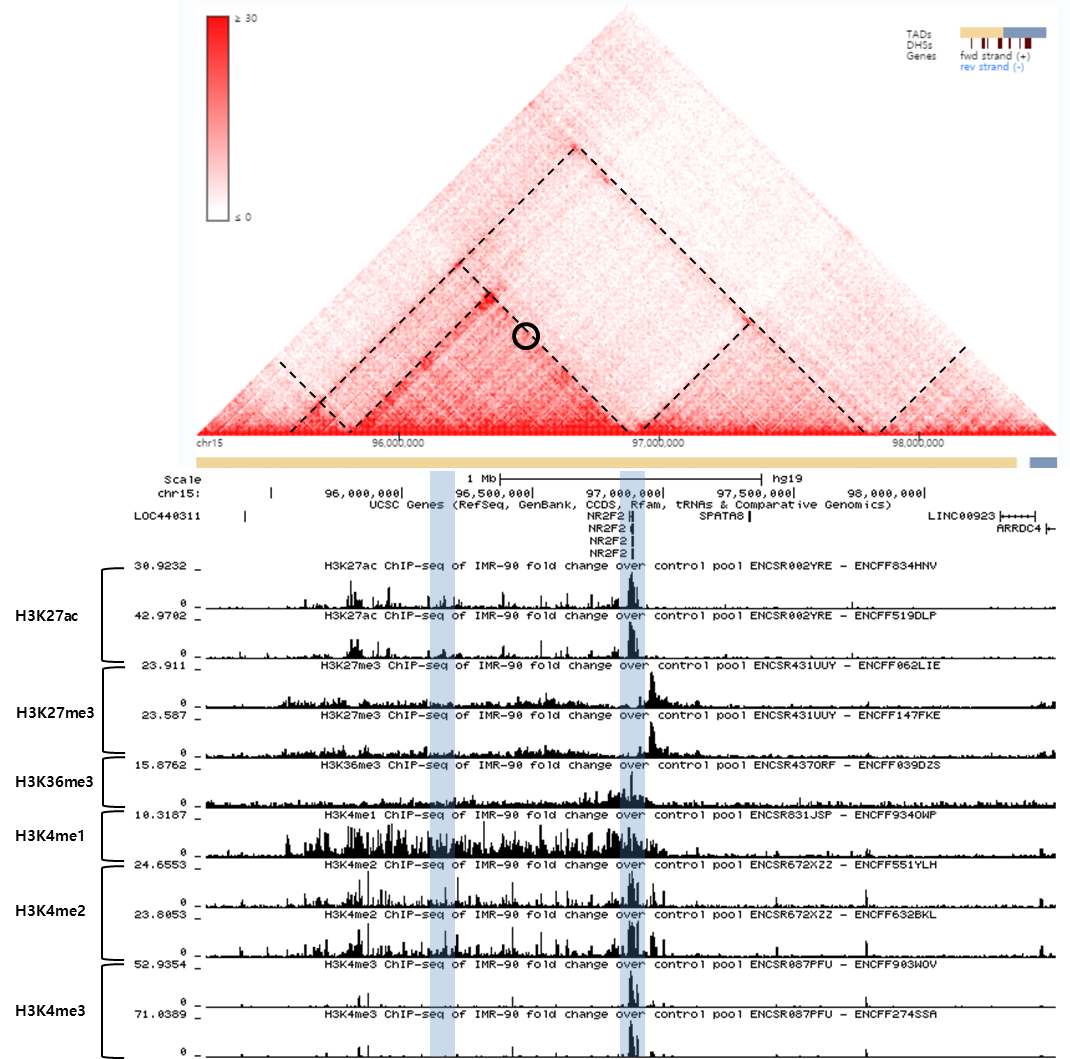
**Supplementary Figure 4. The proportion of phenotypic variance explained by the genotyed SNPs according to disease prevalences ranged from 10-6 to 10-4.**

****

**Supplementary Figure 5. Quantile-quantile and Manhattan plots for the LAM GWAS using the imputed data.** a) The observed distributions of P-values for 5,427,377 genotyped and imputed SNPs are plotted relative to the expected (null) distribution for the Conditional logistic regression (CLR) analysis. The genomic inflation factor was 1.072. b) Manhattan plot. Each dot represents the P-value of a single SNP, plotted on the genome scale at bottom. The Y-axis value is the negative logarithm of the P-value for association between each genotyped SNP and S-LAM. Eighteen SNPs on 15q met genome-wide significance.

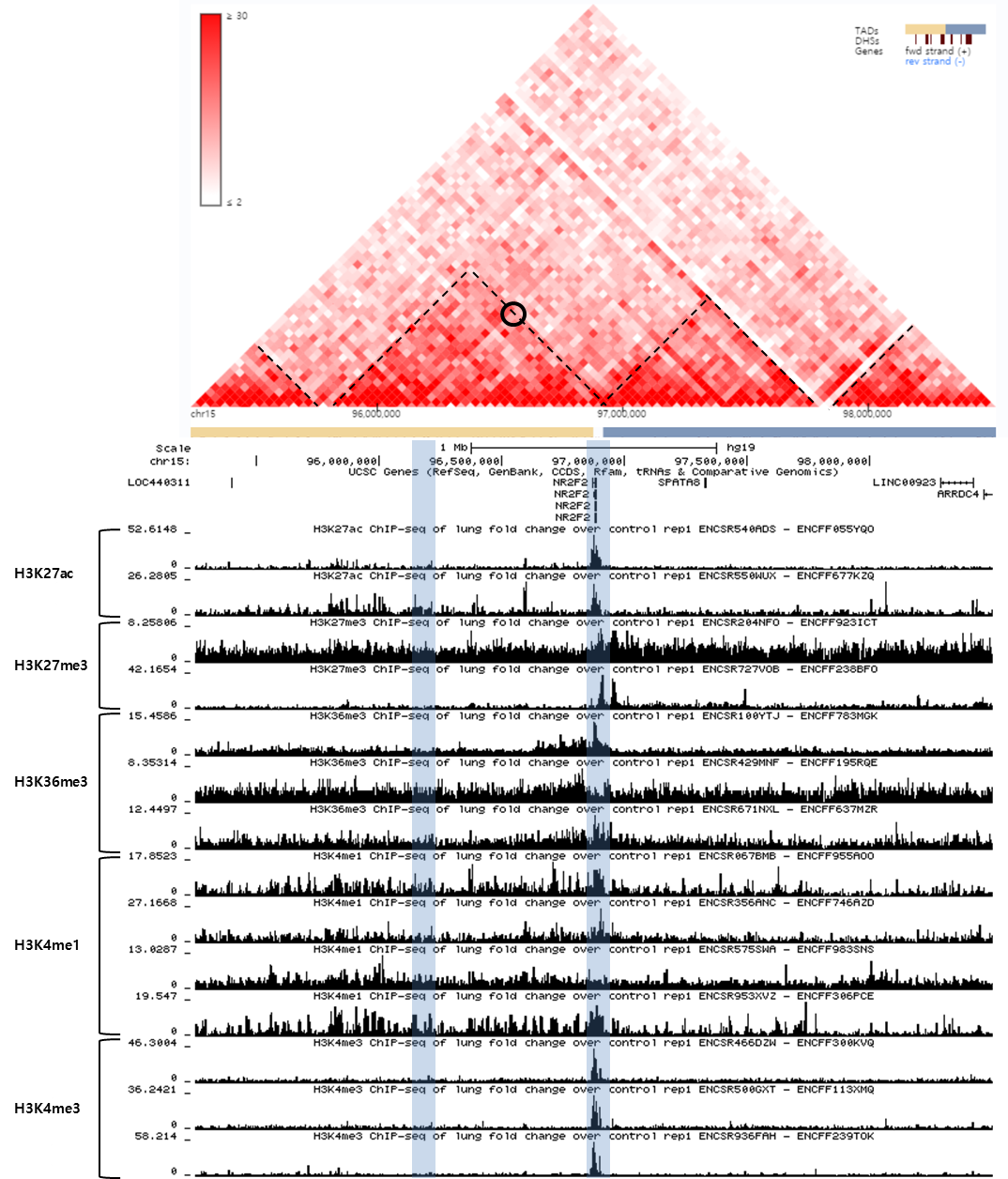
****

**Supplementary Figure 6. Hi-C heatmap and TADs defined in IMR90 cells.** The heatmap shows the degree of physical interaction defined by Hi-C analysis for genomic region pairs from a 3Mb region of chromosome 15q. A deeper red color at the intersection point reflects a greater degree of interaction between the two genomic regions. The dotted lines indicate probable TAD structures in this region. The two blue shaded regions at bottom indicate the genome wide significant SNP region (left) and *NR2F2* (right). The black circle reflects the interaction point between the SNP region and *NR2F2.*

****

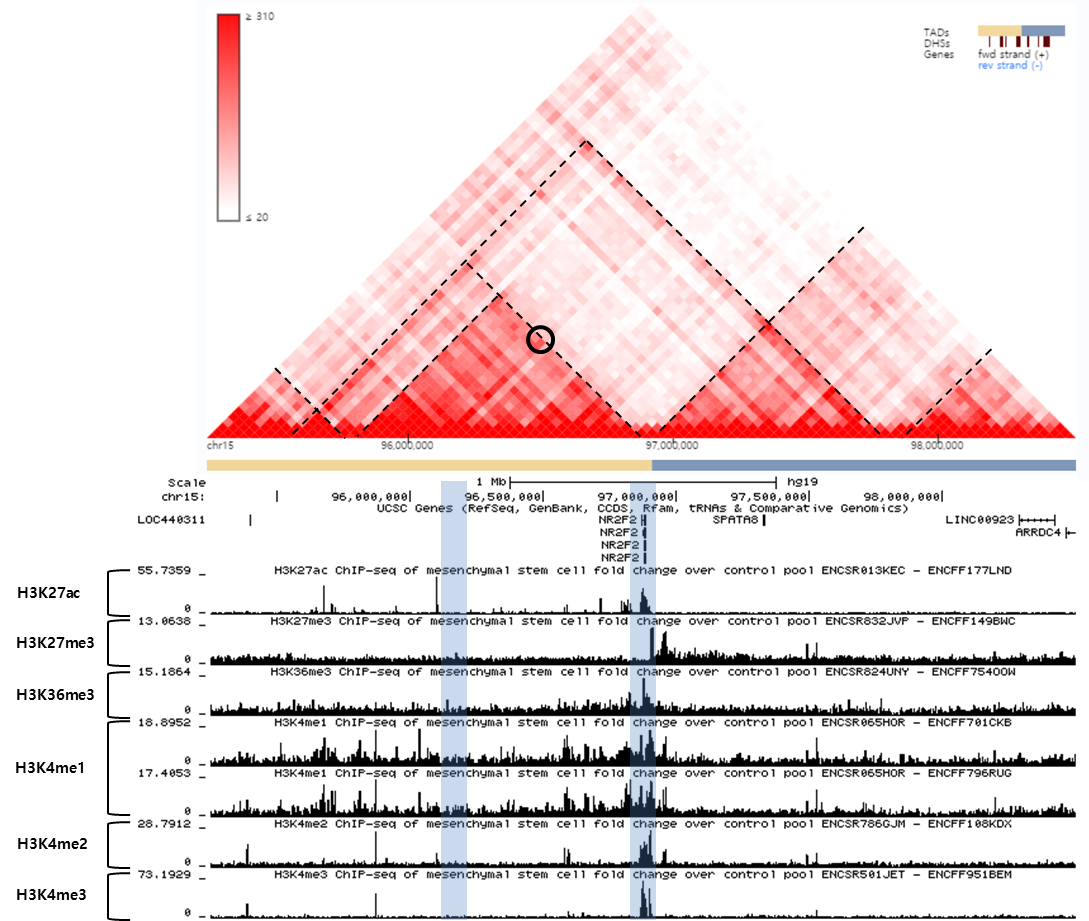
**Supplementary Figure 7. Hi-C heatmap and TADs defined in lung tissue.**

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

****

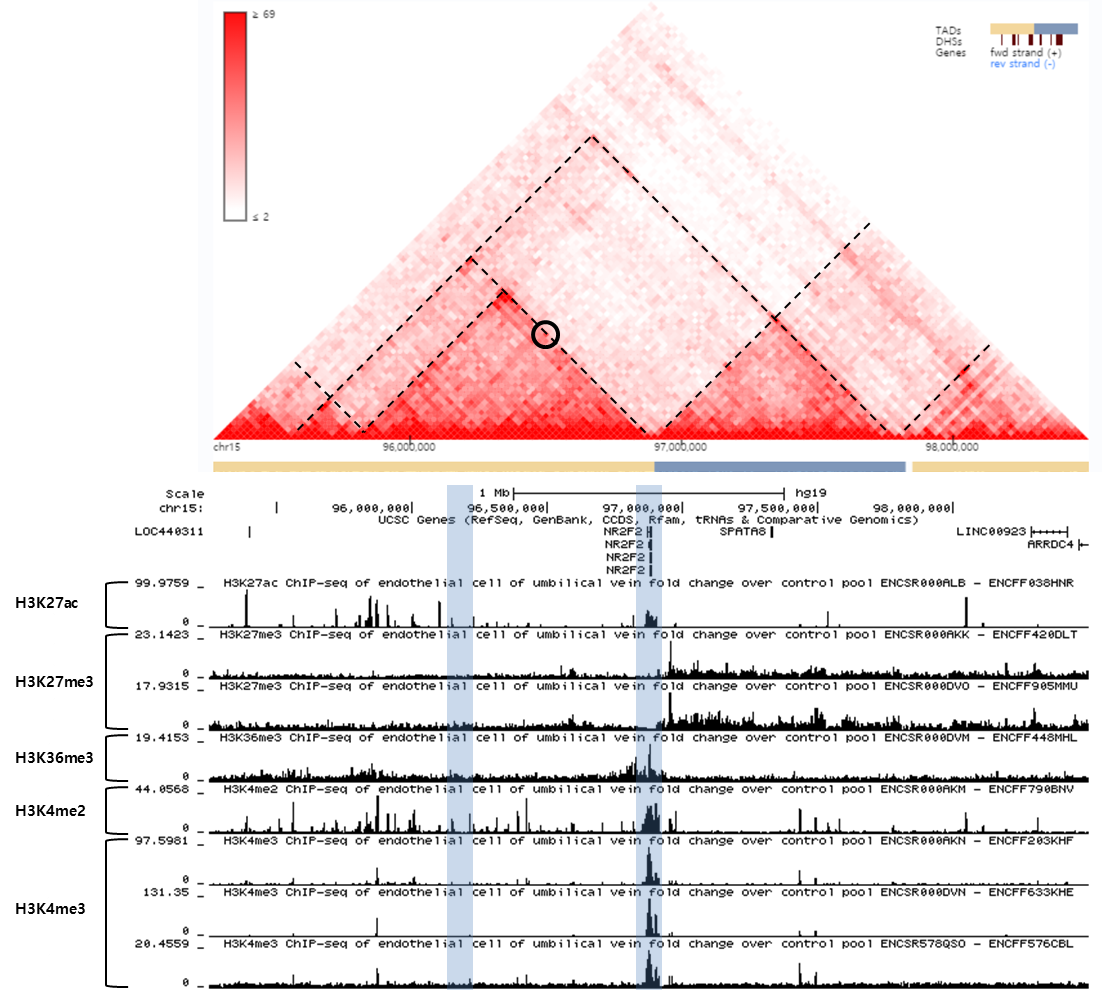
**Supplementary Figure 8. Hi-C heatmap and TADs defined in H1 derived mesenchymal stem cells (h1-MSC) cells.**

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

****

**Supplementary Figure 9. Hi-C heatmap and TADs defined in HUVEC cells.**

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

****

**Supplementary Figure 10. Quantile-quantile plots for classical logistic regression and conditional logistic regression.** We performed classical logistic regression using all cases and controls after QC, including controls dropped due to matching. Two PC scores corresponding to the two largest eigenvalues and age was used as covariates.

