

Low-Frequency Coding Variants at 6p21.33 and 20q11.21 Are Associated with Lung Cancer Risk in Chinese Populations

Guangfu Jin,^{1,2,10} Meng Zhu,^{1,10} Rong Yin,³ Wei Shen,¹ Jia Liu,¹ Jie Sun,¹ Cheng Wang,¹ Juncheng Dai,¹ Hongxia Ma,¹ Chen Wu,⁴ Zhihua Yin,⁵ Jiaqi Huang,⁶ Brandon W. Higgs,⁶ Lin Xu,³ Yihong Yao,⁶ David C. Christiani,⁷ Christopher I. Amos,⁸ Zhibin Hu,^{1,2,3,11} Baosen Zhou,^{5,11} Yongyong Shi,^{9,11} Dongxin Lin,^{4,11} and Hongbing Shen^{1,2,11,*}

Genome-wide association studies have successfully identified a subset of common variants associated with lung cancer risk. However, these variants explain only a fraction of lung cancer heritability. It has been proposed that low-frequency or rare variants might have strong effects and contribute to the missing heritability. To assess the role of low-frequency or rare variants in lung cancer development, we analyzed exome chips representing 1,348 lung cancer subjects and 1,998 control subjects during the discovery stage and subsequently evaluated promising associations in an additional 4,699 affected subjects and 4,915 control subjects during the replication stages. Single-variant and gene-based analyses were carried out for coding variants with a minor allele frequency less than 0.05. We identified three low-frequency missense variants in *BAT2* (rs9469031, c.1544C>T [p.Pro515Leu]; odds ratio [OR] = 0.55, $p = 1.28 \times 10^{-10}$), *FKBP1* (rs200847762, c.410C>T [p.Pro137Leu]; OR = 0.25, $p = 9.79 \times 10^{-12}$), and *BPIFB1* (rs6141383, c.850G>A [p.Val284Met]; OR = 1.72, $p = 1.79 \times 10^{-7}$); these variants were associated with lung cancer risk. rs9469031 in *BAT2* and rs6141383 in *BPIFB1* were also associated with the age of onset of lung cancer ($p = 0.001$ and 0.006 , respectively). *BAT2* and *FKBP1* at 6p21.33 and *BPIFB1* at 20q11.21 were differentially expressed in lung tumors and paired normal tissues. Gene-based analysis revealed that *FKBP1*, in which two independent variants were identified, might account for the association with lung cancer risk at 6p21.33. Our results highlight the important role low-frequency variants play in lung cancer susceptibility and indicate that candidate genes at 6p21.33 and 20q11.21 are potentially biologically relevant to lung carcinogenesis.

Lung cancer is among the most frequently diagnosed cancers and is the leading cause of cancer-related death worldwide.¹ Tobacco smoking is the major cause of lung cancer, whereas genetic factors determine individual predisposition to lung cancer. We and others have identified a subset of loci that are associated with lung cancer risk through genome-wide association studies (GWASs).^{2–10} These variants generally occur at a high frequency (minor allele frequency [MAF] > 0.05) in populations, and the effect of single variants is modest (odds ratios [ORs] = 1.1–1.4 for risk alleles). To date, these known common loci explain only a small fraction of the familial risk of lung cancer, and the remaining missing heritability is uncertain.

GWASs mainly focus on common proxy SNPs that are based on the HapMap Project. Most low-frequency (defined here as a MAF of 0.5%–5%) and rare (MAF < 0.5%) variants were not previously evaluated in most GWASs. An alternative hypothesis is that, unlike common

variants with low penetrance, some low-frequency or rare variants might have strong effects and might contribute to the missing heritability of complex diseases, including cancer. Supporting this hypothesis is that several genes containing known low-frequency or rare missense variants are associated with various cancers: *ATM* (MIM: 607585), *BRIP1* (MIM: 605882), *CHEK2* (MIM: 604373), and *PALB2* (MIM: 610355) for breast cancer;¹¹ *RAD51D* (MIM: 602954) and *BRIP1* for ovarian cancer;^{12,13} and *HOXB13* (MIM: 604607) for prostate cancer.¹⁴ More recently, Wang et al. implicated two large-effect, low-frequency variants—rs11571833 (c.9976A>T [p.Lys3326*]; GenBank: NM_000059) in *BRCA2* (MIM: 600185) and rs17879961 (c.470T>C [p.Ile157Thr]; GenBank: NM_007194.3) in *CHEK2* (MIM: 604373)—in susceptibility to lung cancer in populations of European ancestry on the basis of existing GWAS imputation data;¹⁵ these findings suggest that low-frequency or rare variants in coding regions are important to the missing heritability of lung cancer.

¹Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, Nanjing 211166, China; ²Jiangsu Key Laboratory of Cancer Biomarkers, Prevention, and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, China; ³Jiangsu Key Laboratory of Molecular and Translational Cancer Research, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University Affiliated Cancer Hospital, Nanjing 210009, China; ⁴State Key Laboratory of Molecular Oncology, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China; ⁵Department of Epidemiology, School of Public Health, China Medical University, Shenyang 110001, China; ⁶Medimmune, Gaithersburg, MD 20878, USA; ⁷Department of Environmental Health, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA; ⁸Center for Genomic Medicine, Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH 03755, USA; ⁹Ministry of Education Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Bio-X Institutes, Shanghai Jiao Tong University, Shanghai 200240, China

¹⁰These authors contributed equally to this work

¹¹These authors contributed equally to this work

*Correspondence: hbshen@njmu.edu.cn

<http://dx.doi.org/10.1016/j.ajhg.2015.03.009>. ©2015 by The American Society of Human Genetics. All rights reserved.

Sequencing is an ideal approach for investigating low-frequency or rare variants but has been limited so far because of its cost. The Illumina HumanExome Beadchip (referred to as “exome chip” hereafter) platform has thus been developed to capture low-frequency or rare variants in coding regions on the basis of genetic variants discovered from the whole-exome sequencing of >12,000 individuals. Recently, several groups have validated this platform as an effective complementary approach for determining the genetic basis of complex diseases or traits.^{16–18} To address the role of low-frequency or rare variants in the development of lung cancer, we generated and analyzed exome-chip data for 1,348 lung cancer subjects and 1,998 control subjects and subsequently evaluated promising associations in an additional 4,699 affected subjects and 4,915 control subjects. As a result, we identified three low-frequency missense variants in *BAT2* (MIM: 142580), *FKBP1*, and *BPIFB1*, which are associated with lung cancer risk in Chinese populations.

A three-stage case-control analysis was conducted, and the characteristics of the subjects are summarized in Table S1. In the discovery stage, 1,348 lung cancer subjects and 1,998 control subjects were recruited from Nanjing and the surrounding areas; some of these individuals were also included in our previous GWAS.⁷ In the first replication stage (replication I), 1,115 affected subjects and 1,246 control subjects were recruited according to the same standards as those used in the discovery stage during 2009–2013. In the second replication stage (replication II), a total of 3,584 affected subjects and 3,669 control subjects were recruited from northern China; 2,466 affected subjects and 2,423 control subjects were from Beijing, and 1,118 affected subjects and 1,246 control subjects were from Shenyang. All of these affected subjects were collected from local hospitals and were histopathologically or cytologically confirmed as having lung cancer by at least two pathologists. All control subjects were cancer-free subjects receiving a routine physical examination in a local hospital or participants in a community screening of noncommunicable diseases. All subjects were unrelated ethnic Han Chinese and gave informed consent at recruitment. Smoking information was obtained via interviews; individuals who had smoked an average of one or more cigarettes per day for at least 1 year before recruitment were defined as current smokers, whereas smokers who had quit more than 1 year before recruitment were considered former smokers; otherwise, subjects were considered non-smokers. Current and former smokers were divided into light and heavy smokers according to the median smoking level of 25 pack years (the number of packs of cigarettes smoked per day multiplied by the number of years a person has smoked) among control subjects. This study was approved by the institutional review board of each participating institution.

We successfully genotyped 1,348 lung cancer subjects and 1,998 control subjects by using the Illumina HumanExome Beadchip system and found a total of 247,870 variants. The lung cancer and control subjects

were genotyped together, and the technicians were blinded to the sample status. Genotypes were called by Illumina GenomeStudio software, and the selected variants were re-called by zCall.¹⁹ Systematic quality control of the raw genotyping data was performed to filter unqualified genetic variants and samples (Figure S1). A total of 175,447 variants were excluded from subsequent analysis because they (1) were mitochondrial variants or were located on the X or Y chromosome, (2) had duplicate variants on the chip, (3) were monomorphic in our study subjects, (4) had a call rate of <95%, or (5) presented a p value < 1×10^{-4} in a Hardy-Weinberg equilibrium test among the control subjects. A total of 7 affected subjects and 16 control subjects were excluded because they (1) had an overall genotyping rate of <95%, (2) were duplicates or showed familial relationships ($PI_HAT > 0.25$), or (3) had an extreme heterozygosity rate more than 6 SDs from the mean. Population outliers and stratification were detected with a method based on principal-component analysis. As shown in Figure S2, no individuals were excluded as outliers, and the affected and control subjects were genetically matched. We assessed genotyping consistency on the basis of 37 replicate samples and found an overall concordance rate of 99.98%. Moreover, 1,369 subjects were also scanned with an Affymetrix Genome-Wide Human SNP Array 6.0 in a previous GWAS,⁷ and the concordance rate was 99.93% for 6,660 overlapping variants after quality control. Accordingly, 6 samples and 50 variants with a concordance rate <95% were also excluded. Finally, 72,423 variants in 1,341 affected subjects and 1,982 control subjects were retained for further association analysis.

In this study, we mainly focused on the low-frequency or rare variants with MAFs between 0.1% and 5% when we could call at least six copies of the minor allele in our study samples. On the basis of the following items, we then selected promising variants for further genotyping in the replication stages: (1) variants were in nonsynonymous or splice sites, (2) the single-variant association p value was less than 0.001, (3) variant calling was visually inspected with a clear genotyping cluster, and (4) only one variant was selected when multiple variants were in linkage disequilibrium (LD; $r^2 \geq 0.5$). On the basis of the results from the discovery stage, we genotyped 21 variants in the replication I stage by using SNPscan technology (GeneSky). To obtain positive control samples of minor genotypes for low-frequency or rare variants, we included 192 discovery-stage samples in the replication I stage to ensure the presence of at least two heterozygotes or minor homozygotes, and the concordance rate was 99.4%. In the replication II stage, genotyping was performed with the TaqMan system (Applied Biosystems). Positive and negative control subjects were included in each 384-well plate for quality control. The average concordance rate between duplicate samples was >99%. Genotyping was performed by technicians who were blinded to sample status.

Assuming an additive genetic model, we performed a single-variant association analysis by using a logistic

regression model as implemented in PLINK.²⁰ At the discovery stage, we carried out principal-component analysis with EIGENSOFT²¹ to determine ancestry and population stratification on the basis of 4,604 autosomal ancestry-informative markers included on the exome chips. The top principal component was significant ($p = 0.03$), and we included it (together with age, gender, and smoking level by pack years) in the logistic regression model as a covariate when we estimated ORs and 95% confidence intervals (CIs). We also used the logistic score test²² and the Firth bias-corrected logistic likelihood-ratio test²³ to assess the association results for rare or low-frequency variants. At the replication stages, we used age, gender, and smoking level as covariates. We performed joint analysis to combine the discovery and replication stages and used age, gender, smoking level, and study stage as covariates. We applied conditional analysis to test the independence of genetic variants in each region and used the predefined variant(s) as covariate(s). We performed two gene-based tests using nonsynonymous and splice-site variants with a MAF $< 5\%$ ($n = 43,782$): a simple burden test²⁴ and a sequence kernel association test (SKAT).²⁵ The SKAT was implemented in the sequence kernel association optimal test (SKAT-O).²⁶ We defined statistical significance by using the Bonferroni correction and set the exome-wide significance levels at 2×10^{-7} for single-variant analysis (0.05/250,000 variants) and 2.83×10^{-6} for gene-based analysis (8,840 genes \times 2 tests). The quantile-quantile plot was generated with R v.2.3.1, and regional plots were created with LocusZoom.²⁷ We annotated variants according to GENCODE v.7 coding transcripts,²⁸ dbNSFP v.2.0,²⁹ or documentation files obtained from the Illumina Product Support Files.

We obtained the normalized expression data and clinical information for lung cancer samples from The Cancer Genome Atlas (TCGA) on July 8, 2014. A total of 107 paired samples (lung tumor with adjacent normal tissues) were used in this analysis. The paired t test was used to test whether gene expression differed between tumors and the adjacent normal tissues. Seventy-nine out of the 107 individuals had clinical follow-up information and were included in the survival analysis. The mRNA expression ratio of the tumor and adjacent normal tissues was calculated with read counts normalized to RNA sequencing (RNA-seq) by expectation maximization (RSEM). The individuals were divided into two groups on the basis of the median value of the expression ratio for each gene. The Kaplan-Meier method and the log-rank test were used for evaluating the association between gene expression and survival.

After quality control, 72,423 polymorphic variants were included in the exome chip (29.2% of 247,870 variants) performed on 3,323 Chinese Han subjects. The detailed distributions of these variants are summarized in Table S2. In the single-variant association analysis, the quantile-quantile plot revealed a good match between the distributions of the observed and expected p values (Figure S3).

A small genomic-control inflation factor (λ) of 1.04, which decreased to 1.01 after the removal of variants that showed a cluster of association signals at 6p22.2–6p21.31, indicated a low possibility of false-positive associations resulting from population stratification. However, at the discovery stage, we did not find any variants associated with lung cancer risk at our predefined exome-wide significance level ($p < 2 \times 10^{-7}$) (Figure S4), which was probably due to limited statistical power, especially for low-frequency or rare variants (Figure S5).

We then conducted a two-stage replication study for promising nonsynonymous or splice-site variants with a MAF from 0.1% to 5%. At the replication I stage, we genotyped 21 variants with clear cluster plots (Figure S6) in 1,115 lung cancer subjects and 1,246 control subjects (Table S3). As a result, four variants with a p value < 0.05 at the replication I stage showed consistent associations with variants found at the discovery stage (Table S3). At the replication II stage, we genotyped these four variants and found that the associations were consistent for all four variants and that two of them had a p value < 0.05 (Table S3). When combining the results from the discovery and replication stages, we found three low-frequency, missense variants at *BAT2* (rs9469031, c.1544C>T [p.Pro515Leu]; OR = 0.55, $p = 1.28 \times 10^{-10}$), *FKBP1* (rs200847762, c.410C>T [p.Pro137Leu]; OR = 0.25, $p = 9.79 \times 10^{-12}$), and *BPIFB1* (rs6141383, c.850G>A [p.Val284Met]; OR = 1.72, $p = 1.79 \times 10^{-7}$) to be significantly associated with lung cancer risk and to have p values less than 2×10^{-7} (Table 1). We also found a promising *HIST1H1E* variant (rs2298090, c.455A>G [p.Lys152Arg]; OR = 0.51) with a combined p value of 2.95×10^{-7} . The MAFs of these four variants were also less than 0.05 in other populations, and two of the variants (rs9469031 and rs2298090) were polymorphic but not associated with lung cancer risk according to in silico replication in populations of European ancestry (Table S4).¹⁵

We then analyzed the relationships between the four identified variants and the onset ages of the lung cancer case subjects. We observed that rs9469031 and rs6141383 were significantly associated with onset age after adjusting for gender and smoking level ($p = 0.001$ and 0.006 , respectively; Figure 1). Lung cancer subjects carrying the protective allele (T) of rs9469031 had a higher onset age (62.12 ± 10.56 years) than those without the protective allele (59.52 ± 10.18 years), and those with the risk allele (A) of rs6141383 had a lower onset age (58.63 ± 9.23 years) than those without the risk allele (59.64 ± 10.23 years). In addition, we did not find significantly different associations between the subgroups divided by age, gender, smoking, or histology (Table S5).

We then carefully evaluated genetic variants in the flanking regions (1 Mb upstream or downstream) of rs9469031, rs200847762, rs2298090, and rs6141383. As shown in Figure S7, four variants, including the identified variant (rs200847762) at *FKBP1*, had a lower p value than that of rs9469031, one variant (rs138097862) failed to be

Table 1. The Identified Low-Frequency Variants Associated with Lung Cancer Risk

Chr	Gene	Variant ID	Major/Minor Allele	Variant	Stage	Affected Subjects ^a		Control Subjects ^a		OR (95% CI) ^b	p Value ^c
						MAF	MAF				
6p21.33	BAT2	rs9469031	C/T	c.1544C>T (p.Pro515Leu) (GenBank: NM_004638)	discovery	1,291/50/0	1,840/140/2	0.019	0.036	0.52 (0.37–0.73)	1.54 × 10 ⁻⁴
					replication I	1,066/42/6	1,150/85/10	0.024	0.042	0.61 (0.44–0.83)	1.71 × 10 ⁻³
					replication II	3,429/78/1	3,502/129/0	0.011	0.018	0.62 (0.46–0.84)	1.71 × 10 ⁻³
					combined ^d	–	–	–	–	0.55 (0.46–0.66)	1.28 × 10 ⁻¹⁰
6p21.33	FKBP1	rs200847762	G/A	c.410C>T (p.Pro137Leu) (GenBank: NM_022110)	discovery	1,329/12/0	1,908/73/1	0.004	0.019	0.21 (0.11–0.39)	1.84 × 10 ⁻⁶
					replication I	1,094/6/0	1,206/32/1	0.003	0.014	0.19 (0.08–0.46)	2.24 × 10 ⁻⁴
					replication II	3,566/15/0	3,642/23/1	0.002	0.003	0.66 (0.34–1.28)	0.216
					combined ^d	–	–	–	–	0.25 (0.17–0.37)	9.80 × 10 ⁻¹²
6p22.2	HIST1H1E	rs2298090	A/G	c.455A>G (p.Lys152Arg) (GenBank: NM_005321)	discovery	1,325/16/0	1,904/77/1	0.006	0.020	0.32 (0.19–0.56)	6.16 × 10 ⁻⁵
					replication I	1,073/27/1	1,178/54/3	0.013	0.024	0.56 (0.36–0.87)	9.80 × 10 ⁻³
					replication II	3,385/44/0	3,492/59/0	0.006	0.008	0.67 (0.44–1.02)	6.03 × 10 ⁻²
					combined ^d	–	–	–	–	0.51 (0.39–0.66)	2.95 × 10 ⁻⁷
20q11.21	BPIFB1	rs6141383	G/A	c.850G>A (p.Val284Met) (GenBank: NM_033197)	discovery	1,277/62/2	1,934/48/0	0.025	0.012	2.00 (1.36–2.95)	4.80 × 10 ⁻⁴
					replication I	1,065/42/3	1,209/31/0	0.022	0.013	1.68 (1.07–2.63)	2.30 × 10 ⁻²
					replication II	3,374/133/0	3,368/87/0	0.019	0.013	1.64 (1.24–2.17)	6.20 × 10 ⁻⁴
					combined ^d	–	–	–	–	1.72 (1.40–2.10)	1.79 × 10 ⁻⁷

Abbreviations are as follows: Chr, chromosomal region; MAF, minor allele frequency; CI, confidence interval.

^aMajor homozygote/heterozygote/minor homozygote.

^bDerived from the logistic regression model after adjustment for age, gender, pack years of smoking, and the top principal component (for the discovery stage only) under the assumption of an additive genetic model.

^cDerived from the logistic regression model adjusting for age, gender, pack years of smoking, and the top principal component (for the discovery stage only) under the assumption of an additive genetic model.

^dThe joint analysis was performed to combine the discovery and replication stages with age, gender, smoking level, and study stage as covariates.

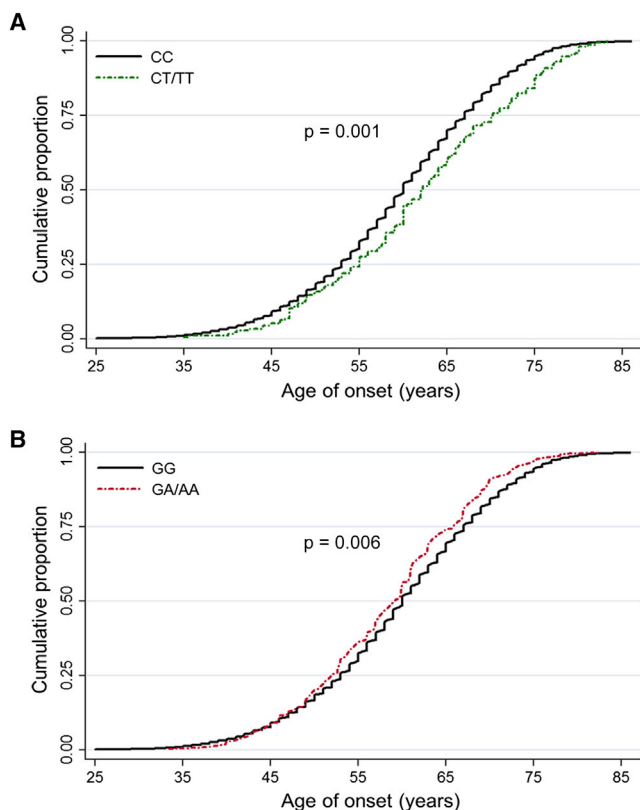


Figure 1. The Relationships between rs9469031 in *BAT2* and rs6141383 in *BPIFB1* and Age of Onset in Individuals with Lung Cancer

Individuals carrying rs9469031 CT/TT genotypes (A), which were associated with a decreased risk of lung cancer, were older at onset (62.12 ± 10.56 years) than those with CC genotypes (59.52 ± 10.18 years, $p = 0.001$ after adjustment for gender and smoking levels). Individuals carrying rs6141383 GA/AA genotypes (B), which were associated with an increased risk of lung cancer, were younger at onset (58.63 ± 9.23 years) than those with GG genotypes (59.64 ± 10.23 years, $p = 0.006$ after adjustment for gender and smoking levels).

replicated at the replication stages, and two variants (rs117160266, c.353A>G [p.Asn28Ser], in *FKBP1L* and rs9469057, c.205G>C [p.Ala8Pro], in *HSPA1L* [MIM: 140559]) were in strong LD with rs9469031 ($r^2 = 0.96$ and 0.99 , respectively) (Table S6). The associations of these two highly correlated variants were abolished after conditioning on rs9469031 (Table S6). We did not find any associations that were more prominent than those of rs200847762, rs2298090, and rs6141383 for their respective regions (Figure S7). There were no other variants in strong LD ($r^2 > 0.5$) with these three variants as genotyped on the exome chip (Table S6).

We further conducted gene-based analysis by using the SKAT-O and burden tests for variants with a MAF < 0.05 and found a significant association between *FKBP1L* and lung cancer risk in both tests ($p = 1.29 \times 10^{-9}$ and 2.0×10^{-10} , respectively; Table 2). As shown in Figure S8, three coding variants of *FKBP1L* were included in the gene-based analysis. In the single-variant analysis, the variants rs200847762 and rs117160266 (tagged by rs9469031 at

BAT2 with $r^2 = 0.96$) were in low LD ($r^2 < 0.1$) and were independently associated with lung cancer risk (Table 1 and Table S6). After we conditioned on either of these two variants, the significance of *FKBP1L* was partially decreased, and the signal was abolished after we conditioned on both variants (Table 2). These results suggest that the gene-based signal of *FKBP1L* is driven by rs200847762 and rs117160266. We also replicated the association between *FKBP1L* and lung cancer risk in the gene-based analyses of the replication stages ($p < 0.001$) by using genotyping data for rs200847762 and rs9469031 (Table 2).

The missense variant rs9469031 in *BAT2* is located at 6p21.33, which is part of the human leukocyte antigen (HLA) region that was initially identified as a lung cancer susceptibility region and was tagged by common variants (rs3117582 and rs3131379) in GWASs involving subjects of European ancestry.⁶ The association was confirmed in some follow-up studies,^{30,31} but not all of these studies involved populations of European ancestry.³² However, studies based on East Asian populations consistently failed to replicate the association^{7,10} because none of the identified variants in European populations are polymorphic in East Asian populations. Notably, although the initially identified variant, rs3117582, was not associated with lung cancer risk in African Americans, the minor allele of missense variant rs2736158 (c.4140G>C [p.Gly1285Ala] in exon 16) in *BAT2* was associated with a decreased risk of squamous cell lung cancer (OR = 0.64, 95% CI = 0.48–0.85).³³ Of interest, this variant was consistently associated with lung cancer risk at our discovery stage (OR = 0.82, 95% CI = 0.70–0.96, $p = 0.011$; Table S6). In the samples examined during the discovery stage, the identified low-frequency variant rs9469031 (MAF = 0.036 in control subjects) was in low LD with the common variant rs2736158 (MAF = 0.139 in control subjects), given that r^2 was 0.19; however, all of the individuals carrying the protective allele of rs9469031 also had the protective allele of rs2736158, yielding a $D' = 1.00$ (Table S6). The association of rs2736158 was abolished after we conditioned on rs9469031 (OR = 0.92, 95% CI = 0.78–1.09, $p = 0.331$), whereas the association of rs9469031 changed modestly (OR = 0.56, 95% CI = 0.39–0.81, $p = 0.002$). Collectively, these findings indicate that the association of the common variant rs2736158 might be driven by association of the low-frequency variant rs9469031. In addition, two independent common variants rs3817963 and rs2395185 at 6p21.32 (837 kb and 772 kb away from rs9469031, respectively) are reported to be associated with lung cancer risk in Japanese¹⁰ and Chinese⁹ populations, respectively. The absence of LD between the low-frequency variant and these two common variants ($r^2 < 0.01$) indicates that the signal of rs9469031 might be independent from the signals reported in Eastern Asian populations.

Because there are at least two variants (rs117160266 in *FKBP1L* and rs9469057 in *HSPA1L*) in strong LD with

Table 2. Association between *FKBPL*, at 6p21.33, and Lung Cancer Risk according to Gene-Based Analysis

Test	Stage	Variants Included	p Value ^a	Conditional Analysis	
				Variants Included (p Value for Single-Variant Test)	p Value for Gene-Based Analysis ^b
SKAT-O	discovery	rs200847762, rs117160266, and rs142997752	1.29×10^{-9}	rs200847762 (1.84×10^{-6})	3.44×10^{-5}
				rs117160266 (4.91×10^{-5})	3.26×10^{-6}
				rs200847762 and rs117160266	0.106
	replication I	rs200847762 and rs117160266 ^c	1.15×10^{-6}	–	–
	replication II	rs200847762 and rs117160266 ^c	5.55×10^{-4}	–	–
Burden	discovery	rs200847762, rs117160266, and rs142997752	2.00×10^{-10}	rs200847762 (1.84×10^{-6})	2.75×10^{-5}
				rs117160266 (4.91×10^{-5})	1.36×10^{-7}
				rs200847762 and rs117160266	0.081
	replication I	rs200847762 and rs117160266 ^c	4.81×10^{-7}	–	–
	replication II	rs200847762 and rs117160266 ^c	9.61×10^{-4}	–	–

^aAfter adjustment for age, gender, pack years of smoking, and the top principal component (for the discovery stage only).
^bAfter additional adjustment for the corresponding lead variant(s).
^cThe genotypes of rs9469031 were used here because rs9469031 was highly correlated with rs117160266 at the discovery stage ($r^2 = 0.96$).

rs9469031 in *BAT2*, it is important to determine which is the causal variant at 6p21.33. *BAT2* (also known as *PRRC2A*) is in a cluster of HLA-B-associated transcripts (*BAT1-5*) in the human major histocompatibility complex class III region.³⁴ *FKBPL* (FK506-binding protein-like), a divergent member of the immunophilin family, is implicated in the regulation of tumor growth and angiogenesis and might act as a cancer prognostic marker and a therapeutic target.^{35,36} *HSPA1L* encodes heat-shock 70-kDa protein 1-like, which in combination with other heat-shock proteins stabilizes existing proteins against aggregation and protects against DNA damage.³⁷ Functional variants in *HSPA1B* have been associated with lung cancer risk and survival.³⁸ RNA-seq data from TCGA indicated that *BAT2* and *FKBPL* were upregulated in 84.1% ($p = 8.50 \times 10^{-16}$) and 91.6% ($p = 2.25 \times 10^{-18}$) of lung tumor tissues, respectively, and *HSPA1L* was significantly downregulated (79.4%, $p = 1.56 \times 10^{-10}$) (Figure S9). We also assessed the clinical relevance of these three genes and did not find significant associations between the mRNA levels and lung cancer survival (Figure S10). Of interest, for rs9469057 at *HSPA1L*, the substitution p.Ala8Pro is predicted to be damaging (Table S7). Nevertheless, gene-based analysis supported the conclusion that *FKBPL*, which has two independent risk-related variants, might be the lung-cancer-associated gene at 6p21.33.

The LD of common variants that were associated with lung cancer risk at chromosome 6 from 26–34 Mb (across 6p22.2–6p21.31) has been reported to extend over long distances.⁶ Genetic variants in this region have been associated with multiple diseases or traits, especially those relating to inflammation and/or immune-related diseases, such as allergies,³⁹ chronic hepatitis B virus infection,⁴⁰ ulcerative colitis,⁴¹ multiple myeloma,⁴² and diffuse large

B cell lymphoma.⁴³ Consistent with this finding, more than 100 low-frequency or rare variants in this region were observed to be associated with lung cancer risk ($p < 0.01$) in our study, yielding an obvious peak on the Manhattan plot (Figure S4). In addition to the two independent 6p21.33 loci that were described above, we also found another promising low-frequency variant: rs2298090 in *HIST1H1E* at 6p22.2. p.Lys152Arg, the amino acid change resulting from this variant, is predicted to be damaging (Table S7). *HIST1H1E* encodes a member of the linker histone H1 family, which interacts with linker DNA between nucleosomes and functions in the compaction of chromatin into higher-order structures.⁴⁴ TCGA RNA-seq data showed that the mRNA levels of *HIST1H1E* were higher in lung tumors than in the paired normal tissues (83.2%, $p = 2.12 \times 10^{-7}$; Figure S9); however, the mRNA levels of *HIST1H1E* were not associated with lung cancer prognosis (Figure S10).

The variant rs6141383 in *BPIFB1* localizes to 20q11.21, a region that has not been reported to be associated with lung cancer susceptibility. As predicted, the substitution p.Val284Met is damaging to *BPIFB1* (Table S7). *BPIFB1* (or *LPLUNC1*) is a secretory protein that is predominantly present in lung tissues and is present at low levels in other organs.⁴⁵ *BPIFB1* has been implicated in host innate immune defenses against pulmonary infection⁴⁶ and in the pathogenesis of chronic lung diseases, such as cystic fibrosis and interstitial lung disease.⁴⁷ Previous studies revealed that *BPIFB1*, which is downregulated in nasopharyngeal carcinoma (NPC),⁴⁸ can inhibit inflammation and NPC growth by downregulating the STAT3 pathway.⁴⁹ *BPIFB1* was also significantly downregulated in lung tumors (67.3%, $p = 0.0025$; Figure S9), which is consistent with the prediction that the lung-cancer-associated risk

allele of rs6141383 damages BPIFB1. However, high mRNA levels were associated with poor prognosis in individuals with lung cancer ($p = 0.013$; Figure S10), possibly suggesting that BPIFB1 has dual roles in the development and progression of lung cancer. Although the role of BPIFB1 in lung carcinogenesis is limited, these findings indicate that BPIFB1 might constitute part of the protective immunity shield that overlies pulmonary inflammation.

In the current study, we identified three low-frequency variants that were associated with lung cancer risk in Chinese populations; together with recent findings based on populations of European ancestry, these results show that low-frequency variants also contribute to lung cancer susceptibility. In particular, our results reveal that lung cancer susceptibility loci across 6p22.2–6p21.31, which have been reported in populations of European ancestry, are also present in East Asian populations and that *FKBP1*, for which two independent risk-related variants were found, might be a lung-cancer-associated gene at 6p21.33. We also observed a relationship between lung cancer and *BPIFB1* at 20q11.21. These genetic associations, together with differences in expression in lung cancer tissues, indicate that genes at 6p21.33 and 20q11.21 might play key roles in lung carcinogenesis.

Supplemental Data

Supplemental Data include ten figures and seven tables and can be found with this article online at <http://dx.doi.org/10.1016/j.ajhg.2015.03.009>.

Acknowledgments

This work was funded by the National Key Basic Research Program (grants 2011CB503805, 2013CB910304, and 2013CB911400), the State Key Program of National Natural Science of China (grant 81230067), the National Distinguished Youth Science Foundation of China (grant 81225020), the National Outstanding Youth Science Foundation of China (grant 81422042), the Science Foundation for Distinguished Young Scholars of Jiangsu (grants BK2012042 and BK20130042), the National Natural Science Foundation of China (grant 81270044), the Jiangsu Specially Appointed Professor Project, the Natural Science Foundation of Jiangsu Province (grant BK20130060), the Key Grant of the Natural Science Foundation of Jiangsu Higher Education Institutions (11KJA330001), the National Program for Support of Top-Notch Young Professionals from the Organization Department of the Communist Party of China Central Committee, the Jiangsu Province Clinical Science and Technology Projects (grant BL2012008), and the Priority Academic Program for the Development of Jiangsu Higher Education Institutions (Public Health and Preventive Medicine). The authors wish to thank all the study participants, research staff, and students who participated in this work.

Received: November 17, 2014

Accepted: March 24, 2015

Published: April 30, 2015

Web Resources

The URLs for the data presented herein are as follows:

dbNSFP, <https://sites.google.com/site/jpopgen/dbNSFP>
Documentation files obtained from Illumina Product Support Files, ftp://ussd-ftp.illumina.com/downloads/ProductFiles/HumanExome-12/HumanExome-12v1-2_A.annotated.txt.
EIGENSOFT, http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html
GENCODE, <http://www.genencodegenes.org/>
LocusZoom, <http://csg.sph.umich.edu/locuszoom/>
OMIM, <http://www.omim.org/>
PLINK, <http://pngu.mgh.harvard.edu/~purcell/plink/>
The Cancer Genome Atlas, <https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm>

References

1. Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* **61**, 69–90.
2. Amos, C.I., Wu, X., Broderick, P., Gorlov, I.P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., et al. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* **40**, 616–622.
3. Hung, R.J., McKay, J.D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., et al. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633–637.
4. McKay, J.D., Hung, R.J., Gaborieau, V., Boffetta, P., Chabrier, A., Byrnes, G., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., et al.; EPIC Study (2008). Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.* **40**, 1404–1406.
5. Thorgeirsson, T.E., Geller, E., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K.P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638–642.
6. Wang, Y., Broderick, P., Webb, E., Wu, X., Vijayakrishnan, J., Matakidou, A., Qureshi, M., Dong, Q., Gu, X., Chen, W.V., et al. (2008). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.* **40**, 1407–1409.
7. Hu, Z., Wu, C., Shi, Y., Guo, H., Zhao, X., Yin, Z., Yang, L., Dai, J., Hu, L., Tan, W., et al. (2011). A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat. Genet.* **43**, 792–796.
8. Dong, J., Hu, Z., Wu, C., Guo, H., Zhou, B., Lv, J., Lu, D., Chen, K., Shi, Y., Chu, M., et al. (2012). Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nat. Genet.* **44**, 895–899.
9. Lan, Q., Hsiung, C.A., Matsuo, K., Hong, Y.C., Seow, A., Wang, Z., Hosgood, H.D., 3rd, Chen, K., Wang, J.C., Chatterjee, N., et al. (2012). Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat. Genet.* **44**, 1330–1335.
10. Shiraishi, K., Kunitoh, H., Daigo, Y., Takahashi, A., Goto, K., Sakamoto, H., Ohnami, S., Shimada, Y., Ashikawa, K., Saito, A., et al. (2012). A genome-wide association study identifies

- two new susceptibility loci for lung adenocarcinoma in the Japanese population. *Nat. Genet.* *44*, 900–903.
11. Stratton, M.R., and Rahman, N. (2008). The emerging landscape of breast cancer susceptibility. *Nat. Genet.* *40*, 17–22.
 12. Loveday, C., Turnbull, C., Ramsay, E., Hughes, D., Ruark, E., Frankum, J.R., Bowden, G., Kalmrzaev, B., Warren-Perry, M., Snape, K., et al.; Breast Cancer Susceptibility Collaboration (UK) (2011). Germline mutations in RADS1D confer susceptibility to ovarian cancer. *Nat. Genet.* *43*, 879–882.
 13. Rafnar, T., Gudbjartsson, D.F., Sulem, P., Jonasdottir, A., Sigurdsson, A., Jonasdottir, A., Besenbacher, S., Lundin, P., Stacey, S.N., Gudmundsson, J., et al. (2011). Mutations in BRIP1 confer high risk of ovarian cancer. *Nat. Genet.* *43*, 1104–1107.
 14. Ewing, C.M., Ray, A.M., Lange, E.M., Zuhlke, K.A., Robbins, C.M., Tembe, W.D., Wiley, K.E., Isaacs, S.D., Johng, D., Wang, Y., et al. (2012). Germline mutations in HOXB13 and prostate-cancer risk. *N. Engl. J. Med.* *366*, 141–149.
 15. Wang, Y., McKay, J.D., Rafnar, T., Wang, Z., Timofeeva, M.N., Broderick, P., Zong, X., Laplana, M., Wei, Y., Han, Y., et al. (2014). Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat. Genet.* *46*, 736–741.
 16. Huyghe, J.R., Jackson, A.U., Fogarty, M.P., Buchkovich, M.L., Stančáková, A., Stringham, H.M., Sim, X., Yang, L., Fuchsberger, C., Cederberg, H., et al. (2013). Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.* *45*, 197–201.
 17. Auer, P.L., Teumer, A., Schick, U., O’Shaughnessy, A., Lo, K.S., Chami, N., Carlson, C., de Denu, S., Dubé, M.P., Haessler, J., et al. (2014). Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. *Nat. Genet.* *46*, 629–634.
 18. Peloso, G.M., Auer, P.L., Bis, J.C., Voorman, A., Morrison, A.C., Stitzel, N.O., Brody, J.A., Khetarpal, S.A., Crosby, J.R., Forrage, M., et al.; NHLBI GO Exome Sequencing Project (2014). Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* *94*, 223–232.
 19. Goldstein, J.I., Crenshaw, A., Carey, J., Grant, G.B., Maguire, J., Fromer, M., O’Dushlaine, C., Moran, J.L., Chambert, K., Stevens, C., et al.; Swedish Schizophrenia Consortium; ARRA Autism Sequencing Consortium (2012). zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* *28*, 2543–2545.
 20. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* *81*, 559–575.
 21. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* *38*, 904–909.
 22. Lin, D.Y., and Tang, Z.Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* *89*, 354–367.
 23. Wang, X. (2014). Firth logistic regression for rare variant association tests. *Front. Genet.* *5*, 187.
 24. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* *83*, 311–321.
 25. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* *89*, 82–93.
 26. Lee, S., Wu, M.C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* *13*, 762–775.
 27. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* *26*, 2336–2337.
 28. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* *22*, 1760–1774.
 29. Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum. Mutat.* *34*, E2393–E2402.
 30. Landi, M.T., Chatterjee, N., Yu, K., Goldin, L.R., Goldstein, A.M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., et al. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am. J. Hum. Genet.* *85*, 679–691.
 31. Broderick, P., Wang, Y., Vijaykrishnan, J., Matakidou, A., Spitz, M.R., Eisen, T., Amos, C.I., and Houlston, R.S. (2009). Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res.* *69*, 6633–6641.
 32. Truong, T., Hung, R.J., Amos, C.I., Wu, X., Bickeböller, H., Rosenberger, A., Sauter, W., Illig, T., Wichmann, H.E., Risch, A., et al. (2010). Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. *J. Natl. Cancer Inst.* *102*, 959–971.
 33. Walsh, K.M., Gorlov, I.P., Hansen, H.M., Wu, X., Spitz, M.R., Zhang, H., Lu, E.Y., Wenzlaff, A.S., Sison, J.D., Wei, C., et al. (2013). Fine-mapping of the 5p15.33, 6p22.1-p21.31, and 15q25.1 regions identifies functional and histology-specific lung cancer susceptibility loci in African-Americans. *Cancer Epidemiol. Biomarkers Prev.* *22*, 251–260.
 34. Spies, T., Blanck, G., Bresnahan, M., Sands, J., and Strominger, J.L. (1989). A new cluster of genes within the human major histocompatibility complex. *Science* *243*, 214–217.
 35. Robson, T., and James, I.F. (2012). The therapeutic and diagnostic potential of FKBPL; a novel anticancer protein. *Drug Discov. Today* *17*, 544–548.
 36. McKeen, H.D., Brennan, D.J., Hegarty, S., Lanigan, F., Jirstrom, K., Byrne, C., Yakkundi, A., McCarthy, H.O., Gallagher, W.M., and Robson, T. (2011). The emerging role of FK506-binding proteins as cancer biomarkers: a focus on FKBPL. *Biochem. Soc. Trans.* *39*, 663–668.
 37. Singh, R., Kolvraa, S., and Rattan, S.I. (2007). Genetics of human longevity with emphasis on the relevance of HSP70 as candidate genes. *Front. Biosci.* *12*, 4504–4513.
 38. Guo, H., Deng, Q., Wu, C., Hu, L., Wei, S., Xu, P., Kuang, D., Liu, L., Hu, Z., Miao, X., et al. (2011). Variations in HSPA1B at 6p21.3 are associated with lung cancer risk and prognosis in Chinese populations. *Cancer Res.* *71*, 7576–7586.

39. Hinds, D.A., McMahon, G., Kiefer, A.K., Do, C.B., Eriksson, N., Evans, D.M., St Pourcain, B., Ring, S.M., Mountain, J.L., Francke, U., et al. (2013). A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. *Nat. Genet.* *45*, 907–911.
40. Hu, Z., Liu, Y., Zhai, X., Dai, J., Jin, G., Wang, L., Zhu, L., Yang, Y., Liu, J., Chu, M., et al. (2013). New loci associated with chronic hepatitis B virus infection in Han Chinese. *Nat. Genet.* *45*, 1499–1503.
41. Juyal, G., Negi, S., Sood, A., Gupta, A., Prasad, P., Senapati, S., Zaneveld, J., Singh, S., Midha, V., van Sommeren, S., et al. (2015). Genome-wide association scan in north Indians reveals three novel HLA-independent risk loci for ulcerative colitis. *Gut* *64*, 571–579.
42. Chubb, D., Weinhold, N., Broderick, P., Chen, B., Johnson, D.C., Försti, A., Vijayakrishnan, J., Migliorini, G., Dobbins, S.E., Holroyd, A., et al. (2013). Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nat. Genet.* *45*, 1221–1225.
43. Cerhan, J.R., Berndt, S.I., Vijai, J., Ghesquières, H., McKay, J., Wang, S.S., Wang, Z., Yeager, M., Conde, L., de Bakker, P.I., et al. (2014). Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. *Nat. Genet.* *46*, 1233–1238.
44. Happel, N., and Doenecke, D. (2009). Histone H1 and its isoforms: contribution to chromatin structure and function. *Gene* *431*, 1–12.
45. Shum, A.K., Alimohammadi, M., Tan, C.L., Cheng, M.H., Metzger, T.C., Law, C.S., Lwin, W., Perheentupa, J., Bour-Jordan, H., Carel, J.C., et al. (2013). BPIFB1 is a lung-specific autoantigen associated with interstitial lung disease. *Sci. Transl. Med.* *5*.
46. Shin, O.S., Uddin, T., Citorik, R., Wang, J.P., Della Pelle, P., Krادين, R.L., Bingle, C.D., Bingle, L., Camilli, A., Bhuiyan, T.R., et al. (2011). LPLUNC1 modulates innate immune responses to *Vibrio cholerae*. *J. Infect. Dis.* *204*, 1349–1357.
47. Bingle, L., Wilson, K., Musa, M., Araujo, B., Rassl, D., Wallace, W.A., LeClair, E.E., Mauad, T., Zhou, Z., Mall, M.A., and Bingle, C.D. (2012). BPIFB1 (LPLUNC1) is upregulated in cystic fibrosis lung disease. *Histochem. Cell Biol.* *138*, 749–758.
48. Zhang, B., Nie, X., Xiao, B., Xiang, J., Shen, S., Gong, J., Zhou, M., Zhu, S., Zhou, J., Qian, J., et al. (2003). Identification of tissue-specific genes in nasopharyngeal epithelial tissue and differentially expressed genes in nasopharyngeal carcinoma by suppression subtractive hybridization and cDNA microarray. *Genes Chromosomes Cancer* *38*, 80–90.
49. Liao, Q., Zeng, Z., Guo, X., Li, X., Wei, F., Zhang, W., Li, X., Chen, P., Liang, F., Xiang, B., et al. (2014). LPLUNC1 suppresses IL-6-induced nasopharyngeal carcinoma cell proliferation via inhibiting the Stat3 activation. *Oncogene* *33*, 2098–2109.

The American Journal of Human Genetics

Supplemental Data

**Low-Frequency Coding Variants
at 6p21.33 and 20q11.21 Are Associated
with Lung Cancer Risk in Chinese Populations**

Guangfu Jin, Meng Zhu, Rong Yin, Wei Shen, Jia Liu, Jie Sun, Cheng Wang, Juncheng Dai, Hongxia Ma, Chen Wu, Zhihua Yin, Jiaqi Huang, Brandon W. Higgs, Lin Xu, Yihong Yao, David C. Christiani, Christopher I. Amos, Zhibin Hu, Baosen Zhou, Yongyong Shi, Dongxin Lin, and Hongbing Shen

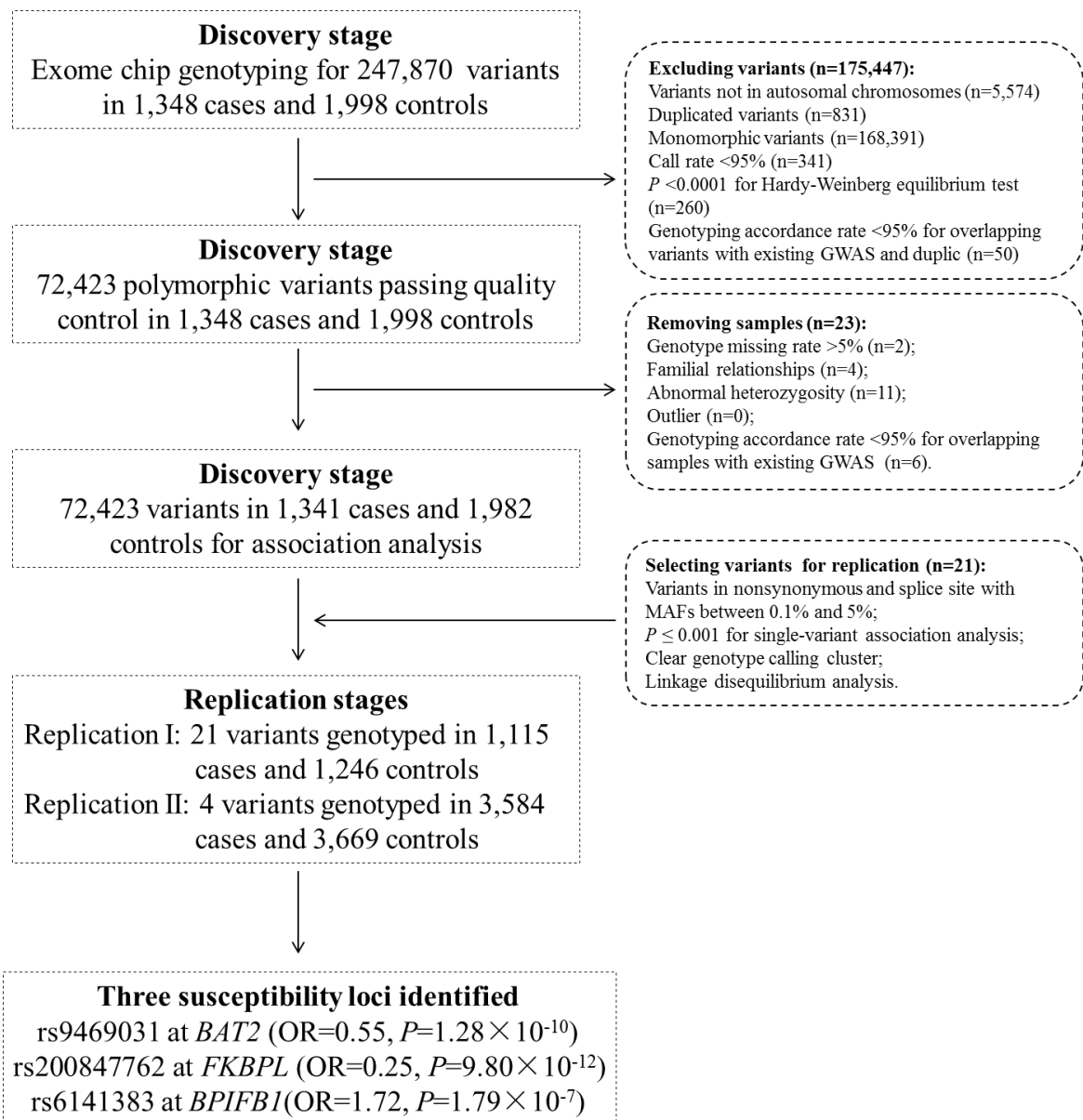
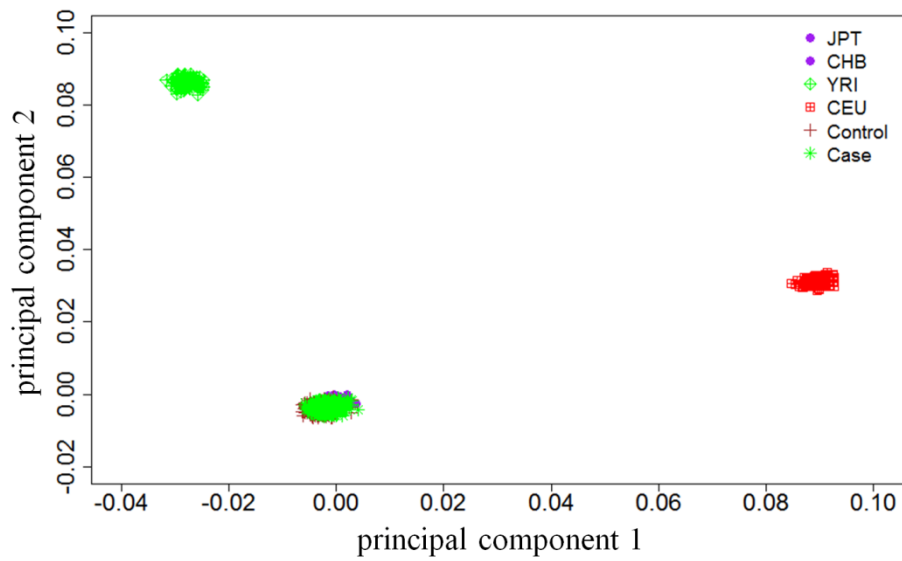


Figure S1. Summary of the study design and work flow.

A:



B:

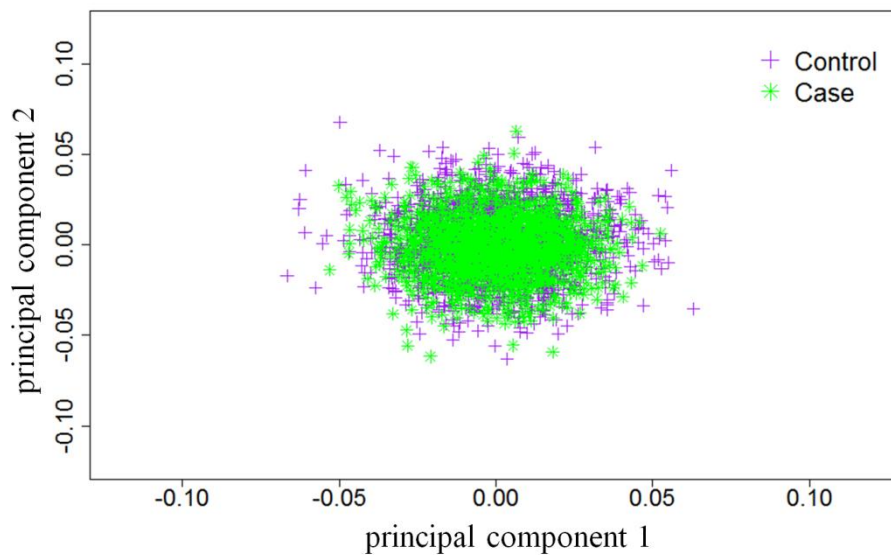


Figure S2. Principal component analysis. The first two principal components for each individual were plotted: A) the relatedness between the studied 1,341 cases and 1,982 controls, together with European (CEU), African (YRI), Chinese (CHB), and Japanese (JPT) data from the HapMap project was analyzed to determine ethnicity; and B) the population structures of the cases and controls.

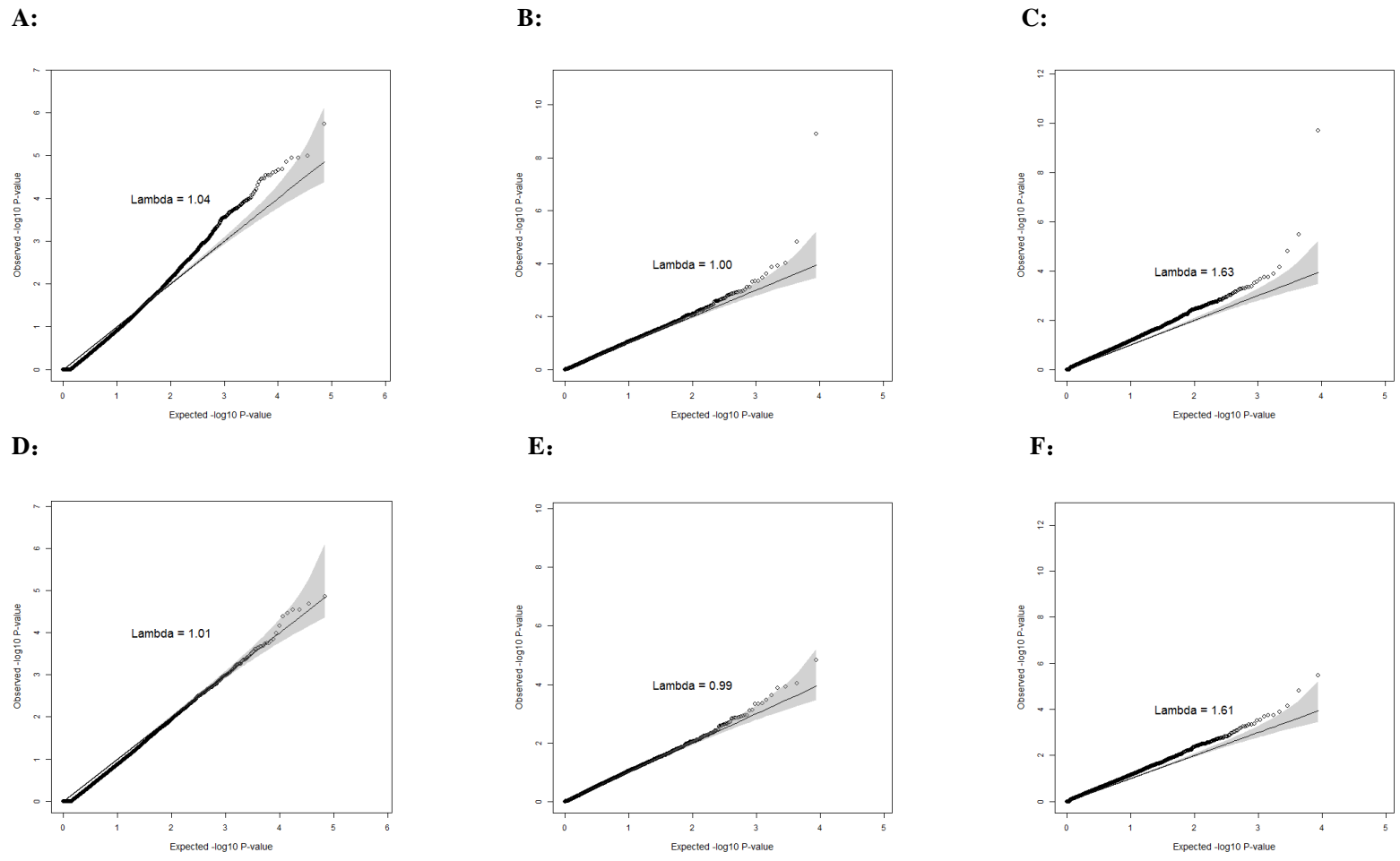


Figure S3. Quantile-quantile plot and genomic inflation factor lambda for associations with lung cancer risk. Observed P values are plotted as a function of theoretical P values: A, single-variant analysis; B, gene-based analysis using the SKAT-O test; C, gene-based analysis using the burden test; D, single-variant analysis after removing variants at 6p22.2-6p21.31; E, gene-based analysis using the SKAT-O test after removing variants at 6p22.2-6p21.31; F, gene-based analysis using the burden test after removing variants at 6p22.2-6p21.31. Gray areas indicate 90% confidence intervals from a null distribution of P values.

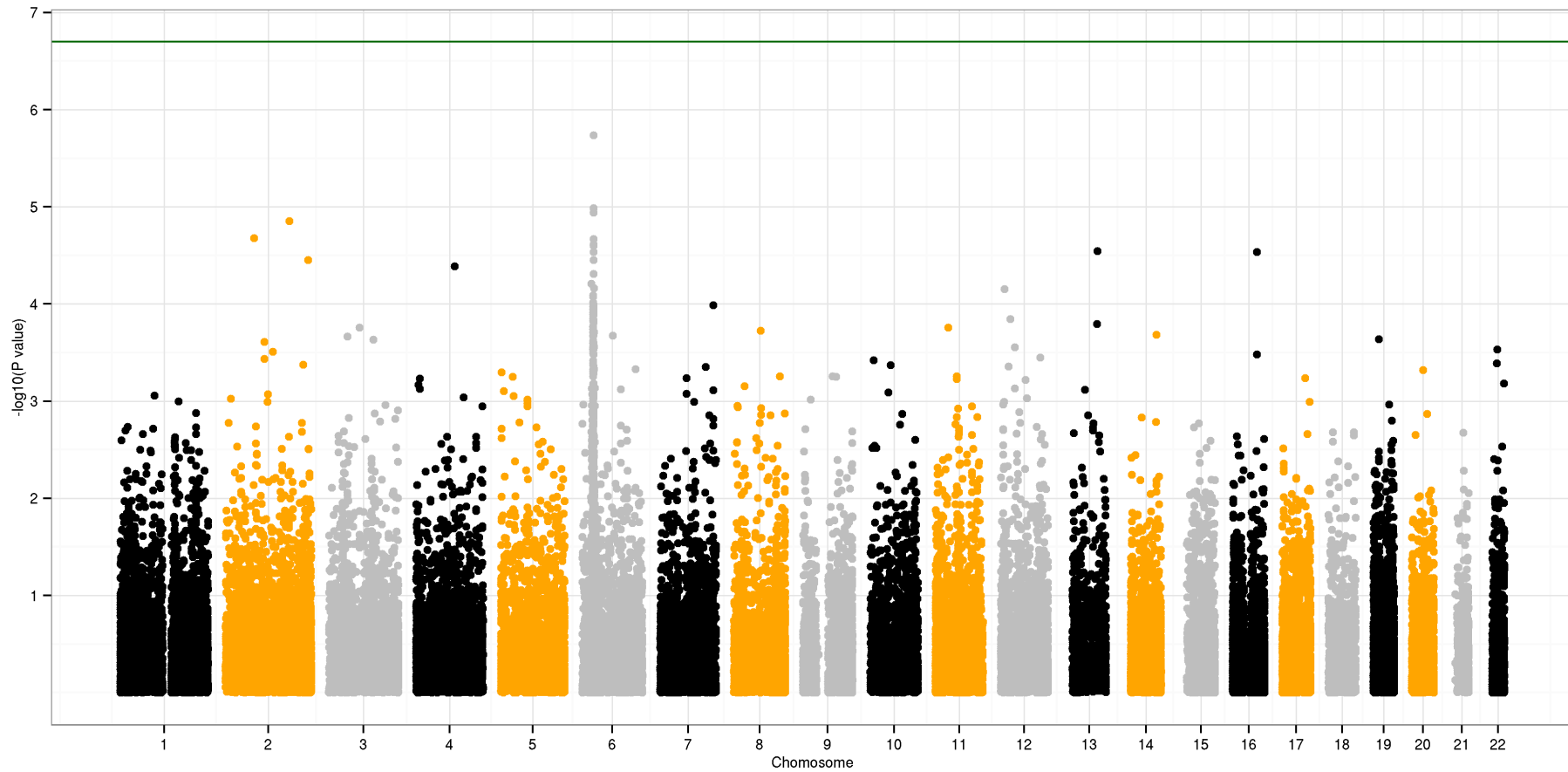
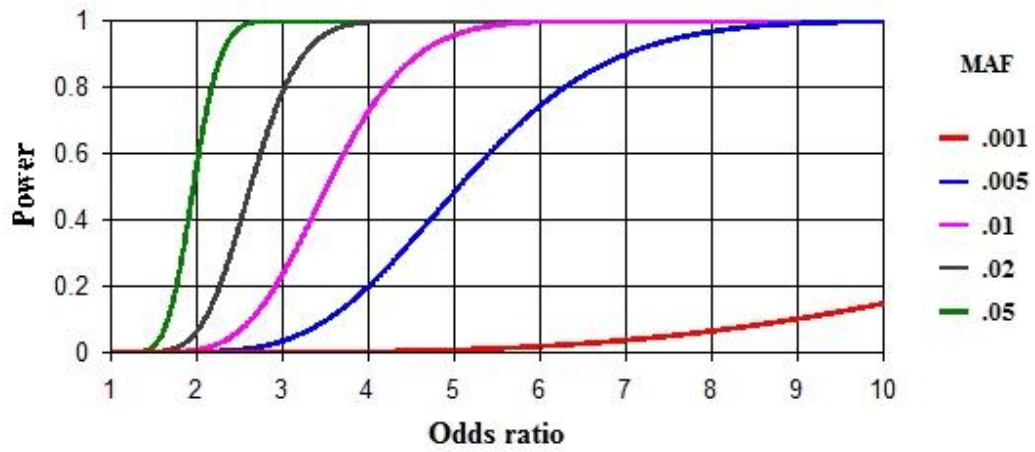


Figure S4. Manhattan plot for associations between genetic variants and lung cancer risk. The associations ($-\log_{10}(P)$ values, Y-axis) are plotted against genomic position (X-axis by chromosome and the chromosomal position of NCBI build 37). The green horizontal line corresponds to a P value threshold of 2.00×10^{-7} .

A:



B:

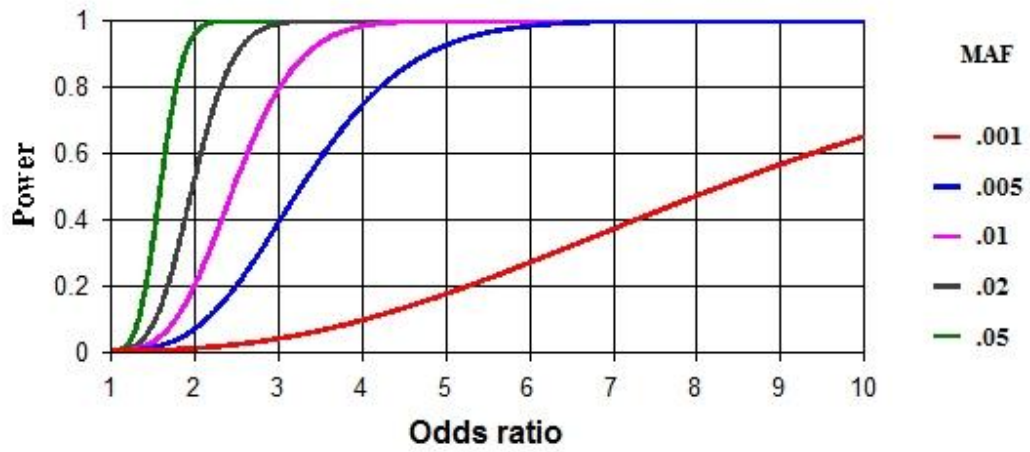
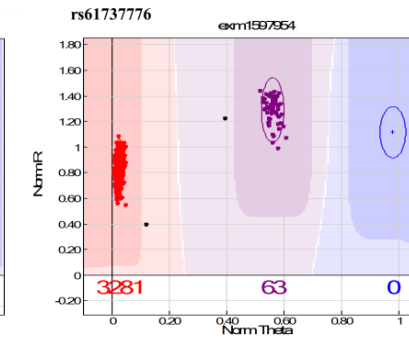
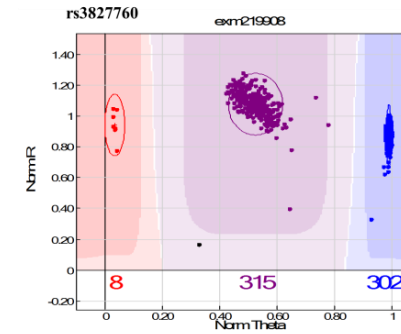
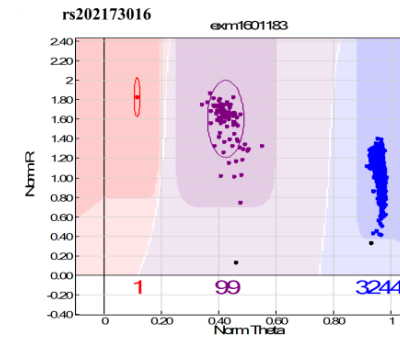
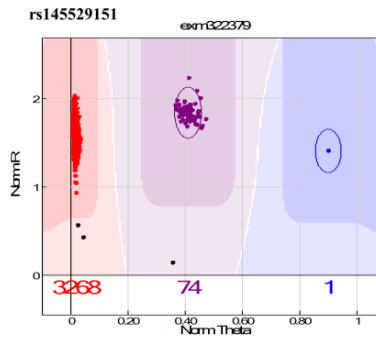
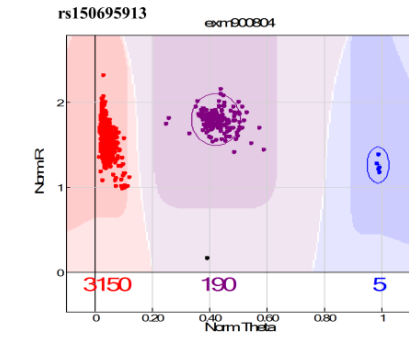
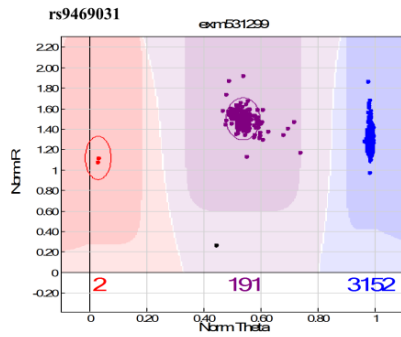
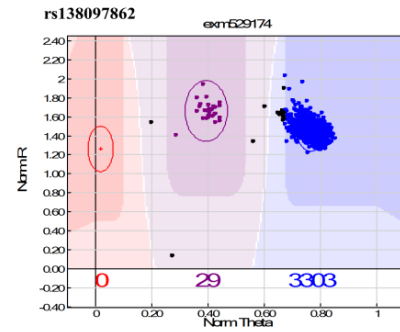
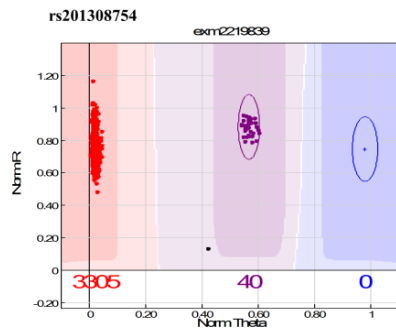
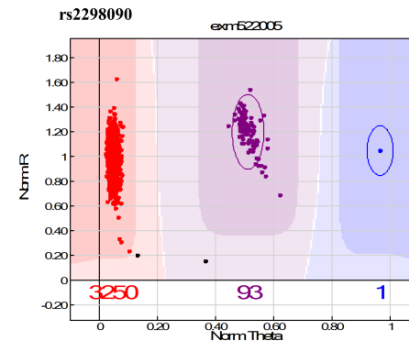
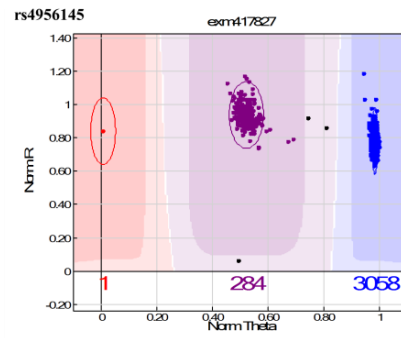
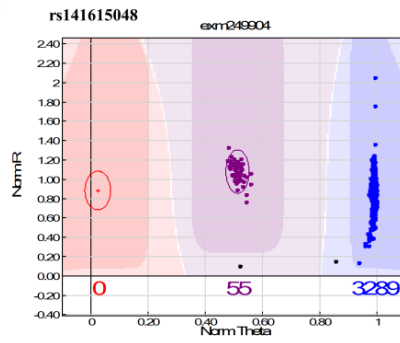
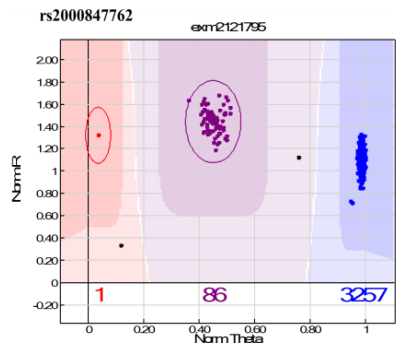


Figure S5. Statistical power estimation based on the sample size at the discovery stage (1,341 cases and 1,982 controls) at alpha levels of 6.90×10^{-7} (A, the predefined exome-wide association significance level) and 0.001 (B, the defined significance level for replication).



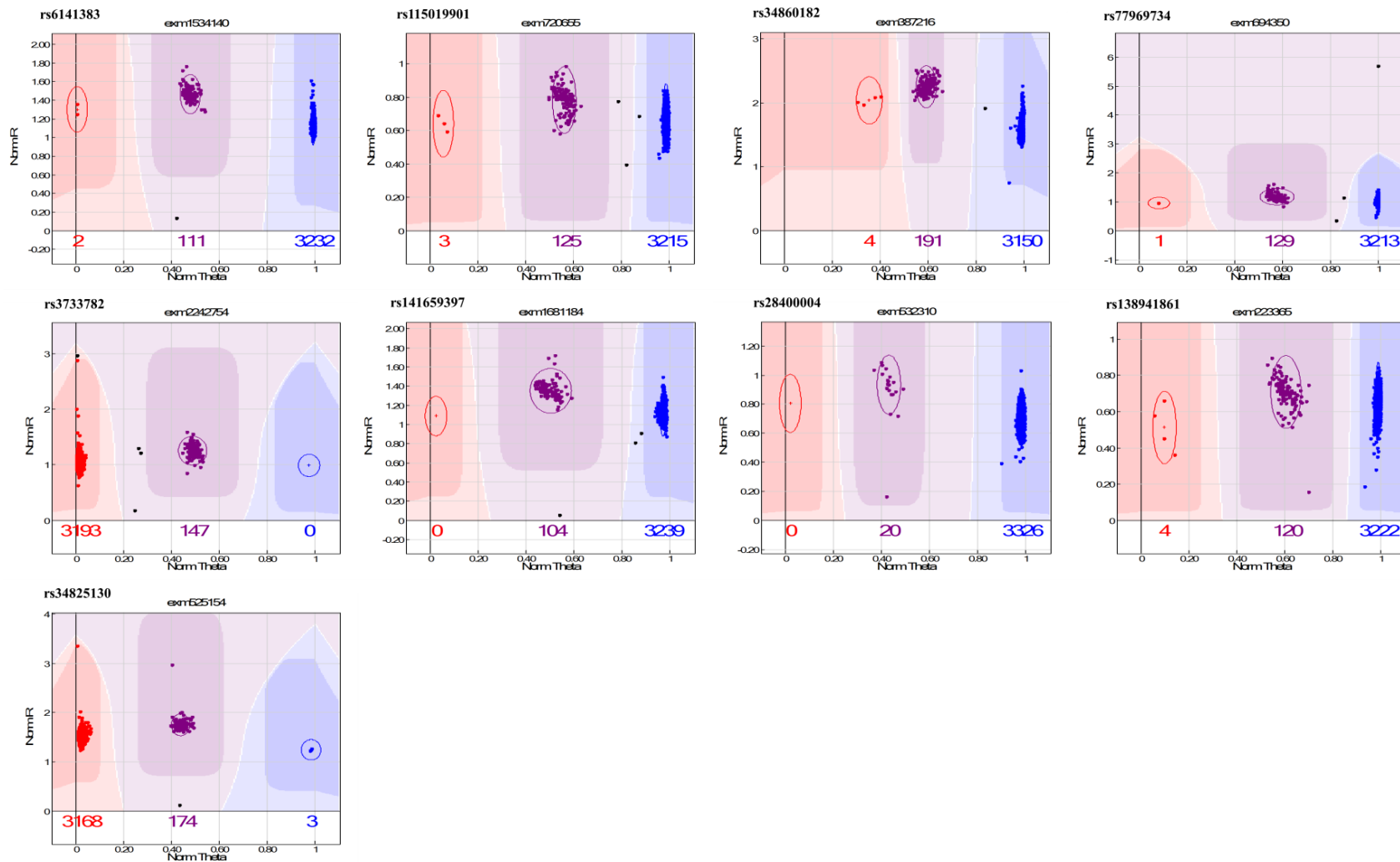
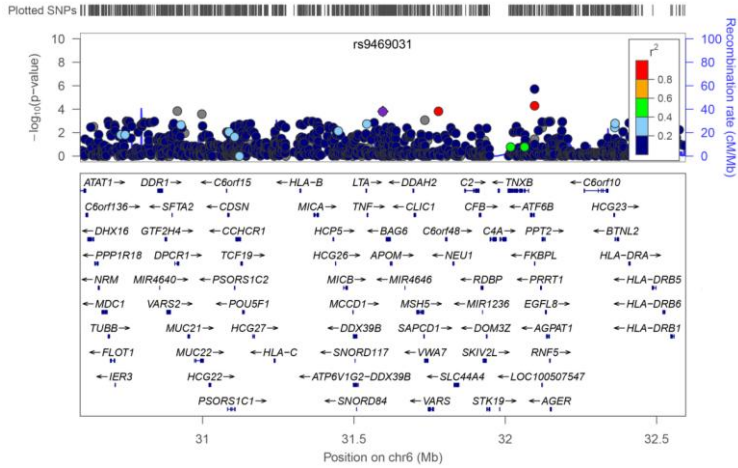
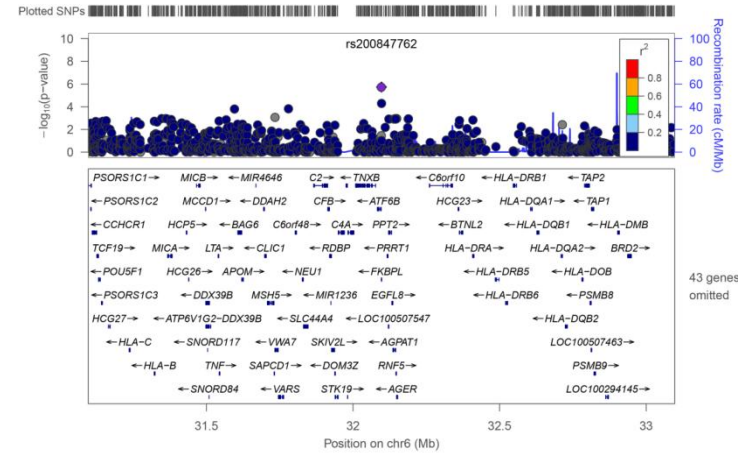


Figure S6. Cluster plots for the 21 variants included at the replication stages.

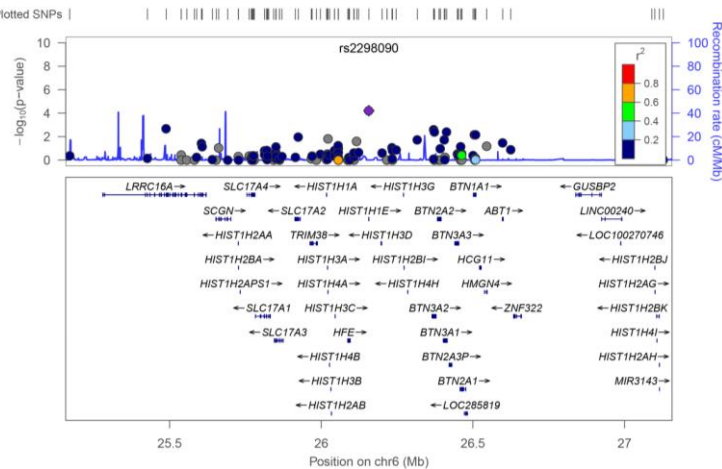
A:



B:



C:



D:

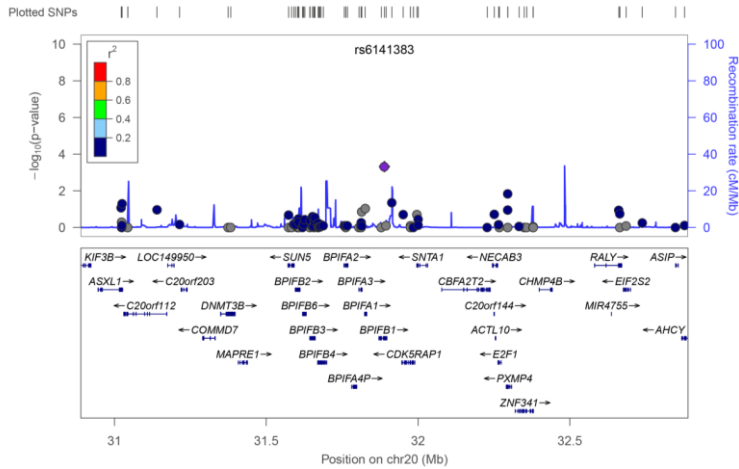


Figure S7. Regional association plots with lead variants showed using purple diamonds at *BAT2* (A), *FKBPL* (B), *HIST1H1E* (C) and *BPIFB1* (D).

Associations of individual variants are plotted as $-\log_{10} P$ against chromosomal position. The right y-axis shows the recombination rate estimated from the 1,000 Genomes Project CHB and JPT data.

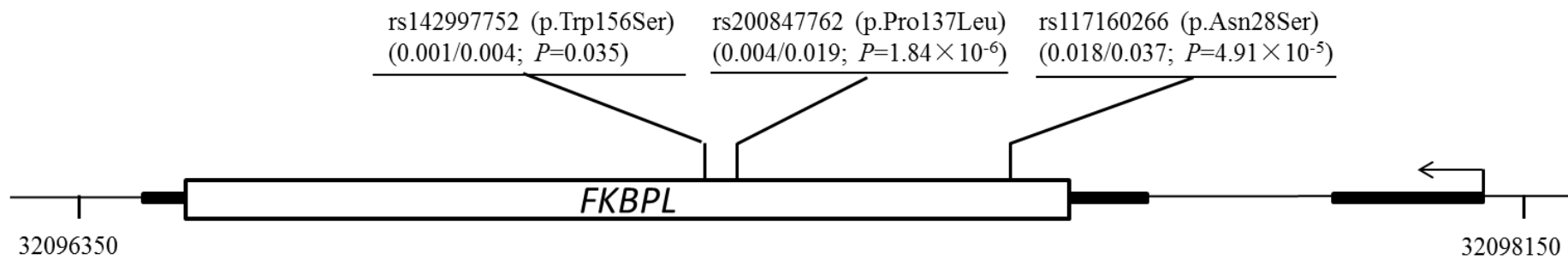
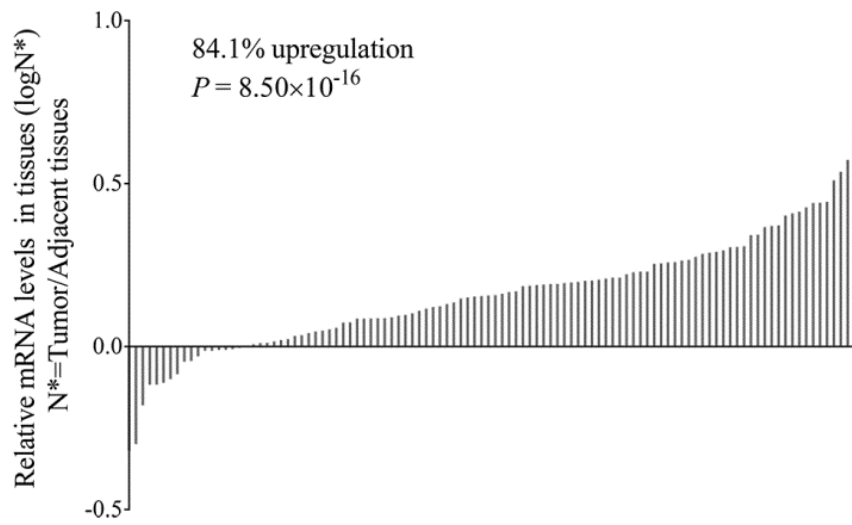
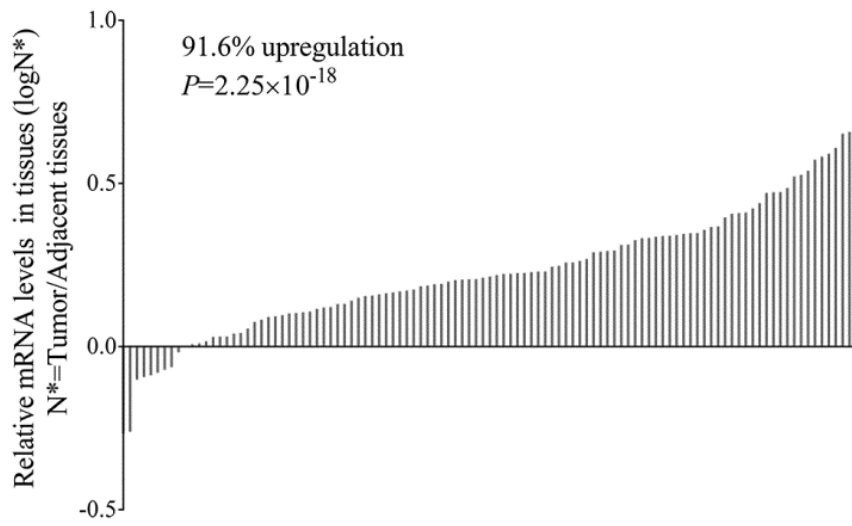


Figure S8. The gene structure of *FKBPL*, at 6p21.33, and three coding variants included in the gene-based analysis. *FKBPL* has two exons, and its coding region is only located in part of exon 2. The minor allele frequency (MAF) between cases and controls and the corresponding P value of the single-variant analysis at the discovery stage are shown for each variant.

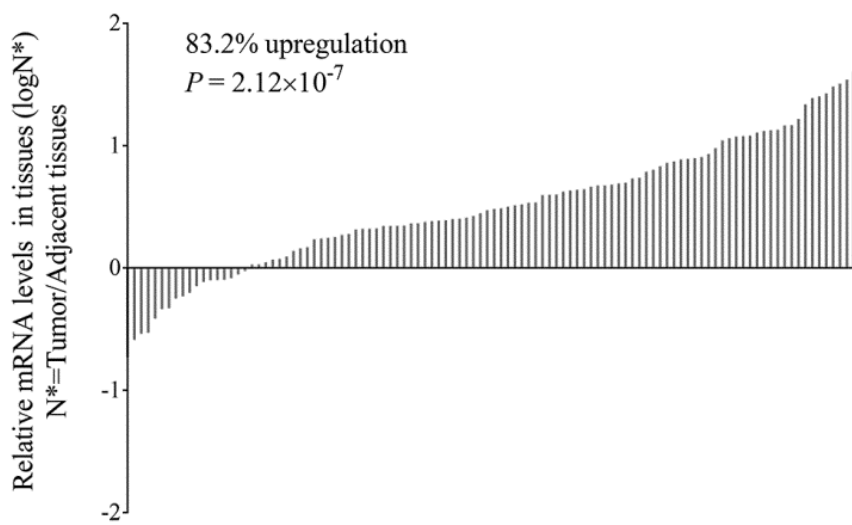
A: *BAT2*



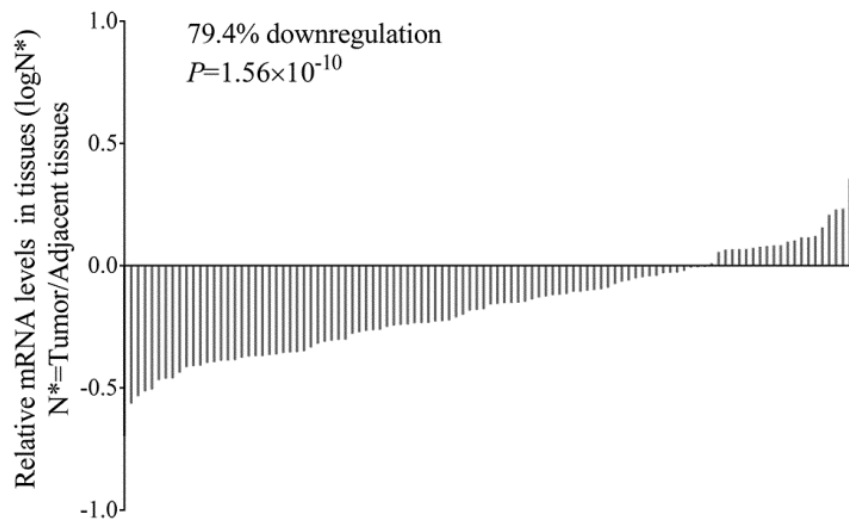
B: *FKBPL*



C: *HIST1H1E*



D: *HSPA1L*



E: *BPIFB1*

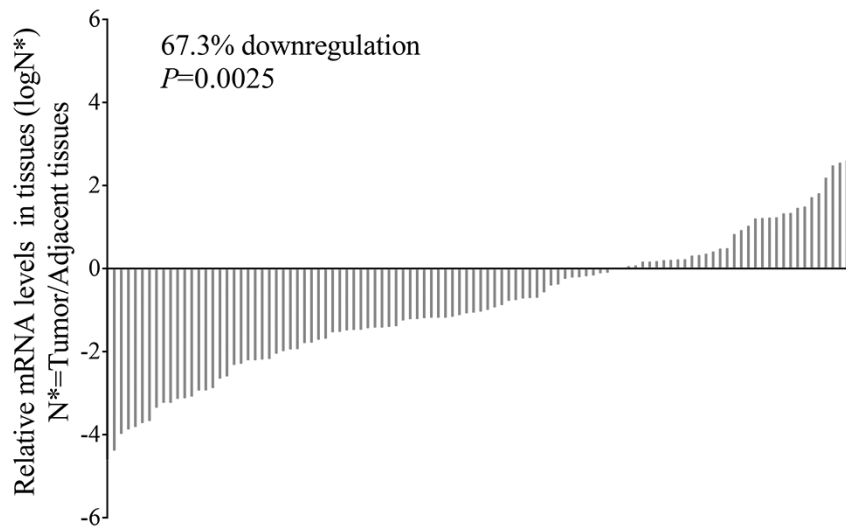
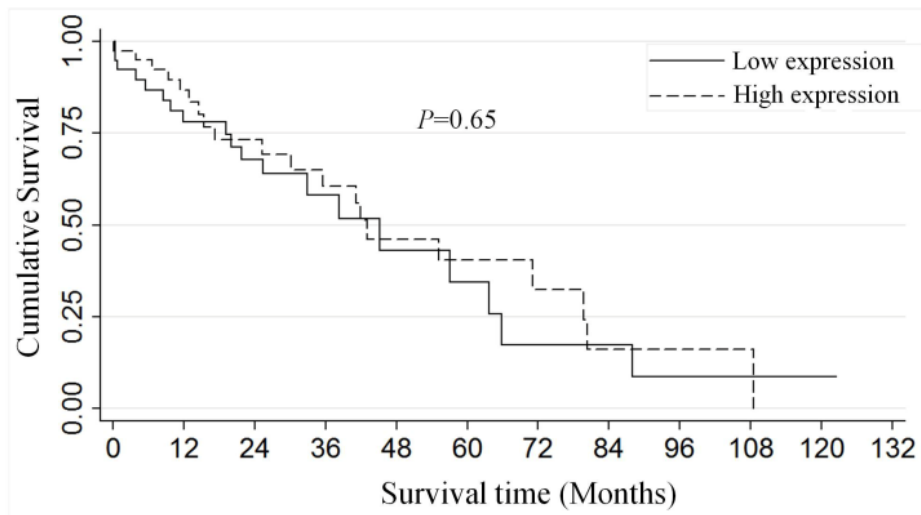
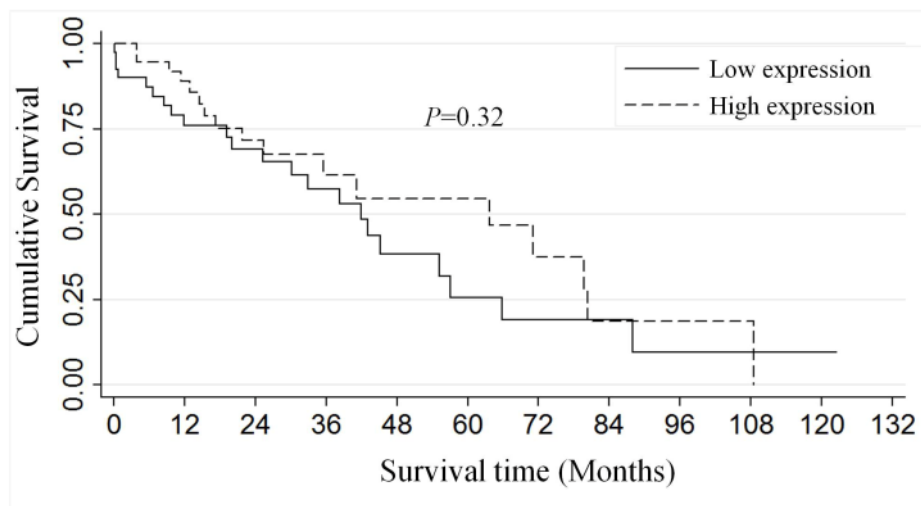


Figure S9. The differential expressions of genes at 6p22.2-6p21.33 and 20q11.21 based on TCGA data. The expression levels of *BAT2* (A: 84.1%, $P=8.50 \times 10^{-16}$) and *FKBPL* (B: 91.6%, $P=2.25 \times 10^{-18}$), both at 6p21.33, and *HIST1H1E* (C: 83.2%, $P=2.12 \times 10^{-7}$), at 6p22.2, were significantly upregulated in most lung tumors, whereas *HSPA1L* (D: 79.4%, $P=1.56 \times 10^{-10}$), at 6p21.33, and *BPIFB1* (E: 67.3%, $P=0.0025$), at 20q11.21, were downregulated.

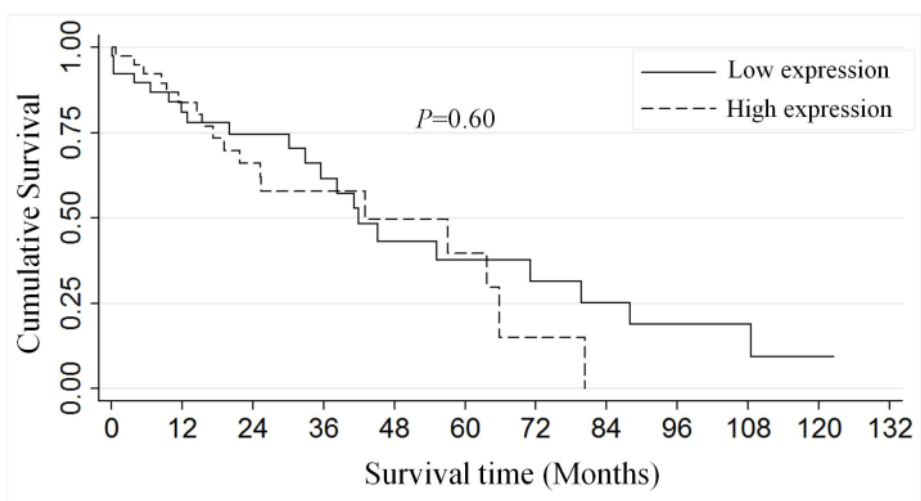
A: BAT2



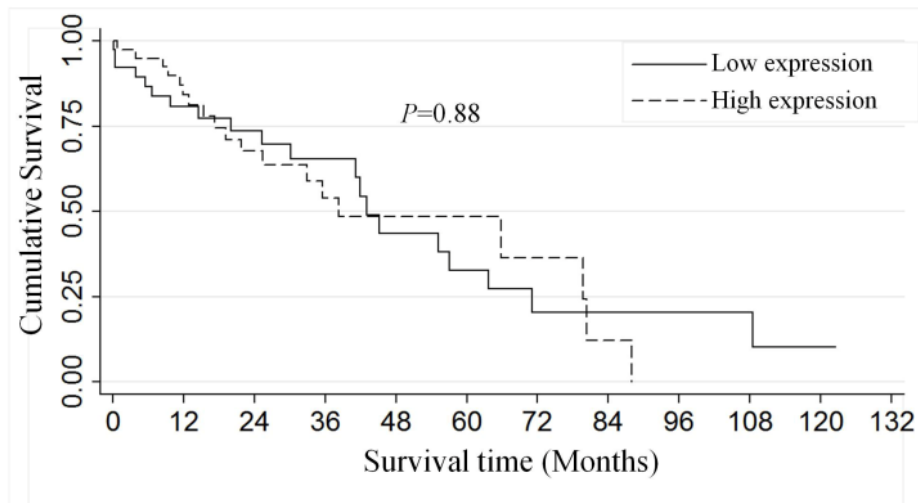
B: FKBPL



C: HIST1H1E



D: *HSPA1L*



E: *BPIFB1*

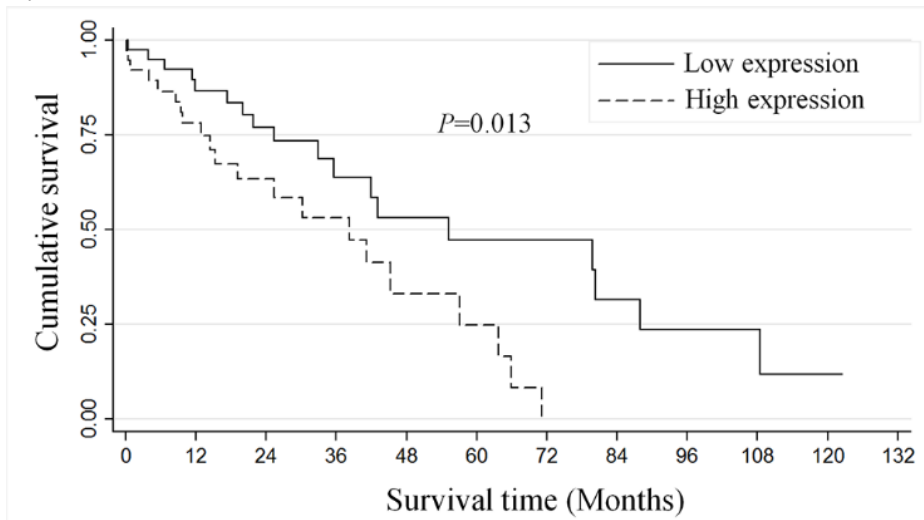


Figure S10. Kaplan-Meier survival curves of lung cancer patients based on gene expression.

The patients were divided into two groups based on the median expression ratios of tumor/adjacent tissues for each gene. The expression of *BAT2* (A), *FKBP1* (B), *HIST1H1E* (C) and *HSPA1L* (D) did not influence the survival time in individuals with TCGA lung cancer.

Individuals with high expression levels of *BPIFB1* (E) exhibited a shorter survival time (median survival time, MST=38.2 months) than individuals with low levels (MST=55.2 months) (log-rank $P=0.013$).

Table S1. Summary of characteristics of study subjects

Variables	Discovery stage		Replication I stage		Replication II stage	
	Case (n=1,348)	Control (n=1,998)	Case (n=1,115)	Control (n=1,246)	Case (n=3,584)	Control (n=3,669)
Age (years), mean \pm S.D.	61.00 \pm 10.22	61.27 \pm 11.07	60.80 \pm 9.58	60.68 \pm 9.78	58.63 \pm 10.29	55.99 \pm 12.74
Gender, n (%)						
Male	954(70.8)	1,369(68.5)	731(65.6)	789(63.3)	2,284(63.7)	2,163(59.0)
Female	394(29.2)	629(31.5)	384(34.4)	457(36.7)	1,300(36.3)	1,506(41.0)
Smoking status, n (%)						
Current	636(47.2)	883(44.2)	437(39.2)	461(37.0)	1,519(42.4)	1,113(30.3)
Former	187(13.9)	87(4.4)	102(9.1)	100(8.0)	457(12.8)	114(3.1)
Never	525(38.9)	1,028(51.4)	576(51.7)	685(55.0)	1,608(44.9)	2,442(66.6)
Smoking levels (pack-years), n (%)	41.42 \pm 28.00	28.95 \pm 20.05	39.11 \pm 25.82	32.59 \pm 21.52	36.86 \pm 23.45	23.67 \pm 16.82
\leq 25	253(30.7)	490(50.5)	171(31.7)	232(41.4)	695(35.2)	780(63.6)
$>$ 25	570(69.3)	480(49.5)	368(68.3)	329(58.6)	1,281(64.8)	447(36.4)
Histology, n (%)						
Squamous cell carcinoma	483(35.8)		332(29.8)		1,287(35.9)	
Adenocarcinoma	865(64.2)		783(70.2)		1,733(48.4)	
Other ^a					564(15.7)	

^a Other includes small-cell lung cancer, large-cell lung cancer and mixed-cell carcinoma.

Table S2. The distributions of variants included on the exome chip according to frequency in a Chinese population

MAF ^a	Lung cancer cases (n=1,341)		Cancer-free controls (n=1,982)		Combined subjects (n=3,323)	
	All variants	Nonsynonymous or splice-site variants	All variants	Nonsynonymous or splice-site variants	All variants	Nonsynonymous or splice-site variants
0	178,923	171,619	177,942	170,678	168,391	161,590
(0, 0.001]	16,154	15,309	17,151	16,180	26,984	25,565
(0.001, 0.005]	9,600	8,930	9,182	8,585	9,391	8,760
(0.005, 0.010]	3,578	3,318	4,132	3,845	3,624	3,377
(0.010, 0.050]	7,297	6,197	7,169	6,093	7,170	6,080
>0.050	25,262	11,421	25,238	11,413	25,254	11,422
Total	240,814	216,794	240,814	216,794	240,814	216,794

^a MAF, minor allele frequency.

Table S3. The results of variants that were found to be potentially associated with lung cancer risk at the discovery stage and that were selected for further evaluation at the replication stages

Chr.	Position (hg19, bp)	ID	Gene	Location	Alleles ^a	Stage	MAF ^b	OR(95%CI) ^c	<i>P</i> ^c	<i>P</i> ^d	<i>P</i> ^e
2	109513601	rs3827760	<i>EDAR</i>	p.Val370Ala	G/A	Discovery	0.060/0.041	1.52(1.21-1.92)	3.66E-04	3.36E-04	3.73E-04
						Replication I	0.052/0.049	1.07(0.83-1.40)	5.89E-01		
2	120005631	rs138941861	<i>STEAP3</i>	p.Arg290His	G/A	Discovery	0.025/0.014	1.86(1.29-2.67)	8.43E-04	6.95E-04	7.97E-04
						Replication I	0.025/0.018	1.33(0.90-1.98)	1.57E-01		
2	180815329	rs141615048	<i>CWC22</i>	p.Asp684Asn	G/A	Discovery	0.014/0.004	3.68(2.04-6.62)	1.40E-05	4.06E-06	5.79E-06
						Replication I	0.009/0.006	1.47(0.75-2.90)	2.65E-01		
3	52568641	rs145529151	<i>NT5DC2</i>	p.Glu10Gly	A/G	Discovery	0.018/0.007	2.46(1.52-3.95)	2.16E-04	1.35E-04	1.67E-04
						Replication I	0.010/0.012	0.84(0.49-1.44)	5.22E-01		
4	6925237	rs34860182	<i>TBC1D14</i>	p.Leu41Val	G/C	Discovery	0.037/0.025	1.66(1.24-2.22)	6.75E-04	5.98E-04	6.94E-04
						Replication I	0.032/0.028	1.16(0.83-1.64)	3.81E-01		
4	108931039	rs4956145	<i>HADH</i>	p.Pro86Leu	G/A	Discovery	0.055/0.035	1.68(1.31-2.16)	4.08E-05	3.54E-05	4.22E-05
						Replication I	0.042/0.036	1.15(0.85-1.55)	3.67E-01		
5	7868188	rs3733782	<i>FASTKD3</i>	p.Leu3Phe	A/T	Discovery	0.013/0.028	0.51(0.35-0.76)	7.88E-04	6.56E-04	5.13E-04
						Replication I	0.023/0.022	1.11(0.75-1.64)	6.11E-01		
6	26157073	rs2298090	<i>HIST1H1E</i>	p.Lys152Arg	A/G	Discovery	0.006/0.020	0.32(0.19-0.56)	6.16E-05	2.82E-05	1.16E-05
						Replication I	0.013/0.024	0.56(0.36-0.87)	9.80E-03		
						Replication II	0.006/0.008	0.67(0.44-1.02)	6.03E-02		
6	28478612	rs34825130	<i>GPX6</i>	p.Tyr53His	A/G	Discovery	0.035/0.021	1.68(1.24-2.28)	8.51E-04	7.50E-04	8.34E-04
						Replication I	0.031/0.023	1.35(0.95-1.93)	9.70E-02		
6	30916645	rs138097862	<i>DPCR1</i>	p.Arg135Gln	G/A	Discovery	0.008/0.002	6.05(2.40-15.29)	1.41E-04	2.15E-05	2.66E-05
						Replication I	0.006/0.004	1.47(0.66-3.25)	3.47E-01		
6	31595795	rs9469031	<i>BAT2</i>	p.Pro515Leu	G/A	Discovery	0.019/0.036	0.52(0.37-0.73)	1.54E-04	1.24E-04	9.09E-05
						Replication I	0.024/0.042	0.61(0.44-0.83)	1.71E-03		
						Replication II	0.011/0.018	0.62(0.46-0.84)	1.72E-03		
6	31734385	rs28400004	<i>VWA7</i>	p.Arg680Gln	G/A	Discovery	0.006/0.001	8.59(2.43-30.32)	8.39E-04	9.49E-05	1.03E-04
						Replication I	0.004/0.003	1.32(0.47-3.65)	5.98E-01		
6	32097148	rs200847762	<i>FKBPL</i>	p.Pro137Leu	G/A	Discovery	0.004/0.019	0.21(0.11-0.39)	1.84E-06	2.22E-07	2.90E-08
						Replication I	0.003/0.014	0.19(0.08-0.46)	2.24E-04		
						Replication II	0.002/0.003	0.66(0.34-1.28)	2.16E-01		
8	30949398	rs77969734	<i>WRN</i>	p.Leu628Val	G/C	Discovery	0.026/0.015	1.86(1.30-2.65)	6.97E-04	5.88E-04	6.83E-04

8	130760802	rs115019901	<i>GSDMC</i>	p.Pro491Leu	G/A	Replication I	0.011/0.011	1.03(0.60-1.77)	9.08E-01		
						Discovery	0.027/0.015	1.86(1.31-2.66)	5.55E-04	4.47E-04	5.26E-04
10	51584682	rs141659397	<i>NCOA4</i>	p.Leu261Phe	G/A	Replication I	0.022/0.026	0.80(0.55-1.16)	2.42E-01		
						Discovery	0.010/0.020	0.45(0.29-0.72)	8.13E-04	6.30E-04	4.93E-04
11	36422712	rs150695913	<i>PRR5L</i>	p.His14Arg	A/G	Replication I	0.010/0.017	0.62(0.37-1.03)	6.20E-02		
						Discovery	0.019/0.036	0.53(0.38-0.74)	1.76E-04	1.41E-04	1.02E-04
12	11061171	rs201308754	<i>TAS2R13</i>	p.Ser243Gly	A/G	Replication I	0.035/0.038	0.92(0.68-1.25)	6.01E-01		
						Discovery	0.012/0.002	4.71(2.19-10.11)	7.00E-05	1.61E-05	1.55E-05
20	31889141	rs6141383	<i>BPIFB1</i>	p.Val284Met	G/A	Replication I	0.006/0.004	1.67(0.71-3.93)	2.41E-01		
						Discovery	0.025/0.012	2.00(1.36-2.95)	4.80E-04	3.85E-04	4.31E-04
						Replication I	0.022/0.013	1.68(1.07-2.62)	2.32E-02		
22	29706927	rs61737776	<i>GAS2L1</i>	p.Arg317Gly	C/G	Replication II	0.020/0.013	1.64(1.24-2.17)	6.20E-04		
						Discovery	0.004/0.013	0.31(0.16-0.59)	4.10E-04	2.16E-04	1.30E-04
22	31687065	rs202173016	<i>PIK3IP1</i>	p.Gly65Se	G/A	Replication I	0.006/0.010	0.61(0.31-1.21)	1.55E-01		
						Discovery	0.022/0.010	2.14(1.42-3.23)	2.93E-04	2.17E-04	2.58E-04
						Replication I	0.014/0.012	1.15(0.68-1.93)	6.09E-01		

^a Major/minor alleles;

^b Minor allele frequencies of cases/controls;

^c Derived from the logistic regression model after adjusting for age, gender, pack-year of smoking and the top principal component (for the discovery stage only) assuming an additive genetic model.

^d Derived from the logistic score test after adjusting for age, gender, pack-year of smoking and the top principal component.

^e Derived from the Firth bias-corrected logistic likelihood ratio test after adjusting for age, gender, pack-year of smoking and the top principal component.

Table S4. The frequencies of the four reported loci among diverse populations and associations with lung cancer risk in populations of European ancestry.

Chr.	Gene	Variant ID	Major/Minor allele	MAF ^a					<i>in silico</i> replication I ^b			<i>in silico</i> replication II ^c		
				ASW	CEU	CHS	JPT	YRI	MAF	OR(95%CI)	<i>P</i>	MAF	OR(95%CI)	<i>P</i>
6p21.33	<i>BAT2</i>	rs9469031	C/T	0.007	0.000	0.005	0.017	0.006	0.003	0.94(0.68-1.31)	0.716	--	--	--
6p21.33	<i>FKBPL</i>	rs200847762	G/A	0.001	0.000	0.000	0.000	0.000	--	--	--	--	--	--
6p22.2	<i>HIST1H1E</i>	rs2298090	A/G	0.007	0.012	0.000	0.034	0.000	0.009	0.94(0.78-1.14)	0.556	0.007	0.80(0.36-1.80)	0.580
20q11.21	<i>BPIFB1</i>	rs6141383	A/G	0.007	0.000	0.015	0.011	0.017	--	--	--	--	--	--

^aMinor allele frequencies among diverse populations based on the 1,000 Genomes Project: ASW, Americans of African Ancestry in SW USA; CEU, Utah Residents (CEPH) with Northern and Western European ancestry; CHS, Southern Han Chinese; JPT, Japanese in Tokyo, Japan; YRI, Yoruba in Ibadan, Nigeria.

^bDerived from a meta-analysis of four lung cancer GWAS in populations of European ancestry (11,348 cases and 15,861 controls): the MD Anderson Cancer Center (MDACC) GWAS, the Institute of Cancer Research (ICR) GWAS, the National Cancer Institute (NCI) GWAS and the International Agency for Research on Cancer (IARC) GWAS (refer to Wang et al. Nat Genet. 2014;46:736-741 for details). rs9469031 and rs2298090 were imputed with high quality based on the 1,000 Genomes Project.

^cDerived from the Harvard Lung Cancer Susceptibility Study in populations of European ancestry (984 cases and 970 controls) (refer to Wang et al. Nat Genet. 2014;46:736-741 for details). rs2298090 was imputed with high quality based on the 1,000 Genomes Project.

Table S5. The associations of four reported low-frequency variants with lung cancer risk in subgroups divided by characteristics

Subgroups	Cases/Controls	rs9469031 at <i>BAT2</i>		rs200847762 at <i>FKBPL</i>		rs2298090 at <i>HIST1H1E</i>		rs6141383 at <i>BPIFB1</i>	
		OR(95%CI) ^a	<i>P</i> ^b	OR(95%CI) ^a	<i>P</i> ^b	OR(95%CI) ^a	<i>P</i> ^b	OR(95%CI) ^a	<i>P</i> ^b
Age (years)			0.377		0.185		0.177		0.091
≤60	3157/3669	0.51 (0.39-0.66)		0.33(0.19-0.56)		0.42(0.29-0.62)		2.01(1.52-2.67)	
>60	2890/3244	0.60 (0.47-0.77)		0.19(0.10-0.34)		0.60(0.42-0.85)		1.42(1.05-1.91)	
Gender			0.626		0.305		0.641		0.487
Male	3969/4321	0.57(0.45-0.72)		0.22(0.14-0.36)		0.49(0.36-0.67)		1.81(1.40-2.34)	
Female	2078/2592	0.52(0.39-0.69)		0.34(0.17-0.67)		0.56(0.35-0.89)		1.56(1.12-2.17)	
Smoking status			0.599		0.984		0.197		0.334
Current	2592/2457	0.49(0.36-0.66)		0.26(0.15-0.46)		0.54(0.37-0.79)		2.00(1.44-2.79)	
Former	746/301	0.66(0.37-1.17)		0.26(0.08-0.79)		0.25(0.11-0.58)		1.35(0.66-2.74)	
Never	2709/4155	0.57(0.45-0.73)		0.28(0.15-0.52)		0.57(0.39-0.82)		1.48(1.12-1.95)	
Smoking levels (pack-years)			0.241		0.724		0.999		0.232
≤25	1119/1502	0.66(0.44-1.00)		0.22(0.08-0.56)		0.47(0.27-0.81)		2.50(1.68-3.72)	
>25	2219/1256	0.48(0.34-0.67)		0.27(0.15-0.49)		0.47(0.30-0.74)		1.72(1.08-2.75)	
Histology			0.198		0.803		0.218		0.693
Squamous cell carcinoma	2102/6913	0.47(0.35-0.64)		0.23(0.12-0.44)		0.51(0.35-0.76)		1.80(1.35-2.40)	
Adenocarcinoma	3381/6913	0.61(0.50-0.75)		0.26(0.16-0.44)		0.55(0.40-0.75)		1.71(1.36-2.15)	
Other	564/6913	0.40(0.23-0.72)		0.16(0.04-0.65)		0.19(0.06-0.60)		1.40(0.85-2.31)	

^a Derived from the logistic regression model after adjusting for age, gender, pack-year of smoking and study stage as appropriate assuming an additive genetic model;

^b Heterogeneity test between subgroups.

Table S7. Protein annotations of low-frequency variants associated with lung cancer risk

Variant ^a	Chr.	Position (hg19, bp)	Gene	Amino acid change	Location	Transcript	Uniprot ID	Polyphen-2 Score ^b	SIFT score ^c
rs9469031	6p21.33	31595795	<i>BAT2</i>	p.Pro515Leu	exon 12	ENST00000376007 ENST00000376033	P48634	0.041 (B)	0.00 (D)
rs9469057	6p21.33	31779728	<i>HSPAIL</i>	p.Ala8Pro	exon 2	ENST00000375654	P34931	1.000 (D)	0.00 (D)
rs200847762	6p21.33	32097148	<i>FKBPL</i>	p.Pro137Leu	exon 2	ENST00000375156	Q9UIM3	0.000 (B)	0.68 (T)
rs117160266	6p21.33	32097475	<i>FKBPL</i>	p.Asn28Ser	exon 2	ENST00000375156	Q9UIM3	0.118(B)	0.49 (T)
rs2298090	6p22.2	26157073	<i>HIST1H1E</i>	p.Lys152Arg	exon 1	ENST00000304218	P10412	0.978 (D)	0.06 (T)
rs6141383	20q11.21	31889141	<i>BPIFB1</i>	p.Val284Met	exon 9	ENST00000253354	Q8TDL5	0.977 (D)	0.00 (D)

^a Each variant is annotated with all isoforms based on the comprehensive set of GENCODE version 7 gene transcripts;

^b Polyphen-2 scores range from 0 (B) to 1 (D). B=benign, P=possibly damaging, D=probably damaging;

^c SIFT scores range from 0 (D) to 1 (T). All scores ≤ 0.05 are predicted to be damaging. T=tolerated, D=damaging.