Published in final edited form as:
Nat Genet. 2008 December ; 40(12): 1407–1409. doi:10.1038/ng.273.

Common 5p15.33 and 6p21.33 variants influence lung cancer risk

Yufei Wang[1], **Peter Broderick**[1], **Emily Webb**[1], **Xifeng Wu**[2], **Jayaram Vijayakrishnan**[1], **Athena Matakidou**[1], **Mobshra Qureshi**[1], **Qiong Dong**[2], **Xiangjun Gu**[2], **Wei Vivien Chen**[2], **Margaret R Spitz**[2], **Timothy Eisen**[3,4], **Christopher I Amos**[2,4], and **Richard S Houlston**[1,4]

1 *Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK*

2 *Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA*

3 *Department of Oncology, University of Cambridge, Cambridge CB2 2RE, UK*

Abstract

We conducted a genome-wide association (GWA) study of lung cancer comparing 511,919 SNP genotypes in 1,952 cases and 1,438 controls. The most significant association was attained at 15q25.1 (rs8042374; $P = 7.75 \times 10^{-12}$), confirming recent observations. Pooling data with two other GWA studies (5,095 cases, 5,200 controls) and with replication in an additional 2,484 cases and 3,036 controls, we identified two newly associated risk loci mapping to 6p21.33 (rs3117582, *BAT3-MSH5;* $P_{\text{combined}} = 4.97 \times 10^{-10}$) and 5p15.33 (rs401681, *CLPTM1L;* $P_{\text{combined}} = 7.90 \times 10^{-9}$).

Support for inherited genetic susceptibility to lung cancer has recently come from genome-wide association studies that have demonstrated that 15q25.1 variation influences lung cancer risk[1–3].

To identify risk variants for lung cancer, we carried out a GWA study. Using Illumina HumanHap550 BeadChips, we genotyped 561,466 SNPs in 1,978 cases (Supplementary Methods online). After application of quality control criteria, genotypes were available for 1,952 cases. We were able to satisfactorily genotype 552,947 SNPs (98.5%) with mean sample call rate 99.7%. For controls, we used publicly accessible HumanHap550 genotype data in 1,438 individuals from the 1958 Birth Cohort[4] (Supplementary Methods). Genotypes were available for 541,327 SNPs (97.5% of 555,352 SNPs typed) and 524,714 SNPs were common to cases and controls. Applying quality control filters, we excluded 8,534 SNPs monomorphic in either cases or controls; 2,744 with call rates < 95%; 770 showing departure from Hardy-Weinberg equilibrium (HWE; $P < 10^{-5}$ in cases or controls) and 747 with minor allele frequency (MAF) <1% in cases or controls; leaving 511,919 informative SNPs for analysis.

Correspondence should be addressed to R.S.H. (E-mail: richard.houlston@icr.ac.uk).
[4] These authors contributed equally to this work.

Note: Supplementary information is available on the Nature Genetics website.

AUTHOR CONTRIBUTIONS

R.S.H. designed the study, with substantial contributions from C.I.A. and T.E. R.S.H. drafted the manuscript, with substantial contributions from E.W. and C.I.A. T.E., R.S.H. and A.M. oversaw sample and data collection of cases for the UK study. P.B. oversaw the genotyping for the UK study. P.B. managed samples for the UK study. J.V. and M.Q. performed sample preparation for the UK study. Y.W. performed statistical and bioinformatics analyses for the UK study. Y.W. and E.W. performed statistical and bioinformatics analyses for the meta-analysis. M.R.S. oversaw sample and data collection of cases for the Texas study. C.I.A. oversaw the GWA genotyping and analyses for the Texas Study. X.W. managed samples for the Texas study. X.G., W.V.C. and Q.D. performed analyses and curated data for the Texas study. All authors contributed to the final paper.

Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

Comparison of observed and expected distributions showed little evidence for inflation of allele test statistics (inflation factor $\lambda = 1.03$; Supplementary Fig. 1 online), excluding the possibility of significant hidden population substructure, cryptic relatedness or differential genotype calling. The distribution of association P values was significantly skewed from the null distribution with 116 SNPs having P value $\leq 10^{-4}$, greater than the 51 expected ($P \sim 10^{-4}$). In keeping with previous studies, 15q25.1 SNPs were most strongly associated; excluding these SNPs, 98 were associated at $P \leq 10^{-4}$.

To replicate associations, we are now genotyping the most strongly associated SNPs in an additional case-control series. In the interim, we have sought to identify new associations by conducting a meta-analysis pooling our UK-GWA study with data from two other studies: the IARC-GWA study of 1,989 cases and 2,625 controls[2], summary data from which is publicly available; and the Texas-GWA study of 1,154 non–small-cell lung cancer cases and 1,137 controls[1]. In both studies genotyping was done using Illumina HumanHap300 Bead-Chips. Pooling was based on the 223,891 autosomal SNPs genotyped in all three GWA studies that had MAFs >1% and no departure from HWE ($P < 10^{-5}$ in cases or controls).

We derived meta-analysis odds ratios (ORs) and confidence intervals (CIs), and associated P values (Supplementary Methods). As expected, the strongest associations were obtained for SNPs mapping to 15q25.1. After exclusion of the 36 SNPs mapping to this locus (76.4–76.8 Mb), there remained evidence of enrichment of associated variants: 77 had P values $\leq 10^{-4}$, compared with 22 expected (Supplementary Table 1 online). Genomic control values for the Texas-GWA and IARC–GWA studies were 0.99 and 1.03, respectively (Supplementary Fig. 1). For the combined dataset λ was 1.04 and 0.92 under fixed and random effects models, respectively (Supplementary Fig. 1), providing little evidence of confounding from population substructure as a source of bias in our meta-analysis.

Aside from 15q25, the strongest evidence for a lung cancer risk loci was found at 6p21.33 (Supplementary Fig. 2 online). Associations were significant for rs3117582 and rs3131379 after adjustment for multiple testing (OR = 1.30, 95% CI = 1.19–1.42; $P = 5.71 \times 10^{-9}$; OR = 1.26, 95% CI = 1.16–1.38; $P = 1.91 \times 10^{-7}$, respectively; Supplementary Table 1), assuming a Bonferroni correction. An additional 12 SNPs mapping to this region also had an association with risk ($P \leq 10^{-5}$).

We observed strong support for an association between rs3117582 and rs3131379 and risk in the UK-GWA and IARC-GWA studies, with P values in each of borderline genome-wide significance ($P = 6.24 \times 10^{-6}$ and 4.41×10^{-6}, respectively). Support was limited in the Texas-GWA study, reflected in the random-effects model ($P = 6.63 \times 10^{-3}$, $P_{\text{het}} = 0.02$ and $P = 0.013$, $P_{\text{het}} = 0.03$ for rs3117582 and rs3131379, respectively). As all three GWA studies were based on subjects with European ancestry, ancestry-related differences are unlikely to underlie between-study heterogeneity, and the nonsignificant association in the Texas-GWA study may reflect study power, especially as these SNPs have low MAFs. To further validate the association between 6p and risk, we genotyped rs3117582 in an additional 2,484 cases and 3,036 controls (UK-Replication series; Supplementary Methods). Genotypes were obtained for 2,448 (98.6%) cases and 2,983 (98.3%) controls. As previously, the C allele was associated with a significantly increased risk (OR = 1.16, 95% CI = 1.04–1.29; $P = 7.30 \times 10^{-3}$). Pooling data from all series provided unequivocal evidence for a relationship between rs3117582 and lung cancer risk ($P = 4.97 \times 10^{-10}$, $P_{\text{het}} = 0.02$, $I^2 = 71\%$; Fig. 1). ORs associated with AC and CC genotypes were 1.20 (95% CI = 1.11–1.29, $P = 7.12 \times 10^{-6}$; $P_{\text{het}} = 0.06$, $I^2 = 60\%$) and 1.80 (95% CI = 1.41–2.30, $P = 2.25 \times 10^{-6}$; $P_{\text{het}} = 0.29$, $I^2 = 19\%$).

rs3117582 (31,728,499 bp) localizes to intron 1 of *BAT3* and rs3131379 (31,829,012 bp) localizes to intron 10 of *MSH5* at 6p21.33 (Fig. 2). Genotypes are highly correlated ($r^2 = 0.99$), hence on the basis of flanking recombination hot spots they define a single locus at 31,676,001–32,303,001 bp. The association could be mediated through LD with a number of transcripts; however, *BAT3* and *MSH5* both represent strong candidates for lung cancer susceptibility. BAT3 is implicated in apoptosis and the protein complexes with E1A-binding protein p300, required for acetylation of p53 in response to DNA damage[5]. MSH5 is involved in DNA mismatch repair (MMR) and meiotic recombination, and deficiency of MMR has been documented to have a role in lung cancer[6–8].

There is some evidence for a risk locus at 6p22.1 (rs9295740; 27,797,481 bp; OR = 1.20, 95% CI = 1.12–1.29; $P = 3.63 \times 10^{-7}$; random effects $P = 3.43 \times 10^{-7}$; Supplementary Table 1 and Fig. 2). LD across 6p21.33–6p22.1 is extensive, and although recombination rates across the region are compatible with an independent susceptibility locus, the moderate LD between rs9295740 and both rs3117582 and rs3131379 (r^2 values of 0.38 and 0.39, respectively) suggests that the associations may be mediated through correlation with the same causal variant.

The most consistent evidence for a new disease locus outside 6p was attained at 5p15.33 (rs401681; OR = 0.88, 95% CI = 0.83–0.93; $P = 4.40 \times 10^{-6}$; $P_{\text{het}} = 0.94$, $I^2 = 0\%$; Supplementary Table 1). rs401681 localizes to intron 13 of *CLPTM1L*[9] within a 60-kb region of LD (1,353,580–1,412,838 bp; Fig. 2 and Supplementary Fig. 3 online) frequently amplified in early-stage NSCLC[10]. Genotyping rs401681 in the UK-Replication series provided further validation of the association. We obtained genotypes for 2,396 (99.7%) cases and 3,001 (98.8%) controls. The A allele was associated with a significantly decreased risk (OR = 0.92, 95% CI = 0.88–0.97; $P = 4.95 \times 10^{-4}$). Pooling data from all series provided unequivocal evidence for a relationship between rs401681 and risk (OR = 0.87, 95% CI = 0.84–0.92; $P = 7.90 \times 10^{-9}$; $P_{\text{het}} = 0.99$, $I^2 = 0\%$; Fig. 1). ORs associated with GA and AA genotypes were 0.86 (95% CI = 0.80–0.92; $P = 2.12 \times 10^{-5}$, $P_{\text{het}} = 0.53$, $I^2 = 0\%$) and 0.77 (95% CI = 0.70–0.84; $P = 3.54 \times 10^{-8}$, $P_{\text{het}} = 0.99$, $I^2 = 0\%$), respectively.

We examined for clinicopathological relationships with rs3117582 and rs401681 in the UK and Texas datasets. The only significant association was between rs3117582 and family history of lung cancer ($P = 0.03$), nonsignificant after adjustment for multiple comparisons (Supplementary Table 2 online). These data suggest that, despite differences in the biology of NSCLC and SCLC, the causal variants affect the risk of all forms of lung cancer, compatible with epidemiological data showing that familial lung cancer risks are not subtype dependent and that intrafamilial histological concordance is poor[11].

The power of our analysis to identify the 15q25.1 and 6p loci at $P = 2.0 \times 10^{-7}$ was high (>80%). In contrast, power to detect alleles with smaller effects and MAFs (for example, those of rs401681) was low. By implication, variants with similar profiles may constitute a larger class of susceptibility loci, whether because of smaller effects or submaximal LD with tagging SNPs.

The 15q25.1 and 6p variants are unlikely to account for >1% of the familial risk, hence a large number of low-risk variants remain to be identified. Further efforts to expand the scale of GWA meta-analyses, in terms of both sample size and SNP coverage, and to increase the number of SNPs taken forward to large-scale replication should identify additional risk variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Cancer Research UK (A1298/A8780) provided principal funding for this study. Additional funding was provided by HEAL and Sanofi-Aventis and US NIH Grants P50CA70907, R01CA121197 and R01CA133996. We would like to thank all individuals who participated in this study. We are grateful to colleagues at UK Clinical Genetics Centres and the UK National Cancer Research Network.

References

1. Amos CI, et al. Nat Genet 2008;40:616–622. [PubMed: 18385676]
2. Hung RJ, et al. Nature 2008;452:633–637. [PubMed: 18385738]
3. Thorgeirsson TE, et al. Nature 2008;452:638–642. [PubMed: 18385739]
4. Power C, Elliott J. Int J Epidemiol 2006;35:34–41. [PubMed: 16155052]
5. Sasaki T, et al. Genes Dev 2007;21:848–861. [PubMed: 17403783]
6. Xu L, et al. Int J Cancer 2001;91:200–204. [PubMed: 11146445]
7. Hirose T, et al. Mol Carcinog 2002;33:172–180. [PubMed: 11870883]
8. Wang YC, Hsu HS, Chen TP, Chen JT. Ann NY Acad Sci 2006;1075:179–184. [PubMed: 17108209]
9. Yamamoto K, Okamoto A, Isonishi S, Ochiai K, Ohtake Y. Biochem Biophys Res Commun 2001;280:1148–1154. [PubMed: 11162647]
10. Kang J, Koo S, Kwon K, Park J, Kim J. Cancer Genet Cytogenet 2008;182:1–11. [PubMed: 18328944]
11. Li X, Hemminki K. Int J Cancer 2004;112:451–457. [PubMed: 15382071]

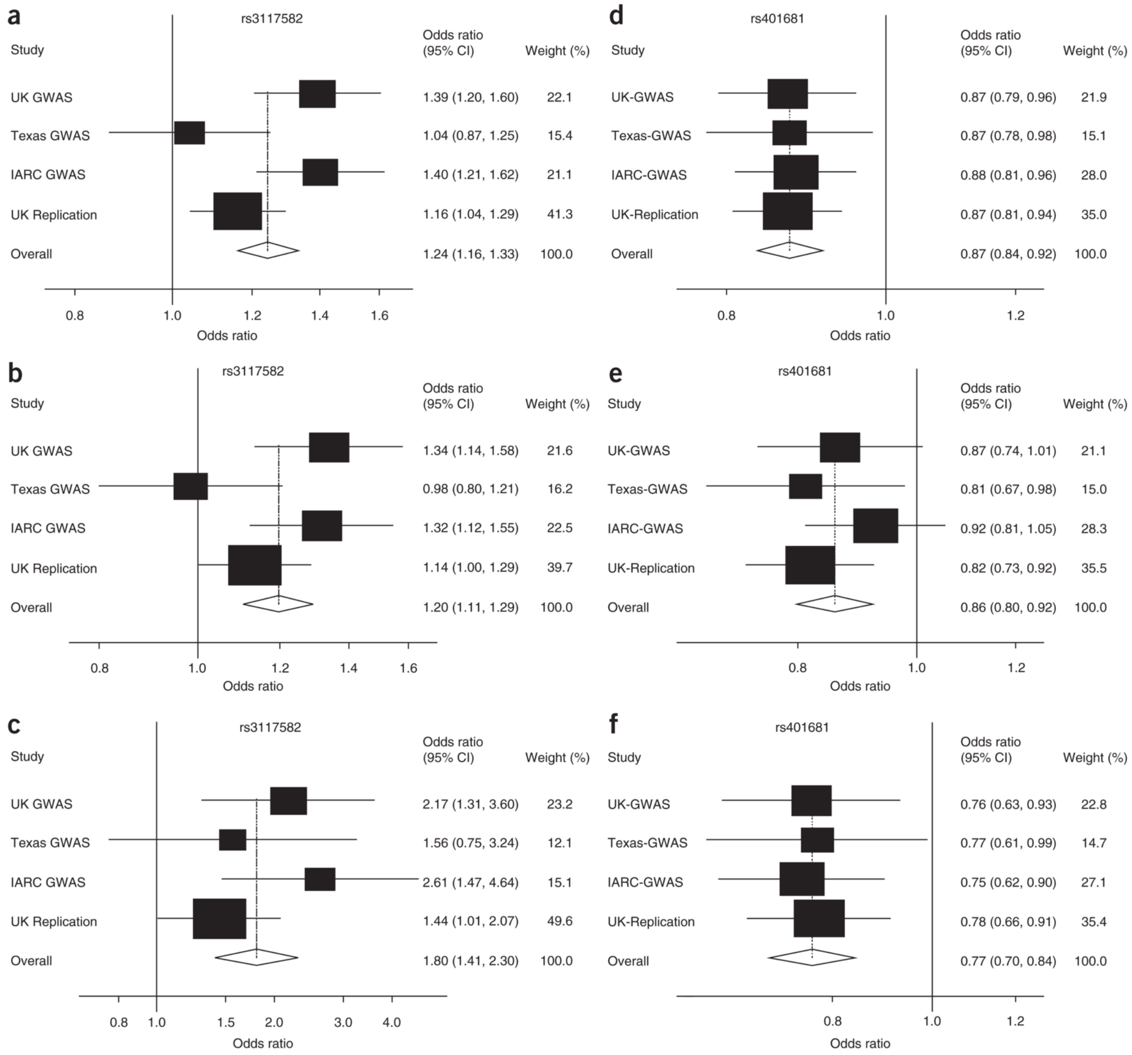

Figure 1.

Summary of association results for rs3117582 and rs401681. (**a**–**f**) Forest plots of rs3117582 per allele ORs (**a**), heterozygote ORs (**b**) and homozygote ORs (**c**); and rs401681 per allele ORs (**d**), heterozygote ORs (**e**) and homozygote ORs (**f**). The *x* axis corresponds to the odds ratio (OR). Horizontal lines represent 95% confidence intervals. Each box represents the OR point estimate and its area is proportional to the statistical weight of the study. The diamonds (and broken lines) represent the summary odds ratios obtained from fixed-effect pooled analysis with 95% confidence intervals given by their widths. The unbroken vertical line is at the null value (OR = 1.0). Frequencies of CC, CA and AA rs3117582 genotypes in cases and controls in the UK-Replication series were 65, 609, 1,774 and 57, 679, 2,247, respectively. Frequencies of AA, AG and GG rs401681 genotypes in cases and controls in the UK-Replication series were 394, 1,134, 868 and 551, 1,506, 994, respectively.

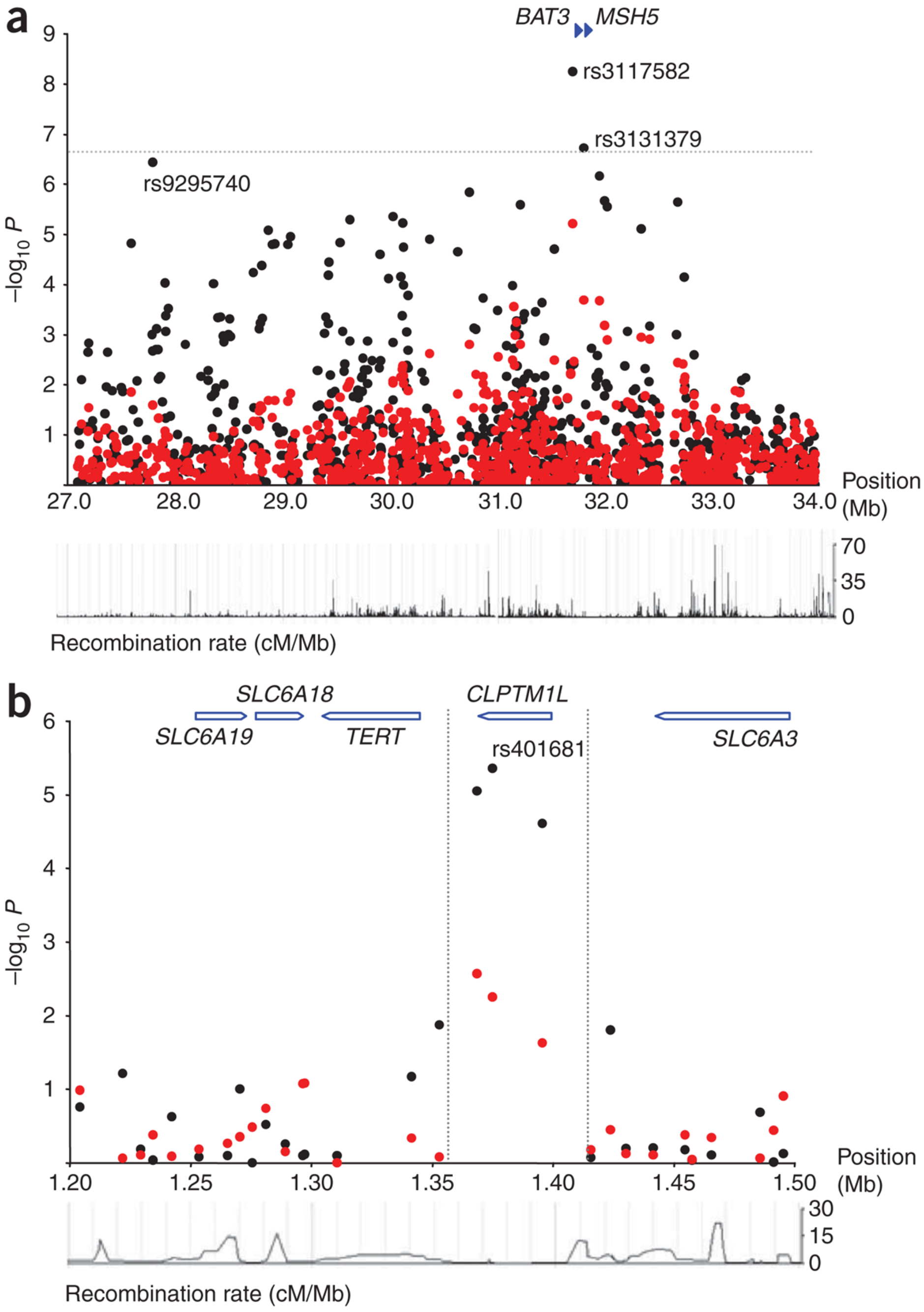

Figure 2.

LD structure and association results for the two confirmed lung cancer–associated regions. (**a**) 6p21.33–6p22.1. (**b**) 5p15.33. Chromosomal positions based on NCBI build 36 coordinates, showing Ensembl (release 48) genes. *P* values (as –$\log_{10}$ values; *y* axis) are shown for SNPs analyzed in UK-GWA study (red circles). Mantel-Haenzel association test *P* values are also shown (black circles). Estimated recombination rates (taken from HapMap) are plotted to reflect the local LD structure around associated SNPs. Annotations of selected genes were taken from the University of California Santa Cruz Genome Browser.