



Published in final edited form as:

Cancer Res. 2009 August 15; 69(16): 6633–6641. doi:10.1158/0008-5472.CAN-09-0680.

Deciphering the impact of common genetic variation on lung cancer risk: A genome-wide association study

Peter Broderick^{1,*}, Yufei Wang^{1,*}, Jayaram Vijayakrishnan¹, Athena Matakidou¹, Margaret R Spitz², Timothy Eisen³, Christopher I. Amos², and Richard S. Houlston¹

¹Section of Cancer Genetics, Institute of Cancer Research, SM2 5NG. UK.

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

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

Abstract

To explore the impact of common variation on the risk of developing lung cancer we conducted a two-phase genome-wide association (GWA) study. In Phase 1, we compared the genotypes of 511,919 tagging single nucleotide polymorphisms (tagSNPs) in 1,952 cases and 1,438 controls; in Phase 2, 30,568 SNPs were genotyped in 2,465 cases and 3,005 controls. SNP selection was based on best supported *P*-values from Phase 1 and two other GWA studies of lung cancer. In the combined analysis of Phases 1 and 2, the strongest associations identified were defined by SNPs mapping to 15q25.1 (rs12914385; $P = 3.19 \times 10^{-16}$), 5p15.33 (rs4975616; $P = 6.66 \times 10^{-7}$), and 6p21.33 (rs3117582; $P = 9.13 \times 10^{-7}$). Variation at 15q25.1, but not 5p15.33 or 6p21.33, was strongly associated with smoking behaviour with risk alleles correlated to higher consumption. Variation at 5p15.33 was shown to significantly influence induction of lung cancer histology. Pooling data from the four series provided 21,620 genotypes for 7,560 cases and 8,205 controls. A meta-analysis provided increased support that variation at 15q25.1 (rs8034191; $P = 3.24 \times 10^{-26}$), 5p15.33 (rs4975616; $P = 2.99 \times 10^{-9}$), and 6p21.33 (rs3117582; $P = 4.46 \times 10^{-10}$) influences lung cancer risk. The next best-supported associations were attained at 15q15.2 (rs748404; $P = 1.08 \times 10^{-6}$) and 10q23.31 (rs1926203; $P = 1.28 \times 10^{-6}$). These data indicate few common variants account for 1% of the excess familial risk underscoring the necessity of having additional large sample series for gene discovery.

Keywords

lung cancer; genome-wide association

Correspondence to: Richard Houlston, Institute of Cancer Research, 15 Cotswold Rd, Sutton, Surrey SM2 5NG, UK, Tel: +44-(0)-208-722-4175, Fax: +44-(0)-208-722-4359, richard.houlston@icr.ac.uk.

*These authors contributed equally to this work

Note: Supplementary information is available on the Cancer Research website.

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

No potential conflicts of interest were disclosed

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests

INTRODUCTION

Lung cancer is a leading cause of cancer death worldwide. While >80% of the population attributable risk of lung cancer can be ascribed to tobacco smoking, several lines of evidence indicate that inherited genetic factors influence the development and progression of lung cancer; in particular epidemiological studies have consistently shown an elevated risk of lung cancer in relatives of lung cancer cases after adjustment for smoking.

Recently genome-wide association (GWA) studies of lung cancer have shown common variation at 15q24–25.1 as a determinant of risk(1–3). Two studies found that the same alleles at this locus increased risk of lung cancer and influenced tobacco smoking behaviour. Genes mapping to this region of association include *CHRNA3*, *CHRNA4*, *CHRNA5*, *PSMA4*, *LOC123688*, and *IREB2*. The *CHRNA* genes encode the nicotinic receptor subunits; in addition to playing a role in development of nicotine dependence, nicotine receptors also influence cell proliferation and apoptosis. Hence these genes represent strong candidates for combined lung cancer susceptibility and predilection to smoking. *PSMA4* encodes the fourth component of the proteasome which plays a role in protein degradation and *IREB2* is involved in iron metabolism which may thus impact on oxidative damage. A second lung cancer locus identified through the GWA studies maps to 5p and includes the genes encoding *TERT* and *CLMPTL1*. In addition to these loci we and others have found statistically significant evidence implicating a third locus at 6p as a risk factor for lung cancer(4,5).

We have previously reported the results of the most extreme hits from Phase 1 of our GWA study with independent replication(4), identifying 15q25, 5p and 6p as disease loci for lung cancer. Here, we report comprehensive findings from our GWA study. In Phase 1 we genotyped 561,466 tagging single nucleotide polymorphisms (SNPs) in 1,978 lung cancer cases, comparing genotype frequencies with 1,438 controls. In Phase 2 we genotyped 33,060 selected SNPs in 2,484 lung cancer cases and 3,036 controls. This analysis in conjunction with a meta-analysis of two other GWA studies(2,4) provides insight into the genetic architecture of inherited susceptibility to lung cancer.

METHODS

Study participants

UK-GWA study Phase 1: Cases (1,182 male, 796 female; mean age at diagnosis 57 years, SD 6) with pathologically confirmed lung cancer were ascertained through the Genetic Lung Cancer Predisposition Study (GELCAPS)(6). All were British residents and self reported to be of European Ancestry. Individuals from the 1958 Birth cohort served as source of controls (7). Comprehensive information on the 1958 Birth Cohort can be obtained through: <http://www.cls.ioe.ac.uk/studies.asp?section=000100020003>.

UK-GWA study Phase 2: An additional 2,484 cases (1,690 male, 794 female; mean age at diagnosis 72 years, SD 7) were ascertained through GELCAPS. Blood samples were obtained from 3,036 healthy individuals (1,497 male, 1,539 female; mean age 61 years, SD 11) recruited to the National Cancer Research Network genetic epidemiological studies, the National Study of Colorectal Cancer (NSCCG; 1999–2006; n = 541), GELCAPS (1999–2004; n = 1,520); and the Royal Marsden Hospital Trust/Institute of Cancer Research Family History and DNA Registry (1999–2004; n = 975). These controls were the spouses or unrelated friends of patients with malignancies. None had a personal history of malignancy at time of ascertainment. All were British residents and self reported to be of European Ancestry.

SNP selection and genotyping

DNA was extracted from samples using conventional methodologies and quantified using PicoGreen (Invitrogen, Carlsbad, USA). Phase 1 of the UK-GWA study was conducted using Illumina Human550 BeadChips according to the manufacturer's protocols (Illumina, San Diego, USA). Phase 2 genotyping was carried out using Illumina Infinium custom arrays according to the manufacturer's protocols. Selection of SNPs were based on a stepwise procedure (Supplementary Figure 1); the majority, 20,000, were chosen by a hypothesis-free (agnostic) strategy, simply on the basis of being most significantly associated with lung cancer risk in Phase 1. The remainder were selected on an alternative basis, briefly: 1,799 additional SNPs (annotated in dbSNP) were included in the 15q25.1, 6p21.33 and 5p15.33 regions, which had been previously reported to be associated with disease risk. 79 SNPs showing an association with lung cancer risk in previously reported GWA studies (IARC-GWA and Texas-GWA studies) at $P < 10^{-4}$, not captured by the 20,000 most significant SNPs in Phase 1 and fine mapping SNPs. 11,182 agnostic SNPs not included in the aforementioned criteria, based on being most significantly associated with lung cancer in a previously reported meta-analysis of Phase 1 and IARC-GWA and Texas-GWA studies(4).

DNA samples with GenCall scores < 0.25 at any locus were considered "no calls". A DNA sample was deemed to have failed if it generated genotypes at $< 95\%$ of loci. A SNP was deemed to have failed if fewer than 95% of DNA samples generated a genotype at the locus. To ensure quality of genotyping, a series of duplicate samples were genotyped and cases and controls were genotyped in the same batches.

Meta-analysis

A meta-analysis pooling both phases of our UK-GWA study with data from two other studies: IARC-GWA study of 1,989 cases and 2,625 controls(2), summary data from which is publicly available; Texas-GWA study of 1,154 non-small cell lung cancer (NSCLC) cases who were all smokers and 1,137 smoking matched controls(4). Comprehensive details of case and control ascertainment and matching criteria, as well as the genotyping of Texas-GWA and IARC-GWA studies have been published previously(2,4).

Ethical approval for the UK study was obtained from the London Multi-Centre Research Ethics Committee (MREC/98/2/67) in accordance with the tenets of the Declaration of Helsinki. All participants provided informed consent.

Statistical analysis

Statistical analysis was undertaken using S Plus v7.0 (Insightful, New York, US), R v2.8.0 and STATA v8.0 (Station College, Texas, US) Software. Genotype data were used to search for duplicates and closely related individuals amongst all samples in Phase 1. Identity by state values were calculated for each pair of individuals on 22,120 SNPs, and for any pair with allele sharing $> 80\%$, the sample generating the lowest call rate was removed from further analysis. In Phase 1, genotyped samples were excluded from further analyses for the following reasons: gender discrepancy ($n = 6$), duplicated ($n = 0$), relatedness ($n = 0$).

The adequacy of case-control matching and possibility of differential genotyping of cases and controls were formally evaluated using Q-Q plots of $-\log_{10}P$ values (based on the 90% least significant SNPs). Deviation of the genotype frequencies in the controls from those expected under Hardy-Weinberg Equilibrium (HWE) was assessed by χ^2 test, or Fisher's exact test where an expected cell count was < 5 . Comparison of the difference in number of associations observed and expected was made using the binomial test.

The association between each SNP and risk was assessed by the allele test. Odds ratios (ORs) and associated 95% confidence intervals (CIs) were calculated by unconditional logistic regression. Associations by histology (NSCLC, small cell lung cancer [SCLC]) were examined by logistic regression in case-only analyses.

Pooling of Phase 1 and Phase 2 data were based on individual genotypes. We imposed stringent criteria for call rates of SNPs and checked for significant disparity of MAFs between series. Only summary data was available for the Texas and IARC-GWA studies. To minimise errors in data harmonisation we examine for deviation in MAF for SNPs in cases and controls across datasets.

Meta-analysis was conducted using standard methods for combining raw data based on the Mantel-Haenszel method and weighted average of study-specific estimates of the ORs(8). Cochran's Q statistic to test for heterogeneity(8) and the I^2 statistic to quantify the proportion of the total variation due to heterogeneity were calculated. This was performed using Metagen module from Meta library for R.

The sibling relative risk attributable to a given SNP was calculated using the formula(9):

$$\lambda^* = \frac{p(pr_2 + qr_1)^2 + q(pr_1 + q)^2}{[p^2r_2 + 2pqr_1 + q^2]^2}$$

where p is the population frequency of the minor allele, $q = 1 - p$, and r_1 and r_2 are the relative risks (estimated as OR) for heterozygotes and rare homozygotes, relative to common homozygotes. Assuming a multiplicative interaction the proportion of the familial risk attributable to a SNP was calculated as $\log(\lambda) / \log(\lambda_0)$, where λ_0 is the overall familial relative risk estimated from epidemiological studies, assumed here to be 1.8(10).

The impact of variants on smoking behaviour was assessed by comparing the prevalence of alleles stratified by cigarette consumption (cigarettes per day, CPD) using the Cochran-Armitage test. We also used the Kruskal-Wallis test to analyse for differences in cigarette consumption stratified by genotype. To test for independent effect of variants on CPD, genotype frequencies in light and medium smokers were compared to frequencies in heavy smokers using multinomial logistic regression.

We estimated power of our analysis to identify associations over a range of MAFs assuming joint analysis and a multiplicative effect model for each SNP.

Bioinformatics

We used Haploview (v3.2) to infer the LD structure of the genome in the regions containing loci associated with disease risk.

RESULTS

UK-GWA study

Our previous publication details numbers of SNPs genotyped and quality controlled in Phase 1(4). Briefly, a total of 552,974 SNPs were satisfactorily genotyped (99.7%). Of the SNPs satisfactorily genotyped, 524,714 were common to both cases and controls. Several quality control processes were sequentially applied to these SNPs (Figure 1a), leaving 511,919 SNPs for which genotype data were informative. For the informative SNPs mean individual sample call rates (the percentage of samples for which a genotype was obtained for each SNP) were 99.8% and 99.6% in cases and controls, respectively. Comparison of the observed and expected

distributions showed little evidence for an inflation of the test statistics (inflation factor $\lambda=1.02$, based on the 90% least significant SNPs), thereby excluding systematic bias. Of the 1978 cases, 1958 were successfully genotyped and following quality control, 1952 cases were used in the analysis (Figure 1a).

In Phase 2 a total of 31,039 SNPs (93.9%) were successfully genotyped, of which 30,568 provided reliable genotypes according to our quality control metrics (Figure 1b). Of the 2,484 cases and 3,036 controls attempted, 2,465 cases and 3,014 controls were successfully genotyped and of these 2,465 cases and 3,005 controls were suitable for analysis (Figure 1b). In the combined analysis of Phase 1 and 2 (Supplementary Table 1), the strongest associations identified were found at polymorphic sites defined by SNPs mapping to 15q25.1 (rs12914385; $P = 3.19 \times 10^{-16}$), 5p15.33 (rs4975616; $P = 6.66 \times 10^{-7}$), and 6p21.33 (rs3117582; $P = 9.13 \times 10^{-7}$), which have been the subject of previous fast-tracking analyses.

On the basis of the combined Phase 1 and Phase 2 GWA data the 15q25.1 association is defined by a 248kb region of strong LD on chromosome 15 extending from 76,499,754bp to 76,747,584bp (Figure 2a). Maximal evidence of a relationship was provided by the SNP rs12914385, which maps at 76,685,778bp (OR = 1.29, 95% CI: 1.21–1.37; $P = 3.19 \times 10^{-16}$; Supplementary Table 1). rs938682 and rs8042374 also provide strong evidence for an association between 15q25 and lung cancer risk (Supplementary Table 2). The relative position of these 3 SNPs to the *CHRNA3*, *CHRNA5*, *CHRNA4*, *IREB2*, *PSMA4*, and *LOC123688* transcripts, which map to the 248kb region of LD within 15q25.1, is shown in Figure 2a. All three SNPs localise to intron 4 of *CHRNA3* strongly favouring variation within this gene as being the basis of 15q25 lung cancer association. rs12914385 and rs938682 / rs8042374 appear to tag different blocks of LD (respective r^2 values for: rs938682- rs12914385, rs938682- rs8042374 rs12914385-rs8042374 are 0.22, 1.00 and 0.22 based on HapMap CEU, and 0.18, 0.99, and 0.17 based on UK-GWA Phase 2 controls). The possibility of two independently acting loci is supported by logistic regression whereby the ORs for rs12914385 were 1.29 ($P_{trend} = 4.79 \times 10^{-16}$) and 1.20 ($P_{trend} = 1.81 \times 10^{-7}$) without and with adjustment for rs8042374. Similarly the ORs for rs8042374 were 0.75 ($P_{trend} = 5.82 \times 10^{-15}$) and 0.82 ($P_{trend} = 2.13 \times 10^{-6}$) without and with adjustment for rs12914385 (Table 1).

At 5p15.33 the best evidence for an association was provided by rs4975616 (OR = 0.86, 95% CI: 0.81–0.91; $P = 6.66 \times 10^{-7}$; Supplementary Table 1) which localizes within a 60kb region of LD (1,353,580–1,412,838bp) between *TERT* and *CLPTMIL* (Figure 2b). It has been proposed that 5p15.33 harbours two independent loci for lung cancer risk(5). The first is defined by rs402710 which maps within intron 16 of *CLPTMIL* and is in strong LD with rs4975616 ($r^2 = 0.53$ based on HapMap CEU; $r^2 = 0.55$ based on UK-GWA Phase 2 controls). The second association signal is defined by rs2736100 which maps within intron 2 of *TERT*. In our combined data series the association between rs2736100 and lung cancer risk was however, weak: OR = 0.96 ($P_{trend} = 0.19$) and OR = 0.97 ($P_{trend} = 0.37$) without and with adjustment for rs4975616 (Supplementary Table 3).

The strongest evidence for a relationship between genetic variation at 6p21.33 and lung cancer risk was attained at rs3117582 and rs1150752 (OR = 1.24, 95% CI: 1.14–1.35; $P = 9.13 \times 10^{-7}$; 11 OR = 1.24, 95% CI: 1.13–1.35; $P = 1.93 \times 10^{-6}$, respectively; Supplementary Table 1). rs3117582 (31,728,499bp) localizes to intron 1 of *BAT3* and rs1150752 (32,172,704bp) localises to exon 3 of *TNXB* (Figure 2c). Genotypes are highly correlated ($r^2 = 0.73$ based on HapMap CEU; $r^2 = 0.91$ based on UK-GWA Phase 2 controls) and on the basis of flanking recombination hotspots define a single locus at 31,676,001–32,303,001bps.

Excluding the SNPs mapping to 15q25.1, 5p15.33 and 6p21.33 loci, the most significant association was provided by rs11264329 and rs2844363 ($P = 1.22 \times 10^{-6}$ and 5.90×10^{-6}

respectively), which map to 153,361,782bp on 1q22 and 37,586,864bp on 3p22.2 (Supplementary Table 1). Whilst suggestive of association none were statistically significant imposing the conventionally accepted threshold for genome-wide significance (i.e. 1×10^{-7}).

15q25.1, 5p15.33 and 6p21.33 variants and lung cancer histology and smoking behavior

In view of the differences in biology of NSCLC and SCLC we examined the relationship between 15q25.1, 5p15.33 and 6p21.33 variants and tumor histology (Table 2). Variation at 15q25.1 defined by rs12914385, rs8042374 or rs9838682 was not associated with any difference in lung cancer histology. At 5p15.33 variation defined by rs4975616 (*CLPMIL*) was not associated with any difference in lung tumor type, however, variation defined by rs2736100 (*TERT*) was shown to influence lung cancer histology. Specifically, there was a significant difference in the allele frequency of rs2736100 between SCLC and NSCLC ($P = 0.0011$). This association was ascribable to a significantly increased frequency of the risk allele in cases with NSCLC-adenocarcinoma. Similarly for variation at 6p21.33 defined by rs3117582, whilst allele frequencies were not significantly different between SCLC and NSCLC cases ($P = 0.15$), a significant difference in allele frequency between adenocarcinoma and squamous disease was shown.

We investigated for an association between 15q25.1, 5p15.33 and 6p21.33 variants and smoking behaviour by studying the relationship with consumption of cigarettes per day (CPD) categorised into different levels of smoking quantity (Table 3). A strong relationship between all 15q25.1 variants (rs12914385, rs8042374 and rs938682) and smoking was observed (Table 3). Statistically significant allele-dependent associations between risk genotype and cigarette consumption was seen in cases. A similar trend was observed in controls for each of the SNPs but was not statistically significant. Adjusting rs8042374 or rs938682 for rs12914385 provided evidence of an independent effect of the two loci on smoking behaviour ($P < 0.05$). No significant association was observed between 5p15.33 (rs4975616, rs2736100) and 6p21.33 (rs3117582) genotypes and smoking.

Meta-analysis of GWA studies

To facilitate the identification of additional risk variants we conducted a meta-analysis pooling our UK-GWA Phase 1 and Phase 2 with two other studies: IARC-GWA and Texas-GWA. Pooling was based on the 21,620 autosomal SNPs genotyped in all three GWA studies which had MAFs $> 1\%$ and no departure from HWE ($P \leq 10^{-5}$ in cases and controls).

As expected the strongest associations were obtained for SNPs mapping to 15q25.1 (rs8034191, OR = 1.29, 95% CI: 1.23–1.35; $P = 3.24 \times 10^{-26}$), 5p15.33 (rs4975616, OR = 0.87, 95% CI: 0.83–0.91; $P = 2.99 \times 10^{-9}$), and 6p21.33 (rs3117582, OR = 1.24, 95% CI: 1.16–1.33; $P = 4.46 \times 10^{-10}$) (Supplementary Table 4). In the meta-analysis the strongest association at 15q25.1 was for rs8034191 which lies 80kb outside of *CHRNA3*. This SNP is in strong LD with rs12914385 ($r^2 = 0.72$ based on HapMap CEU; $r^2 = 0.73$ based on UK-GWA Phase 2 controls). Excluding SNPs mapping to 15q25.1, 5p15.33, 6p21.33, seven SNPs were associated with lung cancer risk at $P < 10^{-5}$ (Supplementary Table 5). The most significant association is provided by rs748404 (OR = 0.87, 95% CI: 0.83–0.92; $P = 1.08 \times 10^{-6}$), mapping to 41,346,523bp on 15q15.2. Two other SNPs associated with lung cancer risk at $P < 10^{-5}$ also map to 15q15.2 (rs504417 and rs11853991, 41,341,518bp and 41,344,841bp respectively) and are in strong LD with rs748404 (respective r^2 values for: rs748404–rs504417, rs748404–rs504417 are 0.65 and 0.68 based on HapMap CEU, and 0.59 and 0.61 based on UK-GWA Phase 2 controls) rs748404 maps 3' to *Transglutaminase 5* (*TGM5*), with rs504417 and rs11853991 mapping to intron 1 of this gene.

Architecture of genetic susceptibility to lung cancer

On the basis of MAFs and associated genotypic risks we estimate the 5p15.33 and 6p21.33 variants individually account for ~1% of the excess familial risk, with the 15q25.1 locus having a much greater impact (~5%). To gain insight into the basis of the inherited risk of lung cancer in general we estimated the power of our analyses to identify disease-associated loci with different MAF which would account for 1% of the familial risk (Figure 3). With the UK-GWA study we had greater than 90% power to harvest variants with similar characteristics to 15q25.1. However, we only had ~30% and 40% power to identify variants such as 5p15.33 and 6p21.33 which have much weaker effects. Using all four datasets our meta-analysis was well powered to identify common variants (MAF >0.15), provided each accounts for $\geq 1\%$ excess risk. Clearly for variants such as 5p15.33 power still remained limited.

DISCUSSION

These analyses provide increased support that variation at 15q25.1, 5p15.33, and 6p21.33 influences the risk of developing lung cancer. Our estimate of the contribution of 15q25.1, 5p15.33 and 6p21.33 loci to the excess familial risk of lung cancer is likely to be conservative as the effect of the causal variant will typically be larger than the association detected through a tag SNP. This is especially relevant with respect to the 15q25.1 association as we provide evidence for two independent loci. Furthermore, since a high proportion of UK-GWA study Phase 2 controls were spouses and unrelated friends of lung cancer cases, over-matching on life-time smoking exposure (i.e., cases and controls may have been more likely to be concordant on smoking status than individuals of the general population) may have impacted on study findings. Hence risk estimates for smoking-related SNPs identified in our analysis may be attenuated. In addition multiple causal variants may exist at each locus including low frequency variants with significantly larger effects on risk. This may impact significantly on the contribution of *CHRNA5-A3* region to the familial risk of lung cancer, especially as our analysis provides evidence to support independent alleles at this locus.

Identification of the causal variants for these loci will be challenging, contingent on resequencing and fine mapping studies. While in part speculative current data provides information on the probable genetic basis of the associations at 15q25.1, 5p15.33, and 6p21.33. While there is strong evidence that a major component (some would assert all) of the lung cancer risk associated with 15q25.1 is mediated through propensity to smoke and hence a higher exposure to smoking-related carcinogens(3,11), it does not exclude the possibility 15q25.1 variation also has a direct effect on lung cancer risk as has also been proposed(2,11).

Whilst our UK study does not support the tenet of two independent loci at 5p15.33 for lung cancer which has recently been proposed (5), data from both IARC and Texas provides strong evidence for an independent locus (Supplementary Table 6). Stratification of UK data by histology does, however provide strong evidence that rs2736100 genotype influences the histology of lung cancer induction favouring development of NSCLC-adenocarcinoma (Table 2). It is therefore noteworthy in that the Texas GWA study was based on analysis of only NSCLC cases and despite its relatively small size a strong association between rs2736100 and risk was detected. Both *CLPTMIL* and *TERT* which map to 5p represent attractive candidates for lung cancer susceptibility *a priori* assuming the causal variant exerts an influence through a *cis* effect. The biology of *TERT* makes it an attractive candidate for a gene that influences lung cancer risk and moreover association between rs401681 risk allele and shorter telomere length has recently been reported(12). High levels of PAH adducts correlate with lung cancer risk and as a major effect of platinum on cells is through adduct formation, *CLPTMIL* (alias *CRR9*) is also an attractive lung cancer susceptibility gene as it encodes a transcript whose over-expression has been linked to cis-platinum resistance(13).

The 6p21.33 association could be mediated through any number of transcripts mapping to the region of LD. *BAT3* represents a strong candidate for lung cancer susceptibility as it is implicated in apoptosis and the protein complexes with E1A-binding protein p300, required for acetylation of p53 in response to DNA damage(14). As the region of LD at 6p21.33 is extensive and contains a large number of transcripts, dissection and elucidation of the causal variant make this the most refractory to interrogation.

Findings from this GWA study provide insight into the allelic architecture of predisposition to lung cancer. Pooling data from our study with two other GWA studies provided a combined analysis of 7,560 cases and 8,205 controls. Nevertheless we have failed to identify additional loci to 5p15.33, 6p21.33 and 15q25.1. As our power to detect the major common loci conferring risks of 1.2 or greater was high we consider that there are unlikely to be many additional SNPs (tagged by the Illumina 550K array) with similar effects for alleles with frequencies > 0.2 in populations of European ancestry. Inevitably we had low power to detect alleles with smaller effects and/or MAFs < 0.1. By implication, variants with such profiles may represent a much larger class of susceptibility loci for lung cancer and current GWA based strategies based on the currently available commercial arrays are not optimally configured to identify low frequency variants with potentially stronger effects. Thus there may a large number of low penetrance variants that remain to be discovered. 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 may thus lead to the identification of additional variants for lung cancer. Twin studies have not however, provided evidence for heritable factors for the risk of lung cancer(15,16). The identification of disease-causing alleles for lung cancer may thus be inherently harder than for other cancers in which familial aggregation of a major lifestyle/environmental risk factor is less likely to be a confounder.

URLs

The R suite can be found at <http://www.r-project.org/>

Detailed information on the tag SNP panel can be found at <http://www.illumina.com/>

dbSNP: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=snp>

HAPMAP: <http://www.hapmap.org/>

Genetic Lung Cancer Predisposition Study (GELCAPS):
<http://pfsearch.ukcrn.org.uk/StudyDetail.aspx?TopicID=1&StudyID=781>

National Study of Colorectal Cancer Genetics (NSCCG):
<http://pfsearch.ukcrn.org.uk/StudyDetail.aspx?TopicID=1&StudyID=1269>

1958 Birth Cohort: <http://www.cls.ioe.ac.uk/studies.asp?section=000100020003>

Central Europe data from IARC-GWAS: <http://www.ceph.fr/cancer>

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

Grant funding: This work was supported by Cancer Research UK (C1298/A8780 and C1298/A8362- Bobby Moore Fund for Cancer Research UK) who provided principal funding for this study. Athena Matakidou was the recipient of a clinical research fellowship from the Allan J Lerner Fund. We are also grateful to NCRN, HEAL and Sanofi-

Aventis. Additional funding was obtained from NIH grants 5R01CA055769, 5R01CA127219, 5R01CA133996, and 5R01CA121197. We would like to thank all individuals that participated in this study and the clinicians who took part in the GELCAPS consortium. This study made use of genotyping data on the 1958 Birth Cohort, this data was generated and generously supplied to us by Panagiotis Deloukas of the Wellcome Trust Sanger Institute. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

REFERENCES

1. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 2008;40(5):616–22. [PubMed: 18385676]
2. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008;452(7187):633–637. [PubMed: 18385738]
3. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008;452(7187):638–642. [PubMed: 18385739]
4. Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* 2008;40(12):1407–1409. [PubMed: 18978787]
5. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet* 2008;40(12):1404–1406. [PubMed: 18978790]
6. Eisen T, Matakidou A, Houlston R. Identification of low penetrance alleles for lung cancer: the Genetic Lung Cancer Predisposition Study (GELCAPS). *BMC Cancer* 2008;8:244. [PubMed: 18715499]
7. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* 2006;35(1):34–41. [PubMed: 16155052]
8. Petitti, D. *Meta-analysis Decision Analysis and Cost-Effectiveness Analysis*. New York, Oxford: Oxford; 1994.
9. Houlston RS, Ford D. Genetics of coeliac disease. *QJM* 1996;89(10):737–743. [PubMed: 8944229]
10. Matakidou A, Eisen T, Houlston RS. Systematic review of the relationship between family history and lung cancer risk. *Br J Cancer* 2005;93(7):825–833. [PubMed: 16160696]
11. Spitz MR, Amos CI, Dong Q, Lin J, Wu X. The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. *J Natl Cancer Inst* 2008;100(21):1552–1556. [PubMed: 18957677]
12. Rafnar T, Sulem P, Stacey SN, et al. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat Genet* 2009;41(2):221–227. [PubMed: 19151717]
13. Yamamoto K, Okamoto A, Isonishi S, Ochiai K, Ohtake Y. A novel gene, CRR9, which was up-regulated in CDDP-resistant ovarian tumor cell line, was associated with apoptosis. *Biochem Biophys Res Commun* 2001;280(4):1148–1154. [PubMed: 11162647]
14. Sasaki T, Gan EC, Wakeham A, Kornbluth S, Mak TW, Okada H. HLA-B-associated transcript 3 (Bat3)/Scythe is essential for p300-mediated acetylation of p53. *Genes Dev* 2007;21(7):848–861. [PubMed: 17403783]
15. Braun MM, Caporaso NE, Page WF, Hoover RN. Genetic component of lung cancer: cohort study of twins. *Lancet* 1994;344(8920):440–443. [PubMed: 7914565]
16. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000;343(2):78–85. [PubMed: 10891514]

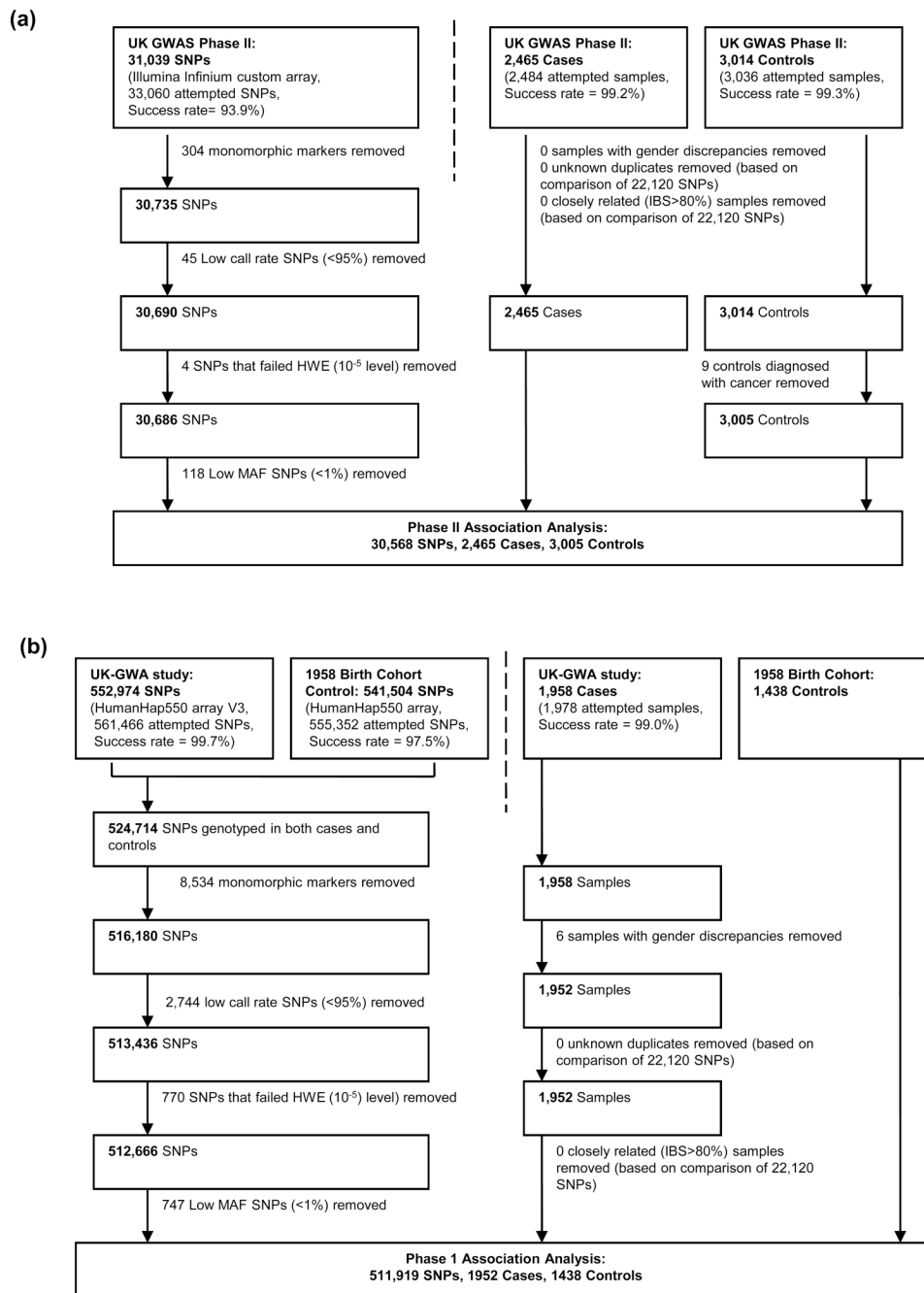


Figure 1. UK-GWA study data quality control of (a) Phase 1, (b) Phase 2

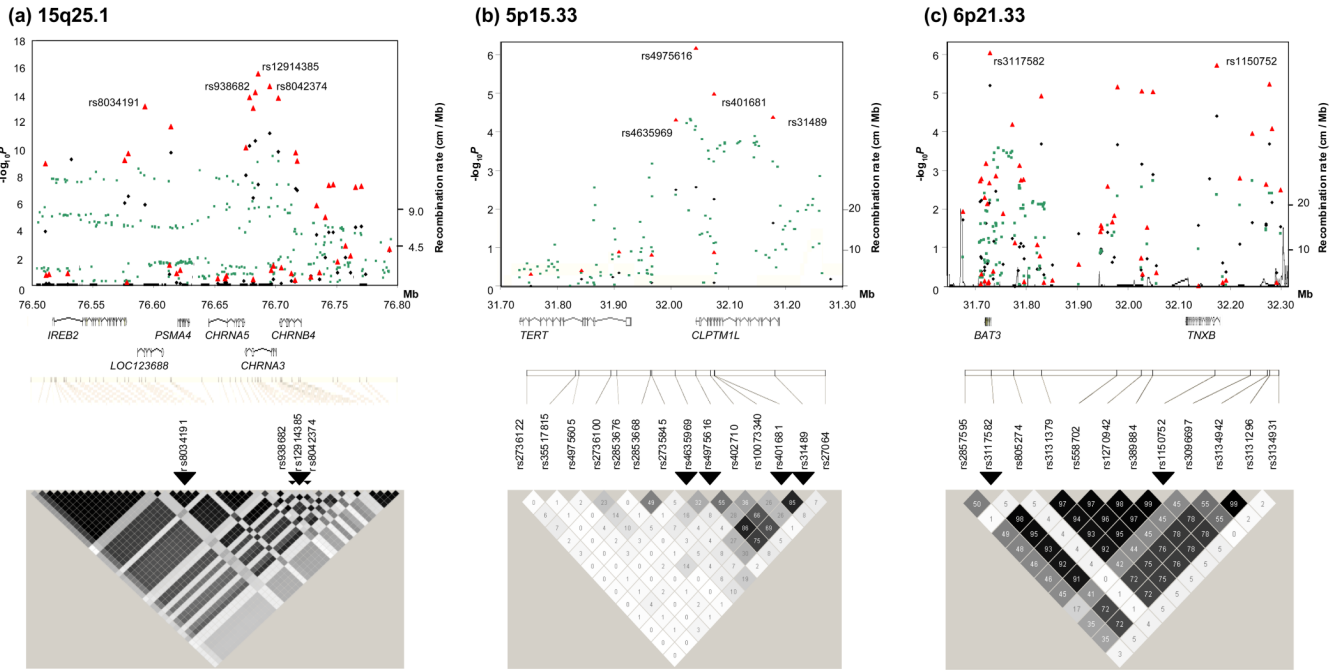


Figure 2. Regional plots of the (a) 15q25.1 (b) 5p15.33 and (c) 6p21.33 associations
 Each panel shows single-marker association statistics (as $-\log_{10}P$) from the analysis of UK-GWA study Phase 1 (diamonds), UK-GWA study Phase 2 (squares), Phases 1 and 2 combined (triangles), as a function of genomic position (NCBI build 36.1). The recombination rate across each region in HapMap CEU is shown in black (right y axis). Also shown for 15q25.1 (a) and 5p15.33 (b) is the relative position of genes mapping to each region of association, there are a large number of genes mapping to the 6p21.33 (c) region so for clarity only *BAT3* and *TNXB* are illustrated. Exons of genes have been redrawn to show the relative positions in the gene, therefore maps are not to physical scale. LD plot was generated using UK-GWA study Phase 2 controls; values and shading show r^2 between each pair of SNPs; the darker the shading, the greater extent of LD.

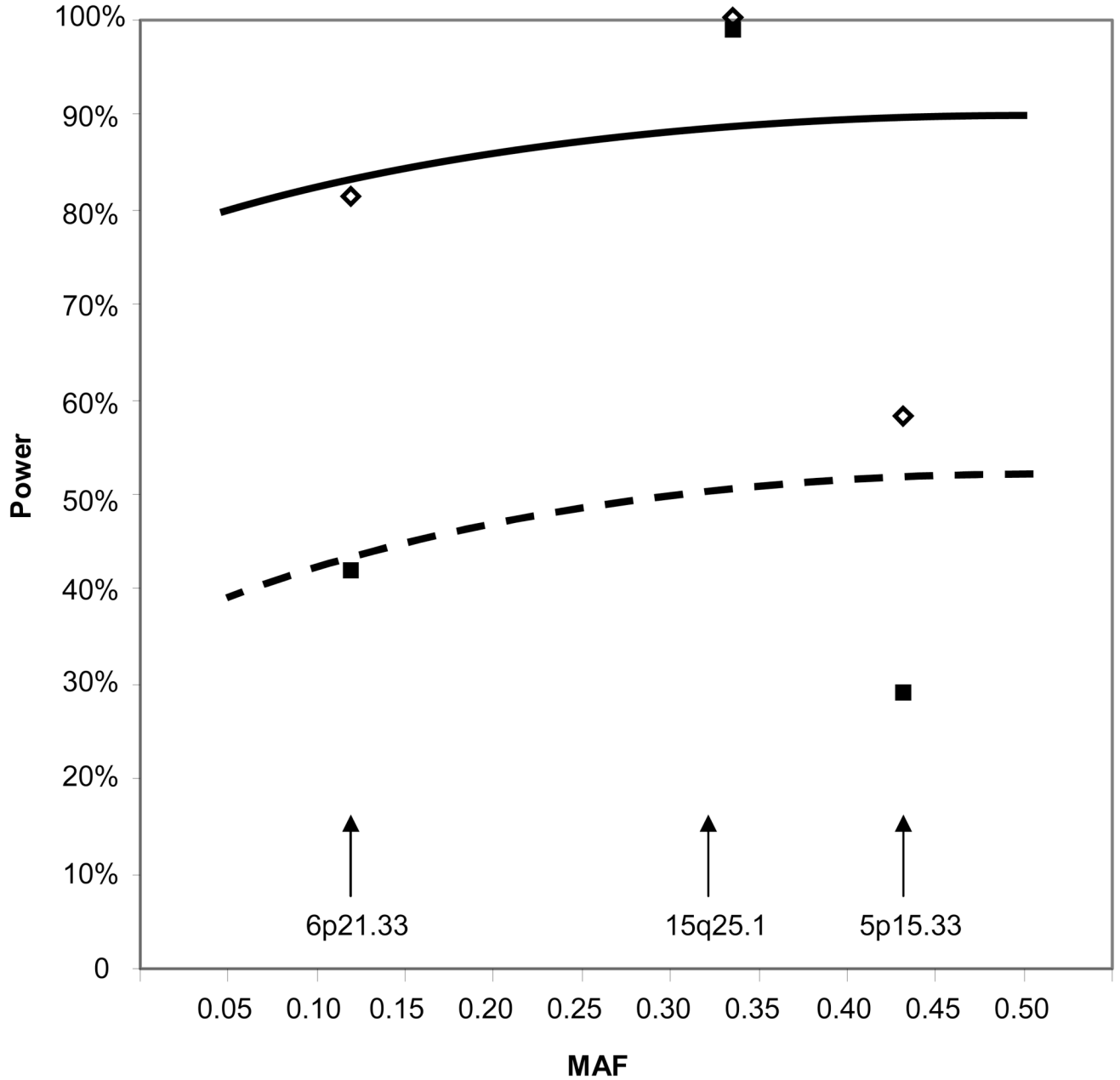


Figure 3. Power to detect lung cancer susceptibility alleles in UK-GWA study (Phases 1 and 2 combined) and the UK-IARC-Texas GWA studies meta-analysis
 The dashed and solid lines show the power of the UK-GWA study (Phases 1 and 2 combined) and the UK (Phases 1 and 2)-IARC-Texas-GWA studies meta-analysis to identify susceptibility alleles with different minor allele frequencies respectively. Power to identify 5p15.33 (rs4975616), 6p21.33 (rs3117582), and 15q25.1 (rs12914385), variants in each analysis denoted by squares and diamonds respectively ($P= 10^{-7}$).

Table 1

Association (in combined analysis of UK-GWA study Phases 1 and 2) between 15q25.1 SNPs (rs938682, rs12914385 and rs8042374) and lung cancer, with and without adjustment for a second variant.


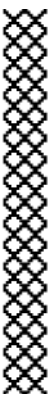

	rs938682	rs12914385	rs8042374
	<i>P_{trend}</i>	<i>P_{trend}</i>	<i>P_{trend}</i>
	OR (95% CI)	OR (95% CI)	OR (95% CI)
	<i>P_{trend}</i>	<i>P_{trend}</i>	<i>P_{trend}</i>
	OR (95% CI)	OR (95% CI)	OR (95% CI)
Unadjusted	1.45 × 10 ⁻¹⁴	4.79 × 10 ⁻¹⁶	5.82 × 10 ⁻¹⁵
Adjusted for:			
rs938682		1.29 (1.21–1.37)	0.75 (0.70–0.81)
rs12914385	4.17 × 10 ⁻⁶		0.94 (0.36–2.48)
rs8042374	0.63	1.20 (1.12–1.28)	0.82 (0.76–0.89)
		1.81 × 10 ⁻⁷	

Table 2

Association between 15q25.1 (rs12914385, rs8042374, rs938682), 5p15.33 (rs4975616, rs2736100), 6p21.33 (rs3117582) and lung cancer histology in combined analysis of UK-GWA study Phases 1 and 2.

SNP	SCLC												NSCLC															
	All				Adeno				Squamous				Other				SCLC vs NSCLC				Adeno vs Squamous							
	N	MAF	N	MAF	N	MAF	N	MAF	N	MAF	N	MAF	N	MAF	N	MAF	N	MAF	OR	(95%CI)	P-value	OR	(95%CI)	OR	(95%CI)	P-value		
15q25.1																												
rs12914385 (CHRNA3)	1035	0.44	3344	0.44	903	0.43	1677	0.44	764	0.44	764	0.71	1.02	(0.92-1.13)	0.63	0.97	(0.87-1.09)											
rs8042374 (CHRNA3)	1033	0.19	3339	0.19	901	0.20	1675	0.18	764	0.18	764	0.76	0.98	(0.86-1.11)	0.20	1.10	(0.95-1.27)											
rs938682 (CHRNA3)	1035	0.19	3345	0.19	903	0.20	1678	0.18	764	0.18	764	0.78	1.02	(0.90-1.16)	0.18	0.91	(0.78-1.05)											
5p15.33																												
rs4975616 (CLPTM1L)	1035	0.40	3340	0.39	901	0.38	1675	0.39	764	0.39	764	0.46	1.04	(0.94-1.15)	0.85	0.99	(0.88-1.11)											
rs2736100 (TERT)	1034	0.51	3343	0.47	903	0.44	1677	0.49	764	0.49	764	0.0011	1.18	(1.07-1.30)	7.2 × 10 ⁻⁴	0.82	(0.73-0.92)											
6p21.33																												
rs3117582 (BAT3)	1034	0.14	3342	0.16	903	0.14	1677	0.17	764	0.17	764	0.15	0.90	(0.79-1.04)	0.02	0.83	(0.71-0.98)											

Table 3

Association between smoking behaviour and (a) genotype, (b) allele frequency

Association between 15q25.1 (rs12914385, rs8042374, rs938682), 5p15.33 (rs4975616, rs2736100), 6p21.33 (rs3117582), and smoking behaviour assessed by studying the relationship with consumption of cigarettes per day (CPD). Complete CPD information was available for 4019 UK-GWA study (Phases 1 and 2) cases and 907 UK-GWA study Phase 2 control samples. CPD were categorised into different smoking quantities: 1 – 18 cigarettes per day, 19 –24 cigarettes per day, 25 or more cigarettes per day; strata defined to ensure number of individuals in each approximately equal.

(a) 15q25.1	Cases			Controls		
	n	Mean CPD	<i>P</i> ¹	n	Mean CPD	<i>P</i> ¹
rs12914385						
CC	1230	21.59		373	18.13	
CT	1973	22.19		413	18.46	
TT*	815	23.91	3.7 × 10 ⁻³	121	18.53	0.17
rs8042374						
Cases						
	n	Mean CPD	<i>P</i> ¹	n	Mean CPD	<i>P</i> ¹
AA*	2649	22.65		541	18.57	
AG	1219	21.96		314	18.00	
GG	144	20.16	2.4 × 10 ⁻³	52	17.86	0.48
rs938682						
Cases						
	n	Mean CPD	<i>P</i> ¹	n	Mean CPD	<i>P</i> ¹
TT*	2648	22.66		542	18.09	
CT	1229	21.98		312	17.89	
CC	142	19.99	1.8 × 10 ⁻³	53	18.61	0.37

(a)	Cases				Controls			
	n	Mean CPD	P^I	P^I	n	Mean CPD	P^I	P^I
<i>I5q25.1</i>								
rs12914385								
<i>5p15.33</i>								
rs4975616								
AA*	1491	22.33			301	17.65		
AG	1893	22.57			455	18.91		
GG	630	21.77	0.78		151	17.97	0.56	
rs2736100								
GG*	1071	22.19			217	18.17		
GT	2009	21.41			457	18.35		
TT	936	22.42	0.95		233	18.45	0.94	
<i>6p21.33</i>								
rs3117582								
AA	2892	22.36			679	18.30		
AC	1008	22.35			209	18.45		

(a)		Cases		Controls		
	n	Mean CPD	P^1	n	Mean CPD	P^1
<i>15q25.1</i>	115	22.27	0.54	19	18.11	0.97
rs12914385						
<hr/>						
(b)		Cases		Controls		
	n	MAF	P^2	n	MAF	P^2
<i>15q25.1</i>						
rs12914385						
Cig / day						
1 - 18	1460	0.42		479	0.34	
19 - 24	1349	0.44		254	0.39	
25+	1209	0.48	6.7×10^{-5}	174	0.37	0.16
rs8042374						
		Cases		Controls		
	n	MAF	P^2	n	MAF	P^2
rs938682						
Cig / day						
1 - 18	1458	0.21		479	0.24	
19 - 24	1348	0.18		254	0.23	
25+	1206	0.17	1.2×10^{-3}	174	0.22	0.46
rs938682						

		Cases			Controls		
	n	MAF	P^2	n	MAF	P^2	
(b)							
<i>15q25.1</i>							
rs12914385							
	n	MAF	P^2	n	MAF	P^2	
Cig / day							
1 - 18	1460	0.21		479	0.24		
19 - 24	1350	0.18		254	0.23		
25+	1209	0.17	9.0×10^{-4}	174	0.22	0.46	
5p15.33							
rs4975616							
	n	MAF	P^2	n	MAF	P^2	
Cig / day							
1 - 18	1458	0.38		479	0.40		
19 - 24	1349	0.40		254	0.45		
25+	1208	0.39	0.55	174	0.41	0.52	
rs2736100							
	n	MAF	P^2	n	MAF	P^2	
Cig / day							
1 - 18	1460	0.52		479	0.49		
19 - 24	1348	0.53		254	0.49		
25+	1208	0.51	0.65	174	0.5	0.78	

	Cases				Controls				
	n	MAF	P^2	n	MAF	P^2	n	MAF	P^2
(b)									
<i>15q25.1</i>									
rs12914385									
6p21.33									
rs3117582									
Cig / day									
1 - 18	1459	0.16		479	0.13				
19 - 24	1348	0.15		254	0.14				
25+	1208	0.15	0.56	174	0.14	0.79			

* risk genotype

¹ From Kruskal-Wallis test

² From trend test