Title:

Pleiotropic Analysis of Lung Cancer and Blood Triglycerides

Authors:

Verena Zuber, Crystal N. Marconett, Jianxin Shi, Xing Hua, William Wheeler, Chenchen Yang, Lei Song, Anders M. Dale, Marina Laplana, Angela Risch, Aree Witoelar, Wesley K. Thompson, Andrew J. Schork, Francesco Bettella, Yunpeng Wang, Srdjan Djurovic, Beiyun Zhou, Zea Borok, Henricus F. M. van der Heijden, Jacqueline de Graaf, Dorine Swinkels, Katja K. Aben, James McKay, Rayjean J. Hung, Heike Bikeböller, Victoria L. Stevens, Demetrius Albanes, Neil E. Caporaso, Younghun Han, Yongyue Wei, Maria Angeles Panadero, Jose I. Mayordomo, David C. Christiani, Lambertus Kiemeney, Ole A. Andreassen, Richard Houlston, Christopher I. Amos, Nilanjan Chatterjee, Ite A. Laird-Offringa, Ian G. Mills*, Maria Teresa Landi*

Author affiliations:

NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine (VZ, AW, FB, YW, OAA), and Prostate Cancer Research Group, Centre for Molecular Medicine Norway, Nordic EMBL Partnership (VZ, IGM), University of Oslo, Oslo, Norway; Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway (VZ, AW, FB, YW, OAA); European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK (VZ); Departments of Surgery and of Biochemistry and Molecular Biology (CNM, CY, IALO) and Will Rogers Institute Pulmonary Research Center and Division of Pulmonary, Critical Care and Sleep Medicine, Department of Medicine (BZ, ZB), Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA (CNM, CY, IALO); Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, US Public Health Service, Bethesda, MD (JS, XH, LS, DA, NEC, NC, MTL); Information Management Services, Inc.; Rockville, MD (WW); Multimodal Imaging Laboratory (AMD, AJS), Center for Human Development (AMD, AJS), Department of Radiology (AMD), Department of Neurosciences (AMD), Department of Psychiatry (WKT), and Cognitive Sciences Graduate Program (AJS), University of California, San Diego, La Jolla, CA; Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ), Heidelberg, Germany (ML, AR); Department of Molecular Biology, University of Salzburg, Salzburg, Austria (AR); Translational Lung Research Center Heidelberg TLRC-H, Member of the German Center for Lung Research DZL, Heidelberg, Germany (AR); Department of Medical Genetics (SD) and Department of Molecular Oncology, Institute of Cancer Research and Department of Urology (IGM), Oslo University Hospital, Oslo, Norway (SD); NORMENT, KG Jebsen Centre for Psychosis Research, Department of Clinical Science, University of Bergen, Bergen, Norway (SD); Radboud University Medical Center, Radboud Institute for Health Sciences, Nijmegen, the Netherlands (HFMvdH, JdG); Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, Nijmegen, the Netherlands (DS); Netherlands Comprehensive Cancer Organization, Utrecht, the Netherlands (KKA); International Agency for Research on Cancer (IARC, World Health Organization [WHO]), Lyon, France (JM); Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario,

Canada (RJH); Zentrum Informatik, Statistik und Epidemiologie, Universit€atsmedizin Go¨ ttingen (HB); Epidemiology Research Program, American Cancer Society, Atlanta, GA (VLS); Center for Genomic Medicine, Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH (YH, CIA); Department of Epidemiology and Environmental Health, Harvard School of Public Health, Boston, MA (YW, MAP, DCC); Division of Medical Oncology, Ciudad de Coria Hospital, Coria, Spain (JIM); Division of Medical Oncology, University Hospital, Zaragoza, Spain (LK); Division of Pulmonary/Critical Care, Department of Medicine Massachusetts General Hospital/ Harvard Medical School, Boston, MA (DCC); Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey, UK (RH); Prostate Cancer UK/ Movember Centre of Excellence for Prostate Cancer Research, Centre for Cancer Research and Cell Biology, Queen's University, Belfast, UK (IGM).

*Authors contributed equally to this work.

Corresponding author:

Maria Teresa Landi, M.D., Ph.D., Senior Investigator, Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, 9609 Medical Center Drive, Room 6E446, MSC 9769, Bethesda, MD 20892-9769 (e-mail: landim@mail.nih.gov)

Ian Geoffrey Mills, M.Biochem. , Ph.D., Centre for Cancer Research and Cell Biology (CCRCB), Queen's University Belfast, 97 Lisburn Road, Belfast, BT9 7BL, United Kingdom (e-mail: I.Mills@qub.ac.uk), Prostate Cancer UK/Movember Centre of Excellence for Prostate Cancer Research, Centre for Cancer Research and Cell Biology, Queen's University, Belfast, UK.

Abstract

Epidemiologically related traits may share genetic risk factors, and pleiotropic analysis could identify individual loci associated with these traits. Because of their shared epidemiological associations, we conducted pleiotropic analysis of genome-wide association studies of lung cancer (12 160 lung cancer case patients and 16 838 control subjects) and cardiovascular disease risk factors (blood lipids from 188 577 subjects, type 2 diabetes from 148 821 subjects, body mass index from 123 865 subjects, and smoking phenotypes from 74 053 subjects). We found that 6p22.1 (rs6904596, ZNF184) was associated with both lung cancer (P ¼ 5.50x10-6) and blood triglycerides (P ¼ 1.39x10-5). We replicated the association in 6097 lung cancer case patients and 204 657 control subjects (P ¼ 2.40 10-4) and in 71 113 subjects with triglycerides data (P ¼ .01). rs6904596 reached genome-wide significance in lung cancer meta-analysis (odds ratio ¼ 1.15, 95% confidence interval ¼ 1.10 to 1.21, Pcombined ¼ 5.20x10-9). The large sample size provided by the lipid GWAS data and the shared genetic risk factors between the two traits contributed to the uncovering of a hitherto unidentified genetic locus for lung cancer.

Genetic heritability of lung cancer is estimated to be 14% (1), but only a few genetic risk loci have been identified to date in genome-wide association studies (GWAS) of lung cancer in Europeans (2). Epidemiological studies have shown associations between lung cancer and cardiovascular disease (CVD) risk factors related to the metabolic syndrome (3,4). There is also substantial evidence that lipid metabolism and innate immunity evolved from common pathways, and consequently genes that influence lipid traits may also influence inflammation and subsequent cancer development (5–7). Lung cancer is also well known to be strongly associated with tobacco smoking. Predicated on the hypothesis that investigating shared genetic risk factors across these traits could enhance the possibility of identifying new genetic loci for lung cancer, we used quantilequantile (Q-Q) plots (8) (Supplementary Methods, available online) to assess potential polygenic enrichment of singlenucleotide polymorphisms (SNPs) associated with lung cancer given association with each CVD risk factor or smoking phenotypes (Figure 1; Supplementary Figure 1, available online).

The analysis was based on the TRICL consortium meta-analysis of lung cancer GWAS, including 12 160 lung cancer case patients and 16 838 control subjects (2) (Supplementary Table 1, available online); the meta-analysis data of blood lipids from the Global Lipids Genetics Consortium (GLGC; including genetic association with triglycerides [TG], and high and low density lipoproteinscholesterol [HDL-C and LDL-C]) from 188 577 subjects (9), of type 2 diabetes (T2D) from 148 821 subjects (10), and of body mass index (BMI) from 123 865 subjects (11); and the metaanalysis of cigarettes per day (CPD) and never vs ever smoking (SMOKER) data from the Tobacco, Alcohol and Genetics (TAG) consortium, including 74 053 subjects (Supplementary Table 2, available online). The Supplementary Materials (available online) contain additional details on the contributing studies, statistical analyses, and functional tests. All statistical

tests were two-sided unless otherwise specified. P values of less than .05 were considered statistically significant.

The Q-Q plots show enrichment between lung cancer and LDL-C and between lung cancer and TG blood lipid traits across multiple P value thresholds, up to 10-5 (Figure 1, A and B), verified by an adaptive permutation procedure (Supplementary Table 3, available online). In contrast, we observed no statistically significant enrichment (Bonferroni-corrected significance threshold P < .0036) between lung cancer and HDL-C, BMI, T2D, or smoking phenotypes (the analysis of smoking excluded the SNP markers mapping to chr15:78,686,690-79,231,478, which are known to be associated with lung cancer and smoking (Supplementary Table 3 and Supplementary Figure 1, available online) (12–13). Thus, we excluded these traits from further analysis.

Cross-phenotype-associated loci between lung cancer and TG and between lung cancer and LDL-C were assessed by conjunction false discovery rate (FDR) (Supplementary Materials, available online) (8). Because controlling FDR is heavily affected by the number of identified SNPs, we pruned SNPs in linkage disequilibrium (LD; r2 > 0.8) and excluded the major histocompatibility complex (MHC; genomic position (hg 19): chr6:29,528,318-33,373,649 [14]), which harbors established lung cancer susceptibility SNPs and is known for long-range LD. By controlling conjunction FDR, we identified one genetic locus at 6p22.1, rs6904596 (A>G, minor allele frequency in caucasians ¼ 0.094), associated with both lung cancer and blood triglycerides (conjunction FDR ¼ 1.24E-02; P ¼ $5.50x10^{-6}$ for lung cancer; P ¼ $1. 39x10^{-5}$ for TG) (this locus and additional genetic loci shared between lung cancer and lipid traits are shown in Supplementary Table 4 and Supplementary Figures 2-6, available online). This locus remained statistically significant as defined by a conjunction FDR of less than .05, also using different thresholds for pruning SNPs in LD (Supplementary Table 5, available online).

We tested this SNP for replication in 6097 lung cancer case patients and 204 657 control subjects from deCODE, Harvard, Holland, and Spain (Supplementary Table 6, available online). This locus was replicated (Preplication ¼ $2.40x10^{-4}$) and attained genome-wide significance for lung cancer risk in the metaanalysis of discovery and replication data (odds ratio ¼ 1.15, 95% confidence interval ¼ 1.10 to 1.21, two-sided Pcombined ¼ $5.20x10^{-9}$, Pheterogeneity ¼ .91) (Table 1). This SNP was also replicated in the association with TG in 71 113 independent samples from deCODE and Holland (two-sided P ¼ .01, Pcombined ¼ 1.34 $x10^{-6}$) (Table 1). The SNP association with lung cancer was mostly driven by the squamous cell carcinoma subtype (P ¼ $2.80x10^{-5}$) and not adenocarcinoma (P ¼ 0.06) (Supplementary Table 7, available online).

rs6904596 localizes to 6p22.1 (27,491,299 bp; hg19) and lies 50 kb 5' of zinc finger protein 184 (ZNF184). It shows expression- QTL in lung tissue (15) with HLA-DRB3 (b ¼ -6.79, P ¼ $1.10x10^{-11}$). Additionally, rs7749305, located on chr6:27,446,566 (r2 ¼ 1 with rs6904596 in HapMap 3 of Caucasian populations), shows suggestive regulatory functions. This SNP showed the strongest association with lung cancer but was not genotyped in the Global Lipids Genetics Consortium GWAS. It lies within a DNaseI hypersensitive region in small airway epithelial cells (SAEC) and A549 adenocarcinoma cells (ENCODE) and lies in a region hypomethylated in primary alveolar epithelial cells (AEC) from our laboratory (Supplementary Figure 7, available online). rs7749305 alternate allele C appears to create ATF3 and HIF1A binding sites. Similar findings are evident in adipocytes (ENCODE), extending the pleiotropic association between lung cancer and lipid traits to their function in respective tissue types.

Additional functional analyses on relevant tissue types would be important to study how the genetic risk variants assert their downstream effect on the traits of interest.

Our study emphasizes that pleiotropic analysis of GWAS data of epidemiologically related traits can uncover hitherto unidentified genetic associations. Moreover, some GWAS of quantitative traits may be much larger than disease-specific GWAS (like in the case of CVD risk factors vs lung cancer) and, thus, may improve the likelihood to identify new loci for the disease with the smaller sample size.

Funding

This work was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, DHHS, Bethesda, MD.

Transdisciplinary Research for Cancer of Lung (TRICL): National Institute of Health U19 CA148127-01 (PI: Amos), Canadian Cancer Society Research Institute (No. 020214, PI: Hung). The Environment and Genetics in Lung Cancer Etiology (EAGLE), Prostate, Lung, Colon, Ovary Screening Trial (PLCO), and Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC) studies were supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute (NCI), Division of Cancer Epidemiology and Genetics. ATBC was also supported by US Public Health Service contracts (N01-CN-45165, N01-RC-45035, and N01-RC-37004) from the NCI.

PLCO was also supported by individual contracts from the NCI to the University of Colorado Denver (NO1-CN- 25514), Georgetown University (NO1-CN-25522), the Pacific Health Research Institute (NO1-CN-25515), the Henry Ford Health System (NO1-CN-25512), the University of Minnesota (NO1-CN-25513), Washington University (NO1-CN-25516), the University of Pittsburgh (NO1-CN-25511), the University of Utah (NO1-CN-25524), the Marshfield Clinic Research Foundation (NO1-

CN-25518), the University of Alabama at Birmingham (NO1-CN-75022), Westat, Inc. (NO1-CN-25476), and the University of California, Los Angeles (NO1-CN- 25404). The Cancer Prevention Study-II (CPS-II) Nutrition Cohort was supported by the American Cancer Society. Funding for the Lung Cancer and Smoking study was provided by the National Institutes of Health (NIH), Genes, Environment and Health Initiative (GEI) Z01 CP 010200, NIH U01 HG004446, and NIH GEI U01HG 004438. For the lung study, the GENEVA Coordinating Center provided assistance with genotype cleaning and general study coordination, and the Johns Hopkins University Center for Inherited Disease Research conducted genotyping. Harvard Lung Study: The Harvard Lung Cancer Study is supported by the National Institutes of Health (R01 CA092824, P50 CA090578, and R01 CA074386).

Global Lipids Genetics Consortium (GLGC): Data on the lipid traits were provided by the Global Lipids Consortium through their access portal (http://csg.sph.umich.edu/abecasis/public/lipids2013/). The full Consortium acknowledgements are available in the supplementary information of Willer et al.

Tobacco, Alcohol and Genetics (TAG) consortium: Data on the smoking traits were provided by the Tobacco, Alcohol and Genetics through their access portal (http://www.broadinstitute.org/mpg/ricopili/). The full Consortium acknowledgements are available in the supplementary infor mation of The Tobacco, Alcohol and Genetics.

Lung eQTL study: The lung eQTL study at Laval University was supported by the Chaire de pneumologie de la Fondation JD Be´gin de l'Universite´ Laval, the Fondation de l'Institut universitaire de cardiologie et de pneumologie de Que´bec, the Respiratory Health Network of the FRQS, the Canadian Institutes of Health Research (MOP - 123369), and the Cancer Research Society and Read for the Cure. Y. Bosse´ is the

recipient of a Junior 2 Research Scholar award from the Fonds de recherche Que´bec – Sante´ (FRQS).

Lung meQTL study: The meQTL study based on the environment and Genetics in Lung Cancer Etiology (EAGLE) study was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Department of Health and Human Services.

Epigenetics: This study was supported by NIH grant 1 R01 HL114094 to IAL-O and ZB, NIH grants 1 P30 H101258 and R37HL062569-13 to ZB. ZB is supported by the Ralph Edgington Chair in Medicine. CNM was supported in part by the department of Surgery, University of Southern California. Generation of epigenetic data was supported in part by a Norris Comprehensive Cancer Center core grant, award number P30CA014089, from the National Cancer Institute.

This work utilized the computational resources of the NIH High-Performance Computing (HPC) Biowulf cluster (http://hpc.nih.gov).

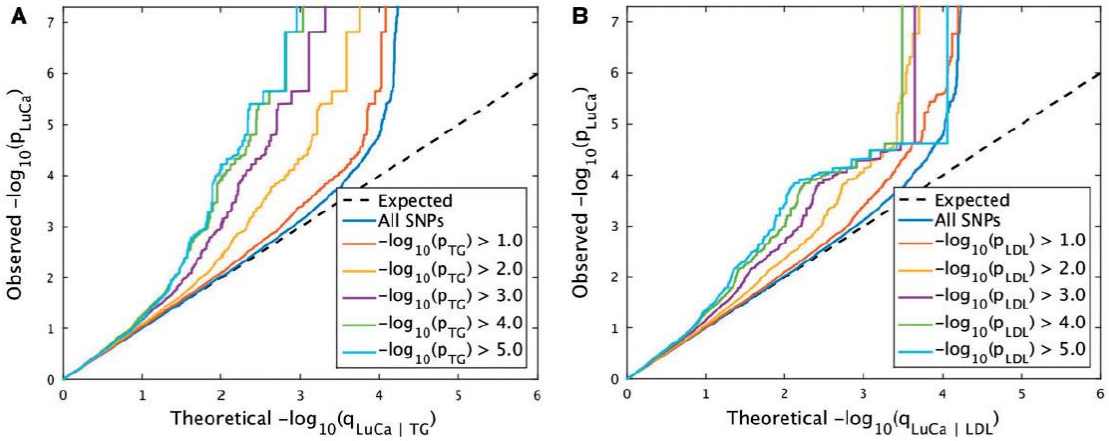
Notes

The study funders had no role in the design of the study; the collection, analysis, or interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. The authors have no conflicts of interest to disclose.

The authors thank deCODE genetics for contributing GWAS data.

Figures

Figure 1. Conditional Q-Q plots: lung cancer (LuCa) j cardiovascular disease (CVD) factors (triglycerides [TG] and low-density lipoproteins-cholesterol [LDL-C]). 'Conditional Q-Q plot' of theoretical vs empirical -log10 P values (corrected for genomic control k) in LuCa below the standard GWAS threshold of -log10 P values equal to 7.3 (equals P values above 5 x 10-8) as a function of statistical significance of association with (A) TG and (B) LDL-C at the levels of P<1.00, P < .10, P<.01, P<.001, P < 1.00x10-4, P<1.00x10-5, respectively. Dotted lines indicate the theoretical line in case of no association. All statistical tests were two-sided.



Tables

Table 1. Association of rs6904596 at 6p22.1 with both lung cancer risk and blood triglyceride levels.

Study	Lung cancer cases/ controls	OR (95% CI)	P*	Study	Triglyceride individuals	β (95% CI)	P*
TRICL	12160/16838	1.15 (1.08 to 1.21)	5.50×10^{-6}	GLGC	188577	0.0244 (0.013 to 0.035)	1.39×10^{-5}
Replication	6097/204657	1.16 (1.07 to 1.25)	2.40×10^{-4}	Replication	71113	0.0290 (0.006 to 0.051)	1.14×10^{-2}
deCODE	3865/196658	1.17 (1.06 to 1.30)	3.05×10^{-3}	deCODE	66027	0.0200 (-0.004 to 0.044)	.10
Harvard	984/ 970	1.18 (0.93 to 1.50)	.17	Holland	5086	0.0891 (0.027 to 0.151)	5.12×10^{-3}
Holland	687/ 5158	1.15 (0.96 to 1.37)	.12				
Spain	561/ 1871	1.10 (0.88 to 1.37)	.40	Combined	259690	0.0253 (0.015 to 0.035)	1.34×10^{-6}
Combined	18257/ 221495	1.15 (1.10 to 1.21)	5.20×10^{-9}				
P_Q-heterogeneity†			.91				.23

*P values were derived from a two-sided Wald test. The reference group for the odds ratio in the lung cancer study was healthy controls without lung cancer. CI = confidence interval; OR = odds ratio.

†Heterogeneity of effect size across studies was evaluated using the Cochran's Q statistic. The test is defined as one-sided.

References

1. Hemminki K, Lonnstedt I, Vaittinen P, et al. Estimation of genetic and environmental components in colorectal and lung cancer and melanoma. Genet Epidemiol. 2001;20(1):107–116.

2. Wang Y, McKay JD, Rafnar T, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet. 2014;46(7):736–741.

3. Braun S, Bitton-Worms K, LeRoith D. The link between the metabolic syndrome and cancer. Int J Biol Sci. 2011;7(7):1003–1015.

4. Kucharska-Newton AM, Rosamond WD, Schroeder JC, et al. HDL-cholesterol and the incidence of lung cancer in the Atherosclerosis Risk in Communities (ARIC) study. Lung Cancer. 2008;61(3):292–300.

5. Kominsky DJ, Campbell EL, Colgan SP. Metabolic shifts in immunity and inflammation. J Immunol. 2010;184(8):4062–4068.

6. Shi J, Chatterjee N, Rotunno M, et al. Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of squamous cell lung carcinoma. Cancer Discov. 2012;2(2):131–139.

7. Yu H, Pardoll D, Jove R. STATs in cancer inflammation and immunity: a leading role for STAT3. Nat Rev Cancer. 2009;9(11):798–809.

8. Andreassen OA, Zuber V, Thompson WK, et al. Shared common variants in prostate cancer and blood lipids. Int J Epidemiol. 2014;43(4):1205–1214.

9. Global Lipids Genetics C, Willer CJ, Schmidt EM, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274–1283.

10. Morris AP, Voight BF, Teslovich TM, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet. 2012;44(9):981–990.

11. Speliotes EK, Willer CJ, Berndt SI, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010; 42(11):937–948.

12. 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):616–622.

13. 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.

14. Shiina T, Hosomichi K, Inoko H, et al. The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet. 2009;54(1):15–39.

15. Hao K, Bosse Y, Nickle DC, et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS Genet. 2012;8(11):e1003029.

Supplementary Material

Supplementary Methods

Data and contributing studies

Input for the genetic epidemiology framework is summary statistics of genome-wide association studies (GWAS). Summary statistics on lung cancer were provided by the TRICL consortium (1) and were generated from a meta-analysis of 12,160 lung cancer cases and 16,838 controls. Further details on the sub-studies contributing to the meta-analysis are given in Supplementary Table 1. The histology-specific analyses were based on 3,718 adenocarcinoma (AD) cases and 3,422 squamous cell carcinoma (SQ) cases from the same TRICL consortium.

Data for the metabolic CVD risk factors was generated from meta-analyses of LDL, HDL, and TG (2), BMI (3), and T2D (4). Information on smoking behavior was measured by cigarettes per day (CPD) and never vs. ever smoking (SMOKER) in a large meta-analysis (5). For more details on these studies we refer to Supplementary Table 2 including references and sample sizes.

We extracted summary statistics (P-values, risk alleles, ORs, and Z-scores) for 2,558,411 common SNPs created as a reference panel from the 1000Genomes data. There was overlap in samples between the lung cancer study and other traits: 2,282 with blood lipids study, 1,959 with the BMI study and 3,179 with the T2D study. We calculated the correlation of two Z-score statistics for each pair of traits due to sample overlap: rLuCa,LDL=-0.0034, r¬¬¬LuCa,TG=-0.0054, r¬¬LuCa,HDL=0.0170, rLuCa,BMI=0.0176, rLuCa,T2D=-0.0327, rLuCa,CPD=-0.0371, and rLuCa,Smoker=-0.0014, suggesting that correlations due to sample overlap had a negligible impact on statistical inference even without explicit adjustment.

The analysis of lung cancer and cardiovascular risk factors included 483,841 independent individuals. The analysis of nicotine dependence included an additional 74,053 individuals. The replication datasets for lung cancer consisted of 6,097 lung cancer cases and 204,657 controls from deCODE, Harvard, Holland, and Spain; the replication for triglycerides and LDL-C consisted of 71,113 and 45,815 subjects, respectively from deCODE and Holland. Thus, we examined overall 885,576 individuals.

Enrichment analysis

We used conditional quantile-quantile (Q-Q) plots (6-9) after randomly pruning SNPs in linkage disequilibrium (LD) ($r2>0.2$) for visualizing the polygenic enrichment patterns of lung cancer by restricting to the SNP sets that showed the strongest associations in a secondary trait (the CVD risk factors and smoking phenotypes). In particular, a conditional Q-Q plot was generated for SNPs with $P<1E-1$, 1E-2, 1E-3, 1E-4, and 1E-5 for the secondary trait. In order to avoid confounding by large LD blocks, we performed a random pruning algorithm to compute the conditional Q-Q plot. Briefly, we defined LD blocks by an r2 threshold of 0.8 and randomly selected one representative SNP from each LD block to compute a Q-Q plot. The final conditional Q-Q plot was generated by averaging the Q-Q plots from 100 random pruning steps.

We performed permutations to test whether high-ranking SNPs for the tested risk factors were enriched in lung cancer GWAS. Suppose there are N SNPs (denoted as S) after LD pruning. We aimed to test whether a specific set of N_1 SNPs (denoted as S_1) were enriched in lung cancer GWAS. The standard Kolmogorov-Smirnov test or Wilcoxon rank sum test was not sensitive to capture the deviation in the tail of the Q-Q plots. Thus, we designed an adaptive permutation procedure to test the deviation in the tails. We

calculated the ranks of all N SNPs according to the association P values. We considered the top α proportion of SNPs in S_1 and calculated the total number of ranks of these SNPs, denoted as (α). We performed permutations to approximate the statistical significance of T(α) as p(α). Because the test was sensitive to the choice of α, we chose a series of α values (α_1, 〖⋯,α〗_m) and derived the corresponding P-values (p(α_1), 〖⋯,p(α〗_m)). The overall one-sided statistic for testing the enrichment was defined as Q=min(p(α_1), 〖⋯,p(α〗_m)) and its statistical significance was evaluated by permutations. In our data analysis we chose (α_1,α_2,α_3,α_4)=(0.05,0.1,0.2,0.3) and ran 10,000 permutations.

Assessing cross-phenotype association by conjunction FDR

In order to identify shared risk loci between two traits we used the conjunction false discovery rate (FDR), a genetic epidemiology framework based on an extension of the FDR into two dimensions. The conditional FDR is an extension of the standard FDR that allows including additional information on the p-value of the same SNP in a secondary trait 2. It is defined as the probability that a given SNP is null given that the p-values for trait 1 and trait 2 are as small or smaller than the observed ones. Low values of conditional FDR can be driven by the first trait only. To detect SNPs associated with both traits at the same time we used the conjunction FDR, which is defined as the probability of being null for either trait, or for both traits simultaneously given that the p-values for the two traits are as small or smaller than the observed ones. Thus, a true discovery is only the case when a SNP is non-null for both traits jointly. For more information on conditional and conjunction FDR and implementation we refer to (6, 8, 9). We aimed to control at a conservative FDR level of 0.05 per pair-wise comparison.

It has been shown that the estimation of the FDR can be impacted by correlation among the summary statistics (10). In order to address LD among SNPs we performed random pruning. First, we defined loci by an r2 threshold of 0.8. Within each block we selected one SNP randomly by chance, where every SNP had equal chance of being selected. This random pruning procedure was repeated 100 times and the final computation of the empirical distribution was averaged over the 100 pruning procedures. Thus, the impact of large LD blocks was reduced on the estimation of the FDR.

eQTL in lung tissue

To identify expression quantitative trait loci we first used data from an analysis of genotype and matched expression data from the Genotype-Tissue Expression Consortium (GTEx) (11)of n=111 healthy lung tissue samples. Additionally, we performed an analysis of genotype and matched expression data on healthy lung tissue in a study of n=1, 111 individuals using methodology previously described by Hao et al.(12).

meQTL in lung tissue

To identify methylation quantitative trait loci (meQTL) we applied a methodological approach on genotype and matched methylation data on n=210 non-tumor lung tissue samples as described previously by Shi et al.,(13), and in addition applied that analytical approach to a second dataset of non-tumor lung tissue samples obtained from The Cancer Genome Atlas (TCGA) lung cancer initiative (14).

Chromatin-level annotation

Publically available chromatin immunoprecipitation (ChIP)-seq data from the ENCODE and ROADMAP consortiums was used to determine relationship between SNPs and chromatin state in small airway epithelial cells (SAEC), adipocytes, and the A549 lung adenocarcinoma cell line, using the UCSC Genome browser, version hg19. Histone marks relating to active promoters were visualized with BigWig density tracks to determine if the identified loci were functionally active in the given cell types. Publically available ChIP-seq data on transcription factor binding was displayed using the UCSC genome browser.

To fully examine DNA methylation in the region surrounding the SNPs, we examined whole genome bisulfite sequencing (WGBS) data that we had obtained from purified primary alveolar epithelial cells (AEC) obtained as previously described (15, 16). Libraries were plated using the Illumina cBot and run on the Hi-Seq 2000 according to manufacturer's instructions using HSCS v 1.5.15.1. Rep 1 underwent Paired End 100 cycling; rep 2 underwent Paired End 75 cycling. Image analysis and base calling were carried out using RTA 1.13.48.0, deconvolution and fastq file generation was carried out using CASAVA_v1.7.1a5. Alignment to the genome was carried out using bsmap V 2.5. Aligned .bam files were visualized using IGViewer V2.3.40 (Broad Institute, Cambridge MA) with alignments colored by Bisulfite mode "CG". AEC WGBS has been made publically available through GEO record GSE65319 along with complete sample preparation description.

References

1. Wang Y, McKay JD, Rafnar T, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet 2014;46(7):736-41.

2. Global Lipids Genetics C, Willer CJ, Schmidt EM, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet 2013;45(11):1274-83.

3. Speliotes EK, Willer CJ, Berndt SI, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 2010;42(11):937-48.

4. Morris AP, Voight BF, Teslovich TM, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 2012;44(9):981-90.

5. Tobacco, Genetics C. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet 2010;42(5):441-7.

6. Schork AJ, Thompson WK, Pham P, et al. All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs. PLoS Genet 2013;9(4):e1003449.

7. Andreassen OA, McEvoy LK, Thompson WK, et al. Identifying Common Genetic Variants in Blood Pressure Due to Polygenic Pleiotropy With Associated Phenotypes. Hypertension 2014; 10.1161/HYPERTENSIONAHA.113.02077.

8. Andreassen OA, Thompson WK, Schork AJ, et al. Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional False Discovery Rate. PLoS Genet 2013;9(4):e1003455.

9. Andreassen OA, Zuber V, Thompson WK, et al. Shared common variants in prostate cancer and blood lipids. Int J Epidemiol 2014; 10.1093/ije/dyu090.

10. Schwartzman A, Lin X. The effect of correlation in false discovery rate estimation. Biometrika 2011;98(1):199-214.

11. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348(6235):648-60.

12. Hao K, Bosse Y, Nickle DC, et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS Genet 2012;8(11):e1003029.

13. Shi J, Marconett CN, Duan J, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat Commun 2014;5:3365.

14. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511(7511):543-50.

15. Ballard PL, Lee JW, Fang X, et al. Regulated gene expression in cultured type II cells of adult human lung. American journal of physiology. Lung cellular and molecular physiology 2010;299(1):L36-50.

16. Marconett CN, Zhou B, Rieger ME, et al. Integrated transcriptomic and epigenomic analysis of primary human lung epithelial cell differentiation. PLoS genetics 2013;9(6):e1003513.

Supplementary Figures

Supplementary Figure 1. Conditional Q-Q plots: LuCa | further CVD factors and smoking traits ‘Conditional Q-Q plot’ of theoretical vs empirical -log10 p-values (corrected for genomic control λ) in lung cancer (LuCa) below the standard GWAS threshold of -log10 p-values equal to 7.3 (equals p-values above 5 x 10-8) as a function of statistical significance of association with (A) HDL cholesterol (HDL-C), (B) body mass index (BMI), (C) type 2 diabetes (T2D), (D) cigarettes per day (CPD), and (E) ever versus never smoking (SMOKER), at the level of $p < 1$, $p < 0.1$, $p < 0.01$, $p < 0.001$, $p < 1.00x10-4$, $p < 1.00x10-5$ respectively. Dotted lines indicate the theoretical line in case of no association. For the analysis of CPD we removed SNPs mapping to the nicotinic acetylcholine receptors (chr15:78,686,690-79,231,478) to exclude this well-established pleiotropic locus between LuCa and CPD. All statistical tests were two-sided.

Supplementary Figure 2. Local Manhattan plot for cross-phenotype association of locus 6p22.1 between lung cancer and triglycerides (TG); top shared variant rs6904596. On display are the log10 p-values for the association of SNPs with lung cancer (upper panel) and TG (lower panel) based on the SNPs genomic position. All variants with a shared association between lung cancer and TG (conjunction FDR < 0.05) are symbolized with diamonds. The leading variant for lung cancer (rs7749305) is symbolized with a triangle (top up) and the leading variant (rs9295740) for TG with a triangle top down. rs7749305 was not included in the Global Lipid Consortium GWAS. Linkage disequilibrium (LD) based on HapMap v.3 in Caucasians is color-coded: variants with r2=1 with the lead variant rs6904596 are colored in red, 0.8=<r2<1 are colored in violet, and 0.5=<r2<0.8 are colored in blue.

Supplementary Figure 3. Local Manhattan plot for cross-phenotype association of locus 1q22 between lung cancer and triglycerides (TG); top shared variant rs2066981. On display are the log10 p-values for the association of SNPs with lung cancer (upper panel) and TG (lower panel) based on the SNPs genomic position. All variants with a shared association between lung cancer and TG (conjunction FDR < 0.05) are symbolized with diamonds. The leading variant for lung cancer (rs914615) is symbolized with a triangle (top up) and the leading variant for TG was the conjunction top variant rs2066981. Linkage disequilibrium (LD) based on HapMap v.3 in Caucasians is color-coded: variants with r2=1 with the lead variant rs2066981 are colored in red, 0.8=<r2<1 are colored in violet, and 0.5=<r2<0.8 are colored in blue.

Supplementary Figure 4. Local Manhattan plot for cross-phenotype association of locus 6p22.2 between lung cancer and high-density lipoprotein (HDL); top shared variant rs9393692. On display are the log10 p-values for the association of SNPs with lung cancer (upper panel) and HDL (lower panel) based on the SNPs genomic position. All variants with a shared association between lung cancer and HDL (conjunction FDR < 0.05) are symbolized with diamonds. The leading variant for lung cancer (rs3999544) is symbolized with a triangle (top up) and the leading variant (rs9358913) for HDL with a triangle top down. Linkage disequilibrium (LD) based on HapMap v.3 in Caucasians is

color-coded: variants with r2=1 with the lead variant rs9393692 are colored in red, 0.8=<r2<1 are colored in violet, and 0.5=<r2<0.8 are colored in blue.

Supplementary Figure 5. Local Manhattan plot for cross-phenotype association of locus 11q12.2 between lung cancer and triglycerides (TG); top shared variant rs2851682. On display are the log10 p-values for the association of SNPs with lung cancer (upper panel) and TG (lower panel) based on the SNPs genomic position. All variants with a shared association between lung cancer and TG (conjunction FDR < 0.05) are symbolized with diamonds. The leading variant for lung cancer is the conjunction variant rs2851682; the leading variant for TG (rs1535) is symbolized with a triangle top down. Linkage disequilibrium (LD) based on HapMap v.3 in Caucasians is color-coded: variants with r2=1 with the lead variant rs2851682 are colored in red, 0.8=<r2<1 are colored in violet, and 0.5=<r2<0.8 are colored in blue.

Supplementary Figure 6. Local Manhattan plot for cross-phenotype association of locus 15q15.2 between lung cancer and triglycerides (TG); top shared variant rs554001. On display are the log10 p-values for the association of SNPs with lung cancer (upper panel) and TG (lower panel) based on the SNPs genomic position. All variants with a shared association between lung cancer and TG (conjunction FDR < 0.05) are symbolized with diamonds. The leading variant for lung cancer is the conjunction variant rs554001; the leading variant for TG (rs580469) is symbolized with a triangle top down. Linkage disequilibrium (LD) based on HapMap v.3 in Caucasians is color-coded: variants with r2=1 with the lead variant rs554001 are colored in red, 0.8=<r2<1 are colored in violet, and 0.5=<r2<0.8 are colored in blue.

Supplementary Figure 7. rs7749305 lies in a regulatory element in lung and fat tissues/cell lines. Epigenetic annotation of rs7749305 on chromosome 6 in a frame of 10KB up and downstream using peaks from tissues relevant for lung cancer and blood lipid traits. At the top, the UCSC Genome browser image shows the ZNF184 locus and to the right below that the position of rs7749305. Below that we show: in small airway epithelial cells (SAEC), DNase hypersensitive sites; in adipocytes, duplicate lanes of ChIP-seq marks for H3K4me1, H3K4me3 and H3K9ac; in A459 cells, duplicate DNase hypersensitive sites, and ChIP-seq marks for H3K4me1, H3k4me3 and H3K9ac; and in our own human alveolar epithelial cells (AEC), duplicate whole genome bisulfite sequencing (WGBS; red=methylated CpGs, blue=unmethylated CpGs). At the bottom, the position of the SNP and its effect on ATF3 and HIF1A binding sites is indicated.

Supplementary Tables

Supplementary Table 1. TRICL lung cancer studies in the discovery dataset.

Supplementary Table 2. Discovery dataset.

Supplementary Table 3. Testing for enrichment by permutation.

Supplementary Table 4. Summary data for cross-phenotype associated loci for lung cancer and lipid traits.

Supplementary Table 5. Sensitivity analysis: p-values for conjunction FDR for lung cancer and triglyceride levels of the 6p22.1 locus using different r2 thresholds for pruning SNPs in LD.

Supplementary Table 6. Replication dataset.

Supplementary Table 7. Summary data for genetic loci and lead SNP rs6904596 with reference allele A and alternative allele G at 6p22.1 in lung adenocarcinoma and squamous cell carcinoma.