HMG Advance Access published August 29, 2016

TITLE: A Whole-Blood Transcriptome Meta-Analysis Identifies Gene Expression Signatures of Cigarette Smoking

Tianxiao Huan[1,2†], Roby Joehanes[1,2,3†], Claudia Schurmann[4,5,6†], Katharina Schramm[7,8†], Luke C. Pilling[9†], Marjolein J. Peters[10,11†], Reedik Mägi[12†], Dawn DeMeo[13], George T O'Connor[14], Luigi Ferrucci[15], Alexander Teumer[16], Georg Homuth[6], Reiner Biffar[17], Uwe Völker[6], Christian Herder[18, 19], Melanie Waldenberger[20], Annette Peters[20,21], Sonja Zeilinger[20], Andres Metspalu[12], Albert Hofman[11,22], André G. Uitterlinden[10,11,22], Dena G. Hernandez[23], Andrew B. Singleton[23], Stefania Bandinelli[24], Peter J. Munson[25], Honghuang Lin[14], Emelia J. Benjamin[1,14], Tõnu Esko[12,26,27*], Hans J. Grabe[28,29*], Holger Prokisch[7,8*], Joyce B.J. van Meurs[10,11*], David Melzer[9*], Daniel Levy[1,2*]

1 The National Heart, Lung, and Blood Institute's and Boston University's Framingham Heart Study, 73 Mt. Wayte Avenue, Framingham, MA, United States;
2 The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Bethesda, MD, United States;
3 Hebrew SeniorLife, Harvard Medical School, Boston, MA, United States;
4 The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York City, New York, United States;
5 Genetics of Obesity & Related Metabolic Traits Program, Icahn School of Medicine at Mount Sinai, New York, New York, United States;
6 Interfaculty Institute of Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany;
7 Institute of Human Genetics, Helmholz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany;
8 Institute of Human Genetics, Technical University Munich, Munich, Germany;
9 Epidemiology and Public Health Group, Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, United Kingdom;
10 Department of Internal Medicine, Erasmus Medical Centre Rotterdam, The Netherlands;
11 The Netherlands Genomics Initiative-sponsored Netherlands Consortium for Healthy Aging (NGI-NCHA), Leiden / Rotterdam, the Netherlands;
12 Estonian Genome Center, University of Tartu, Tartu, Estonia;
13 Harvard Medical School, Boston, MA, United States;
14 Boston University School of Medicine and School of Public Health, Boston, MA, United States;
15 Clinical Research Branch, National Institute on Aging, Baltimore, MD, United States;
16 Institute for Community Medicine, University of Greifswald, Greifswald, Germany;
17 Department of Prosthetic Dentistry, Gerostomatology and Dental Materials, Center of Oral Health, University Medicine Greifswald, Greifswald, Germany;
18 Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany;
19 German Center for Diabetes Research (DZD), München-Neuherberg, Germany;
20 Research Unit of Molecular Epidemiology, Helmholz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany;
21 Institute of Epidemiology II, Helmholz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany;
22 Department of Epidemiology, Erasmus Medical Center Rotterdam, the Netherlands;
23 Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, United States;
24 Geriatric Unit, Azienda Sanitaria di Firenze, Florence, Italy.
25 Mathematical and Statistical Computing Laboratory, Center for Information Technology, National Institutes of Health, United States;
26 Division of Endocrinology, Boston Children's Hospital, Boston, MA, United States;
27 Broad Institute of MIT and Harvard, Cambridge, MA, United States;
28 Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany;
29 German Center for Neurodegenerative Diseases DZNE, Site Rostock/ Greifswald, Germany;

† These authors contribute equally.

*** Denoted corresponding authors contribute equally; and the correspondence should be addressed to:**

Daniel Levy, MD
Framingham Heart Study
Population Sciences Branch
National Heart, Lung, and Blood Institute
73 Mt. Wayte Avenue, Suite 2
Framingham, MA 01702
Email: Levyd@nih.gov
Phone: 508-935-3458
Fax: 508-872-2678

Abstract

Cigarette smoking is a leading modifiable cause of death worldwide. We hypothesized that cigarette smoking induces extensive transcriptomic changes that lead to target-organ damage and smoking-related diseases. We performed a meta-analysis of transcriptome-wide gene expression using whole blood-derived RNA from 10,233 participants of European ancestry in six cohorts (including 1421 current and 3955 former smokers) to identify associations between smoking and altered gene expression levels. At a false discovery rate (FDR) <0.1, we identified 1270 differentially expressed genes in current vs. never smokers, and 39 genes in former vs. never smokers. Expression levels of 12 genes remained elevated up to 30 years after smoking cessation, suggesting that molecular consequence of smoking may persist for decades. Gene ontology analysis revealed enrichment of smoking-related genes for activation of platelets and lymphocytes, immune response, and apoptosis. Many of the top smoking-related differentially expressed genes, including *LRRN3* and *GPR15*, have DNA methylation loci in promoter regions that were recently reported to be hypomethylated among smokers. By linking differential gene expression with smoking-related disease phenotypes, we demonstrated that stroke and pulmonary function show enrichment for smoking-related gene expression signatures. Mediation analysis revealed expression of several genes (e.g. *ALAS2*) to be putative mediators of the associations between smoking and inflammatory biomarkers (IL6 and C-reactive protein levels). Our transcriptomic study provides potential insights into the effects of cigarette smoking on gene expression in whole blood and their relations to smoking-related diseases. The results of such analyses may highlight attractive targets for treating or preventing smoking-related health effects.

Introduction

For more than a half century, numerous studies have characterized the deleterious health effects of cigarette smoking including cancers, cardiovascular disease (CVD), and chronic obstructive pulmonary disease (1). Cigarette smoking is the leading cause of preventable death in the United States, accounting for more than 443,000 deaths each year (2). Cessation campaigns have had an effect; since 2002, the number of former smokers in the United States has exceeded the number of current smokers (3), which is estimated at 43.8 million, or 19% of all adults aged 18 years or older (4).

Research also has characterized persisting long-term health risks of cigarette smoking, even decades after cessation (5). While the risks for some diseases may quickly return to those of never-smokers, risks for some diseases remain elevated for years, including risks for lung cancer (5), many other cancers (6), and stroke (7).

Previous studies have reported a genetic predisposition (8-11) to cigarette smoking. Other studies have reported smoking-related DNA methylation patterns (12-18). Gene expression is under strong genetic and epigenetic control (19, 20). Transcriptomic analyses may expand our understanding of molecular mechanisms affected by smoking. Several previous studies investigated the associations between cigarette smoking and transcriptomic changes in lung tissues (21-23), monocytes (24, 25), and peripheral whole blood (26-29). These studies, however, had small sample sizes (<300 current smokers) that limited their statistical power to detect modest transcriptomic changes due to tobacco exposure. In addition, most of the previously identified smoking-related gene expression signatures have not replicated. For example, only four genes (<1%) overlapped between two published studies of smoking-related gene expression signatures in lung tissues (21, 23), and 18 genes (~3%) overlapped in any two of the four published studies in whole blood (26-29). In addition, the long-term effects of cigarette smoking on the transcriptome remain poorly characterized.

We conducted a meta-analysis of the associations of cigarette smoking with gene expression in whole blood-derived RNA in over 10,000 individuals across six cohort studies, including 1421 current smokers and 3955 former smokers. We sought to characterize both the short-term and long-term impact of

smoking on the transcriptome, and to identify affected pathways. We also sought to link the transcriptomic changes associated with smoking to smoking-related diseases. Understanding the long-term molecular consequences of cigarette smoking may identify targets for the treatment and primary prevention of smoking-related diseases.

Results

Study sample characteristics

Characteristics of the study participants in each cohort are provided in **Supplementary Table 1**. Among the 10,233 participants in the six cohorts, 14% were current smokers (n=1421, mean age ranging from 34 to 68, 50% were men), 39% were former smokers (n=3955, mean age ranging from 43 to 74, 53% were men), 47% were never smokers (n=4860, mean age ranging from 38 to 70, 40% were men). The average white blood cell counts were 7.2, 6.1, and 5.9 cells per cubic mm in current, former, and never smokers respectively.

Identify and replicate gene expression signatures of cigarette smoking

At FDR <0.1, we identified 1270 differentially expressed genes in current vs. never smokers (**Supplementary Table 2**). Of the 1270 smoking-related gene expression signatures, 717 (56%) were up-regulated and 553 (44%) were down-regulated (**Figure 1A**). The top 25 (by *P*-value) differentially expressed genes are listed in **Table 1**. Secondary analyses of pack-years smoked yielded about the same genes as those for current vs. never smokers (**Supplementary Figure 1**). Adjustment for body mass index (BMI), coronary heart disease (CHD), forced expiratory volume in 1 second (FEV1), physical activity, and alcohol consumption did not significantly alter the differentially expressed genes associated with smoking in the FHS. **Supplementary Figure 2** shows the correlations of T statistic values with and without additional covariate adjustment (Pearson correlation coefficient >0.99).

In order to evaluate the replicability of our results, we split the overall samples into discovery (N= 4610) and replication (N=5623) sets. Samples in the discovery and replication sets were from independent studies (see **Methods**). The T statistics of each gene in the smoking-related gene expression signatures

were highly consistent between the discovery and replication sets (Pearson correlation is 0.87, **Supplementary Figure 3**). Sixty-four percent of genes identified in the discovery set at FDR<0.1 replicated in the replication set at FDR<0.1.

By comparing our results with previously reported results, we found that 68 genes identified in our study overlapped with previously identified smoking-related gene expression signatures in whole blood (enrichment P=5.06 x 10^{-7}) (26-29), 137 genes overlapped with previously identified genes in monocytes (enrichment P<1 x 10^{-32}) (24, 25), and 31 genes including *CYP1B1*, *SCNA*, and *CX3CR1* overlapped with previously identified genes in lung tissues (21-23) (enrichment P=0.65, **Supplementary Table 3**). We also found 92 genes with adjacent DNA methylation sites (CpGs) that were reported to be differentially methylated in relation to smoking in previous studies (30) (enrichment P=3.4 x 10^{-4}, **Supplementary Table 4**). For example, *LRRN3*, the top gene in our results, has three DNA methylation loci (cg09837977, cg05221370, and cg11556164) in its 5'UTR region that were recently reported to be associated with smoking in two studies (12, 14). Another example is *GPR15*, which has a DNA methylation site (cg19859270) that was reported to be associated with smoking in four studies (12, 14-16).

In turn, we checked the adjacent genes of previously reported smoking-induced DNA methylation loci (30) for differential expression effects at a nominal P<0.05. **Supplementary Table 5** provides the full list of genes and CpGs.

Long-term effects of cigarette smoking on whole blood gene expression levels

Thirty-nine genes were statistically significant (FDR<0.1) in analyses contrasting former vs. never smokers (**Supplementary Table 6**, and **Table 2** shows the top 25 genes), including 14 up-regulated and 25 down-regulated genes (**Figure 1B**). As shown in **Supplementary Figure 4**, 35 of the 39 gene expression signatures (87%) in analysis of former vs. never smokers show the same directionality in analysis of current vs. never smokers (i.e. when the gene was upregulated in former smokers vs. never smokers is was also upregulated in current smokers vs. never smokers.). Of these 35 overlapping genes, 19 including *LRRN3*, *GPR15*, and *CLDND1,* were statistically significant (FDR<0.1) in analyses of current vs. never smokers. Of the 39 genes, one (*GNLY*) showed differential expression in relation to smoking in lung tissues (21), and five genes (*LRRN3*, *GPR15*, *CLDND1*, *STAT3*, and *PTGDR*) harbor CpGs that were previously reported to

be differentially methylated in relation to smoking in whole blood (30). To further investigate the long-term effects of these genes, we performed an in-depth analysis of 39 gene transcript levels in relation to the time since smoking cessation among former smokers. Twelve genes including *LRRN3*, *GPR15*, and *CLDND1*, remained differentially expressed in former vs. never-smoker 30 years following smoking cessation (see **Methods** section and **Figure 2**).

Coexpression network analysis of smoking genes

To understand the molecular mechanisms by which cigarette smoking are associated with the whole blood transcriptome, we performed a coexpression network analysis of the 1290 smoking gene expression signatures (unique set of 1270 genes for current vs. never smokers plus 39 genes for former vs. never smokers). We discovered five major coexpression network modules (coEMs for short; coEMs named using different colors; **Supplementary Figure 5**). Genes in each coEM formed a tightly co-regulated network structure that we hypothesize is functionally related to tobacco exposure. Gene ontology enrichment analyses were then performed on each coEM (**Table 3**).

Three coEMs are enriched for genes involved in immune response-related pathways, including the Turquoise coEM (for platelet activation; corrected *P*=3.1e-3, and inflammatory response, corrected *P*=3.7e-2), the Blue coEM (for lymphocyte activation, corrected *P*=3.9e-7), and the Brown coEM (for immune cell mediated cytotoxicity, corrected *P*=3.9e-8). The Green coEM is enriched for genes involved in protein biosynthesis (corrected *P*=6.3e-3).

Smoking-related gene expression signatures in association with human complex diseases and traits

Cigarette smoking has been recognized as a key causal risk factor for multiple complex diseases and traits (1). Our results suggest that smoking may disturbs the expression levels of many genes across multiple critical pathways in whole blood that may relate to many disease phenotypes. To test this hypothesis, we further determined if the identified smoking-related gene expression signatures in whole blood are enriched GWAS SNPs associated with smoking-related diseases and traits.

We linked the 1290 smoking-related gene expression signatures with whole blood gene expression-associated SNPs (eSNPs) (31) (Joehanes R, PhD, unpublished data, 2016), and then cross referenced the eSNPs with NHGRI GWAS Catalog SNPs (32). We identified 536 smoking-related gene

expression signatures having at least one eSNP associated with human complex diseases or traits reported in the NHGRI GWAS Catalog (**Supplementary Table 7**). Recent research suggests that using eSNPs and GWAS mapping may permit the linking of gene transcripts with diseases or traits (33). Therefore, the 536 genes having blood eSNPs linked with GWAS SNPs for diseases or traits can be considered a set of putative blood gene expression signatures of the diseases or traits even though the associations of these genes with diseases or traits were not directly measured. Smoking-related gene expression signatures as a set show enrichment for disease- and trait-associated genes (enrichment $P<1x10^{-32}$, by Fisher's exact test), indicating that smoking-induced gene expression changes may be associated with a wide range of clinical traits.

We further focused the search on diseases and traits known to be associated with cigarette smoking, including cardiovascular diseases, obesity-related traits, inflammatory biomarkers, pulmonary function, and various lung diseases including chronic obstructive pulmonary disease and asthma. Enrichment tests were performed for the traits or diseases that overlapped with smoking-related gene expression signatures for at least five genes. As shown in **Table 4**, the smoking-related gene expression signatures as a set were enriched for genes having *cis*-eSNPs that were also GWAS SNPs for stroke (enrichment *P*=4.5e-5) and pulmonary function (enrichment *P*=3.7e-3), and for BMI-related traits and asthma (enrichment *P*<0.05). Smoking-related gene expression signatures were also enriched for genes having *trans*-eSNPs that were also GWAS SNPs for weight, asthma, and coronary heart disease (enrichment *P*<0.05; details regarding correlated eSNPs are provided in **Supplementary Table 7**).

We analyzed the association of smoking-related gene expression signatures with two inflammatory biomarkers (serum concentrations of IL6 and CRP) and with pulmonary function (FEV1, FVC, and the FEV1/ FVC ratio) in FHS participants. IL6, CRP, and FEV1, were significantly associated with smoking status (**Supplementary Table 8**). We identified 3, 55, and 7 smoking gene-expression signatures that were differentially expressed in relation to IL6, and CRP, and FEV1, respectively, at Bonferroni corrected *P*<0.05. The overlapping genes that were significantly associated with smoking and with IL6, CRP, and FEV1, were further tested to determine if their gene expression levels mediated the association of smoking on these phenotypes (IL6, CRP, and FEV1). At Bonferroni corrected *P*<0.05 (by

the Sobel test), we identified 1 gene (*ALAS2*) that appears to be a mediator of the association between smoking and IL6, and seven genes including *ALAS2* that were mediators for CRP (**Table 5**).

Discussion

By meta-analyzing gene expression data from 10,233 individuals from six cohort studies, we identified 1270 genes that were differentially expressed in current vs. never cigarette smokers, and 39 genes that were differentially expressed in former vs. never smokers, including 12 genes with persistent gene expression changes up to 30 years following smoking cessation.

In contrast to previous smaller studies of smoking-related gene expression signatures (21-29), we were able to replicate our findings by splitting the overall study samples into discovery and replication sets. The samples in discovery (n=4610) and replication (n=5623) sets were from independent cohorts and used different microarray platforms. We found that sixty-four percent of smoking-related differentially expressed genes identified in the discovery set replicated in the replication set.

Pathway and coexpression network analysis identified four coEMs related to smoking representing many critical pathways including platelets activation, lymphocyte activation, inflammatory response, and protein biosynthesis. Smoking induces aberrant platelet activation (34, 35), which may increase the risk of thrombotic events including atherothrombotic cardiovascular disease (36). Three coEMs are enriched for immune function-related genes, including *DUSP1* and *FOS* (**Supplementary Table 9**), consistent with the findings that serum concentrations of CRP and IL6 are significantly higher in current vs. never smokers(37, 38)(**Supplementary Table 8**) and pointing toward putative mechanisms by which smoking may cause systemic inflammation. Two of the three smoking-related immune function coEMs were significantly associated with CRP (e.g., Turquoise coEM at P=0.02 and Blue coEM at P=0.03). Based on these findings, we hypothesize that the association of smoking on inflammation is mediated by gene expression changes, although further functional validation is required to establish precise mechanisms. Previous studies showed effects of nicotine on protein biosynthesis in muscle (39), brain, and liver (40). We identified one smoking-related coEM that was enriched for protein biosynthesis, providing evidence that smoking may affect protein biosynthesis in whole blood.

Epigenetic studies have shown that smoking is an important epigenetic modifier that affects the DNA methylation pattern of thousands of CpGs (30). By overlapping our transcriptomic results with previous epigenetic findings (30), we found 92 genes with altered expression and DNA methylation in relation to smoking status (**Supplementary Table 4**). Most notable among these are *LRRN3* and *GPR15*, which were upregulated in current and former smokers (vs. never smokers) and displayed long-term persistent associations of smoking with mRNA expression levels. The differential expression of *LRRN3* and *GPR15* in smokers was also reported by Tsaprouni et al (12). These two genes have nearby CpGs that were reported to be significantly hypomethylated in cigarette smokers (12, 14, 16, 41). Three CpGs, cg09837977, cg05221370, and cg11556164, located in the 5'UTR region of *LRRN3* and cg19859270 located in the first exon of *GPR15* are located in active gene promoter regions. This is consistent with the concept that DNA methylation in gene promotor regions may inhibit gene transcription (42). Therefore, we speculate that many of the identified smoking-related gene expression signatures are mediated by smoking-induced epigenetic changes. However, we cannot exclude the possibility that the overlap of gene expression and DNA methylation change in relation to smoking may be due to changes in white blood cell types. A recent study by Bauer et al suggested that smoking-related differential methylation and expression of *GPR15* results from the enrichment of a smoking-induced lymphocyte population (43).

Smoking is one of the most important causal lifestyle risk factors for a wide range of diseases, although the molecular underpinnings of smoking-related risks remain largely unknown. In an attempt to link smoking-related gene expression signatures to disease phenotypes, we used GWAS results from the NHGRI GWAS Catalog (32). By cross referencing eSNPs of genes that are differentially expressed in relation to smoking status with GWAS SNPs associated with various smoking-related diseases, we sought to obtain insights into the potential roles of smoking-related differentially expressed genes in a variety of smoking-related health outcomes. We observed that (as a set) gene expression signatures of smoking show enrichment for *cis*- and *trans*- eSNPs that are also GWAS SNPs for smoking-related diseases and clinical traits such as stroke and pulmonary function (**Table 4**), suggesting that smoking-induced transcriptomic changes are linked to smoking-related diseases. Without any experimental validation, however, we cannot prove causal mechanistic links of smoking to gene expression and smoking-related diseases.

We further tested if any NHGRI GWAS Catalog SNPs showed an interaction with smoking that may affect gene expression levels in FHS samples. We did not find any significant *cis*-eSNP showing SNP-by-smoking interaction on corresponding transcripts levels, but several *trans*-eSNPs (**Supplementary Table 10)** displayed interactions. The *trans*-eSNP results need to be replicated. We acknowledge that our study may still lack power to identify SNP-by-smoking interaction on gene expression levels.

One limitation of our study is that we used whole blood for expression profiling. Whole blood is easy to obtain in population-based studies but may not be the primary tissue for many smoking-related diseases, such as lung cancer and chronic obstructive pulmonary disease. By comparing our results with previously reported lung tissue-based results, we found 31 smoking gene signatures that also showed differential expression in relation to cigarette smoking in lung tissue. For example, *CYP1B1* was significantly upregulated in whole blood of current smokers at FDR=7.6e-7 in our study and was reported to be significantly upregulated in non-tumor lung tissue (21) and in the bronchial mucosa of smokers(44). This finding suggests that whole blood may partially capture smoking-induced pathological molecular changes occurring in targeted tissues. In addition, peripheral whole blood expression patterns can be linked to many other diseases including systemic inflammatory and immune-related disorders(45) and metabolic and cardiovascular diseases (46, 47), which are smoking-related. We explored the relationship of smoking to two inflammatory biomarkers, serum concentration of IL6 and CRP. Previous studies showed that smoking induces systemic inflammation, which is reflected in elevated levels of IL6 (37) and CRP (38). We similarly observed that IL6 and CRP were significantly higher in current smokers (**Supplementary Table 8**). We further identified three smoking-related gene expression signatures in association with IL6 and 55 with CRP, even after adjusting for smoking status. Among these genes, we detected one that was a mediator of the association of smoking with IL6 concentration, and seven genes mediating the association of smoking with CRP. *ALAS2* emerged as a gene mediator for both IL6 and CRP. *ALAS2* (5'-aminolevulinate synthase 2) codes for a mitochondrial enzyme that is erythroid-specific. We speculate that *ALAS2* expression might be related to smoking-induced inflammation, but experimental validation is needed to support this hypothesis.

Another limitation of our study is its cross-sectional nature. We cannot prove causal relations between smoking and gene expression. We were able, however, to include longitudinal analyses of time

since smoking cessation. Further longitudinal studies of smoking-induced gene expression effects on downstream disease phenotypes are warranted. Last, our study included six large epidemiologic studies that all rely on questionnaire-reported ascertainment of smoking status. Self-reported smoking status is imperfect as subjects may not report their status correctly.

In conclusion, we identified transcriptomic signatures of cigarette smoking in a well-powered population-based meta-analysis. Our results suggest that smoking induces global gene expression changes that may involve multiple critical pathways. By linking gene expression signatures with multiple smoking-related diseases, we demonstrated that smoking-related gene expression changes are associated with many smoking-related diseases. Our list of smoking-related gene expression signatures may serve as a compelling resource for future studies.

Materials and Methods

Study participants

Our study included samples from six studies: the Framingham Heart Study (FHS) (48-50), the Rotterdam Study (RS) (51), the Cooperative Health Research in the Region of Augsburg (KORA F4) Study (52), the InCHIANTI Study (53), the Study of Health in Pomerania (SHIP-TREND) (54), and the Estonian Biobank (EGCUT) (55). Each of the six studies followed the recommendations of the Declaration of Helsinki. Informed consent was obtained from each study participant.

Smoking status was ascertained by questionnaire. Current smoking was defined as smoking on average at least one cigarette per day during the past 12 months. Former smoking was defined previously having smoked on average at least one cigarette per day, but having quit for at least 12 months. Never smokers were those who reported having never smoked on average a least one cigarette per day for at least one year. Smoking pack-years was computed by multiplying the average number of cigarettes smoked per day by the number of years smoked, divided by 20. For studies with longitudinal data and with missing or inconsistent pack-years data, pack-years were calculated based on the mean of the reported average number of cigarettes smoked per day using data from all available examinations.

Gene expression profiling

RNA was isolated from whole blood samples. FHS, RS, KORA F4, InCHIANTI and SHIP-TREND collected RNA using PaxGene tubes (Becton Dickinson, Breda, the Netherlands; PreAnalytiX, Hombrechtikon, Switzerland). EGCUT collected RNA using Blood RNA Tubes (Life Technologies, NY, USA). Gene expression in the FHS samples used the Affymetrix Exon Array ST 1.0. RS, KORA F4, InCHANTI, SHIP-TREND, and EGCUT used the Illumina HumanHT12 v3 (KORA F4, InCHANTI, SHIP-TREND, and EGCUT) or v4 (RS) array. The details of sample collection, microarrays, and data processing and normalization in each cohort are provided in the **Supplementary Materials**.

Identification of differentially expressed genes associated with cigarette smoking

Linear regression models were used to test the associations of gene expression with smoking status in each cohort respectively. Smoking status was coded as current=1, never=0, and former=1, and never=0; smoking status was the independent variables and expression of each gene was the outcome. Analyses were conducted for current vs. never and former vs. never smokers. For cohorts without pedigree information, we performed statistical analysis using the *lme4* (56) package of R version 3.0.1, adjusting for age, sex, blood cell counts, and applicable technical covariates (e.g., batch). For cohorts with pedigree information, we performed statistical analysis using the pedigreemm package(57) of R, accounting for the reported familial relationship in addition to the aforementioned factors.

Measured blood cell counts (billion cells/L) including white blood cells, neutrophils, lymphocytes, monocytes, eosinophils, and basophils were available in EGCUT, RS (only white blood cell, lymphocytes, and monocytes available), InCHIANTI, KORA F4 (only white blood cell available), and SHIP-TREND. In FHS, blood cell counts were measured in 2138 FHS Third Generation cohort participants, but not in the Offspring cohort. We estimated the cell counts in all FHS samples by partial least squares regression (58) based on mRNA levels using a model based on the 2138 subjects with both gene expression profiling and differential cell counts. The estimated cell counts values were highly consistent with the measured cell counts (details in **Supplementary Methods**). We collected the effect estimate (β), standard errors, T-statistics, R^2, and *P*-values. We performed a dose-response analysis by using pack-years of cigarette

smoking as an independent variable and gene expression as the outcome. Covariates and a statistical model for the dose-response analysis were the same as those described above.

Evaluate the reproducibility of smoking-related gene transcripts

We conducted meta-analysis of all six cohorts to assess smoking-related gene expression signatures (See Methods, *Meta-analysis*). In order to evaluate the reproducibility of smoking-related gene transcripts, we split the overall sample into independent discovery and replication sets. Our overall analysis framework is presented in **Supplementary Figure 6**. The meta-analysis results from RS, EGCUT, InCHIANTI, KORA F4, and SHIP-TREND samples (N=4610) were used as the discovery set. Results from FHS samples (N= 5623) were used for replication purposes. Because discovery and replication sets used different gene profiling platforms, this analysis evaluated the reproducibility of gene expression signatures in independent cohorts and for different expression array platforms. We at first identified differentially expressed genes for smoking in the discovery set at FDR<0.1, and then attempted replication in the replication set. The replication ratio is defined as the proportion of differentially expressed genes for smoking in the discovery set at FDR<0.1 that could be replicated in the replication set at FDR<0.1.

Meta-analysis

We estimated the heterogeneity of each gene across the six studies. Since, we found <5% of genes with heterogeneity I^2>75%, we performed a meta-analysis using a fixed effect restricted maximum likelihood model (rma method, using default weighting) provided by the metafor package(59) of R. To overcome expression platform differences, the meta-analysis was performed on all transcripts with matching gene Entrez IDs (16,866 unique genes). Meta-analysis was performed across all six studies. For discovery and replication purposes, meta-analysis was also performed for the five Illumina cohorts (RS, EGCUT, InCHIANTI, KORA F4, and SHIP-TREND). We compared the meta-analysis results of the Illumina cohorts with the results of the FHS. We computed the Benjamini-Hochberg false discovery rate (FDR) (60) on the resulting *P*-values by correcting for the number of transcripts that were present in all gene expression microarray platforms (n=16,866). The significant threshold for the identification of smoking-related gene expression signatures was FDR<0.1.

Supplementary Table 11 reports smoking-related gene expression signatures from the final meta-analysis that were associated with measured blood cell types, PC1, and Batch_lump using 1298 never-smokers whose cell types were measured in the FHS.

Identification of long-term persistent associations of cigarette smoking with gene expression levels

A long-term gene expression persistence analysis was performed on FHS participants since it is the only cohort with longitudinal data on smoking cessation status for 35 years. The analysis was performed on a series of six dichotomous variables indicating smoking cessation of at least 5, 10, 15, 20, 25, and 30 years, using a linear mixed model in the pedigreemm package (57) with the same set of covariates as in the primary analysis. We used the T-statistics value that defined statistical significance in the current vs. never smoker analysis ($|T|>3.0$, corresponding to $P<0.002$). Transcripts with $|T|>3.0$ across all six time points are deemed to be statistically significant compared to never-smoker levels.

Gene coexpression network analysis

For the smoking-related gene expression signatures (current vs. never, and former vs. never at FDR<0.1), we performed a gene coexpression network analysis using FHS gene expression data. Gene coexpression networks were constructed using weighted gene coexpression network analysis (WGCNA) [61, 62]. The WGCNA R package uses a fitting index to evaluate a scale-free network structure built upon Pearson gene-gene correlations from gene expression variance among individuals [61]. Genes were grouped based on the topological overlap of their connectivity using average linkage hierarchical clustering [61], followed by a dynamic cut-tree algorithm to dynamically cut the clustering dendrogram branches into gene coexpression network modules (coEMs) [63].

We first adjusted for sex, age, blood count proportions, and technical covariates from the expression data using linear mixed models (lme4 package [56] in R) in order to minimize confounding of other smoking-related covariates. The residuals were kept for the coexpression network construction. First, we built weighted gene coexpression networks and identified coEMs that fit a scale-free topological structure by fitting the index $R^2>0.8$ of the linear model that regressed *log(p(k))* on *log(k)*, where *k* is the connectivity of every node (gene) in the network and *p(k)* is the frequency distribution of connectivity. The fitting index of a perfect scale-free network is 1.

We tested the association of each smoking-related coEM (using the first principle components of each coEM) with each cell types. As shown in **Supplementary Figure 7** the coEMs were not associated with cell types. We further tested the associations of each coEM with two inflammatory biomarkers (i.e., serum concentrations of Interleukin 6 [IL6] and C-reactive protein [CRP]). We calculated the first principal component (eigengene) of each coEM, then used a linear mixed model implemented in the kinship package in R (64), to test the association between a module's eigengene and IL6 and CRP, modeling covariates (including BMI and smoking status). IL6 and CRP related coEMs were identified at $P<0.05$.

Gene ontology enrichment analysis

Each smoking-related gene coexpression network module was classified using Gene Ontology - biology process (GO-BP) categories to define biological process enrichment (65). Fisher's exact test was used to calculate enrichment *P* values. The *P* value was further corrected by the number of unique GO-BP terms (N=825). A threshold of $P<6e\text{-}5$ (0.05/825) was considered significant.

Linking smoking-related gene expression signatures to complex diseases and traits

We looked up the relations of smoking-related gene expression signatures to disease phenotypes and traits using two resources. First, we used the NHGRI GWAS catalog (assessed July, 2015)(32), which collected the associations of SNPs with hundreds of disease and trait phenotypes ($P<1e\text{-}5$). We linked smoking-related gene expression signatures to gene expression-associated SNPs (eSNPs), and then cross-referenced the eSNPs with NHGRI GWAS catalog SNPs. In doing so, we were able to explore the associations of smoking-related gene expression signatures with a large disease-related GWAS SNP sets. *cis-/trans-* eSNPs (i.e. SNPs associated with expression level of a gene) were identified in whole blood based on expression in the FHS using Affymetrix exon array (n~5600) (Joehanes R, PhD, unpublished data, 2016) and a meta-analysis of seven cohorts using Illumina arrays (n~5300) (31). A *cis*-eSNP was defined as a SNP residing within 1Mb of the transcript start site (TSS) for the corresponding gene. eSNPs that were remote from the TSS were defined as *trans*-eSNPs. All eSNPs used in this study passed FDR<0.1. Enrichment analysis was performed using Fisher's exact test by testing enrichment of the intersecting number of genes (i.e., $M \cap N$) in the NHGRI GWAS catalog having eSNPs (M) and smoking-related gene expression signatures having eSNPs (N) with a background of the total number of genes having eSNPs (T).

Second, we looked at the associations of smoking-related gene expression signatures with pulmonary function and two inflammatory biomarkers (CRP and IL6, nature log-transformation) that are related to cigarette smoking(66, 67) and were measured in FHS at the same visit as the gene expression blood sample collection. Pulmonary function measures included forced expiratory volume at one second (FEV1), forced vital capacity (FVC), and the FEV1/FVC ratio. FEV1 and FVC were measured on the FHS Offspring cohort at Examination 8 and on the FHS Third Generation cohort at Examination 2 using a Collins CPL dry rolling-seal spirometer and Collins 2000 Plus/SQL software (Collins Medical, Inc., Braintree, MA). The highest value among acceptable efforts was used, as per the American Thoracic Society-European Respiratory Society guidelines(68). CRP was measured on the FHS Offspring cohort at Examination 8 and the Third Generation cohort at Examination 2 using a high sensitivity Dade-Behring BN 100 nephelometer. Serum IL6 was measured in FHS Offspring cohort participants at Examination 8 using the Quantikine HS IL6 Immunoassay kit (R/D Systems, Minneapolis, MN). Intra-assay coefficients of variation for inflammatory marker measurements were <9.2%.

Residuals for genes after adjusting for technical covariates (as independent variables) were used to identify differentially expressed genes associated with CRP, IL6, and pulmonary function phenotypes (as outcomes) using linear mixed models implemented in the kinship R package (64). The covariates for analyzing pulmonary function include age, sex, height, weight, smoking statues, imputed differential white blood cell proportions, and family structure. The covariates for analyzing CRP and IL6 included age, sex, BMI, smoking status, imputed differential white blood cell proportions, and family structure. Statistical significance was based on Bonferroni correction (P<0.05/1270) for the number of smoking-related gene signatures (n=1270).

Mediation analysis

For the overlapping gene expression signatures of smoking and smoking-related phenotypes (including natural log transformed [log-transformed] CRP, log-transformed IL6, and FEV1), we tested if gene expression signatures mediated the associations of smoking with the smoking-related phenotypes. Mediation analysis was performed in individuals whose gene expression and phenotype data were both

available (n=5615 for CRP analysis, n=2422 for IL6, and n= 5199 for FEV1). **Supplementary Figure 8** shows the distribution of log-transformed CRP, log-transformed IL6, and FEV1 values.

Mediation was considered to be present when there was a significant decrease in the association of smoking with downstream phenotypes (Model 1) after adjusting for gene expression signatures (Model 2). The Sobel test was used to evaluate mediation effects, and the significance level was a Bonferroni corrected $P<0.05$ (0.05/ the number of genes in the overlapping gene expression signatures for smoking and smoking-related phenotypes).

Model 1: Outcome ~ β1 (Smoking) + covariates

Model 2: Outcome ~ β1' (Smoking) + β2 (Gene) + covariates

In Model 1 and Model 2, the "outcomes" were log-transformed CRP, log-transformed IL6, and FEV1. For genes showing mediation, the mediation proportion was defined as (β1- β1') / β1. Covariates included age, sex, technical covariates, blood cell counts and family structure. The analysis for FEV1, additional adjustment included height and weight.

Test SNP-by-Smoking interaction on gene expression levels

SNP-by-smoking interaction in relation to expression of each gene was tested in ~5300 FHS participants with gene expression and genotyping data by utilizing the following model using the kinship package in R(64):

geneExp ~ β1 (SNP) + β2 (Smoking) + β3 (SNP x Smoking) + covariates

where β1 and β2 are the regression coefficients for the SNP (additive model) and smoking status (current vs. never), respectively. β3 is the regression coefficients for the SNP-by-smoking status interaction. geneExp is the gene expression residual after adjusting for technical covariates (age, sex and white blood cell types as fixed effects, and family structure as a random effect). The interaction tests were limited to the SNPs in the NHGRI GWAS catalog that overlapped with SNPs genotyped or imputed in the FHS at minor allele frequency >1% (15,579 SNPs), and smoking gene expression signatures (1290 unique genes from among 1270 genes for current vs. never smokers plus 39 genes for former vs. never smokers). The Benjamini-Hochberg method(60) was used to calculate false discovery rate (FDR).

Smoking-related genes and DNA methylation loci from published literatures

To compare the smoking-related gene expression signatures identified in this study with previous studies, we collected previously reported smoking-related gene expression signatures in whole blood (529 genes) (26-29), monocytes (311 genes) (24, 25), and lung tissues (479 genes) (21-23). In addition, a list of genes whose CpGs were reported to be differentially methylated in relation to smoking was download from a review article (30). In the review article, the authors collected the gene list by reviewing 14 published epigenome-wide association studies including 1460 unique CpGs for 939 unique genes. All the previously identified genes and CpGs were available within the results of our analysis of current vs. never smokers. Fisher's exact test was used to test if the smoking-related gene expression signatures identified in our study were enriched for previously identified smoking-related gene expression signatures or DNA methylation signatures.

Data Availability

Raw gene expression profiling data are available online (FHS [http://www.ncbi.nlm.nih.gov/gap; accession number phs000007], RS [GSE33828], KORA F4 [E-MTAB-1708], InCHIANTI [GSE48152], SHIP-TREND [GSE36382] and EGCUT [GSE48348]).

Acknowledgements

The **Framingham Heart Study** is funded by National Institutes of Health contract N01-HC-25195 HHSN268201500001I; 1RO1 HL64753; R01 HL076784; 1 R01 AG028321; Dr. Benjamin is funded by 1P50HL120163. The laboratory work for this investigation was funded by the Division of Intramural Research, National Heart, Lung, and Blood Institute, National Institutes of Health. The analytical component of this project was funded by the Division of Intramural Research, National Heart, Lung, and Blood Institute, and the Center for Information Technology, National Institutes of Health, Bethesda, MD.

The **InCHIANTI study** was supported as a "targeted project" (ICS110.1/RF97.71) by the Italian Ministry of Health at baseline (1998 –2000), with subsequent visits supported in part by the U.S. National Institute on Aging (Contracts: 263 MD 9164 and 263 MD 821336).

The **Rotterdam Study** is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Netherlands Organisation of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012), the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists. The generation and management of RNA-expression array data for the Rotterdam Study was executed and funded by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. We thank Marjolein Peters, MSc, Ms. Mila Jhamai, Ms. Jeannette M. Vergeer-Drop, Ms. Bernadette van Ast-Copier, Mr. Marijn Verkerk and Jeroen van Rooij, BSc for their help in creating the RNA array expression database.

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the grant 03IS2061A of the Federal Ministry of Education and Research.

Generation of whole-blood transcriptome data was funded by the grant no. 03ZIK012 of the Federal Ministry of Education and Research. The University of Greifswald is a member of the Caché Campus program of the InterSystems GmbH.

The **KORA** research platform (KORA, Cooperative Research in the Region of Augsburg) was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The German Diabetes Center is funded by the German Federal Ministry of Health (BMG) and the Ministry of Innovation, Science, Research and Technology (MIWF) of the State North Rhine-Westphalia. This study was supported in part by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.). We are indebted to all study participants. Furthermore, we are grateful to the field staff in Augsburg and Munich who were involved in the conduct of the KORA studies, and the staff of the Genome Analysis Center at the Helmholtz Zentrum München involved in the omics measurements.

EGCUT work was supported through the Estonian Genome Center of University of Tartu by the Targeted Financing from the Estonian Ministry of Science and Education [SF0180142s08]; the Development Fund of the University of Tartu (grant SP1GVARENG); the European Regional Development Fund to the Centre of Excellence in Genomics (EXCEGEN; grant 3.2.0304.11-0312); and through FP7 grant 313010.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

Conflict of Interest Statement

None declared.

References

1 Adler, I. (1912) *Primary malignant growths of the lungs and bronchi*. Longmans, Green, and Company.

2 Control, C.f.D. and Prevention. (2008) Smoking-attributable mortality, years of potential life lost, and productivity losses--United States, 2000-2004. *MMWR. Morbidity and mortality weekly report*, **57**, 1226.

3 Control, C.f.D. and Prevention. (2005) Cigarette smoking among adults--United States, 2004. *MMWR. Morbidity and mortality weekly report*, **54**, 1121.

4 Control, C.f.D. and Prevention. (2011) Vital signs: current cigarette smoking among adults aged≥ 18 years--United States, 2005-2010. *MMWR. Morbidity and mortality weekly report*, **60**, 1207.

5 Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2004) Mortality in relation to smoking: 50 years' observations on male British doctors. *Bmj*, **328**, 1519.

6 Tabuchi, T., Ito, Y., Ioka, A., Nakayama, T., Miyashiro, I. and Tsukuma, H. (2013) Tobacco smoking and the risk of subsequent primary cancer among cancer survivors: a retrospective cohort study. *Annals of oncology*, **24**, 2699-2704.

7 Lu, M., Ye, W., Adami, H.-O. and Weiderpass, E. (2008) Stroke incidence in women under 60 years of age related to alcohol intake and smoking habit. *Cerebrovascular Diseases*, **25**, 517-525.

8 Vink, J.M., Smit, A.B., de Geus, E.J., Sullivan, P., Willemsen, G., Hottenga, J.-J., Smit, J.H., Hoogendijk, W.J., Zitman, F.G. and Peltonen, L. (2009) Genome-wide association study of smoking initiation and current smoking. *The American Journal of Human Genetics*, **84**, 367-379.

9 Argos, M., Tong, L., Pierce, B.L., Rakibuz-Zaman, M., Ahmed, A., Islam, T., Rahman, M., Paul-Brutus, R., Rahaman, R. and Roy, S. (2014) Genome-wide association study of smoking behaviours among Bangladeshi adults. *Journal of medical genetics*, **5**, 327-333.

10 Tobacco and Consortium, G. (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics*, **42**, 441-447.

11 David, S., Hamidovic, A., Chen, G., Bergen, A., Wessel, J., Kasberger, J., Brown, W., Petruzella, S., Thacker, E. and Kim, Y. (2012) Genome-wide meta-analyses of smoking behaviors in African Americans. *Translational psychiatry*, **2**, e119.

12 Guida, F., Sandanger, T.M., Castagné, R., Campanella, G., Polidoro, S., Palli, D., Krogh, V., Tumino, R., Sacerdote, C. and Panico, S. (2015) Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Human molecular genetics*, **24**, 2349-2359.

13 Breitling, L.P., Yang, R., Korn, B., Burwinkel, B. and Brenner, H. (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *The American Journal of Human Genetics*, **88**, 450-457.

14 Wan, E.S., Qiu, W., Baccarelli, A., Carey, V.J., Bacherman, H., Rennard, S.I., Agusti, A., Anderson, W., Lomas, D.A. and DeMeo, D.L. (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Human molecular genetics*, **21**, 3073-3082.

15 Wan, E.S., Qiu, W., Carey, V.J., Morrow, J., Bacherman, H., Foreman, M.G., Hokanson, J.E., Bowler, R.P., Crapo, J.D. and DeMeo, D.L. (2015) Smoking Associated Site Specific Differential Methylation in Buccal Mucosa in the COPDGene Study. *American journal of respiratory cell and molecular biology*, **2**, 246-254.

16 Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., Weidinger, S., Lattka, E., Adamski, J. and Peters, A. (2013) Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One*, **8**, e63812.

17 Shenker, N.S., Ueland, P.M., Polidoro, S., van Veldhoven, K., Ricceri, F., Brown, R., Flanagan, J.M. and Vineis, P. (2013) DNA methylation as a long-term biomarker of exposure to tobacco smoke. *Epidemiology*, **24**, 712-716.
18 Zhu, J., Sova, P., Xu, Q., Dombek, K.M., Xu, E.Y., Vu, H., Tu, Z., Brem, R.B., Bumgarner, R.E. and Schadt, E.E. (2012) Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biol*, **10**, e1001301.
19 Huan, T., Liu, C., Joehanes, R., Zhang, X., Chen, B.H., Johnson, A.D., Yao, C., Courchesne, P., O'Donnell, C.J. and Munson, P.J. (2015) A systematic heritability analysis of the human whole blood transcriptome. *Human genetics*, **134**, 343-358.
20 Jones, P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, **13**, 484-492.
21 Landi, M.T., Dracheva, T., Rotunno, M., Figueroa, J.D., Liu, H., Dasgupta, A., Mann, F.E., Fukuoka, J., Hames, M. and Bergen, A.W. (2008) Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS one*, **3**, e1651.
22 Staaf, J., Jönsson, G., Jönsson, M., Karlsson, A., Isaksson, S., Salomonsson, A., Pettersson, H.M., Soller, M., Ewers, S.-B. and Johansson, L. (2012) Relation between smoking history and gene expression profiles in lung adenocarcinomas. *BMC medical genomics*, **5**, 22.
23 Boelens, M.C., van den Berg, A., Fehrmann, R.S., Geerlings, M., de Jong, W.K., te Meerman, G.J., Sietsma, H., Timens, W., Postma, D.S. and Groen, H.J. (2009) Current smoking‐specific gene expression signature in normal bronchial epithelium is enhanced in squamous cell lung cancer. *The Journal of pathology*, **218**, 182-191.
24 Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K. and Rossmann, H. (2010) Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PloS one*, **5**, e10693.
25 Charlesworth, J.C., Curran, J.E., Johnson, M.P., Göring, H.H., Dyer, T.D., Diego, V.P., Kent, J.W., Mahaney, M.C., Almasy, L. and MacCluer, J.W. (2010) Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. *BMC medical genomics*, **3**, 1.
26 van Leeuwen, D.M., van Agen, E., Gottschalk, R.W., Vlietinck, R., Gielen, M., van Herwijnen, M.H., Maas, L.M., Kleinjans, J.C. and van Delft, J.H. (2006) Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs. *Carcinogenesis*, **28**, 691-697.
27 Beineke, P., Fitch, K., Tao, H., Elashoff, M.R., Rosenberg, S., Kraus, W.E. and Wingrove, J.A. (2012) A whole blood gene expression-based signature for smoking status. *BMC medical genomics*, **5**, 1.
28 Paul, S. and Amundson, S.A. (2014) Differential effect of active smoking on gene expression in male and female smokers. *Journal of carcinogenesis & mutagenesis*, **5**.
29 Na, H.-K., Kim, M., Chang, S.-S., Kim, S.-Y., Park, J.Y., Chung, M.W. and Yang, M. (2015) Tobacco smoking-response genes in blood and buccal cells. *Toxicology letters*, **232**, 429-437.
30 Gao, X., Jia, M., Zhang, Y., Breitling, L.P. and Brenner, H. (2015) DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clinical epigenetics*, **7**, 1.
31 Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K. and Powell, J.E. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics*, **45**, 1238-1243.
32 Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide

association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9362-9367.
33 Gusev, A., Ko, A., Shi, H., Bhatia, G., Chong, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma, D.I. and Wright, F.A. (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, **48**, 245-252.
34 Schmidt, K. and Rasmussen, J. (1984) Acute platelet activation induced by smoking. In vivo and ex vivo studies in humans. *Thrombosis and haemostasis*, **51**, 279-282.
35 Nowak, J., Murray, J.J., Oates, J.A. and FitzGerald, G.A. (1987) Biochemical evidence of a chronic abnormality in platelet and vascular function in healthy individuals who smoke cigarettes. *Circulation*, **76**, 6-14.
36 Davì, G. and Patrono, C. (2007) Platelet activation and atherothrombosis. *New England Journal of Medicine*, **357**, 2482-2494.
37 Śliwińska-Mossoń, M., Milnerowicz, H., Jabłonowska, M., Milnerowicz, S., Nabzdyk, S. and Rabczyński, J. (2012) The effect of smoking on expression of IL-6 and antioxidants in pancreatic fluids and tissues in patients with chronic pancreatitis. *Pancreatology*, **12**, 295-304.
38 Ohsawa, M., Okayama, A., Nakamura, M., Onoda, T., Kato, K., Itai, K., Yoshida, Y., Ogawa, A., Kawamura, K. and Hiramori, K. (2005) CRP levels are elevated in smokers but unrelated to the number of cigarettes and are decreased by long-term smoking cessation in male smokers. *Preventive medicine*, **41**, 651-656.
39 Petersen, A.M.W., Magkos, F., Atherton, P., Selby, A., Smith, K., Rennie, M.J., Pedersen, B.K. and Mittendorfer, B. (2007) Smoking impairs muscle protein synthesis and increases the expression of myostatin and MAFbx in muscle. *American Journal of Physiology-Endocrinology and Metabolism*, **293**, E843-E848.
40 Sershen, H., Reith, M., Lajtha, A. and Gennaro, J. (1981) Effect of cigarette smoke on protein synthesis in brain and liver. *Neuropharmacology*, **20**, 451-456.
41 Wan, E.S., Qiu, W., Carey, V.J., Morrow, J., Bacherman, H., Foreman, M.G., Hokanson, J.E., Bowler, R.P., Crapo, J.D. and DeMeo, D.L. (2015) Smoking-Associated Site-Specific Differential Methylation in Buccal Mucosa in the COPDGene Study. *American journal of respiratory cell and molecular biology*, **53**, 246-254.
42 Kass, S.U., Landsberger, N. and Wolffe, A.P. (1997) DNA methylation directs a time-dependent repression of transcription initiation. *Current Biology*, **7**, 157-165.
43 Bauer, M., Linsel, G., Fink, B., Offenberg, K., Hahn, A.M., Sack, U., Knaack, H., Eszlinger, M. and Herberth, G. (2015) A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. *Clinical epigenetics*, **7**, 1.
44 Port, J.L., Yamaguchi, K., Du, B., De Lorenzo, M., Chang, M., Heerdt, P.M., Kopelovich, L., Marcus, C.B., Altorki, N.K. and Subbaramaiah, K. (2004) Tobacco smoke induces CYP1B1 in the aerodigestive tract. *Carcinogenesis*, **25**, 2275-2281.
45 Heller, R.A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D.E. and Davis, R.W. (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proceedings of the National Academy of Sciences*, **94**, 2150-2155.
46 Joehanes, R., Ying, S., Huan, T., Johnson, A.D., Raghavachari, N., Wang, R., Liu, P., Woodhouse, K.A., Sen, S.K. and Tanriverdi, K. (2013) Gene expression signatures of coronary heart disease. *Arteriosclerosis, thrombosis, and vascular biology*, **33**, 1418-1426.
47 Huan, T., Zhang, B., Wang, Z., Joehanes, R., Zhu, J., Johnson, A.D., Ying, S., Munson, P.J., Raghavachari, N. and Wang, R. (2013) A systems biology framework identifies molecular underpinnings of coronary heart disease. *Arteriosclerosis, thrombosis, and vascular biology*, **33**, 1427-1434.
48 Feinleib, M., Kannel, W.B., Garrison, R.J., McNamara, P.M. and Castelli, W.P. (1975) The Framingham offspring study. Design and preliminary data. *Preventive medicine*, **4**, 518-525.

49 Kannel, W.B., Feinleib, M., McNAMARA, P.M., Garrison, R.J. and Castelli, W.P. (1979) An investigation of coronary heart disease in families The Framingham offspring study. *American journal of epidemiology*, **110**, 281-290.
50 Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Fox, C.S., Larson, M.G. and Murabito, J.M. (2007) The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *American journal of epidemiology*, **165**, 1328-1335.
51 Hofman, A., Brusselle, G.G., Murad, S.D., van Duijn, C.M., Franco, O.H., Goedegebure, A., Ikram, M.A., Klaver, C.C., Nijsten, T.E. and Peeters, R.P. (2015) The Rotterdam Study: 2016 objectives and design update. *European journal of epidemiology*, **30**, 661-708.
52 Schurmann, C., Heim, K., Schillert, A., Blankenberg, S., Carstensen, M., Dörr, M., Endlich, K., Felix, S.B., Gieger, C. and Grallert, H. (2012) Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PloS one*, **7**, e50938.
53 Ferrucci, L., Bandinelli, S., Benvenuti, E., Iorio, A., Macchi, C., Harris, T.B. and Guralnik, J.M. (2000) Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *Journal of the American Geriatrics Society*, **48**, 1618-1625.
54 Volzke, H., Alte, D., Schmidt, C.O., Radke, D., Lorbeer, R., Friedrich, N., Aumann, N., Lau, K., Piontek, M., Born, G. *et al.* (2011) Cohort profile: the study of health in Pomerania. *Int J Epidemiol*, **40**, 294-307.
55 Hense, H., Kuulasmaa, K., Zaborskis, A., Kupsc, W. and Tuomilehto, J. (1989) Quality assessment of blood pressure measurements in epidemiological surveys. The impact of last digit preference and the proportions of identical duplicate measurements. WHO Monica Project [corrected]. *Revue d'épidémiologie et de santé publique*, **38**, 463-468.
56 Pinheiro, J. and Bates, D. (2006) *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.
57 Vazquez, A., Bates, D., Rosa, G., Gianola, D. and Weigel, K. (2010) Technical note: an R package for fitting generalized linear mixed models in animal breeding. *Journal of animal science*, **88**, 497-504.
58 Boardman, A.E., Hui, B.S. and Wold, H. (1981) The partial least squares-fix point method of estimating interdependent systems with latent variables. *Communications in statistics-theory and methods*, **10**, 613-639.
59 Viechtbauer, W. (2010) Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, **36**, 1-48.
60 Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, in press., 289-300.
61 Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, **4**, Article17.
62 Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
63 Langfelder, P., Zhang, B. and Horvath, S. (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, **24**, 719-720.
64 Abecasis, G.R., Cardon, L.R., Cookson, W., Sham, P. and Cherny, S.S. (2000) Association analysis in a variance components framework. *Genetic epidemiology*, **21**, S341-346.
65 Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.
66 Yanbaeva, D.G., Dentener, M.A., Creutzberg, E.C., Wesseling, G. and Wouters, E.F. (2007) Systemic effects of smoking. *Chest Journal*, **131**, 1557-1566.

67 Levitzky, Y.S., Guo, C.-Y., Rong, J., Larson, M.G., Walter, R.E., Keaney, J.F., Sutherland, P.A., Vasan, A., Lipinska, I. and Evans, J.C. (2008) Relation of smoking status to a panel of inflammatory markers: the framingham offspring. *Atherosclerosis*, **201**, 217-224.
68 Miller, M.R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., Crapo, R., Enright, P., Van der Grinten, C. and Gustafsson, P. (2005) Standardisation of spirometry. *Eur respir J*, **26**, 319-338.

Figure Legends

Figure 1: Volcano plots of meta-analysis of differentially expressed for smoking. A) Current vs. Never Smokers; B) Former vs. Never Smokers.

Figure 2: Twelve genes do not return to never smoker levels at 35 years after cessation. X axis denotes the time after cessation. Y axis denotes the T-statistics between former smokers that quit at a certain period or longer vs. never smokers. Dotted lines indicate threshold of never smoker levels.

Table 1: Top 25 smoking gene signatures for current vs. never smokers based on meta-analysis of six studies

Entrez Gene ID	Gene Symbol	Chr	Beta	Std.Err	P.Value	FDR
54674	*LRRN3*	7	0.64	0.02	1.17E-281	2.94E-277
23328	*SASH1*	6	0.18	0.01	1.09E-98	1.36E-94
56650	*CLDND1*	3	0.18	0.01	1.00E-66	8.40E-63
55022	*PID1*	2	0.25	0.02	1.67E-53	1.05E-49
10462	*CLEC10A*	17	0.09	0.01	6.00E-51	3.01E-47
4118	*MAL*	2	0.17	0.01	9.34E-35	3.91E-31
149628	*PYHIN1*	1	-0.10	0.01	2.42E-34	8.68E-31
1524	*CX3CR1*	3	-0.16	0.01	9.49E-33	2.98E-29
2838	*GPR15*	3	0.09	0.01	3.15E-30	7.91E-27
55020	*TTC38*	22	-0.16	0.01	2.86E-30	7.91E-27
5729	*PTGDR*	14	-0.11	0.01	2.15E-29	4.91E-26
51176	*LEF1*	4	0.15	0.01	5.41E-25	1.13E-21
1028	*CDKN1C*	11	-0.20	0.02	2.28E-24	4.40E-21
53637	*S1PR5*	19	-0.24	0.02	5.09E-24	9.12E-21
9788	*MTSS1*	8	-0.10	0.01	9.61E-24	1.61E-20
154075	*SAMD3*	6	-0.07	0.01	2.52E-22	3.72E-19
4050	*LTB*	6	0.10	0.01	4.31E-21	6.01E-18
23178	*PASK*	2	0.13	0.01	6.34E-21	8.37E-18
2359	*FPR3*	19	0.09	0.01	8.30E-21	1.04E-17
2517	*FUCA1*	1	0.13	0.01	6.30E-20	7.53E-17
389	*RHOC*	1	-0.11	0.01	1.53E-19	1.75E-16
51348	*KLRF1*	12	-0.14	0.02	1.99E-19	2.17E-16
114879	*OSBPL5*	11	-0.11	0.01	2.09E-19	2.19E-16
146330	*FBXL16*	16	0.12	0.01	4.82E-19	4.84E-16
5243	*ABCB1*	7	-0.08	0.01	6.33E-19	6.11E-16

Table 2: Top 25 smoking gene signatures for former vs. never smokers based on meta-analysis of six studies

Entrez Gene ID	Gene Symbol	Chr	Beta	Std.Err	P.Value	FDR
54674	*LRRN3*	7	0.100	0.013	3.27E-14	8.21E-10
11186	*RASSF1*	3	-0.025	0.004	6.13E-09	7.69E-05
284207	*METRNL*	17	-0.026	0.005	1.99E-08	1.66E-04
55020	*TTC38*	22	-0.059	0.012	7.76E-07	4.87E-03
4118	*MAL*	2	0.060	0.012	9.85E-07	4.94E-03
10578	*GNLY*	2	-0.053	0.011	1.94E-06	8.14E-03
7102	*TSPAN7*	X	0.036	0.008	3.61E-06	0.01
29992	*PILRA*	7	-0.034	0.007	5.49E-06	0.02
10023	*FRAT1*	10	-0.038	0.009	1.07E-05	0.03
25829	*TMEM184B*	22	-0.054	0.012	1.30E-05	0.03
6352	*CCL5*	17	-0.056	0.013	1.66E-05	0.04
27202	*C5AR2*	19	-0.029	0.007	3.50E-05	0.06
5729	*PTGDR*	14	-0.037	0.009	3.77E-05	0.06
4145	*MATK*	19	-0.025	0.006	4.08E-05	0.06
8745	*ADAM23*	2	0.033	0.008	4.16E-05	0.06
6774	*STAT3*	17	-0.029	0.007	4.28E-05	0.06
7462	*LAT2*	7	-0.025	0.006	4.30E-05	0.06
56979	*PRDM9*	5	0.019	0.005	4.46E-05	0.06
51176	*LEF1*	4	0.050	0.012	4.72E-05	0.06
56650	*CLDND1*	3	0.038	0.009	5.68E-05	0.07
25996	*REXO2*	11	0.044	0.011	7.55E-05	0.09
10331	*B3GNT3*	19	-0.032	0.008	8.98E-05	0.09
3257	*HPS1*	10	0.036	0.009	9.41E-05	0.09
51339	*DACT1*	14	0.015	0.004	9.85E-05	0.09
5329	*PLAUR*	19	-0.032	0.008	0.000103	0.09

Table 3: Gene ontology enrichment analysis of smoking-related gene coexpression network modules

CoEM	Ontology category	Overlap	Fold Change	P Value	Corrected P Value
Turquoise	Response to wounding	35	2.35	1.40E-06	1.15E-03
	Platelet activation	9	7.28	3.75E-06	3.09E-03
	Integrin-mediated signaling pathway	11	4.99	1.26E-05	0.01
	Blood coagulation	12	4.39	1.86E-05	0.01
	Inflammatory response	23	2.50	4.53E-05	0.04
Blue	T cell activation	19	6.10	2.67E-10	2.20E-07
	Lymphocyte activation	23	4.75	4.68E-10	3.86E-07
	Transmembrane receptor protein tyrosine kinase signaling pathway	17	3.53	6.35E-06	5.24E-03
	T cell proliferation	8	6.76	2.32E-05	0.02
Brown	Immune cell mediated cytotoxicity	9	25.35	4.71E-11	3.89E-08
	Cellular defense response	15	6.69	6.01E-09	4.96E-06
	Positive regulation of apoptosis	21	3.52	3.95E-07	3.26E-04
	Cell migration	23	3.09	9.95E-07	8.21E-04
	Regulation of cytokine biosynthesis	11	5.34	6.52E-06	5.38E-03
Green	Protein biosynthesis	10	5.29	7.58E-06	6.26E-03

Table 4: Smoking-related diseases and traits enriched for smoking-related gene expression signatures

Trait	Overlap Gene NO	Total Genes having eSNPs in GWAS Catalog	Fold Change	P-val	Gene List
--link by cis-eSNP					
Stroke	5	11	6.3	4.5E-5	*ALDH2; CAMTA1; SH2B3; TMEM116; ERP29*
Pulmonary function	16	116	1.9	3.7E-3	*C6orf48; CAMK1D; CSNK2A2; GYPE; HLA-DPA1; HLA-DRA; LST1; LTA; NCR3; SEC61A2; TAP2; TNS1; TRIM10; RPL10A; NAP1L5; WDR11*
Weight	6	40	2.1	0.02	*C6orf48; GTF3A; HLA-DRA; LST1; NCR3; PRKCA*
Asthma	11	89	1.7	0.02	*AGPAT1; C6orf48; CDC25B; HLA-DPA1; HLA-DRA; IL2RB; LST1; MYL6B; TAP2; BLK; VAV3*
-- link by trans-eSNP					
Asthma	6	40	2.1	0.02	*BTN3A2; CCL5; LIMS1; MYC; SSR4; TRIM10*
Coronary heart disease	6	45	1.9	0.04	*BTN3A2; FOS; GBP1; GBP2; GBP4; GZMH*

Table 5: Mediation analysis examining the indirect association of smoking with IL6 and CRP through gene expression

Gene	β1	P for β1	β1'	P for β1'	β2	P for β2	Mediation Prop	Z-Val	Sobel P	Sobel P (corrected)
-- Mediation Analyses: Smoking → Gene → IL6										
ALAS2	0.3	3.39E-07	0.27	3.49E-06	-0.26	1.36E-05	0.09	2.77	5.66E-03	0.017
-- Mediation Analyses: Smoking → Gene → CRP										
ALAS2	0.11	0.03	0.06	0.24	-0.45	1.07E-18	0.47	5.32	1.04E-07	5.73E-06
PLAUR	0.11	0.03	0.06	0.23	0.50	6.54E-08	0.44	4.78	1.74E-06	9.58E-05
DARS	0.11	0.03	0.06	0.23	0.21	4.80E-07	0.44	4.58	4.66E-06	2.56E-04
MFGE8	0.11	0.03	0.08	0.12	0.63	5.88E-10	0.29	4.33	1.52E-05	8.34E-04
RPS2P8	0.11	0.03	0.08	0.11	0.43	3.07E-09	0.28	4.20	2.70E-05	1.48E-03
SNORD48	0.11	0.03	0.07	0.14	0.35	3.12E-05	0.31	3.81	1.37E-04	7.52E-03
CDC25B	0.11	0.03	0.08	0.10	0.49	1.07E-06	0.24	3.81	1.37E-04	7.55E-03

* Model 1: phenotype ~ β1 Smoking + covariates
& Model 2: phenotype ~ β1' Smoking + β2 gene + covariates
§ Bonferroni correction for IL6 was to correct for 3 genes; for CRP was to correct for 55 genes.

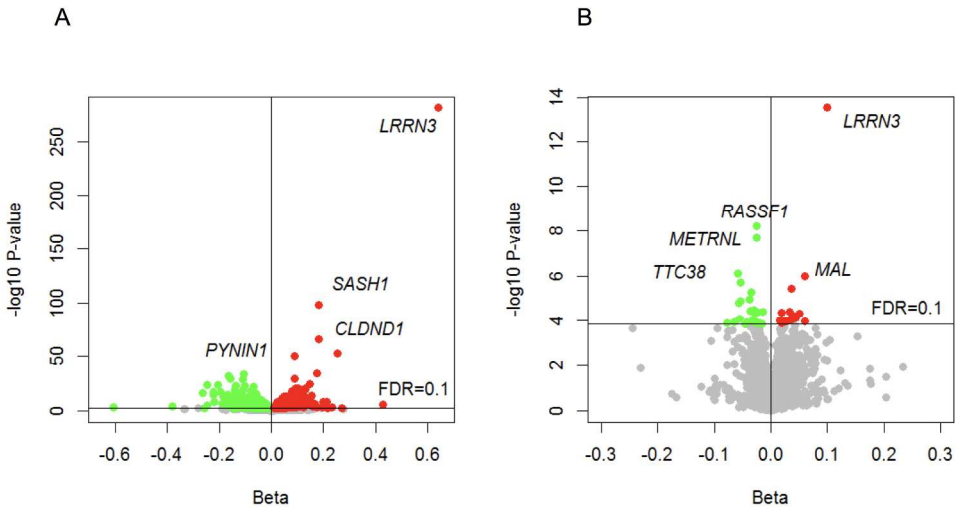
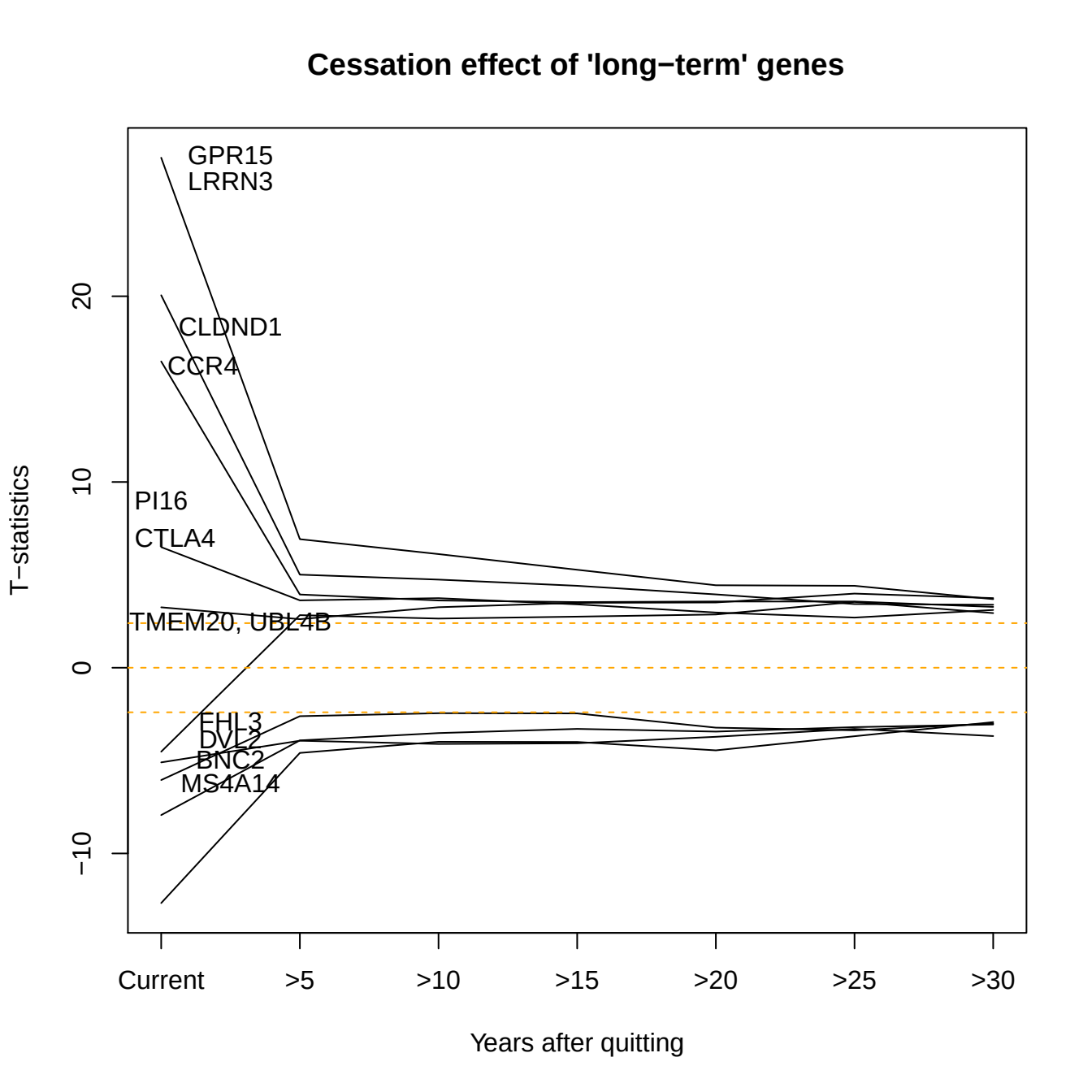


Figure 1: Volcano plots of meta-analysis of differentially expressed for smoking. A) Current vs. Never Smokers; B) Former vs. Never Smokers.

250x133mm (300 x 300 DPI)


Cessation effect of 'long−term' genes
GPR15
LRRN3
CLDND1
CCR4
PI16
CTLA4
TMEM20, UBL4B
FHL3
DVL2
BNC2
MS4A14
20
10
0
−10
T−statistics
Current
>5
>10
>15
>20
>25
>30
Years after quitting