Addiction Biology

SSA SOCIETY FOR THE STUDY OF ADDICTION

HUMAN GENETIC STUDY

doi:10.1111/adb.12322

Differential gene expression patterns between smokers and non-smokers: cause or consequence?

Jacqueline M. Vink[1,2,5,†], Rick Jansen[2,3†], Andy Brooks[4], Gonneke Willemsen[1,5], Gerard van Grootheest[3], Eco de Geus[1,5], Jan H. Smit[3], Brenda W. Penninx[2,3,5] & Dorret I. Boomsma[1,2,5]

Department of Biological Psychology, VU University, Amsterdam, The Netherlands[1], Neuroscience Campus Amsterdam, VU University Medical Center, Amsterdam, The Netherlands[2], Department of Psychiatry/GGZ in Geest, VU University Medical Center, Amsterdam, The Netherlands[3], Department of Genetics, Rutgers University, New Brunswick, New Jersey, USA[4] and EMGO Institute for Health and Care Research, Amsterdam, The Netherlands[5]

ABSTRACT

The molecular mechanisms causing smoking-induced health decline are largely unknown. To elucidate the molecular pathways involved in cause and consequences of smoking behavior, we conducted a genome-wide gene expression study in peripheral blood samples targeting 18 238 genes. Data of 743 smokers, 1686 never smokers and 890 ex-smokers were available from two population-based cohorts from the Netherlands. In addition, data of 56 monozygotic twin pairs discordant for ever smoking were used. One hundred thirty-two genes were differentially expressed between current smokers and never smokers ($P < 1.2 \times 10^{-6}$, Bonferroni correction). The most significant genes were G protein-coupled receptor 15 ($P < 1 \times 10^{-150}$) and leucine-rich repeat neuronal 3 ($P < 1 \times 10^{-44}$). The smoking-related genes were enriched for immune system, blood coagulation, natural killer cell and cancer pathways. By taking the data of ex-smokers into account, expression of these 132 genes was classified into reversible (94 genes), slowly reversible (31 genes), irreversible (6 genes) or inconclusive (1 gene). Expression of 6 of the 132 genes (three reversible and three slowly reversible) was confirmed to be reactive to smoking as they were differentially expressed in monozygotic pairs discordant for smoking. *Cis*-expression quantitative trait loci for *GPR56* and *RARRES3* (downregulated in smokers) were associated with increased number of cigarettes smoked per day in a large genome-wide association meta-analysis, suggesting a causative effect of *GPR56* and *RARRES3* expression on smoking behavior. In conclusion, differential gene expression patterns in smokers are extensive and cluster in several underlying disease pathways. Gene expression differences seem mainly direct consequences of smoking, and largely reversible after smoking cessation. However, we also identified DNA variants that may influence smoking behavior via the mediating gene expression.

Keywords gene expression, genome wide, smoking.

Correspondence to: Jacqueline Vink, Department of Biological Psychology, VU University, Amsterdam, The Netherlands. E-mail: jm.vink@vu.nl

† Both authors contributed equally to this work.

INTRODUCTION

The molecular mechanisms causing smoking-induced health decline are largely unknown. Associations between gene expression and smoking behavior have been observed in multiple tissues (airway epithelial cells, alveolar macrophages, leucocytes, lymphocytes, B cells, monocytes and in whole blood). With one exception ($n = 1240$) (Charlesworth *et al.* 2010), all studies had relatively small sample sizes ($n < 200$) (Spira *et al.* 2004; Heguy *et al.* 2006; Beane *et al.* 2007; Lodovici *et al.* 2007; Charlesworth *et al.* 2010; Pan *et al.* 2010; Beineke *et al.* 2012; Paul & Amundson 2014). Gene expression studies can help elucidating the molecular pathways involved in the etiology and consequences of smoking behavior.

Most studies explored differential gene expression patterns for smoking by comparing gene expression in current smokers with never smokers. Reversibility of gene expression due to smoking can be addressed when ex-smokers are also included. Reversible genes are differentially expressed between current and ex-smokers, but not between ex-smokers and never smokers (Beane *et al.* 2007). If a gene shows a reversible gene expression

pattern, this might suggest that the gene expression pattern is a reaction to smoking (a *reactive gene expression*). This does not have to be the case: the higher expression in the current smokers compared with non-smokers could be the result of a higher genetic liability to smoking behavior, making a person more vulnerable to start smoking and continue smoking (*causative gene expression*).

When studying gene expression associations with smoking behavior in a cross-sectional design, reactive and causative gene expression cannot easily be distinguished. By studying monozygotic (MZ) twin pairs discordant for smoking, reactive genes may be detected, as differential gene expression between a smoking MZ twin and the genetically identical non-smoking co-twin cannot be caused by differences in genetic liability, and the differential expression is therefore likely to be reactive to smoking (van Leeuwen *et al.* 2007). Studying expression quantitative trait loci (eQTLs) may identify causative gene expression. When differential expression between smokers and non-smokers is associated with DNA variants that also influence smoking behavior, this influence may be mediated by gene expression (which is therefore likely to be causative for smoking).

In the present study, we analyze micro-array gene expression measurements in peripheral blood in two Dutch cohorts (Jansen *et al.* 2014; Wright *et al.* 2014). First, we aim to identify genes with differential expression between 743 current smokers and 1686 never smokers. Second, we aim to determine the reversibility of the identified genes by comparing the gene expression levels of current and never smokers with the gene expression levels of 890 ex-smokers. The third aim is to classify the expression of the identified genes as reactive or causative for smoking, using the MZ twins and eQTL lookups in the Tobacco and Genetic Consortium (TAG Consortium 2010) genome-wide association (GWA) study for smoking behavior. This is the first large-scale gene expression study for smoking in peripheral blood combining these different approaches.

METHODS

Subjects

Two projects supplied data for this study: the Netherlands Twin Register (NTR) (Willemsen *et al.* 2010) and the Netherlands Study of Depression and Anxiety (NESDA) (Spijker *et al.* 2004). Both studies were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Center, Amsterdam [Institutional Review Board (IRB) number IRB-2991 under Federalwide Assurance 3703; IRB/institute codes, NESDA 03-183; NTR 03-180], and all subjects provided written informed consent. The sample consisted of 3109 participants from NTR and 1962 from NESDA. The age ranged from 18 to 88 years (mean 38, standard deviation 13), and 65 percent of the sample was female. Data were used of 2892 individuals (from 1433 families) from the NTR and 427 unrelated participants from NESDA. NESDA participants without data on blood counts ($n = 1535$) were excluded. In addition, data of 56 MZ twin pairs discordant for smoking were available from the NTR. eQTL lookup was performed using data of 5071 subjects (3109 NTR and 1962 NESDA participants) for which both genotype and gene expression data were available.

Measures

Information on smoking behavior was obtained during the home visit for blood collection as part of the NTR biobank project (2004–2008). Participants were asked whether they were current smokers, ex-smokers or never smokers. For current smokers, the number of cigarettes smoked per day was obtained, and for ex-smokers, the time since quitting (in years). Subjects who exclusively smoked pipe or cigar were excluded. Data were verified by linking this information to survey data available from the longitudinal survey study of the NTR (Supporting Information Method A). For NESDA subjects, information on smoking was obtained during the day of blood sampling as part of the face-to-face baseline interview. Sex, age at the time of blood draw, body mass index (weight/height squared in kg/m^2) at the time of blood draw, lymphocyte count, monocyte count, neutrophil count, eosinophil count, basophil count and blood hemoglobin (mmol/l) were extracted from NESDA and the NTR databases (Jansen *et al.* 2014; Wright *et al.* 2014). In addition, cotinine levels were available for 612 of 743 current smokers.

RNA extraction

The blood sampling and RNA extraction procedures have been described in detail previously (Willemsen *et al.* 2010; Jansen *et al.* 2014; Wright *et al.* 2014). In short, for NTR, venous blood samples were drawn between 7 and 10 AM after an overnight fast. Within 20 minutes of sampling, heparinized whole blood was transferred into PAXgene Blood RNA tubes (Qiagen, Valencia, Florida, USA) and stored at −20°C. The PAXgene tubes were shipped to the Rutgers University Cell and DNA Repository (http://www.rucdr.org). RNA was extracted using Qiagen Universal liquid handling system (PAXgene extraction kits from manufacturer's protocol). Total RNA was measured by spectroscopy (Trinean DropSense, GentBrugge, Belgium) to determine purity and concentration; RNA fidelity was measured by the (Agilent Bioanalyzer, Santa Clara, California, USA) analysis.

From NESDA subjects, venous whole-blood samples were obtained between 8 and 10 AM, after overnight

fasting. Between 10 and 60 minutes after blood draw, 2.5 ml of heparinized blood was transferred into PAXgene tubes. This tube was kept at room temperature for a minimum of 2 hours and then stored at −20°C. Average time between blood sampling and RNA extraction was 113 weeks. Total RNA was extracted according to the manufacturer's protocol (Qiagen) (Spijker *et al.* 2004).

Gene expression

Gene expression was assessed at the Rutgers University Cell and DNA Repository. Samples were randomly assigned to plates. For cDNA synthesis, 50 ng of RNA was reverse transcribed and amplified in a plate format on a Biomek FX liquid handling robot (Beckman Coulter, Brea, California, USA) using Ovation Pico WTA reagents per the manufacturer's protocol (NuGEN, San Carlos, California, USA). Products purified from single-primer isothermal amplification were then fragmented and labeled with biotin using Encore Biotin Module (NuGEN). Prior to hybridization, the labeled cDNA was analyzed using electrophoresis to verify the appropriate size distribution (Caliper AMS90, HT DNA 5K/RNA LabChip). Samples were hybridized to Affymetrix U219 array plates (GeneTitan, Affymetrix, Santa Clara, California, USA). The U219 array contains 530 467 probes for 49 293 transcripts. All probes are 25 bases in length and designed to be 'perfect match' complements to a designated transcript. Array hybridization, washing, staining and scanning were carried out in an Affymetrix GeneTitan System per the manufacturer's protocol.

Gene expression quality control

Gene expression data were required to pass standard Affymetrix quality control metrics (Affymetrix expression console). Probes were removed when their location was uncertain or if their location intersected a polymorphic single-nucleotide polymorphism (SNP), leaving 44 241 probe sets for analysis. Expression values were obtained using robust multi-array average normalization implemented in Affymetrix Power Tools (v 1.12.0). Data for samples that displayed an average Pearson correlation below 0.8 with the probe set expression values of other samples and samples with incorrect sex chromosome expression were removed.

Statistical analyses

1a. Differential gene expression between current smokers and never smokers

Linear mixed models were used to explore differences in gene expression patterns between current smokers ($n = 743$) and never smokers ($n = 1686$) while correcting for family relatedness (Visscher, Benyamin, & White 2004). For each of the 44 241 probe sets, a mixed model was fitted with gene expression as dependent variable and smoking status (current versus never) as independent variable. Fixed-effect covariates included in the final model were sex, age at the time of blood draw, body mass index, lymphocyte count, monocyte count, neutrophil count, eosinophil count, basophil count, average correlation between arrays, hemoglobin (mmol/l), cohort (NTR or NESDA), time of blood sampling, month of blood sampling, time between blood sampling and RNA extraction and the time between RNA extraction and RNA amplification. Random effects were plate, well, family ID and zygosity. Non-significant covariates were not included in the final model: depression status (DSM-IV depression yes/no), psychotropic medication (yes or no), alcohol use (≥12 glasses of alcohol per week versus <12 glasses per week or not drinking), education level (low, moderate or high), time between RNA amplification and RNA fragmentation and time between RNA fragmentation and RNA hybridization. Mixed models and *P*-values were computed using the R function lmer from the package lme4. To correct for multiple testing, a Bonferroni correction was applied ($P < 0.05/44\,241 = 12^{-06}$). The genes differentially expressed between current and never smokers are called 'smoking-related genes' throughout the rest of the manuscript.

1b. Pathway analyses of smoking-related genes. We used Fisher's exact test to explore whether gene categories were enriched among the identified smoking-related genes. We used gene categories from Gene Ontology Biological Process (GOBP), Kyoto Encyclopedia of Genes and Genome (KEGG) and Ingenuity Pathway Analysis (IPA). KEGG, GOBP and IPA have overlapping but also unique gene categories. GOBP does not contain disease gene categories, but KEGG and IPA do. GOBP covers many biological processes not covered by KEGG and IPA. KEGG contains strongly curated pathways, whereas IPA provides more broadly defined categories. IPA provides *P*-values and corresponding false discovery rates (FDRs) for enrichment. For the GOBP and KEGG pathways, the statistical software package R was used. GOBP categories were retrieved using the R package org.Hs.eg.db (version 2.10.1). KEGG pathways were downloaded from http://www.broadinstitute.org/gsea/downloads.jsp (c2.cp.kegg.v4.0.entrez.gmt). The reference set and the gene set were defined separately for KEGG and GOBP analysis. In total, 9902 GOBP and 186 KEGG categories contained one or more genes measured by the U219 microarrays. The gene set with smoking-related genes was tested for enrichment (separately for upregulated and downregulated genes) in categories with more than one gene overlap. The *P*-value from the exact test is the chance that the overlap between the gene set and the gene category is not larger than for random gene sets of this size within the reference set. For the *P*-values derived from these tests, the FDR was computed.

2. Reversibility of the associations between smoking and gene expression

2a. Smoking status. To explore reversibility of associations between smoking and gene expression, for the smoking-related genes, pairwise comparisons of gene expression levels were made using mixed models as described in item 1a. The pairwise comparisons (in addition to current smokers versus never smokers) were current ($n = 743$) versus ex-smokers ($n = 890$) and ex-smokers versus never smokers ($n = 1686$). For significance, we used the threshold FDR <5 percent (only correcting for tests performed for smoking-related genes).

Reversibility was considered:

Reversible genes (gene expression current > ex = never): if genes differed in gene expression level between current and non-smokers (ex-smokers or never smokers) but were similar between ex-smokers and never smokers

Slowly reversible genes (gene expression current > ex > never): if the expression pattern was different between current and ex-smokers but also between ex-smokers and never smokers

Irreversible genes (gene expression current = ex > never): if gene expression was the same in current and ex-smokers but differed between ever (current and ex) and never smokers

2b. Time since quitting smoking. Gene expression levels of ex-smokers who quit smoking more than 5 years ago (long-term quitters) were compared with ex-smokers who quit 5 years ago or less. We expect the gene expression levels of long-term quitters to be closer to non-smokers than for short-term quitters. Time since quitting was also analyzed as a continuous variable, to check for a linear relationship between level of gene expression and time since quitting.

2c. Cigarettes per day. To explore the association between number of cigarettes per day and gene expression levels, gene expression levels of three groups of current smokers were compared: smokers who smoked more than 20 cigarettes per day (heavy smokers), smokers who smoked 10–19 cigarettes per day (moderate smokers) and smokers who smoked less than 10 cigarettes per day (light smokers). Cigarettes per day was also analyzed as a continuous variable, to check for a linear relationship between level of gene expression and time since quitting. In addition, cotinine levels were analyzed as biomarker for quantity smoked.

3. Reactive versus causative expression of smoking-related genes?

3a. Comparison of gene expression in MZ twin pairs discordant for smoking. Discordant MZ twin pairs were selected: one twin never smoked, and the other was a current smoker at the time of blood sampling ($n = 56$ pairs). First gene expression was residualized using all covariates also used for the mixed models (excluding sex and age as MZ twin pairs have the same sex and age). The residuals were used in a paired *t*-test comparing smoking twins with their non-smoking co-twins for the 132 smoking-related genes.

3b. eQTL analysis. Previous eQTL analysis using NESDA and NTR sample subsets were described elsewhere (Jansen *et al.* 2014; Wright *et al.* in press). For each of the smoking-related genes (result from 1a), the local DNA variant with the strongest association (the top *cis*-eQTL SNP, FDR < 0.05) and strongest global association (the top *trans*-eQTL SNP, FDR < 0.05) were selected. For these SNPs, the *P*-values for association with ever smoking and cigarettes per day in meta-GWA analysis of the TAG Consortium (2010) were retrieved and checked for significance (FDR correction based on the selected *P*-values only).

RESULTS

Characteristics of the sample

In the ex-smoking and never smoking groups, the percentage of women and the mean age is higher compared with that of the group of current smokers. Importantly, neutrophil ($P = 1.1 \times 10^{-40}$), monocyte ($P = 1.7 \times 10^{-34}$), lymphocyte ($P = 1.2 \times 10^{-30}$) and eosinophil ($P = 9.4 \times 10^{-14}$) counts differed between current smokers and never smokers. Blood subcell constitution has a major impact on whole-blood gene expression, and thus, blood cell counts are major confounders when identifying smoking-associated gene expression differences. The characteristics of the participants are summarized in Supporting Information Table S1.

1a. Differential gene expression levels in smokers compared with never smokers. We identified 220 probe sets targeting 132 genes (from 44 241 probe sets targeting 18 238 genes) differentially expressed between current smokers and never smokers ($P < 1.13 \times 10^{-06}$, Bonferroni correction). For each gene, we selected the probe set most significantly associated with smoking for further comparisons. In total, 66 percent of the 132 genes were downregulated in smokers. The most significant results were found for G protein-coupled receptor 15 (*GPR15*, $P < 10^{-150}$) and leucine-rich repeat neuronal 3 (*LRRN3*, $P < 10^{-44}$), which were both upregulated in smokers compared with never smokers (Fig. 1).

The top 25 genes (Table 1) differentially expressed in current smokers compared with never smokers is shown in Table 2, and the complete list of the 132 smoking-related genes can be found in Supporting Information Figure S1. The expression of the 132 identified genes were not significantly associated with age * smoking interaction effects. All subsequent analyses were only carried out with the 132 smoking-related genes.

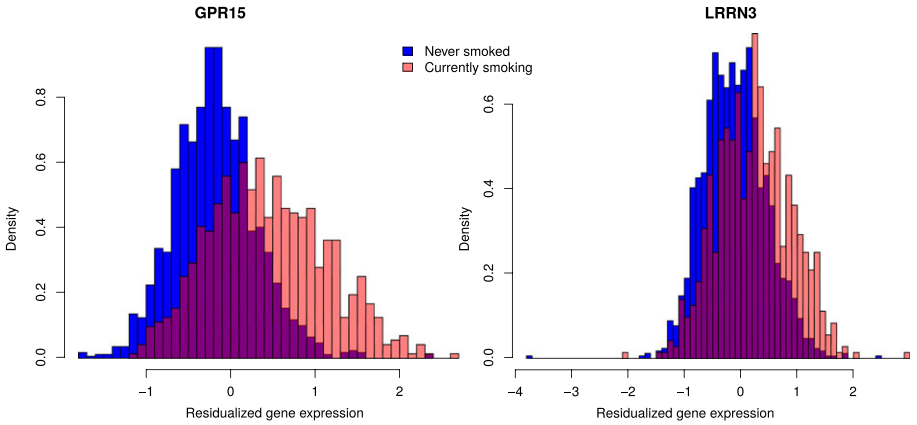

Figure 1 Gene expression distribution in current smokers (pink) and never smokers (blue) for the most significant genes: *GPR15* and *LRRN3*. [Colour figure can be viewed at wileyonlinelibrary.com]

Table 1 Top 25 genes differentially expressed between current smokers and never smokers.

Gene	*Full gene name*	*Location*	*Type*	*P-value*	*Beta*
GPR15	G protein-coupled receptor 15	Plasma membrane	G protein-coupled receptor	2.10e−151	*−0.7289*
LRRN3	Leucine-rich repeat neuronal 3	Extracellular space	Other	4.70e−45	*−0.4167*
S1PR5	Sphingosine-1-phosphate receptor 5	Plasma membrane	G protein-coupled receptor	2.20e−29	**0.3186**
PID1	Phosphotyrosine interaction domain containing 1	Cytoplasm	Other	9.20e−24	*−0.2221*
FGFBP2	Fibroblast growth factor binding protein 2	Extracellular space	Other	3.50e−22	**0.3856**
TBX21	T-box 21	Nucleus	Transcription regulator	4.50e−22	**0.2564**
TRDC	T-cell receptor delta constant	Unknown	Unknown	1.80e−20	**0.3153**
PRSS23	Protease, serine, 23	Extracellular space	Peptidase	1.30e−19	**0.3032**
NKG7	Natural killer cell group 7 sequence	Plasma membrane	Other	2.00e−19	**0.2380**
CCR4	Chemokine (C–C motif) receptor 4	Plasma membrane	G protein-coupled receptor	6.40e−19	*−0.2091*
SPON2	Spondin 2, extracellular matrix protein	Extracellular space	Other	2.10e−18	**0.2885**
GPR56	G protein-coupled receptor 56	Plasma membrane	G protein-coupled receptor	2.30e−18	**0.2756**
KLRD1	Killer cell lectin-like receptor subfamily D, member 1	Plasma membrane	Transmembrane receptor	3.10e−18	**0.2704**
CHST2	Carbohydrate (*N*-acetylglucosamine-6-*O*) sulfotransferase 2	Cytoplasm	Enzyme	1.40e−17	**0.2472**
CLDND1	Claudin domain containing 1	Plasma membrane	Other	2.30e−17	*−0.1460*
PRF1	Perforin 1 (pore forming protein)	Cytoplasm	Other	2.80e−17	**0.2192**
GZMH	Granzyme H (cathepsin G-like 2, protein h-CCPX)	Cytoplasm	Peptidase	1.80e−16	**0.3975**
GPR55	G protein-coupled receptor 55	Plasma membrane	G protein-coupled receptor	2.00e−16	*−0.2000*
FCRL6	Fc receptor-like 6	Plasma membrane	Other	1.90e−15	**0.3178**
MAL	Mal, T-cell differentiation protein	Plasma membrane	Transporter	2.70e−15	*−0.1524*
TGFBR3	Transforming growth factor, beta receptor III	Plasma membrane	Kinase	4.70e−15	**0.2847**
GZMA	Granzyme A (granzyme 1, cytotoxic T lymphocyte-associated serine esterase 3)	Cytoplasm	Peptidase	4.80e−15	**0.2388**
LGR6	Leucine-rich repeat containing G protein-coupled receptor 6	Plasma membrane	G protein-coupled receptor	9.50e−15	**0.1952**
PYHIN1	Pyrin and HIN domain family, member 1	Nucleus	Other	3.60e−14	**0.1606**
TNFRSF13B	Tumor necrosis factor receptor superfamily, member 13B	Plasma membrane	Transmembrane receptor	1.30e−13	*−0.3235*

Beta = effect size, a negative beta (italics) reflects a gene that is downregulated in smokers compared with never smokers, while a positive beta (bold) reflects a gene that is upregulated; Location = the location in the cell where the gene product is present; *P*-value = *P*-value of the test comparing gene expression level in current smokers with never smokers; Type = type of gene product.

Table 2 Gene ontology (GO) pathway analysis in Panther for the 132 genes ($n = 124$ available in Panther) that are differentially expressed in current smokers compared with never smokers.

Biological process	*n Genes total*	*n 124 genes*	*n Expected*	*P-value*
Immune response	653	31	4.81	1.34e−14
Immune system process	2283	49	16.80	1.17e−10
Response to stimulus	1610	40	11.85	4.25e−10
Natural killer cell activation	110	9	0.81	2.66e−05
Response to interferon-gamma	108	8	0.79	2.77e−04
Signal transduction	3861	52	28.41	3.33e−04
Cell communication	4063	53	29.90	6.74e−04
Blood coagulation	261	11	1.92	7.14e−04
Response to external stimulus	261	11	1.92	7.14e−04
Cell adhesion	1236	24	9.10	1.86e−03

Biological process = GO category; *n* Expected = number of expected genes out of the 84 smoking-related genes to be member of this GO category if based on chance; *n* Genes total = total number of genes that are member of this GO category; *n* 124 genes = number of genes out of the 84 smoking-related genes that are member of this GO category; *P*-value = *P*-value for enrichment.

1b. Smoking-related genes enrich several biological pathways. The 132 smoking-related genes were tested for enrichment of GOBP, KEGG or IPA pathways. The upregulated and downregulated genes were explored separately.

The 84 downregulated genes were enriched for 63 GOBP categories (FDR < 0.05) including immune system process (26 genes overlap with this category, FDR = 6.2e−4) and blood coagulation (11 genes overlap, FDR = 5.3e−3) and for 2 KEGG pathways (natural killer cell-mediated cytotoxicity) (6 genes overlap, FDR = 1e−2) and graft-versus-host disease (3 genes overlap, FDR = 3.2e−02, Supporting Information Table S3). Using IPA, we identified 15 canonical pathways enriching the downregulated genes (FDR < 0.1), including natural killer cell signaling (7 genes overlap, FDR = 0.01) and sperm motility (4 genes overlap, FDR = 0.07), 379 disease and biofunctions (FDR < 0.05), including asthma (8 genes overlap, FDR = 1.23e−4) and rheumatic disease (21 genes overlap, FDR = 1.01e−4), and 3 tox functions (FDR < 0.05), including liver cirrhosis (5 genes overlap, FDR = 4.88e−02) and cardiac infarction (5 genes overlap, FDR = 4.88e−02).

The 48 upregulated genes were enriched for 239 GOBP categories (FDR < 0.05) including immune system process (23 genes overlap with this category, FDR = 1.3e−7) and leukocyte differentiation (11 genes overlap, FDR = 1.3e−6) and for 7 KEGG pathways including colorectal cancer [4 genes overlap (*BCL2*, *LEF1*, *TCF7* and *MYC*), FDR = 3.8e−3] and acute myeloid leukemia [3 genes over lap (*LEF1*, *TCF7* and *MYC*), FDR = 3.8e−3, Supporting Information Table S3]. Also, 8 canonical pathways from IPA were enriched (FDR < 0.1, including thyroid cancer and acute myeloid leukemia signaling), 461 disease and biofunctions (FDR < 0.05, including quantity of leukocytes and development of tumor) and 25 tox functions (including inflammation of liver, Supporting Information Table S3).

The GOBP categories enriched for downregulated genes overlap for 33 percent with the GOBP pathways enriched for upregulated genes (21 GOBP categories overlap), indicating that upregulated and downregulated genes are involved in unique (e.g. blood coagulation and cancer) but also shared pathways (like immune system processes).

2a. Smoking status. Of the 132 smoking-related genes, gene expression levels were significantly different in 125 genes when comparing current smokers and ex-smokers, while 37 of the 132 genes were differentially expressed between ex-smokers and never smokers (FDR < 0.05, corrected for 132 tests, Supporting Information Table S2).

In the total group of 132 smoking-related genes, the gene expression levels of 94 genes were *reversible* (gene expression current > ex = never), 31 genes were found to be *slowly reversible* (gene expression current > ex > never) and 6 *irreversible* (gene expression current = ex > never): *LEF1*, *ADAMTS1*, *CST7*, *CCR7*, *GNB2L1* and *SFXN1*. The reversibility of one gene was inconclusive.

The mean gene expression levels of ex-smokers tend to be in between those of never and current smokers (Fig. 2). In general, gene expression levels of ex-smokers were closer to the levels of never smokers than to the levels of current smokers.

2b. Time since quitting smoking. The gene expression levels of long-term quitters (>5 years ago) are often more similar to those of the non-smokers than short-term quitters (≤5 years ago, Fig. 3). We also analyzed the time since quitting as a continuous variable in the ex-smokers. Of the 132 smoking-related genes, the gene expression levels of 15 genes were significantly associated with the time since quitting (FDR < 5 percent, 132 tests, Supporting Information Table S1).

2c. Cigarettes per day. Most of the 132 smoking-related genes show a stronger effect (Fig. 3) on gene expression

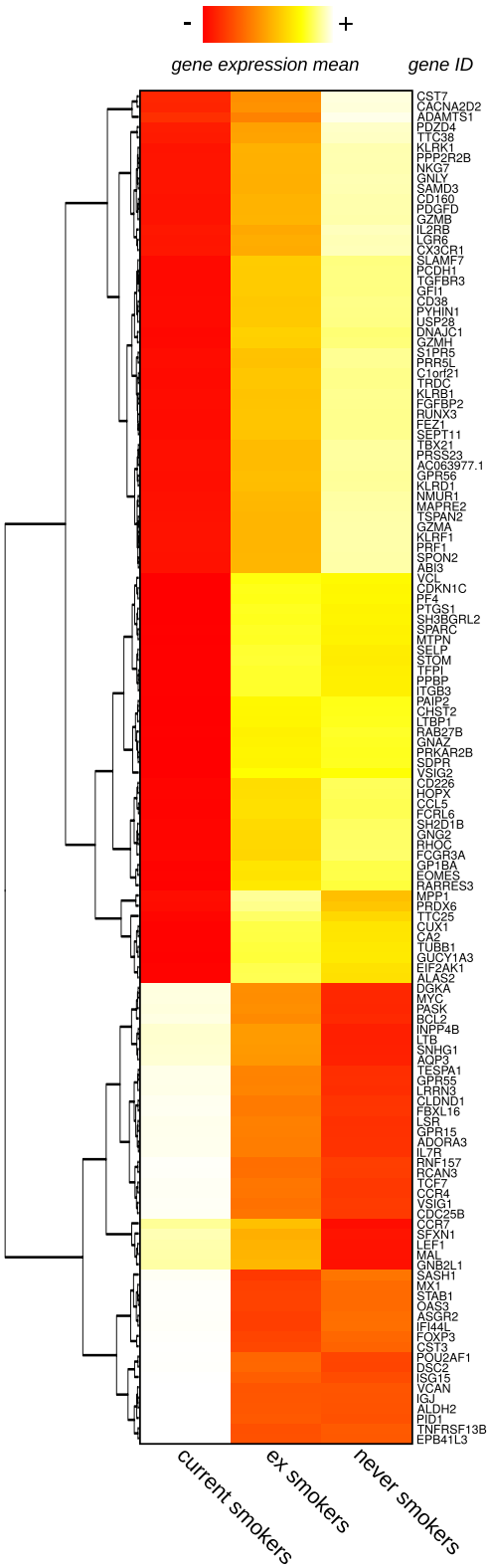

Figure 2 Mean expression levels in smokers, ex-smokers and never smokers for the 132 smoking-related genes (listed on the right) that are differentially expressed between current and never smokers. Gradation from red through orange and yellow to white reflects the amount of gene expression, with red reflecting high gene expression and white reflecting low gene expression. [Colour figure can be viewed at wileyonlinelibrary.com]

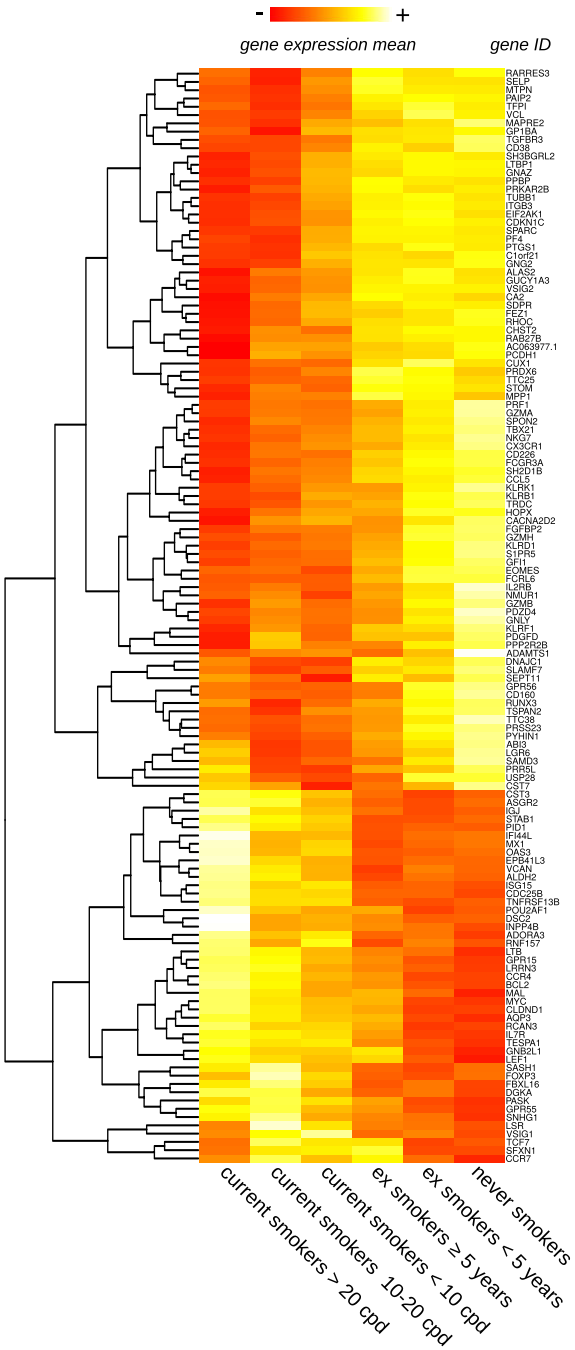

Figure 3 Mean expression in six groups (never smokers, ex-smokers who quit more than 5 years ago, ex-smokers who quit 5 years ago or less, current smokers who smoke less than 11 cigarettes per day, current smokers who smoke 10–20 cigarettes per day and current smokers who smoke more than 20 cigarettes per day) for the 132 candidate genes that are differentially expressed between current and never smokers. Gradation from red through orange and yellow to white reflects the amount of gene expression, with red reflecting high gene expression and white reflecting low gene expression. [Colour figure can be viewed at wileyonlinelibrary.com]

in the heavy smokers (20+ cigarettes per day) compared with the moderate (10–19 cigarettes per day) or light smokers (<10 cigarettes per day), suggesting an association

between dose and gene expression levels. To quantify this association, we analyzed the number of cigarettes per day as a continuous variable. Of the 132 smoking-related genes, the gene expression levels of 28 genes (including the top hits *GPR15* and *LRRN3*) were significantly associated with the number of cigarettes per day. In addition, 10 genes (six overlapping with the cigarettes per day analyses) were significantly associated with cotinine levels, including the top hits *GPR15* and *LRRN3* (Supporting Information Table S2).

3a. Gene expression reactive to smoking identified in MZ discordant twin pairs Of the 132 smoking-related genes, six genes were differentially expressed in the MZ pairs discordant for current smoking: *GPR15*, *PF4*, *TTC38*, *ACO63977.1*, *ALAS2* and *EIF2AK1*. Those genes can be considered to show reactive gene expression. The effects in the discordant MZ twin pairs are in the same direction for 75 percent of the genes, compared with the effects in the total population, which suggests more genes may likely be reactive genes (Supporting Information Fig. S2C).

3b. eQTLs underlying smoking-related genes whose expression is associated with smoking behavior. For each of the 132 smoking-related genes that were differentially expressed between smokers and never smokers, the top *cis*-eQTL SNP and the top *trans*-eQTL were selected. Of the 132 smoking-related genes, 106 have a local *cis*-eQTL and 60 a *trans*-eQTL. For these SNPs, the *P*-values for association with ever smoking and cigarettes per day in the meta-GWA analysis of the TAG Consortium (2010) were gathered and checked for significance. We identified two top *cis*-eQTL SNPs that were significantly associated with the number of cigarettes smoked per day (rs8058865 at *GPR56*, $P = 6.475 \times 10^{-5}$, and rs10897430 at *RARRES3*, $P = 9.3 \times 10^{-4}$) in TAG.

In our study, the T allele of rs8058865 in *GPR56* is associated with decreased expression of *GPR56* ($P = 9.6 \times 10^{-28}$). *GPR56* expression is lower in smokers compared with never smokers ($P = 2 \times 10^{-18}$).

The TAG meta-analyses results showed that the T allele is associated with increased number of cigarettes per day (ß = 0.347, $P = 6.475 \times 10^{-5}$). Together, this suggests that the association between this allele and smoking quantity may be mediated by decreased expression of *GPR56*. However, we did not observe an association between *GPR56* gene expression levels and cigarettes smoked per day in current smokers ($P = 0.54$).

Likewise, the C allele of rs10897430 in *RARRES3* is associated with decreased expression of *RARRES3* (ß = −0.062, $P = 5.6 \times 10^{-8}$). The TAG results show that this SNP is associated with increased number of cigarettes per day (ß = 0.0075, $P = 0.0009$). This suggests that the association between the *RARRES3* allele and smoking quantity is mediated by decreased expression of *RARRES3*. However, we did not observe an association between *RARRES3* gene expression levels and number of cigarettes per day in our sample of current smokers.

None of the other eQTL SNPS from the set of 132 smoking-related genes were significantly associated with smoking behavior in TAG (Supporting Information Table S4).

DISCUSSION

The present study is the largest investigation of gene expression levels in peripheral blood of smokers and non-smokers to date. The results revealed 132 differentially expressed genes in current smokers compared with non-smokers. These smoking-related genes were enriched for several disease-related pathways. Expression of most of these genes is likely to be reactive to smoking as can be derived from integrating data from MZ twin and ex-smokers. However, using eQTL analysis, we identified two SNPs (rs8058865 in *GPR56* and rs10897430 in *RARRES3*) that may influence smoking behavior through intermediate gene expression.

Most previous studies investigating gene expression patterns in relation to smoking were carried out in relatively small samples (Spira *et al.* 2004; Heguy *et al.* 2006; Beane *et al.* 2007; Lodovici *et al.* 2007; Charlesworth *et al.* 2010; Pan *et al.* 2010; Beineke *et al.* 2012; Wright *et al.* in press). So far, 1 larger study is published including 1240 individuals (Charlesworth *et al.* 2010). They detected 323 genes (FDR 5 percent) whose expression levels in lymphocytes were significantly correlated with smoking behavior. In total, 54 of the 323 genes were also present in our list of 132 smoking-related genes (=41 percent of our genes). In a smaller study ($n = 209$ subjects + 180 subjects for replication), a five-gene predictive model for smoking status in whole blood was developed (Beineke *et al.* 2012), and three of the five genes (*LRRN3*, *CLDND1* and *LEF1*) were also found in our list of 132 smoking-related genes. This confirms that the differential gene expression in smokers is a robust finding. In contrast, there was no overlap between the genes expressed in airway epithelial cells of smokers versus non-smokers and the 132 smoking-related genes detected in our study using peripheral blood. This might point to tissue-specific gene expression: smoking a cigarette might influence the expression of other genes in lung tissue compared with blood cells.

To rule out the possibility that the effects we found are merely due to difference in general health between current smokers and non-smokers, we added a variable reflecting overall health status to the model. Betas and *P*-values for

smoking status were highly correlated between models with and without the health variable (Spearman's rho = 0.99 and 0.91, respectively; data not shown), indicating that health variables are unlikely to explain the observed associations between smoking status and gene expression.

Pathway analyses of the 132 smoking-related genes showed that both the upregulated and downregulated genes are involved in the immune response, the immune system and natural killer cell activation. It is well known that inflammatory cells produce a variety of mediators in response to smoking (Sopori 2002; Rom *et al.* 2013). It has been speculated that many of the health consequences of chronic inhalation of cigarette smoke might be due to its adverse effects on the immune system (reviewed by Rom *et al.* 2013). In addition, the downregulated genes were involved in blood coagulation, asthma and cardiac infarction, while the upregulated genes were part of cancer pathways. The expression of some well-known cancer genes, like *MYC* and *LEF1*, was upregulated in smokers compared with never smokers. *LEF1* codes for a lymphoid enhancer-binding factor 1, which is located in the nucleus, and the protein encoded by *MYC* is a multifunctional, nuclear phosphoprotein that plays a role in cell cycle progression, apoptosis and cellular transformation. Both *MYC* and *LEF1* act as transcription regulators and are involved in many types of cancers including lung cancer (Nguyen *et al.* 2009; Dang 2012).

Reversibility of smoking–gene expression associations

Most of the genes were reversible (71 percent) or slowly reversible (24 percent) after smoking cessation, and only six genes (4.5 percent) were irreversible. Of the six genes that were classified as irreversible, *LEF1* is also associated with smoking in a previous study (Beineke *et al.* 2012). The other five genes are not reported in previous studies on smoking but need further investigation in order to explain the differential gene expression between ever smokers and never smokers. The genes classified as 'slowly reversible' included our top hits *GPR15* and *LRRN3*. For *GPR15*, this is in line with a previous study, while *LRRN3* was classified as both rapidly reversible (Wan *et al.* 2012) and slowly reversible (Beineke *et al.* 2012) in previous studies. *GPR15* is a G protein-coupled receptor that acts as a chemokine receptor for human immunodeficiency virus types 1 and 2. *LRRN3* is a gene coding for a leucine-rich repeat neuronal 3, and it is highly expressed in lymphocytes. Both genes were associated with smoking status in peripheral blood data sets (Wan *et al.* 2012; Paul & Amundson 2014), and *GPR15* has been associated with cigarette smoking in three different epigenetic studies focusing on changes in DNA methylation (Breitling *et al.* 2011; Wan *et al.* 2012; Sun *et al.* 2013). More evidence could be obtained with longitudinal data investigating differences in gene expression levels within individuals who quit or start smoking.

The fact that we also observed a relation between the number of cigarettes smoked per day or cotinine levels and the level of gene expression suggests a dose–response relationship. The pattern was observed for most genes but was significant for the expression of 28 genes with cigarettes per day and 10 genes with cotinine. This dose–response relationship is confirmed in the ex-smokers in whom the gene expression levels of long-term quitters were often closer to those of non-smokers than those of short-term quitters.

In order to assess the relative importance of environmental versus genetic sources, we compared the gene expression patterns in MZ twin pairs discordant for smoking. The MZ twin design is a powerful design. A paper of Haque, Gottesman & Wong (2009) reviewed several studies that showed substantial epigenetic differences in MZ twin pairs discordant for psychiatric phenotypes. In smoking research, only one small study with nine MZ twin pairs discordant for smoking reported several genes in which the expression differed in smokers and their non-smoking co-twins, but none of these genes overlapped with our top results. Of the 132 smoking-related genes, six genes were differentially expressed in the MZ pairs consisting of a current smoking twin and a twin who never smoked: *GPR15*, *PF4*, *TTC38*, *ACO63977.1*, *ALAS2* and *EIF2AK1* (FDR < 5 percent for 132 tests). Gene expression of the six genes was categorized as reversible ($n = 3$) or slowly reversible ($n = 3$). Those genes can be considered to show reactive gene expression. The other 126 genes might also have reactive gene expression, but considering the small sample size, the power might be limited. For most of the 132 smoking-related genes (75 percent), the gene expression effects (upregulated or downregulated) in the MZ twin pairs were in the same direction as the effects in the total population. This suggests that cigarette smoking influences gene expression.

In addition, we identified two SNPs (eQTLs of the smoking-related genes *GPR56* and *RARRES3*), which were positively associated with the number of cigarettes per day in the large meta-analyses of the TAG Consortium. Both *cis*-eQTLs were also found in another eQTL study in whole blood (Westra *et al.* 2013). In our sample, these SNPs were both significantly downregulated in current smokers. *GPR56* is a member of the G protein-coupled receptor family. *GPR56* has been shown to have numerous roles in cell guidance/adhesion as demonstrated by its roles in tumor inhibition and neuron development (Fève *et al.* 2014). *RARRES3* codes for the retinoic acid receptor responder protein 3. *RARRES3* is thought to act as a tumor suppressor or growth regulator. A recent study showed that loss of function of *RARRES3* in estrogen receptor-negative breast cancer cells stimulates their invasive capacity and promotes metastasis to the lung. It is

remarkable that both genes are associated with tumor inhibition. Previous studies have demonstrated that the well-known association between cancer and smoking is largely explained by causal effects of smoking (Lee & Hashibe in press), but these results surprisingly suggest an additional role for pleiotropy (same genes influencing both smoking and cancer risk). The TAG consortium GWA study *P*-values for these two SNPS were not very low and need replication in future studies.

Our study was carried out in blood. Although effects of smoking on gene expression might be present in other tissues too, like airway epithelial cells, peripheral blood appears to be a good surrogate tissue for investigating the effect of smoking on gene expression as it expresses a large proportion of the genes encoded in the human genome. Comparison of peripheral blood transcriptome with genes expressed in nine different human tissue types revealed that over 80 percent of gene expression was shared with any given tissue (Liew *et al.* 2006). The current study showed that it is possible to identify gene expression sites and eQTLs with demonstrable criterion validity for smoking behavior in peripheral blood draws.

Another important note is that the current study focused on cigarette smoking; we did not include questions on other ways to take in nicotine (like e-cigarettes or water pipe) or on cannabis use (often smoked) in the biobank study. It should be noted that this information is available from the longitudinal survey study of the NTR for a subsample, but especially with gene expression studies, it is crucial that the information is collected on the same time as the blood collection.

Lastly, it should be noted that persons who have quit a long time ago might be healthier than those who quit recently. We did not ask for the reasons of quitting, so this might represent a mix of health and non-health related reasons. However, when we corrected the models for self-reported health, the health variable did not explain the observed associations between smoking and gene expression.

In conclusion, our results suggest that cigarette smoke causes differential gene expression. The differentially expressed genes play a role in several disease pathways including cancer. Most smoking-related gene expression seem reversible after smoking cessation. In addition, we found two genetic variants influencing gene expression and making subjects vulnerable for smoking behavior. The current results are an important step to provide insights into the association between smoking behavior and differential gene expression.

Acknowledgements

This work was supported by the Netherlands Organization for Scientific Research (MagW/ZonMW grants 904-61-090, 985-10-002, 904-61-193, 480-04-004, 400-05-717 and 912-100-20; Spinozapremie 56-464-14192; Geestkracht program grant 10-000-1002); the Center for Medical Systems Biology (NWO Genomics), Biobanking and Biomolecular Resources Research Infrastructure, VU University's Institutes for Health and Care Research and Neuroscience Campus Amsterdam, NBIC/BioAssist/RK (2008.024); the European Science Foundation (EU/QLRT-2001-01254); the European Community's Seventh Framework Program (FP7/2007-2013); ENGAGE (HEALTH-F4-2007-201413); and the European Research Council (ERC 284167 and ERC 230374). Gene expression data were funded by the US National Institute of Mental Health (RC2 MH089951) as part of the American Recovery and Reinvestment Act of 2009. R.J. was supported by the Biobank-based Integrative Omics Study (BIOS) consortium, which is funded by the Biobanking and Biomolecular Research Infrastructure (BBMRI-NL, NWO project 184.021.007).

References

Beane J, Sebastiani P, Liu G, Brody J, Lenburg M, Spira A (2007) Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. Genome Biol 8:R201.

Beineke P, Fitch K, Tao H, Elashoff M, Rosenberg S, Kraus W, Wingrove J, Investigators P (2012) A whole blood gene expression-based signature for smoking status. BMC Med Genomics 5:58.

Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet 88:450–457.

Charlesworth J, Curran J, Johnson M, Goring H, Dyer T, Diego V, Kent J, Mahaney M, Almasy L, MacCluer J, *et al.* (2010) Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. BMC Med Genomics 3:29.

Dang CV (2012) MYC on the path to cancer. Cell 149:22–35.

Fève M, Saliou J-M, Zeniou M, Lennon S, Carapito C, Dong J, Van Dorsselaer A, Junier M-P, Chneiweiss H, Cianférani S, *et al.* (2014) Comparative expression study of the endo-G protein coupled receptor (GPCR) repertoire in human glioblastoma cancer stem-like cells, U87-MG cells and non malignant cells of neural origin unveils new potential therapeutic targets. PLoS One 9e91519.

Haque FN, Gottesman II, Wong AH (2009) Not really identical: epigenetic differences in monozygotic twins and implications for twin studies in psychiatry. Am J Med Genet C Semin Med Genet 151C:136–141.

Heguy A, O'Connor TP, Luettich K, Worgall S, Cieciuch A, Harvey BG, Hackett NR, Crystal RG (2006) Gene expression profiling of human alveolar macrophages of phenotypically normal smokers and nonsmokers reveals a previously unrecognized subset of genes modulated by cigarette smoking. J Mol Med 84:318–328.

Jansen R, Batista S, Brooks A, Tischfield J, Willemsen G, van Grootheest G, Hottenga J-J, Milaneschi Y, Mbarek H, Madar V, *et al.* (2014) Sex differences in the human peripheral blood transcriptome. BMC Genomics 15:33.

Lee Y-CA, Hashibe M (in press) Tobacco, alcohol, and cancer in low and high income countries. Annals of Global Health 80:378–383.

van Leeuwen DM, van Agen E, Gottschalk RW, Vlietinck R, Gielen M, van Herwijnen MH, Maas LM, Kleinjans JC, van Delft JH (2007)

Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs. Carcinogenesis 28:69107.

Liew C-C, Ma J, Tang H-C, Zheng R, Dempsey AA (2006) The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. J Lab Clin Med 147:126–132.

Lodovici M, Luceri C, De Filippo C, Romualdi C, Bambi F, Dolara P (2007) Smokers and passive smokers gene expression profiles: correlation with the DNA oxidation damage. Free Radic Biol Med 43:415–422.

Nguyen DX, Chiang AC, Zhang XHF, Kim JY, Kris MG, Ladanyi M, Gerald WL, Massagué J (2009) WNT/TCF signaling through LEF1 and HOXB9 mediates lung adenocarcinoma metastasis. Cell 138:51–62.

Pan F, Yang T-L, Chen X-D, Chen Y, Gao G, Liu Y-Z, Pei Y-F, Sha B-Y, Jiang Y, Xu C, *et al.* (2010) Impact of female cigarette smoking on circulating B cells *in vivo*: the suppressed *ICOSLG*, *TCF3*, and *VCAM1* gene functional network may inhibit normal cell function. Immunogenetics 62:237–251.

Paul S, Amundson SA (2014) Differential effect of active smoking on gene expression in male and female smokers. J Carcinog. & Mutagen. 5:1000198.

Rom O, Avezov K, Aizenbud D, Reznick AZ (2013) Cigarette smoking and inflammation revisited. Respir Physiol Neurobiol 187:5–10.

Sopori M (2002) Effects of cigarette smoke on the immune system. Nat Rev Immunol 2:372–377.

Spijker S, van de Leemput JC, Hoekstra C, Boomsma DI, Smit AB (2004) Profiling gene expression in whole blood samples following an *in-vitro* challenge. Twin Res Hum Genet 7:564–570.

Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody JS (2004) Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci U S A 101:10143–10148.

Sun Y, Smith A, Conneely K, Chang Q, Li W, Lazarus A, Smith J, Almli L, Binder E, Klengel T *et al.* (2013) Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. Hum Genet 132:1027–37.

Tobacco and Genetic Consortium (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet 42:441–447.

Visscher PM, Benyamin B, White I (2004) The use of linear mixed models to estimate variance components from data on twin pairs by maximum likelihood. Twin Res Hum Genet 7:670–674.

Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, Agusti A, Anderson W, Lomas DA, DeMeo DL (2012) Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet 21:3073–3082.

Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, ... Franke L (2013). Systematic identification of *trans* eQTLs as putative drivers of known disease associations. Nat Genet 45: 1238–1243. doi: 10.1038/ng.2756

Willemsen G, de Geus EJC, Bartels M, van Beijsterveldt CEM, Brooks AI, Estourgie-van Burk GF, Fugman DA, Hoekstra C, Hottenga JJ, Kluft K, *et al.* (2010) The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. Twin Res Hum Genet 13:231–245.

Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, Madar V, Jansen R, Chung W, Zhou YH, Abdellaoui A, Batista S, Butler C, Chen G, Chen TH, D'Ambrosio D, Gallins P, Ha MJ, Hottenga JJ, Huang S, Kattenberg M, Kochar J, Middeldorp CM, Qu A, Shabalin A, Tischfield J, Todd L, Tzeng JY, van Grootheest G, Vink JM, Wang Q, Wang W, Wang W, Willemsen G, Smit JH, de Geus EJ, Yin Z, Penninx BW, Boomsma DI (2014) Heritability and genomics of gene expression in peripheral blood. Nat Genet 46:430–7. doi: 10.1038/ng.2951.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Figure S1 (A) The effect (beta) of the comparison between current and never on the *x*-axis, with the effect (beta) of the comparison between current and ex on the *y*-axis for the 132 candidate genes. (B) The effect (beta) of the comparison between current and never on the *x*-axis, with the effect (beta) of the comparison between ex and never on the *y*-axis for the 132 candidate genes. Effect is in the same direction when dots are located in the upper right square of the plot (both positive) or in the lower left square of the plot (both negative), which is the case for 100 percent of the genes in (A) and for 80 percent of the genes in (B).
Figure S2 (A) The effect (beta) of the regression of CPD on gene expression levels on the *y*-axis, with the effect (beta) of the comparison of gene expression levels between current and never on the *x*-axis for the 132 smoking-related genes. (B) The effect of the regression of time since quitting on gene expression levels on the *y*-axis with the effect (beta) of the comparison of gene expression levels between current and never on the *x*-axis for the 132 smoking-related genes. (C) The effect of the comparison of gene expression levels of the 132 candidate genes between monozygotic twin pairs discordant for current smoking on the *x*-axis, with the effect of the comparison between current and never smokers in the total population on the *y*-axis.
Nextline:
Table S1 Characteristics of the study sample consisting of smokers, ex-smokers and never smokers.
Table S2 Overview of the 132 smoking-related genes.
Table S3 Pathway analyses for the 132 smoking-related genes.