Supplementary material for 'Increased genetic vulnerability to smoking at *CHRNA5* in early-onset smokers', *Archives of General Psychiatry* (in press).

Hartz, SM, S Short, NL Saccone, R Culverhouse, L Chen, T Schwantes-An, H Coon, Y Han, SH Stephens, J Sun, X Chen, F Ducci, N Dueker, N Franceschini, J Frank, F Geller, D Guðbjartsson, N Hansel, C Jiang, K Keskitalo-Vuokko, J Liu, LP Lyytikäinen, M Michel, R Rawal, A Rosenberger, P Scheet, JR Shaffer, A Teumer, JR Thompson, J Vink, N Vogelzangs, A Wenzlaff, W Wheeler, X Xiangjun , BZ Yang, SH Aggen, A Balmforth, S Baumeister, T Beaty, S Bennett, A Bergen, H Boyd, U Broms, H Campbell, N Chatterjee, J Chen, YC Cheng, S Cichon, D Couper, F Cucca, D Dick, T Foroud, H Furberg, I Giegling, F Gu, A S Hall, J Hällfors, S Han, AM. Hartmann, C Hayward, K Heikkilä, JK. Hewitt, J Hottenga, M Jensen, P Jousilahti, M Kaakinen, S Kittner, B Konte, T Korhonen, MT Landi, T Laatikainen, M Leppert, SM. Levy, R Mathias, DW. McNeil, S Medland, G Montgomery, T Muley, T Murray, M Nauck, K North, M Pergadia, O Polasek, E Ramos, S Ripatti, A Risch, I Ruczinski, I Rudan, V Salomaa, D Schlessinger, U Styrkársdóttir, A Terracciano, M Uda, G Willemsen, X Wu, G Abecasis, K Barnes, H Bickeböller, E Boerwinkle, DI. Boomsma, N Caporaso, J Duan, HJ. Edenberg, C Francks, PV. Gejman, J Gelernter, HJ Grabe, H Hops, MR Jarvelin, V Jorma, M Kähönen, K. Kendler, T Lehtimäki, DF. Levinson, ML. Marazita, J Marchini, M Melbye, B Mitchell, JC. Murray, MM. Nöthen, BW. Penninx, O Raitakari, M Rietschel, D Rujescu, NJ Samani , AR. Sanders, A Schwartz, S Shete, J Shi, M Spitz, K Stefansson, G Swan, T Thorgeirsson, H Völzke, Q Wei, H.-E Wichmann, C Amos, N Breslau, D Cannon, M Ehringer, R Grucza, D Hatsukami, A Heath, E Johnson, J Kaprio, P Madden, N G. Martin, V Stevens, JA. Stitzel, RB. Weiss, P Kraft, and LJ. Bierut.

Supplementary Methods

Heterogeneity in the meta-analysis was assessed using the goodness of fit statistic Q':

$$Q' = \sum_{j=1}^{J} \sum_{i=1}^{K_j} w_{ij} \left(\beta_{ij} - \bar{\beta}_i\right)^2$$

Here, β_{ij} refers to the observed β estimate for rs16969968-A from each of K strata defined by category i (defined by age of onset) and study j, w_{ij} refers to the weights used in the meta-analysis for each stratum (inverse variance of β_{ij}), and $\bar{\beta}_i$ refers to the estimated effect for rs16969968-A in exposure level i from the meta-analysis (using fixed effects only). This statistic has a χ^2 distribution with N-m degrees of freedom, where N is the total number of substrata used ($N = \sum_{j=1}^{J} K_j$), and m is the number of parameters used in the model to estimate $\bar{\beta}_i$. This differs from the standard Cochran's Q statistic by comparing each observed β_{ij} to the meta-analysis estimated parameter $\bar{\beta}_i$, rather than the overall study mean $\bar{\beta}$, allowing for the effect of the tested SNP to differ across exposure strata.

Dataset Descriptions

The American Cancer Society (ACS) Cancer Prevention Study-II Nutrition Cohort (ACS_COPD; ACS_LCA)

The American Cancer Society (ACS) Cancer Prevention Study-II (CPS-II) Nutrition Cohort is a prospective study of cancer incidence and mortality among 86,404 men and 97,786 women. The Nutrition Cohort, which is described in detail elsewhere[1], was initiated in 1992 as a subgroup of CPS-II, a prospective study of cancer mortality involving approximately 1.2 million Americans begun in 1982. Participants in the Nutrition Cohort were recruited from CPS-II members who resided in 21 states and were between the ages of 50 and 74 years. At enrollment in 1992/1993, participants completed a self-administered questionnaire that included demographic, medical, dietary, and lifestyle information. Follow-up questionnaires were sent to all living Nutrition Cohort members in 1997, and every two years after this to update exposure information and to ascertain newly diagnosed cancers. All aspects of the CPS-II Nutrition Cohort study are approved by the Emory University Institutional Review Board.

For the smoking population, all subjects were required to have smoked more than 100 cigarettes lifetime. A detailed description of this population appears elsewhere[2]. Cases were required to have smoked at least 30 cigarettes per day for at least five years. Controls smoked for at least one year during their lifetime and, in 1982 and 1992, reported having smoked fewer than 5 cigarettes per day, and in 1997, fewer than 10 cigarettes per day. Cases were selected to be balanced for gender whereas controls were predominantly female even after all available males were selected. For this meta-analysis, the CPS-II Nutrition cohort smoking cohort (CPS-II_CPD) contributed 2844 unrelated European-Americans (1454 heavy smokers and 1385 light smokers).
DNA was obtained from either a buffy coat or buccal cell sample collected from participants between 1998 and 2002. Genotyping was carried out using Illumina GoldenGate and Sequenom MassArray iPLEX technology. SNPs with a call rate of less than 95% and those for which Hardy-Weinberg equilibrium (HWE) was rejected ($p<0.05$) were excluded. DNA samples with call rate <90% were also excluded.

Participants who developed lung cancer between enrollment in 1992 and 2006 were identified either through self report on a follow-up questionnaire or through linkage with the National Death Index. The lung cancer diagnosis of the self reported cases was verified through medical records or linkage with state cancer registries. Controls were selected from a group of CPS-II participants for whom extensive genotyping had already been completed and who were cancer-free at the time of diagnosis of their matched case. Controls were matched to cases on age (± 2.5 years), gender, and sample type for DNA (buffy coat or buccal cell). All cases and controls were of European descent. For this meta-analysis, the CPS-II Nutrition cohort lung cancer population (CPS-II_LCA) contributed 1006 unrelated European-American smokers.
DNA was obtained from either a buffy coat or buccal cell sample collected from CPS-II Nutrition cohort participants between 1998 and 2002. Genotyping of the DNA samples was carried out using Illumina HumanHap550K, HumanHap610, or HumanHap 1 Million technologies. Genotypes with a call rate of less than 85%, more than 1 HapMap replicate error, more than a 3% (autosomal) or 5% (X chromosome) difference in call rate between genders, or more than 0.5% male AB frequency for the X chromosome, were excluded.

References:

1. Calle EE, Rodriguez C, Jacobs EJ, et al. (2002) The American Cancer Society Cancer Prevention Study II Nutrition Cohort. *Cancer* 94:2490-2501.

2. Stevens VL, Bierut LJ, Talbot JT, et al. (2008) Nicotinic receptor gene variants influence susceptibility to heavy smoking. *Cancer Epidemiol Biomarkers Prev* 17:3517-3525.

The National Longitudinal Study of Adolescent Health (Add Health)

The National Longitudinal Study of Adolescent Health (Add Health) is a longitudinal study of adolescents in grades 7-12 in the United States during the 1994-1995 school year. A sample of 80 high schools and 52 middle schools was systematically selected to ensure the sample was representative of United States schools with respect to region, urbanicity, school size, school type, and ethnicity. The Add Health cohort has been followed into young adulthood with four in-home interviews (corresponding to Waves I, II, III, IV) http://www.cpc.unc.edu/projects/addhealth/projects/addhealth. The most recent interview occurred in 2008, when the sample was aged 24-32. Survey data include respondents' social, economic, psychological and physical well-being with contextual data on family, neighborhood, community, school, friendships, peer groups, and romantic relationships. The Add Health genetic pairs sample includes pairs of individuals with varying genetic similarity including monozygotic twins, dizygotic twins, full siblings, half siblings, and unrelated siblings who were raised in the same household[1].

The number of cigarettes-per-day (CPD) was taken from the *Tobacco, Alcohol, Drugs* section of the in-home questionnaire for Waves I, II and III (as Wave IV data are not yet public). CPD was assessed for the period of heaviest smoking (Wave III questionnaire) and from current use for Waves I, II, and III; the maximum of these values was used to define the CPD trait for analyses. Age of first regular smoking was ascertained from the in-home questionnaire for Wave IV. Individuals were asked whether they had ever smoked cigarettes regularly. If they answered no or did not respond, they were excluded from analysis. Both FTND score and educational attainment were derived from questions asked during the Wave IV questionnaire. A total of 1478 Caucasian non-Hispanic sibling

pairs were genotyped from this community sample. For this meta-analysis, Add Health contributed a sample of 862 unrelated, Caucasian, non-Hispanic subjects (self-reported) from the genetic pairs sample (1 randomly extracted individual from each family). Of this subsample, 501 reported smoking. Before the start of the interview, the interviewer described the interview and obtained consent for participation.
DNA was derived from buccal cells collected from the genetic pairs sample. Genomic DNA was preamplified with the method of Zheng[2]. Taqman assays for allelic discrimination (Applied Biosystems, Foster City, CA) were used to determine SNP genotypes. QC performed on the genotyped sample (by sample and by SNP) excluded individuals with less than 50% genotypes (assumed poor quality DNA sample). All SNPs had greater than 95% genotype calling after exclusion of individuals with low quality DNA samples. All genotypes were called by two independent individuals.

References:

1. Harris KM, Halpern CT, Smolen A, Haberstick BC (2006). The National Longitudinal Study of Adolescent Health (Add Health) Twin Data. Twin Research and Human Genetics 9(6): 988-997.

2. Zheng S, Ma X, Buffler PA, Smith MT, Wiencke JK (2001). Whole genome amplification increases the efficiency and validity of buccal cell genotyping in pediatric populations. Cancer Epidemiol Biomarkers Prev 10: 697–700.

Acknowledgements: This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

Atherosclerosis Risk Communities Study (ARIC)

The ARIC study is a population-based, prospective cohort study of cardiovascular disease and its risk factors sponsored by National Heart, Lung and Blood Institute (NHLBI)[1]. ARIC included 15,792 individuals aged 45-64 years at baseline (1987-89), chosen by probability sampling from four US communities[2]. Cohort members completed four clinic examinations, conducted three years apart between 1987 and 1998. Follow-up for clinical events was annual. The current analysis included 8330 males and females of European ancestry on whom baseline smoking information was available.

References:

1. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am J Epidemiol* 129, 687-702 (1989).

2. Newton-Cheh, C. et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* (2009).

University of Bonn and CIMH Mannheim (BoMa)

The BoMa sample (n=1050) included in this meta-analysis is comprised of three subsamples: patients with a DSM-IV diagnosis of: i) major depression (n=249), ii) bipolar affective disorder (n=389), iii) schizophrenia (n=412) included in GWASs[1,2,3] on the above mentioned disorders and originating from larger samples collected for association studies on the respective phenotypes. Patients were recruited from consecutive admissions to the Department of Psychiatry of the University of Bonn, and the Central Institute of Mental Health Mannheim, Germany and were all of self reported German ancestry/ethnicity. All subjects included here were current or former smokers. The number of cigarettes per day (CPD) was assessed for the period of heaviest smoking. Written informed consent was obtained from all the participants. The studies were approved by the appropriate institutional review boards.

Genomic DNA was prepared from whole blood according to standard procedures. Genotyping of the patients was performed using Illumina HumanHap550v3 bead chips. Stringent quality control procedures were followed, briefly: DNA samples with call rates < 98% were dropped, and SNPs were required to pass the following filters: call rate ≥ 98%, minor allele frequency ≥ 0.01 and conformity with HWE ($p \geq 1e\text{-}6$). Self reported ancestry was verified using EIGENSOFT[4].

The following variables were included in the analysis: gender, cigarettes per day ("Think about the year in your life when you smoked most. During that time, about how many cigarettes did you usually have per day?"), age of onset ("How old were you when you started smoking regularly the first time in your life"), educational attainment ("What is the highest educational degree or diploma you hold?")

References:

1. Rietschel M, Mattheisen M, Frank J, et al. (2010) Genome-wide association-, replication-, and neuroimaging study implicates HOMER1 in the etiology of major depression. Biol Psychiatry 68(6):578-85

2. Cichon S, Mühleisen TW, Degenhardt FA, et al (2011) Genome-wide Association Study Identifies Genetic Variation in Neurocan as a Susceptibility Factor for Bipolar Disorder. Am J Hum Genet 88(3):372-81.

3. Rietschel M, Mattheisen M, Degenhardt F, et al. (resubmitted) Association between genetic variation in a region on chromosome 11 and schizophrenia in large samples from Europe.

4. Price AL, Patterson NJ, Plenge RM, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904-9.

Collaborative Study on the Genetics of Alcoholism (COGA)

This is a case-control study of alcoholism, in which the subjects have been drawn from the Collaborative Study on the Genetics of Alcoholism (COGA), a large, ongoing family-based study that includes subjects from seven sites around the US [1]. COGA has gathered detailed, standardized data on study participants, including diagnostic and neurophysiological assessments. This sample has already proved successful in identifying several genes that influence the risk for alcoholism and neurophysiological endophenotypes, which have been independently replicated [2,3].

Alcoholic probands were recruited from treatment facilities, assessed by personal interview, and after securing permission, other family members were also assessed. A set of comparison families was drawn from the same communities as the families recruited through an alcoholic proband. Assessment involved a detailed personal interview developed for this project, the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA), which gathers detailed information on alcoholism related symptoms along with other drugs and psychiatric symptoms.

References:

1. Edenberg, H. J. (2002) The Collaborative Study on the Genetics of Alcoholism: an update. Alcohol Res Health 26, 214-218.,

2. Bierut, LJ, NL Saccone, JP Rice, A Goate, T Foroud, HJ Edenberg, L Almasy, PM Conneally, R Crowe, V Hesselbrock, T-K Li, JI Nurnberger, Jr, B Porjesz, MA Schuckit, J Tischfield, H Begleiter, and T Reich (2002) Defining alcohol-related phenotypes in humans: The Collaborative Study on the Genetics of Alcoholism. Alcohol Res Health 26, 208-213.

3. Edenberg HJ and Foroud T (2006) The genetics of alcoholism: identifying specific genes through family studies. Addiction Biology 11, 386-396.

Acknowlegements: The Collaborative Study on the Genetics of Alcoholism (COGA), Principal Investigators B. Porjesz, V. Hesselbrock, H. Edenberg, L. Bierut, includes ten different centers: University of Connecticut (V. Hesselbrock); Indiana University (H.J. Edenberg, J. Nurnberger Jr., T. Foroud); University of Iowa (S. Kuperman, J. Kramer); SUNY Downstate (B. Porjesz); Washington University in St. Louis (L. Bierut, A. Goate, J. Rice, K. Bucholz); University of California at San Diego (M. Schuckit); Rutgers University (J. Tischfield); Southwest Foundation (L. Almasy), Howard University (R. Taylor) and Virginia Commonwealth University (D. Dick). Other COGA collaborators include: L. Bauer (University of Connecticut); D. Koller, S. O'Connor, L. Wetherill, X. Xuei (Indiana University); Grace Chan (University of Iowa); N. Manz, M. Rangaswamy (SUNY Downstate); A. Hinrichs, J. Rohrbaugh, J-C Wang (Washington University in St. Louis); A. Brooks (Rutgers University); and F. Aliev (Virginia Commonwealth University). A. Parsian and M. Reilly are the NIAAA Staff Collaborators. We continue to be inspired by our memories of Henri Begleiter and Theodore Reich, founding PI and Co-PI of COGA, and also owe a debt of gratitude to other past organizers of COGA, including Ting-Kai Li, currently a consultant with COGA, P. Michael Conneally, Raymond Crowe, and Wendy Reich, for their critical contributions. This national collaborative study is supported by NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA). Funding support for GWAS genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the National Institute on Alcohol Abuse and Alcoholism, the NIH GEI (U01HG004438), and the NIH contract "High throughput genotyping for studying the genetic contributions to human disease" (HHSN268200782096C). The authors thank Kim Doheny and Elizabeth Pugh from CIDR and Justin Paschall from the NCBI dbGaP staff for valuable assistance with genotyping and quality control in developing the dataset available at dbGaP.

Collaborative Genetic Study of Nicotine Dependence (COGEND)

The Collaborative Genetic Study of Nicotine Dependence (COGEND) is a United States multi-site project. Subjects were recruited from St. Louis, Detroit, and Minneapolis through community-based telephone screening to determine eligibility for the study. Cases were required to have current Fagerström Test for Nicotine Dependence (FTND) ≥ 4 and controls were required to have a lifetime maximum FTND of 0 or 1, even during the period of heaviest smoking. The number of cigarettes per day (CPD) was assessed for the period of heaviest smoking as well as for current and other time points; the maximum of these values was used to define the CPD trait for analysis.

For this meta-analysis, COGEND contributed a sample of 2062 unrelated European-Americans (641 nicotine dependent cases and 1011 non-dependent smoking controls). All subjects were smokers and reported smoking ≥ 100 cigarettes lifetime. The study obtained informed consent from participants and approval from the appropriate institutional review boards.

DNA was derived from whole blood maintained by the Rutgers University Cell and DNA Repository following stringent quality control and assurance procedures (www.rucdr.org). Genotyping of the DNA samples was carried out using Perlegen, Illumina GoldenGate, and Sequenom MassArray iPLEX technology. Cleaning procedures have been detailed[1,2]. Briefly, DNA samples with call rates < 90% were dropped; SNPs were required to pass a call rate threshold of 98%.

References:

1. Saccone NL, Saccone SF, Hinrichs AL, Stitzel JA, Duan W, Pergadia ML, Agrawal A, Breslau N, Grucza RA, Hatsukami D, Johnson EO, Madden PAF, Swan GE, Wang JC, Goate AM, Rice JP, Bierut LJ. Multiple distinct risk loci for nicotine dependence identified by dense coverage of the complete family of nicotinic receptor subunit (*CHRN*) genes (2009). *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 150B:453-466.

2. Saccone NL, Wang JC, Breslau N, Johnson EO, Hatsukami D, Saccone SF, Grucza RA, Sun L, Duan W, Budde J, Culverhouse RC, Fox L, Hinrichs AL, Steinbach JH, Wu M, Rice JP, Goate AM, Bierut LJ. The *CHRNA5-CHRNA3-CHRNB4* nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans (2009). *Cancer Research* 69: 6848-6856.

Acknowlegements: This work was supported by the NIH grant 5 P01CA89392 from the National Cancer Institute. Investigators directing data collection: Laura Bierut, Naomi Breslau, Dorothy Hatsukami, and Eric Johnson. Data management is organized by Nancy Saccone and John Rice. Laboratory analyses are led by Alison Goate. Tracey Richmond is in charge of project administration.

Financial Disclosures: Drs. LJ Bierut, AM Goate, AJ Hinrichs, JP Rice, SF Saccone, and JC Wang are listed as inventors on a patent (US 20070258898) covering the use of certain SNPs in determining the diagnosis, prognosis, and treatment of addiction. Dr. Bierut has acted as a consultant for Pfizer, Inc. in 2008.

Danish National Birth Cohort (DNBC):

The Danish National Birth Cohort (DNBC) is a well-established, prospective cohort that enrolled women early in pregnancy, prior to any adverse pregnancy outcomes, to minimize bias in data collection and sampling. See details at: http://www.ssi.dk/English.

The DNBC followed over 100,000 pregnancies beginning in the first trimester and has extensive biological material and epidemiologic data on health outcomes in both mother and child. The current study contains data from a genome-wide case/control study using approximately 1,000 mothers of preterm children from the DNBC most with spontaneous onset of labor or preterm premature rupture of membranes (PPROM), along with 1,000 mothers whose child was born at ~40 weeks' gestation. After data cleaning some small changes in case/control status and other variables resulted in minor changes in numbers of cases or controls in certain categories. Environmental variables are being used as covariates in the analysis. The sample was genotyped on ILLUMINA Human660W-Quad.

This study is part of the Gene Environment Association Studies initiative (GENEVA, http://www.genevastudy.org) funded by the trans-NIH Genes, Environment, and Health Initiative (GEI). The overarching goal is to identify novel genetic factors that contribute to prematurity and its complications through large-scale genome-wide association studies of a well-characterized cohort of Danish mothers and babies. Genotyping was performed at the Johns Hopkins University Center for Inherited Disease Research (CIDR). Data cleaning and harmonization were done at the GEI-funded GENEVA Coordinating Center at the University of Washington.

References:

1. Olsen J, Melbye M, Olsen SF, Sorensen TI, Aaby P, Andersen AM, Taxbol D, Hansen KD, Juhl M, Schow TB, Sorensen HT, Andresen J, Mortensen EL, Olesen AW, Sondergaard C. “The Danish National Birth Cohort- its background, structure and aim” Scand J Public Health 2001; 29: 300-307

Acknowlegements: Support for the Danish National Birth Cohort was obtained from the Danish National Research Foundation, the Danish Pharmacists' Fund, the Egmont Foundation, the March of Dimes Birth Defects Foundation, the Augustinus Foundation and the Health Fund of the Danish Health Insurance Societies. The generation of GWAS genotype data for the DNBC samples was carried out within the GENEVA consortium with funding provided through the NIH Genes, Environment and Health Initiative (GEI) (U01HG004423). Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01HG004446). Genotyping was performed at Johns Hopkins University Center for Inherited Disease Research, with support from the NIH GEI (U01HG004438).

deCODE

Written informed consent was obtained from all subjects in the seven participating populations (Iceland, Spain, The Netherlands, Sweden, Italy, Austria and New Zealand). Inclusion in the study required the availability of genotypes from either GWA studies performed at deCODE genetics or follow-up genotyping of rs1051730 in additional subjects.

References:

1. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008;452(7187):638-42.

Dental Caries: GENEVA Genome Wide Association Study ("Caries GENEVA Study")

The Dental Caries: GENEVA Genome Wide Association Study (Caries GENEVA Study) is a multi-site collaboration that is part of the GENEVA consortium (www.genevastudy.org). Subjects were recruited from Pennsylvania, West Virginia and Iowa through three on-going population-based family cohort studies of factors related to oral health: the Center for Oral Health Research in Appalachia (COHRA; Polk et al 2008, Wang et al, 2010), the Iowa Fluoride Study (IFS; Marshall et al 2007) and the Iowa Bone Development Study (IBDS; Marshall et al 2008). COHRA ascertained families from rural counties in Pennsylvania and West Virginia, the IFS and IBDS ascertained families from primarily central Iowa. The families in each of the cohort studies were not selected based on any phenotype; they were meant to be representative of their respective communities. For this meta-analysis, the Caries GENEVA Study contributed a sample of 639 unrelated European-Americans $\geq$ 25 years of age who had ever smoked. The number of cigarettes per day (CPD) was provided. Each cohort study obtained informed consent from participants and approval from the appropriate institutional review boards.

DNA was extracted from whole blood, saliva samples or buccal samples from each participant. Genotyping was done at the Johns Hopkins Center for Inherited Disease Research (www.cidr.jhmi.edu). Data was released for 4,073 study samples (99.4% of attempted samples), including the 639 included here. Genotyping was performed using Illumina Human610-Quadv1_B BeadChips (Illumina, San Diego, CA, USA) and the Illumina Infinium II assay protocol (Gunderson et al. 2006). General data cleaning procedures have been detailed elsewhere (Laurie et al, 2010). Briefly, DNA samples with call rates < 85% were dropped as were SNPs with more than 1 HapMap replicate error, more than a 2% (autosomal) or 10% (X) difference in call rate between genders, more than 1.8% male AB frequency (X), or more than a 7% (autosomal) or 5% (XY) difference in AB frequency. Self-reported race was verified.

References:

1. Gunderson KL, Steemers FJ, Ren H, Ng P, Zhou L, Tsan C, Chang W, Bullis D, Musmacker J, King C, Lebruska LL, Barker D, Oliphant A, Kuhn KM, Shen R. Whole genome genotyping. Methods Enzymol. 2006;410:359-76.

2. Marshall TA, Eichenberger-Gilmore JM, Broffitt BA, Warren JJ, Levy SM. Dental caries and childhood obesity: roles of diet and socio-economic status. Community Dent Oral Epidemiol, 35(6):449-458, 2007.

3. Marshall TA, Eichenberger-Gilmore JM, Stumbo PJ, Levy SM. Relative validation in children of targeted nutrient and commercial questionnaires: beverage, calcium and vitamin D intakes. J Am Diet Assoc 2008; 108(3):465-72. doi:10.1016/j.jada.2007.12.002

4. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS; for the GENEVA Investigators (2010): "Quality control and quality assurance in genotypic data for genome-wide association studies," Genet Epidemiol. 2010 Aug 17. [Epub ahead of print]

5. Polk DE, Weyant RJ, Crout RJ, McNeil DW, Tarter RE, Thomas JG, Marazita ML. Study protocol of the Center for Oral Health Research in Appalachia (COHRA) etiology study. BMC Oral Health 2008, 8:18 (03Jun2008). PMID: 18522740

6. Wang X, Shaffer JR, Weyant RJ, T.Cuenco K, DeSensi RS, Crout R, McNeil DW, Marazita ML. Genes and their effects on dental caries (tooth decay) may differ between primary and permanent dentitions. Caries Research 44:277-284, 2010.

Acknowlegements: Funding support for the GWAS was provided by the National Institutes of Dental and Craniofacial Research (NIDCR) as part of the trans-NIH Genes, Environment and Health Initiative [GEI] (U01-DE018903). Data and samples were provided by (1) the Center for Oral Health Research in Appalachia, a collaboration of the University of Pittsburgh and West Virginia University funded by NIDCR R01-DE 014899; (2) the Iowa Fluoride Study and the Iowa Bone Development Study, funded by NIDCR R01-DE09551and R01-DE12101, respectively. Genotyping was done by the Johns Hopkins University Center for Inherited Disease Research (CIDR) which is fully funded through a federal contract from the National Institutes of Health (NIH) to The Johns Hopkins University, contract number HHSN268200782096C. Funds for this project's genotyping were provided by the NIDCR through CIDR's NIH contract. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01-HG004446)

EAGLE/PLCO

The 5955 participants derive from EAGLE (3899) and PLCO (2056). All subjects from EAGLE and PLCO were genotyped on the Illumina 550K chip. The GenCall software, part of Illumina's Bead Studio Suite, was used to automatically cluster probe intensity values, call genotypes and assign confidence scores. The overall completion rate was 99.61%. Other quality control included checks of sample heterozygosity, gender check, Hardy-Weinberg proportion, concordance/discordance rates, relatedness check, and assessment of population structure using the STRUCTURE program. See Landi 2009 for details. Some details of the studies follow.

EAGLE (Environment and Genetics in Lung Cancer Etiology study)

EAGLE is a large population-based case-control study designed and conducted to investigate the genetic and environmental determinants of lung cancer and smoking persistence using an integrative approach that allows combined analysis of genetic, environmental, clinical, and behavioral data. (Landi et al, 2007, 2009).

The study includes over 2,101 incident lung cancer cases, both males and females of Italian nationality, ages 35 to 79 years old, with verified lung cancer of any histological type, and over 2,120 healthy population-based controls matched to cases by age, gender, and residence. The participation rate was high: 85% and 73% in cases and controls, respectively. The age distribution of the subjects: 224 (< 50), 283 (51-55), 524 (56-60), 61-65 (752), 66-70 (883), 71-75 (807), and >75 (426).

Lung cancer cases were enrolled from 13 hospitals within the Lombardy region of Italy. The healthy controls were randomly selected from the same residential area of the lung cancer cases. The study setting, the Lombardy region of Italy, is served by a network of modern hospitals, medical schools, and a regional health service. Within the Lombardy region, the catchment's area includes 5 cities (Milan, Monza, Brescia, Pavia, and Varese) 216 surrounding municipalities, encompassing, in the selected age range, over 3.0 million people.

Extensive epidemiological data have been collected through both an interview-based computer-assisted questionnaire and a self-administered questionnaire. Available data includes demographical characteristics, detailed smoking history (active and passive), family history of lung cancer and other cancers, previous lung diseases, medications, diet, alcohol, attempts at quitting smoking, anxiety, depression, personality scores, occupation, reproductive and residential history.

Clinical data (stage, grade, histology, imaging and pathology reports, spirometry, and routine laboratory studies) were recorded. All study subjects donated a blood sample (or, rarely, a buccal rinse sample), which was processed to obtain cryopreserved lymphocytes, RBC, granulocytes, DNA, RNA, whole blood, buffy coat, serum, plasma, and blood cards. Lung tissue paraffin blocks and slides were collected from the cases who underwent surgery, biopsy or cytological examination of the lung tumor. Multiple fresh "normal" lung tissue and tumor samples, frozen in liquid nitrogen within 20 minutes of excision, were also collected from over 436 surgical cases. Paraffin-embedded tissue blocks (656) or tissue slides (1192) are obtained on a substantial subset of the cases.

All data and biospecimen information are stored in a secure relational database. Quality control procedures were implemented to ensure accuracy, completeness, and privacy of the data collected. Epidemiological data and DNA specimens were collected from 98.4% and 97.3% of the cases, respectively. Extensive epidemiological data was collected through both a Computer Assisted Personal interview (CAPI) and a self-administered questionnaire.

Data on tobacco smoking included: information on number of cigarettes and other tobacco products per day, averaged over each smoking period of life and during the last year, age at first cigarette, at initiation (i.e., at least once per week) and quitting; the number of quitting attempts and time between attempts, inhalation habits, passive smoking during childhood, at

home and in the workplace, self-reported willingness to quit smoking. Smoking status (classification into never, ever and current smoking status) was established by review of smoking data. Ever smokers all had 'smoked greater than 100 cigarettes during their lifetime' with a frequency of one or more cigarettes per week, establishing their status as smokers. Former smokers had indicated that in addition 'during the last 6 months' they had not been smoking, i.e. (not at all or less than one cigarette per week'). Current smokers, in addition to their status as smokers, indicated that during the last six months they smoked at least one cigarette per week. This smoking information was cross validated through checks for concordance with the other smoking information listed above.

The Prostate, Lung, Colon, Ovary Clinical Trial (PLCO)

The Prostate, Lung, Colon, Ovary Clinical Trial randomized 150,000 individuals aged 55-74 years from 10 US study centers between 1992-2001 to undergo a specific cancer screening regime or receive routine medical care to evaluate the effects of screening on disease-specific mortality (Prorok CP et al, 2000).

Study participants from PLCO include 2056 from the lung cancer study comprised of 1174 males and 882 females. The age distribution of the participants was 51-54 (3), 56-60 (65), 61-65 (255), 66-70 (485), 71-75 (579), >75 (669). Smoking behaviors were measured by baseline questionnaire (BQ) in the PLCO (administered from 1994-2001). Former smokers were defined as ever-smokers who did not smoke regularly at BQ and were asked to report the age at which they stopped smoking regularly. Ever smokers were asked to provide information on the number of cigarettes they smoked per day, in categories (1-10, 11-20, 21-30, 31-40, 41-60, 61-80, over 80). For continuous analyses we assigned subjects to the midpoint of their category (or 90 cigarettes per day for over 80). Smoking status was revisited in a follow-up Supplemental Questionnaire (SQX) conducted between April 2006 and May 2007. Among 134,992 eligible not deceased PLCO participants, 103,643 forms were returned, keyed, and processed. In subjects from the intervention arm (all participants in the current study derive from this group) the never, former and current categories assessed in the original questionnaire remained very stable. For example, among 21,111 former smokers, only 571 'relapsed' (2.7%), returned to current smoking status. 485 (2.3%) relapsed but indicated that they had now quit again.

References:

1. Landi MT, Consonni D, Rotunno M, Bergen AW, Goldstein AM, Lubin JH et al. Environment and Genetics in lung Cancer Etiology (EAGLE) study: An integrative population-based case-control study of lung cancer. BMC Public Health. 2008; 8:203
2. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, Bergen AW, Li Q, Consonni D, Pesatori AC, Wacholder S, Thun M, Diver R, Oken M, Virtamo J, Albanes D, Wang Z, Burdette L, Doheny KF, Pugh EW, Laurie C, Brennan P, Hung R, Gaborieau V, McKay JD, Lathrop M, McLaughlin J, Wang Y, Tsao MS, Spitz MR, Wang Y, Krokan H, Vatten L, Skorpen F, Arnesen E, Benhamou S, Bouchard C, Metsapalu A, Vooder T, Nelis M, Välk K, Field JK, Chen C, Goodman G, Sulem P, Thorleifsson G, Rafnar T, Eisen T, Sauter W, Rosenberger A, Bickeböller H, Risch A, Chang-Claude J, Wichmann HE, Stefansson K, Houlston R, Amos CI, Fraumeni JF Jr, Savage SA, Bertazzi PA, Tucker MA, Chanock S, Caporaso NE. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet. 2009 Nov;85(5):679-91.

3. Prorok PC, Andriole GL, Bresalier RS, Buys SS, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Controlled Clinical Trials 21: 273S-309S, 2000.

FINRISK Study

Description of the study

The National FINRISK Study is a population risk factor survey on non-communicable diseases carried out in Finland every fifth year since 19721. For every survey round an age and sex stratified random sample has been drawn from the population register. The sample size for each survey round has varied between 8000 and13500. The survey includes a self-administered questionnaire, physical measurements carried out by trained survey nurses and a draw of blood samples. For this meta-analysis, the National FINRISK Study data since 1992, including DNA sample collection, have been used including surveys carried out in 1992, 1997, 2002 and 2007.

Smoking info from the study

Altogether a sample of 7864 current daily smokers was available out of a total of 32070 subjects, of which 26980 were genotyped for the SNPs. However, due to missing data on cigarettes per day (CPD) for 138 participants, the valid sample for phenotype analyses included 7726 smokers (42% women). All participants with CPD information were current daily smokers, 41% (n=3194) reported smoking of ≤10 CPD, 46% (n=3532) of >10 to 20 CPD, 10% (n=781) more than >20 to 30 CPD, and 3% (n=219) more than 30 CPD.

Age of onset of regular smoking information was from 7782 participants ranging from 6 years to 60 years (mean 18.3 years, SD 5.3). Forty two percent of them had started regular smoking at 16 years old or younger.

Informed consent/IRB approval was current

All FINRISK surveys have been approved by the appropriate ethics committees. An informed consent has been received from all participants.

How DNA was obtained

DNA was derived from whole blood samples. Samples were transferred to the laboratory of molecular genetics in the National Institute of Health and Welfare (earlier National Public Health Institute, KTL), where the DNA was extracted. Genotyping of the DNA samples was carried out using iPlex assay on the MassARRAY System (Sequenom, San Diego, CA, USA) standard protocols. DNA samples were missing/not given for some participants.

SNP info

The SNPs included in the analyses, rs16969968, rs578776, rs588765 andrs6265all had genotyping success of nearly 98% (37--51 missing genotypes per SNP) and Hardy-Weinberg equilibrium test p-values of 0.59, 0.02, 0.15, and 0.17respectively. The minor (MAF) and major alleles were A (32%) and G ; T (32%) and C ; T (37%) and C ;and A (15%) and G , respectively. rs16969968_A (n=6584), rs578776_T (n=6587), rs588765_T (n=6578), rs6265_A (n=6573)

Variable definitions

Age varied 25-74 years (mean age 44.6 years, SD 12.0) at the time of assessment. Participants were born between 1923 and 1982, and most of them (90%) between 1923 and 1972.

Educational attainment information was from 7650 participants with CPD data; 53% had terminal degree of high school or less while 47% had terminal degree greater than high school.

Age of onset of regular smoking information was from 7542 participants (mean 21.6 years, SD 12.1); 39.6% of them had started regular smoking at 16 years old or younger.

Sex and birth year were available for all subjects from the Central Population Register of Finland.
Age of onset of regular smoking information was from 7782 participants ranging from 6 years to 60 years (mean 18.3 years, SD 5.3). Forty two percent of them had started regular smoking at 16 years old or younger.

References

1. Vartiainen E, Laatikainen T, Peltonen M, Juolevi A, Männistö S, Sundvall J, Jousilahti P, Salomaa V, Valsta L, Puska P. Thirty-five-year trends in cardiovascular risk factors in Finland. Int J Epidemiol 2010;39(2):504-18.

GenMetS

Subjects were drawn from a Health2000 study that includes 8028 subjects aged 30 or over and is a nationally representative sample of the adult Finnish population1. GenMetS is a subcohort of 2124 individuals selected for a case-control genome-wide association study on metabolic syndrome2; 918 cases were selected according the International Diabetes Federation Worldwide Definition of the Metabolic Syndrome and 1206 controls were selected for not carrying the trait. The subjects participated in a health interview conducted by Statistic Finland's interview staff at the home of the participants. During the interview respondents were handed an information leaflet and an informed consent form that was returned after signing. The interview extensively examined factors influencing health, including a one-page questionnaire on smoking behavior. The number of cigarettes smoked daily (CPD) was queried by a question "How much do you daily smoke currently or did prior to quitting?" The respondent was asked to indicate the number of cigarettes smoked as an open-ended question. The genotyped sample included 1134 smokers who answered yes to the question "Have you smoked at least 100 times during your lifetime?". Current daily smokers were identified by asking whether they smoked daily or almost daily.

DNA was extracted from a venous blood sample and genotyped by the Illumina 610 Quad V1 BeadChip at the Sanger Wellcome Trust Institute. This chip provides whole-genome SNP genotyping information with 598,203 SNP markers. The SNPs and samples

were screened for the following: SNP clustering probability for each genotype > 95%, Call rate > 95% for both individuals and markers, MAF > 1%, and HWE p > 1*10-6. In addition, heterozygosity, gender checks and relatedness checks were performed and any discrepancies were removed. The SNPs included in the analyses, rs8034191, rs578776, rs621849, and rs8192475, all had genotyping success of nearly 100% (0-3 missing genotypes per SNP) and Hardy-Weinberg equilibrium test p-values of 0.41, 0.11, 0.43, and 0.63, respectively. The minor (MAF) and major alleles were G (34%) and A; A (32%) and G; G (37%) and A; and A (2.4%) and G, respectively.

Sex and birth year were available for all subjects from the Central Population Register of Finland.

References Cited:

1. Aromaa A, Koskinen,S. (eds) (2004) Health and functional capacity in Finland. Publications of the National Public Health Institute, KTL B12: Helsinki, Finland. http://www.terveys2000.fi/julkaisut/baseline.pdf

2. Keskitalo K, Broms U, Heliövaara M, Ripatti S, Surakka I, Perola M, Pitkäniemi J, Peltonen L, Aromaa A, Kaprio J. (2009) Association of serum cotinine level with a cluster of three nicotinic acetylcholine receptor genes (CHRNA3/CHRNA5/CHRNB4) on chromosome 15. Human Molecular Genetics 18:4007-12.

Acknowlegements

The GenMetS study would like to acknowledge the significant contributions of Leena Peltonen to the study.

University of Maryland: The Genetics of Early Onset Stroke (GEOS) Study:

The Genetics of Early Onset Stroke (GEOS) study is a population-based case-control study designed to identify the genetic determinants of ischemic stroke. Subjects were recruited from the greater Baltimore-Washington area between 1992 and 2008. Cases were defined as young adults 15-49 with first-ever ischemic stroke identified through discharge surveillance from 1 of 59 participating hospitals and direct physician referral. Abstracted medical records were reviewed and adjudicated for ischemic stroke subtype by two neurologists according to previously published procedures[1,2], with discrepancies resolved by a third neurologist. The ischemic stroke subtype classification system retains information on all probable and possible causes, and is reducible to the more widely used TOAST[3] system that assigns each case to a single category. Controls had no history of ischemic stroke and were identified through random digit dialing. Controls were matched to cases based on sex, age, race, and geographic location.

Ischemic strokes with the following characteristics were excluded from participation: stroke occurring as an immediate consequence of trauma; stroke within 48 hours after a hospital procedure, stroke within 60 days after the onset of a nontraumatic subarachnoid hemorrhage, and cerebral venous thrombosis. Additional exclusions for these genetic analyses were known single-gene or mitochondrial disorder recognized by a distinctive phenotype (e.g., cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), mitochondrial encephalopathy with lactic acidosis and stroke-like episodes (MELAS), homocystinuria, Fabry disease, or sickle cell anemia); mechanical aortic or mitral valve at the time of index stroke; untreated or actively treated bacterial endocarditis at the time of the index stroke; neurosyphilis or other CNS infections; neurosarcoidosis; severe sepsis with hypotension at the time of the index stroke; cerebral vasculitis by angiogram and clinical criteria; post-radiation arteriopathy; left atrial myxoma; major congenital heart disease; and cocaine use in the 48 hours prior to their stroke. This list is based on published proposals for exclusion criteria for genetic studies of ischemic stroke[4] but includes additional exclusions based on the experience of phenotyping a large number of ischemic strokes in young adults.

All participants were asked about their smoking history. Ever smokers were defined as individuals having smoked more than 100 cigarettes in their lifetime. Ever smokers were asked questions about the number of cigarettes-per-day (CPD) in the thirty days prior to their stroke (cases) or interview (controls) and the number of CPD during their year(s) as a smoker; the maximum of these values was used to define the CPD trait for the analysis.

For this meta-analysis, GEOS contributed a sample of 497 unrelated European-Americans (277 ischemic stroke cases and 220 controls) and 391 unrelated African-Americans (216 ischemic stroke cases and 175 controls). All subjects are smokers and report smoking ≥100 cigarettes in their lifetime. For the CPD phenotype analyzed here, GEOS contributed 374 smokers with CPD ≤ 10 cigarettes, 357 with 11 ≤ CPD ≤ 20, 84 with 21 ≤ CPD ≤30, and 73 with CPD ≥ 31. The study obtained written informed consent from all participants and approval from the appropriate institutional review boards.

DNA was derived from whole blood (n=389), mouthwash samples (n=6), cell lines (n=481), and WGA (n=12). Genotyping was performed using the Illumina Omni 1-Quad 1M beadchip. Genotype cleaning included removal of unexpected duplicate samples, samples that were unexpectedly related and gender mismatch samples. The sample call rate ranged from 98.4% to 99.9%. SNPs were

required to have a minor allele frequency > 0.01, call rate ≥ 98% and Hardy Weinberg Equilibrium p-value ≥ 0.01. No SNP imputation was performed.

This analysis included the following variables: gender, cigarettes per day (maximum of “What was the average number of cigarettes per day you smoked 30 days prior to your stroke?” or “When you smoke(d), what is (was) the average number of cigarettes you smoked per day?”) age of onset ("How old were you when you first started to smoke cigarettes?"), educational attainment ("Including grade school, high school, and college, business, vocational, professional, and postgraduate schooling, what is the highest grade or year of school you have completed?").

References:

1. Johnson CJ, Kittner SJ, McCarter RJ, Sloan MA, Stern BJ, Buchholz D, et al. Interrater reliability of an etiologic classification of ischemic stroke. Stroke 1995;26:46-51.

2. Kittner SJ, Stern BJ, Wozniak M, Buchholz DW, Earley CJ, Feeser BR, et al. Cerebral infarction in young adults: the Baltimore-Washington Cooperative Young Stroke Study. Neurology 1998;50:890-4.

3. Adams HP, Jr., Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke 1993;24:35-41.

4. Meschia,JF, Brown,RD, Jr., Brott,TG, Chukwudelunzu,FE, Hardy,J, Rich,SS: The Siblings With Ischemic Stroke Study (SWISS) protocol. BMC.Med.Genet 3:1, 2002

HGF-Study

The HGF GWA study (center 1: Wichmann, Sauter, center 2: Bickeböller, Rosenberger, center 3: Risch, Chang-Claude) was made up of three independent German studies as detailed below: In total 506 incident lung cancer cases (LUCY-study: n=305, Heidelberg lung cancer case-control study: n=201) were compared to 480 popula-tion controls (KORA surveys KORA). After excluding never smokers, individuals with missing values, non-Caucasians and related individuals 496 cases (LUCY-study: n=287, Heidelberg lung cancer case-control study: n=182) and 260 controls entered the data analysis for the GEMINI project.

1. LUCY-study (Helmholtz Zentrum Muenchen, PIs Wichmann, Bickeböller).

LUCY (LUng Cancer in the Young) is an multicenter study with 31 recruiting hospitals in Germany. The study is conducted by the Institute of Epidemiology, Helmholtz Zentrum Muenchen, and the Department of Genetic Epidemiology, Medical School, University of Göttingen). The LUCY-study provides access to a nation wide, population based family and a case-control sample (control population KORA, described below) of lung cancer patients aged 50 years or younger at diagnosis. Detailed epidemiologic data have been collected including data on medical history, education, family history of cancer and smoking exposure by phase assessment. Blood samples are taken and DNA and lymphoblastoid cell lines are prepared of all cases and controls and of parts of the relatives. Phenotype data of 847 young patients with primary lung cancer and 5524 relatives have been collected. LUCY is a member of the International Lung Cancer COnsortium (ILCCO), which aims sharing comparable data from ongoing lung cancer case-control and cohort studies to achieve greater power, especially for subgroup analyses.

References:

1. Rosenberger A, Illig T, Korb K, Klopp N, Zietemann V, Wölke G, Meese E, Sybrecht G, Kronenberg F, Cebulla M, Degen M, Drings P, Gröschel A, Konietzko N, grosse Kreymborg K, Häußinger K, Höffken G, Jilge B, Ko JD, Morr H, Schmidt C, Schmidt EW, Täuscher D, Bickeböller H, Wichmann HE (2008), "Do genetic factors protect for early onset lung cancer? – A case-control study before the age of 50 years", BMC Cancer, 8, 60.

2. Sauter W, Rosenberger A, Beckmann L, Kropp S, Mittelstrass K, Timofeeva M, Wölke G, Steinwachs A, Scheiner D, Meese E, Sybrecht G, Kronenberg F, Dienemann H; LUCY-Consortium, Chang-Claude J, Illig T, Wichmann HE, Bickeböller H, Risch A. Matrix Metalloproteinase 1 (MMP1) Is Associated with Early-Onset Lung Cancer. Cancer Epidemiol Biomarkers Prev. 2008 May;17(5):1127-35.

Funding: This work was funded in part by the National Genome Research Network (NGFN), the DFG (BI 576/2-1; BI 576/2–2), the Helmholtzgemeinschaft (HGF) and the Federal office for Radiation Protection (BfS: STSch4454). Genotyping was performed in the Genome Analysis Center (GAC) of the Helmholtz Zentrum Muenchen.

2. Heidelberg lung cancer case-control study (German Cancer Research Center, PI Risch).

As part of an ongoing hospital based case-control study, the DKFZ has recruited over 2000 LC cases at and in collaboration with the Thoraxklinik Heidelberg, including 300 LC cases with onset of disease at the age of ≤ 50. Approximately 750 hospital-based controls have also been recruited. Data on occupational exposure, tobacco smoking, educational status, and for a subgroup also on family history of lung cancer, assessed by a self-administered questionnaire is available. Blood samples have been taken, and DNA has been extracted. This study also is a member of ILCCO.

References:

1. Dally H, Edler L, Jager B, et al. The CYP3A4*1B allele increases risk for small cell lung cancer: effect of gender and smoking dose. Pharmacogenetics 2003;13:607 – 18.

2. Sauter W, Rosenberger A, Beckmann L, Kropp S, Mittelstrass K, Timofeeva M, Wölke G, Steinwachs A, Scheiner D, Meese E, Sybrecht G, Kronenberg F, Dienemann H; LUCY-Consortium, Chang-Claude J, Illig T, Wichmann HE, Bickeböller H, Risch A. Matrix Metalloproteinase 1 (MMP1) Is Associated with Early-Onset Lung Cancer. Cancer Epidemiol Biomarkers Prev. 2008 May;17(5):1127-35

Funding: This work was in part supported by grant 70-2919 from the Deutsche Krebshilfe and a Helmholtz-DAAD fellowship (A/07/97379) to Maria Timofeeva.

3. KORA surveys (PI Wichmann) – "sample of population controls"

The Helmholtz Center Munich has established the population-based KORA platform (Coop-erative health research in the Region of Augsburg In total, four population based health surveys have been conducted during 1984/85-1999/2001 with altogether 18000 participants in the age range between of 25 and 74 years at first recruitment with several phenotypical, medical, laboratory and interview data as well as DNA from 16000 probands are available. Furthermore EBV immortalized cell lines have been established from 1600 subjects. KORA is a well established population, genetically representative for the Caucasian German population. It was already used, especially for GWA studies, in several national and international networks. Major population stratification between KORA (Southern Germany) and two other cohorts from Northern Germany could not be detected.

References:

1. Wichmann HE, Gieger C, Illig T, MONICA/KORA Study Group: KORA-gen - resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 2005, 67 (Suppl 1):S26-S30.

Funding: The KORA Surveys were financed by the GSF, which is funded by the German Federal Ministry of Education, Science, Research and Technology and the State of Bavaria.

Acknowlegements: We thank Dr. M. Timofeeva, Dr. H. Dally, Mrs. A. Seidel, Dr. T. Muley, Dr. S. Thiel, Dr. H. Wikman, Ms. U. von Seydlitz-Kurzbach, Ms. D. Bodemer, Dr. C. Klappenecker, Mr. M. Hoffmann and Mr. C. Beynon for help with sample and/or data collection and archiving for the Heidelberg lung study. We are grateful to all patients and staff at the Thoraxklinik Heidelberg, who participated in the Heidelberg lung cancer study and particularly Prof. Peter Drings and Prof. Hendrik Dienemann for making it possible.

We gratefully acknowledge the KORA study group especially G. Fischer, H. Grallert, N. Klopp, C. Gieger, R. Holle, C. Hanrieder and A. Steinwachs, Institute of Epidemiology, Helmholtz Centre Munich, Neuherberg, Germany and all individu-als, who participated as cases or controls in this study and the KORA Study Center and their co-workers for organizing and conducting the data collection.

We also gratefully acknowledge the LUCY-consortium especially G. Wölke and V. Zietemann for coordinating recruitment, all patients and staff of the participating hospitals: Aurich (Dr. Heidi Kleen); Bad Berka (Dr. med. R. Bonnet, Klinik für Pneumologie, Zentralklinik Bad Berka GmbH); Bonn (Prof. Ko, Dr. Geisen, Innere Medizin I, Johanniter Krankenhaus); Bonn (Dr. Stier, Medizinische Poliklinik, Universität Bonn); Bremen (Prof. Dr. D. Ukena, Dr. Penzl, Zentralkrankenhaus Bremen Ost, Pneumologische Klinik); Chemnitz (PD Dr. Schmidt, OA Dr. (Jielge, Klinikum Chemnitz, Abteilung Innere Medizin); Coswig (Prof. Höffken, Dr. Schmidt, Fachkrankenhaus Coswig); Diekholzen (Dr. Hamm, Kreiskrankenhaus Diekholzen, Klinik für Pneumologie); Donaustauf (Prof. Pfeifer, Dr. v. Bültzingslöwen, Fr. Schneider, Fachklinik für Atemwegserkrankungen); Essen (Prof. Teschler, Dr. Fischer, Ruhrlandklinik - Universitätsklinik, Abt. Pneumologie); Gauting (Prof. Häußinger, Prof. Thetter, Dr. Düll, Dr. Wagner, Pneumologische Klinik München-Gauting); Gera (CA MR Dr. Heil, OÄ Dr. Täuscher, OA Dr. Lange, II. Medizinische Klinik, Wald-Klinikum Gera); Göttingen (Prof. Trümper, Prof. Griesinger, Dr. Overbeck, Abteilung Onkologie, Hämatologie); Göttingen (Prof.

Schöndube, Dr. Danner, Abteilung Thorax-, Herz-, und Gefäßchirurgie); Göttingen/Weende (Dr. med. Fleischer, Ev. Krankenhaus Göttingen-Weende e.V., Abteilung Allgemeinchirurgie); Greifenstein (Prof. Morr, Dr. M. Degen, Dr. Matter, Pneumologische Klinik, Waldhof Elgershausen); Greifswald (Prof. Ewert, Dr. Altesellmeier, Universitätsklinik Greifswald, Klinik für Innere Medizin B); Hannover (Prof. Schönhofer, Dr. Kohlaußen, Klinikum Hannover Oststadt, Medizinische KlinikII, Pneumologie); Heidelberg (Prof. Drings, Dr. Herrmann, Thoraxklinik-Heidelberg GmbH, Abt. Innere Medizin-Onkologie); Hildesheim (Prof. Kaiser, St. Bernward Krankenhaus, Medizinische Klinik II); Homburg (Prof. Sybrecht, OA Dr. Gröschel, Dr. Mack, Uniklinik des Saarlandes, Innere Medizin V); Immenhausen (Prof. Andreas, Dr. Rittmeyer, Fachklinik für Lungenerkrankungen); Köln (Priv. Doz. Dr. Stölben, Kliniken der Stadt Köln, Lungenklinik Krankenhaus Merheim); Köln (Prof. Wolf, Dr. Staratschek-Jox, Klinikum der Universität Köln, Klinik I für Innere Medizin); Leipzig (Prof. Gillisen, OA Dr. Cebulla, Städt. Klinikum St. Georg, Robert-Koch-Klinik); Leipzig (Kreymborg, Universitätsklinikum Leipzig, Medizinische Klinik I, Abteilung Pneumologie); Lenglern (Prof. Criée, Dr. Körber, Dr. Knaack, Ev. Krankenhaus Weende e.V., Standort Lenglern, Abt. Pneumologie; München (Prof. Huber, Dr. Borgmeier, Klinikum der LMU-Innenstadt, Abt. Pneumologie); Neustadt a. Harz (Dr. Keppler, Schäfer, Evangelisches Fachkrankenhaus für Atemwegserkrankungen); Rotenburg (Prof. Schaberg, Dr. Struß, Diakoniekrankenhaus Rotenburg, Lungenklinik Unterstedt); St. Pölten – Österreich (OA Dr. M. Wiesholzer, Zentralklinikum St. Pölten, I. Medizinische Kli-nik)

Family Health Study and the Women's Epidemiology of Lung Disease Study (KCI-WSU):

Data are from two population-based, case-control studies of lung cancer detailed in a recent publication: the Family Health Study (FHS studies I, II and III) and the Women's Epidemiology of Lung Disease (WELD) Study[1]. Only Caucasian subjects with DNA from sources other than tissue blocks were included in these analyses. All studies were conducted by the same study staff using identical procedures, with cases ascertained through the population-based Metropolitan Detroit Cancer Surveillance System, an NCI-funded SEER registry. Studies differed only in the eligibility of cases, with the FHS focused on never smokers and cases diagnosed before age 50 years and the WELD study focusing on women. Only non small cell lung cancer (NSCLC) histology cases were included in the WELD study. Population-based controls were chosen using random digit dialing methods. All study controls were frequency matched to cases by 5-year age group, sex and race. Institutional Review Board approval was obtained for all studies, and informed consent was obtained from all participants.

Individuals who had smoked at least 100 cigarettes in their lifetime were designated as smokers. These subjects were also asked for the average number of cigarettes per day they smoked and the total number of smoking years. All subjects were asked if they had ever been diagnosed by a physician as having emphysema, chronic obstructive pulmonary disease (COPD) or chronic bronchitis. Individuals reporting one or more of these conditions were considered to have COPD.

DNA was extracted from whole blood or buccal cells (buccal swab or mouthwash sample). DNA was isolated from blood using a Gentra AutoPure Kit (Qiagen, Valencia, CA), buccal swabs with the BuccalAmp DNA Extraction Kit (Epicentre Technologies, Madison, WI) and mouthwash samples with the Gentra Puregene Kit (Qiagen). TaqMan Genotyping Assays (Applied Biosystems, Foster City, CA) were used to detect polymorphisms. DNA isolated from buccal cells was pre-amplified in an outer PCR reaction for added sensitivity. Either 25 ng DNA or 1 µl of the outer nest was amplified, with primers designed using Primer Express software (Applied Biosystems), and detected using an AB 7900 Sequence Detection System (Applied Biosystems). For quality control, 5% of the products were sequenced and 10% were directly repeated.

References:

1. Schwartz AG, Cote ML, Wenzlaff AS, Amos, CI. (2009) Racial differences in the association between SNPs on 15q25.1, smoking behavior, and risk of non-small cell lung cancer *J Thorac Oncol* 4(10):1195-201.

KORCULA study:

The KORCULA study included healthy volunteers aged 18 and over from the villages of Lumbarda, Žrnovo, and Račišće on the Island of Korcula, Croatia[1]. A comprehensive set of phenotypic measurements was performed for each subject, further supplemented with a number of traits measured from plasma and serum. Smoking information was obtained from a survey designed for the purpose of reporting general lifestyle and behavior. Ethical approval was obtained from the Committees of the Medical School of the University of Zagreb and Medical School of the University of Split. DNA extraction was made from the leukocyte samples, using Nucleon kits. Samples were genotyped with Illumina CNV370 panel, imputed to 2.4 million SNPs.

References:

1. Rudan I, Marusić A, Janković S, Rotim K, Boban M, Lauc G, et al. "10001 Dalmatians:" Croatia launches its national biobank. Croat Med J. 2009 Feb;50(1):4-6.

Lung Health Study (LHS), and LHS-Utah Subsample (LHS-Utah):

The Lung Health Study (LHS) cohort was drawn from a multi-site longitudinal study of COPD sponsored by the Division of Lung Disease of the National Heart, Lung and Blood Institute[2]. All LHS participants had COPD as determined by pulmonary function testing, and all were smoking at the time of recruitment. All participants were of European descent, and all had smoked more than 100 cigarettes lifetime. Cigarettes per day (CPD) was based on period of heaviest smoking lifetime. Study procedures were approved by the local IRB.

DNA was isolated from peripheral blood lymphocytes collected by the LHS Study Investigators supported by the NHLBI[2]. SNP genotyping methods were previously described[1] and used either the SNPlex assay (Applied Biosystems) or TaqMan assay (Applied Biosystems). The call rates were 100% in the LHS cohort. SNPs genotyped by both the TaqMan and SNPlex methods in 236 individuals had a concordance rate > 99.7%.

References:

1. Weiss RB, Baker TB, Cannon DS, von Niederhausern A, Dunn DM, Matsunami N, et al. (2008). A candidate gene approach identifies the CHRNA5-A3-B4 region as a risk factor for age-dependent nicotine addiction. *PLoS Genet,* 4(7), e1000125.
2. Anthonisen NR, Connett JE, Kiley JP, Altose MD, Bailey WC, Buist AS, et al. (1994). Effects of smoking intervention and the use of an inhaled anticholinergic bronchodilator on the rate of decline of FEV1. The Lung Health Study. *JAMA,* 272(19), 1497-1505.

LOLIPOP:

LOLIPOP is an ongoing population based study of ~30 000 Indian Asian and European white men and women recruited from the lists of 58 general practitioners in west London. Assessment of participants was performed by a trained nurse using a standard protocol including questions on medical history, family history, cardiovascular risk factors, alcohol intake, physical activity and drug history (verified from the practice computerised records). Subsequently 2293 Indian Asian and European white subjects, aged 35–74 years and free from clinical CVD, were selected at random and enrolled into the LOLIPOP atherosclerosis cohort substudy. Participants were defined as Indian Asian if all four grandparents were born in the Indian subcontinent (India, Pakistan or Bangladesh) and European white if all four grandparents were born in northern Europe.

References:

1. Chahal, N.S. *et al*. Ethnicity-related differences in left ventricular function, structure and geometry: a population study of UK Indian Asians and European whites. Heart 96, 466–471 (2009).

MD Anderson Lung Cancer (MDACC-LCA):

Study subjects (all self-reported Caucasians and African Americans) from the U.T. M.D. Anderson Cancer Center (MDACC) are derived from a lung cancer case-control study that has been ongoing since 1991[1]. Cases were newly diagnosed, histologically-confirmed patients presenting at M.D. Anderson Cancer Center with the diagnosis of non-small cell lung cancer and who had not previously received treatment other than surgery. Controls were healthy individuals seen for routine care at Kelsey-Seybold Clinics; the largest physician group-practice plan in the Houston Metropolitan area[2]. Controls were frequency matched to cases according to their smoking behavior, age in 5 year categories, ethnicity, and sex. Former smoking controls were further frequency matched to former smoking cases according to the number of years since smoking cessation (in 5 year categories). The study protocols were approved by the Institutional Review Board of the U.T. M.D. Anderson Cancer Center. Informed consent was obtained from all patients. Epidemiologic data including smoking status were collected during an in-person interview using a structured questionnaire. Genomic DNA was extracted from peripheral blood samples using the Human Whole Blood Genomic DNA Extraction Kit (Qiagen, Valencia, CA). Genotypes were generated by the Center for Inherited Disease Research for 317,498 polymorphic tagging SNPs using Illumina HumanHap300 v1.1 BeadChips and the Illumina Infinium II assay[3]. For this meta-analysis, MDACC contributed a sample of 2291 unrelated European-Americans (including 1136 cases and 250 controls). Ever smokers were defined as those who smoked more than 100 cigarettes over their lifetime. Former smokers had quit a year before diagnosis (cases) or interview (controls).

References:

1. Spitz MR, Wei Q, Dong Q, Amos CI, Wu X (2003) Genetic susceptibility to lung cancer: the role of DNA damage and repair. *Cancer Epidemiol. Biomarkers Prev*. 12, 689-698.

2. Hudmon KS, et al. (1997) Identifying and recruiting healthy control subjects from a managed care organization: a methodology for molecular epidemiological case-control studies of cancer. *Cancer Epidemiol. Biomarkers Prev*. 6, 565-571.

3. Amos CI, Wu X, Broderick P, et al. (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet.* 40:616–622.

4. Amos CI, Gorlov IP, Dong Q, Wu X, Zhang H, Lu EY, Scheet P, Greisinger AJ, Mills GB, Spitz MR. Nicotinic acetylcholine receptor region on chromosome 15q25 and lung cancer risk among African Americans: a case-control study. J Natl Cancer Inst. 2010 Aug 4;102(15):1199-205.

MD Anderson Melanoma (MDACC-Melanoma):

This research builds upon an extensive resource of melanoma cases and hospital based controls collected over several years at the U.T. M.D. Anderson Cancer Center. The goal of this research is to identify novel susceptibility and outcome-related genes for melanoma using a systematic genome-wide association-based approach. Our goal is to conduct high-density SNP association and outcome studies. This dbGaP study contains samples from 2000 European ancestry cases and 1000 European ancestry controls using the Illumina OMNI1-Quad SNP chip. As a part of an ongoing R01 project, we have epidemiological data together with candidate gene results for 1000 of the melanoma cases and the controls. With regard to the outcome aspect of our design, as part of our melanoma Specialized Program of Research Excellence (SPORE) grant, our MelCore database contains comprehensive, prospectively maintained clinical information from all melanoma patients included in the study cohort, including primary tumor histopathology and staging information, standard and investigational blood tumor markers, details of surgical and systemic therapies, and extensive follow-up information, including time to relapse or recurrence, pattern of recurrence and survival duration. Finally, we intend to collaborate with the GenoMEL collaboration so we can jointly evaluate each other's findings. The goal of our analysis will be to identify novel genetic factors predisposing the development of melanoma, as well as genetic factors controlling melanoma stage at presentation, recurrence and progression.

This study is part of the Gene Environment Association Studies initiative (GENEVA, http://www.genevastudy.org) funded by the trans-NIH Genes, Environment, and Health Initiative (GEI). The overarching goal is to identify novel genetic factors that contribute to melanoma through large-scale genome-wide association studies of 2000 European ancestry cases and 1000 European ancestry controls. Genotyping was performed at the Johns Hopkins University Center for Inherited Disease Research (CIDR). Data cleaning and harmonization were done at the GEI-funded GENEVA Coordinating Center at the University of Washington.

References:

1. Li C, Liu Z, Wang LE, Gershenwald JE, Lee JE, Prieto VG, Duvic M, Grimm EA, Wei Q. Haplotype and genotypes of the VDR gene and cutaneous melanoma risk in non-Hispanic whites in Texas: a case-control study. Int J Cancer. 2008 May 1; 122(9):2077-84.

2. Li C, Zhao H, Hu Z, Liu Z, Wang LE, Gershenwald JE, Prieto VG, Lee JE, Duvic M, Grimm EA, Wei Q. Genetic variants and haplotypes of the caspase-8 and caspase-10 genes contribute to susceptibility to cutaneous melanoma. Hum Mutat. 2008 Dec; 29(12):1443-51.

Molecular Genetics of Schizophrenia (MGS) Collection:

Ascertainment, consent, assessment, phlebotomy, diagnosis, ancestry, sample and genotypic quality control (QC), and sharing of biomaterials and data for the Molecular Genetics of Schizophrenia (MGS) case-control sample have been previously described in detail [1-3]. European ancestry (EA) and African American (AA) samples (both non-Hispanic) were collected, although for this study only subjects of European ancestry were used. Most DNA specimens (~87%) were extracted from Epstein-Barr virus (EBV) transformed lymphoblastic cell lines (LCLs), but later in the study, the NIMH repository extracted some DNAs (~13%) from whole blood (primarily for cases for whom fewer access requests were expected). For the MGS genome wide association study (GWAS) of schizophrenia, we genotyped 2,838 EA cases and 2,817 EA control samples with the Affymetrix 6.0 array, and 95% case and control samples passed stringent sample and genotypic quality control (QC) [2], briefly described below. Samples were excluded for high missing data rates (autosomal genotyping call rate less than 97%), outlier proportions of heterozygous genotypes (approach varied for EA versus AA; see [2] for details), incorrect sex, or unexpected (cryptic) genotypic relatedness to other subjects (duplicates detected by identity-by-descent analyses, relatives detected by identity-by-state analyses; thus the resultant sample consisted of unique unrelated cases and controls). SNPs were excluded for minor allele frequencies less than 1%, high missing data rates (call rate <95%), Hardy-Weinberg equilibrium (HWE) deviation (HWE $p<10^{-6}$ in controls), excessive Mendelian errors (more than 2 in the QC trios), discordant genotypes (more than 1 in QC duplicate samples), or large allele frequency differences among DNA plates. Principal component scores reflecting continental and within-Europe ancestries of each subject were computed and outliers were excluded. The

genotypes provided for the current study were all directly genotyped (i.e., none were imputed). Genotypes and phenotypic data for these controls are available by application to dbGaP (database of Genotypes and Phenotypes, dbgap.ncbi.nlm.nih.gov, Study Accessions: phs000021.v2.p1 and phs000167.v1.p1), and DNA, lymphoblastoid cell lines (LCLs), and additional phenotypic data are available through the NIMH repository (nimhgenetics.org). The samples used in the current study are those cases and controls over 25 years old, ever smokers, and with cigarettes per day (CPD) data available, totaling 1,671 EA cases, 841 AA cases, 1,317 EA controls, and 447 AA controls. Phenotypic aspects of case and control samples are briefly described separately below.

MGS Controls

A survey company (Knowledge Networks, under MGS guidance) recruited self-identified non-Hispanic adult control subjects from the United States, either EA or AA, from their nationwide panel of survey participants, which had been assembled by random digit dialing (~59% of the AA controls were recruited through a subcontract to Survey Sampling International by internet banner ad recruitment). Institutional review board approval was obtained at NorthShore University HealthSystem. The order of procedures for control subjects was recruitment, online consent (identical hard-copy consent signed at venipuncture), online questionnaire completion, venipuncture for DNA extraction and establishment of LCLs at Rutgers University Cell and DNA Repository (RUCDR), and full anonymization of data and biomaterials. The questionnaire (available at nimhgenetics.org) was primarily comprised of the Composite International Diagnostic Interview – Short Form (CIDI-SF) [4-6], modified to screen for lifetime diagnoses (alcohol dependence, drug dependence, major depressive episode/s, generalized anxiety disorder, specific phobia, social phobia, agoraphobia, panic attacks, and obsessive compulsive disorder). The questionnaire also included other components assessing various traits and disorders: Fagerström Test for Nicotine Dependence (FTND) [7]; Eysenck brief neuroticism and extraversion scales [8]; sexual identity; height and body mass index (BMI); psychosis and mania screens; ancestry (race/ethnicity) [9] for each grandparent; and basic demographics. We scored the dichotomous presence/absence of individual disorders according to the CIDI-SF [6] scoring memo [10]. Subjects answering anything but "no" to three psychosis screening questions were excluded. Besides demography questions (ancestry, age, sex, birth year, and education), the phenotypic data – CPD and resultant smoking quantity (SQ) bin – for the current study were collected from the pertinent subset of all items asked for generating a Fagerström Test for Nicotine Dependence [7] score in these control subjects. The specific question was: "Please think back to the time in your life when you smoked the most. How many cigarettes did you smoke on a typical day?" For the CPD phenotype analyzed here, MGS contributed 1,317 EA control daily smokers: 234 with CPD 1-10, 520 with CPD 11-20, 240 with CPD 21-30, and 323 with CPD >30, and 447 AA control daily smokers: 212 with CPD 1-10, 160 with CPD 11-20, 46 with CPD 21-30, and 29 with CPD >30.
MGS Schizophrenia Cases.

The schizophrenia case sample (~10% with schizoaffective disorder) was collected from the United States and Australia by the MGS collaboration, and ascertainment is described in detail elsewhere [1, 2, 11, 12]. Cases were assessed with the Diagnostic Interview for Genetic Studies (97%) [9], collateral informant data (usually a family member; 33%), and review of medical records (90%), and had a consensus diagnosis of schizophrenia or schizoaffective disorder (DSM-IV) [13]. Subjects with more than moderate mental retardation or those deemed to have psychosis only due to psychotogenic substance use or medical conditions were excluded. Cases gave written informed consent, and each collecting site's institutional review board approved the human subjects protocol. Besides demography questions (ancestry, age, sex, birth year, and education), the phenotypic data – CPD and resultant smoking quantity (SQ) bin – for the current study were collected from the DIGS medical history section (packs per day were converted to CPD by multiplying by 20 cigarettes in a pack). The specific question was: "If you ever smoked cigarettes on a daily basis, estimate the number of packs per day?" For the CPD phenotype analyzed here, MGS contributed 1,671 EA schizophrenia case daily smokers: 251 with CPD 1-10, 826 with CPD 11-20, 225 with CPD 21-30, and 369 with CPD >30, and 841 AA schizophrenia case daily smokers: 270 with CPD 1-10, 432 with CPD 11-20, 44 with CPD 21-30, and 95 with CPD >30.

References:

1. APA (2000). *Diagnostic and statistical manual of mental disorders : DSM-IV-TR*. Washington, DC: American Psychiatric Association.

2. C. R., Kaufmann, C. A., Faraone, S. V., Malaspina, D., Svrakic, D. M., Harkavy-Friedman, J., et al. (1998). Genome-wide search for schizophrenia susceptibility loci: the NIMH Genetics Initiative and Millennium Consortium. *Am J Med Genet,* 81, 275-281.

3. Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Pers Individ Dif,* 6, 21-29.

4. Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., & Fagerstrom, K. O. (1991). The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *Br J Addict,* 86, 1119-1127.

5. Kessler, R. C., Andrews, G., Mroczek, D., Ustun, B., & Wittchen, H. U. (1998a). The World Health Organization Composite International Diagnostic Interview short-form (CIDI-SF). *Int J Methods Psychiatr Res,* 7, 171-185.

6. Kessler, R. C., Wittchen, H. U., Abelson, J. M., Kendler, K. S., Knauper, B., McGonagle, K. A., et al. (1998b). Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. *Int J Methods Psychiatr Res,* 7, 33–55.

7. Nelson, C. B., Kessler, R. C., & Mroczek, D. (2001). Scoring the World Health Organization's Composite International Diagnostic Interview Short Form (CIDI-SF; v1.0 NOV98 for all disorders except OCD which is from v1.1 MAR99) Retrieved from http://www.hcp.med.harvard.edu/ncs/ftpdir/cidisf_readme.pdf

8. Nurnberger, J. I., Jr., Blehar, M. C., Kaufmann, C. A., York-Cooler, C., Simpson, S. G., Harkavy-Friedman, J., et al. (1994). Diagnostic interview for genetic studies. Rationale, unique features, and training. NIMH Genetics Initiative. *Arch. Gen. Psychiatry,* 51, 849-859.

9. Sanders, A. R., Duan, J., Levinson, D. F., Shi, J., He, D., Hou, C., et al. (2008). No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics. *Am. J. Psychiatry,* 165, 497-506.

10. Sanders, A. R., Levinson, D. F., Duan, J., Dennis, J. M., Li, R., Kendler, K. S., et al. (2010). The Internet-Based MGS2 Control Sample: Self Report of Mental Illness. *Am J Psychiatry*.

11. Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'er, I., et al. (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature,* 460, 753-757.

12. Suarez, B. K., Duan, J., Sanders, A. R., Hinrichs, A. L., Jin, C. H., Hou, C., et al. (2006). Genomewide Linkage Scan of 409 European-Ancestry and African American Families with Schizophrenia: Suggestive Evidence of Linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the Combined Sample. Am. J. Hum. Genet., 78, 315-333.
13. Wittchen, H. U. (1994). Reliability and validity studies of the WHO--Composite International Diagnostic Interview (CIDI): a critical review. J Psychiatr Res, 28, 57-84.

Acknowledgements: This study was supported by NIH R01 grants (**MH67257** to Nancy G. Buccola, **MH59588** to Bryan J. Mowry, MH59571 to Pablo V. Gejman, **MH59565** to Robert Freedman, **MH59587** to Farooq Amin, **MH60870** to William F. Byerley, **MH59566** to Donald W. Black, **MH59586** to Jeremy M. Silverman, MH61675 to Douglas F. Levinson, **MH60879** to C. Robert Cloninger, and **MH81800** to Pablo V. Gejman), NIH U01 grants (**MH46276** to C. Robert Cloninger, **MH46289** to Charles Kaufmann, **MH46318** to Ming T. Tsuang, MH79469 to Pablo V. Gejman, and MH79470 to Douglas F. Levinson), the Genetic Association Information Network (GAIN), and by The **Paul Michael Donovan Charitable Foundation**. Genotyping was carried out by the Center for Genotyping and Analysis at the Broad Institute of Harvard and MIT (Stacy Gabriel and Daniel B. Mirel), which is supported by grant U54 **RR020278** from the National Center for Research Resources. Genotyping of half of the EA sample and almost all the AA sample was carried out with support from GAIN. The GAIN quality control team (Gonçalo R. Abecasis and Justin Paschall) made important contributions to the project. We thank Shaun Purcell for assistance with PLINK.

Munich Germany (MUC12SCS; MUC12SCTL; MUCMDCS; MUCMDCTL)

For this meta-analysis, the German sample contributed unrelated European Caucasians (1052 healthy controls and 641 schizophrenia patients for MUCMD, and 235 controls and 421 schizophrenia patients for MUC12S).
Healthy unrelated volunteers of German descent (i.e., both parents German) were randomly selected from the general population of Munich, Germany, and contacted by mail. To exclude subjects with central neurological diseases and psychotic disorders or subjects who had first-degree relatives with psychotic disorders, several screenings were conducted before the volunteers were enrolled in the study. First, subjects who responded were screened by phone for the absence of neuropsychiatric disorders. Second, detailed medical and psychiatric histories were assessed for subjects and their first-degree relatives by using a semi-structured interview. Third, if no exclusion criteria were fulfilled, they were invited to a comprehensive interview including the Structured Clinical Interview for DSM-IV (SCID I and SCID II)[1,2] to validate the absence of any lifetime psychotic disorder. Additionally, the Family History Assessment Module[3] was conducted to exclude psychotic disorders among first-degree relatives. Furthermore, a neurological examination was conducted to exclude subjects with current CNS impairment. In the case that the volunteers were older than 60 years, the Mini Mental Status Test[4] was performed to exclude subjects with possible cognitive impairment.

Individuals with schizophrenia were ascertained from the Munich area in Germany. Of the samples MUC12S and MUCMD, 70.1% and 71.0% were of German descent and 29.9% and 30.0% were Caucasian middle Europeans, respectively. (No evidence for ethnic

stratification was observed after testing with the software STRUCTURE)[5]. Case participants had a DSM-IV and ICD-10 diagnosis of schizophrenia with the following subtypes (MUC12S/MUCMD): paranoid 78.1%/79.2%, disorganized 16.9%/15.9%, catatonic 0.5%/1.3% and undifferentiated 4.5%/3.6%). Detailed medical and psychiatric histories were collected, including a clinical interview using the SCID, to evaluate lifetime Axis I and II diagnoses. Four physicians and one psychologist rated the SCID interviews, and all measurements were double-rated by a senior researcher. Exclusion criteria included a history of head injury or neurological diseases. All case participants were outpatients or stable inpatients. Further details can be found in previous reports[6].

Smoking behavior was grouped into current, former and never smokers. The number of cigarettes per day (CPD) as well as FTND was assessed for the period of heaviest smoking and for average use. All subjects were smokers and reported smoking 100 cigarettes lifetime. The study obtained informed consent from participants and approval from the appropriate institutional review boards.

DNA was obtained from peripheral blood. DNA concentration was adjusted using the PicoGreen quantitation reagent (Invitrogen, Karlsruhe, Germany), and 1 ng was genotyped using the iPLEX assay on the MassARRAY MALDI-TOF mass spectrometer (SEQUENOM, Hamburg, Germany). Genotyping call rates in cases and controls were all >97%. Allele frequencies were similar to CEU sample frequencies. A subsample of

SNPs and DNA was genotyped twice to check for genotyping errors.
For this meta-analysis, MUC12S and MUCMD were each separated into schizophrenic cases (CS) and controls (CTL) prior to running association analyses, because schizophrenic patients are known to have different, heavy patterns of smoking compared to normal controls.

References:

1. First MB, Spitzer RL, Gibbon M, Williams BW, Benjamin L (1990) Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II). New York: Biometrics Research Department, New York State Psychiatric Institute.

2. First MB, Spitzer RL, Gibbon M, Williams JB (1995) Structured Clinical Interview for DSM-IV Axis I Disorders - Patient Edition (SCID - I/P, Version 2.0). New York: Biometrics Research Department, New York State Psychiatric Institute.
3. Rice JP, Reich T, Bucholz KK, Neuman RJ, Fishman R, Rochberg N, Hesselbrock VM, Nurnberger JI, Jr., Schuckit MA, Begleiter H (1995) Comparison of direct interview and family history diagnoses of alcohol dependence. *Alcohol Clin Exp Res* 19:1018-23.

4. Folstein MF, Folstein SE, McHugh PR. (1990) Mini-Mental-Status-Test. German Version: Kessler J, Folstein SE, Denzler P. Weinheim: Beltz.

5. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 65(1):220-8.

6. Van den Oord EJ, Rujescu D, Robles JR, Giegling I, Birrell C et al. (2006) Factor structure and external validity of the PANSS revisited. *Schizophr Res* 82: 213-23.

National Youth Survey – Family Study (NYSFS; originally "National Youth Survey")

The National Youth Survey began in 1976. At that time 1,725 adolescents between the ages of 11 and 17 years old as well as one of their parents were interviewed. Participants were chosen by a scientific method designed to select individuals who were representative of the national population. It was a sample of households with all children between 11 and 17 within a chosen household recruited. It is a longitudinal study, with 12 waves of interviews conducted so far. DNA was collected as part of wave 10 interviews[1-3]. Individuals were asked whether they had ever smoked cigarettes regularly (at least once per month). If they answered no or did not respond, they were excluded from the analysis. CPD was defined as the number of cigarettes smoked per day when smoking the most (over their lifetime), reported in the wave 10 interview. Age of first regular smoking was ascertained from the in-home questionnaire for Wave 10. Educational attainment was derived from questions asked in the Wave 10 questionnaire. Subjects were all adults at the time of the wave 10 interviews. Educational attainment was also derived from questions asked during the Wave 10 interviews. DNA was collected for 1071 individuals, 20 of whom have mostly missing phenotype information and were thus excluded, so genotypes and phenotypes were used for 1051 individuals. Total recruited was 1725, but there has been a low level of attrition thru time (there is no evidence for any systematic trends in either attrition or DNA collection refusal). After selection for unrelated individuals, 548 smokers were phenotyped/genotyped. All research protocols and consent forms were approved by institutional review boards of the University of Colorado.

DNA was derived from buccal cells. Taqman assays for allelic discrimination (Applied Biosystems, Foster City, CA) were used to determine SNP genotypes. QC performed on the genotyped sample (by sample and by SNP) excluded individuals with less than 50%

genotypes (assumed poor quality DNA sample). All SNPs had greater than 95% genotype calling after exclusion of individuals with low quality DNA samples. All genotypes were called by two independent individuals.

References:

1. Elliott DS, Huizinga D, Ageton SS (1985) Explaining Delinquency and Drug Use. Beverly Hills, CA: Sage Publications.

2. Elliott DS, Huizinga D, Menard S (1989) Multiple Problem Youth: Delinquency, Drugs and Mental Health Problems. New York, NY: Springer.

3. Hoft NR, Corley RP, Schlaepfer IR, McQueen MB, Huizinga D, Menard S, Ehringer MA. (2009) Genetic association of the CHRNA6 and CHRNB3 genes with tobacco dependence in a nationally representative sample. Neuropsychopharmacology. 34(3):698-706.

Nicotine Addiction Genetics Project and Australian Big Sibship Projects (NAG-Aus; NAG-Finland)

Description of the study

The study participants for the Nicotine Addiction Genetics Project (NAG) were enrolled at two different sites: the Queensland Institute of Medical Research (QIMR) in Australia and the University of Helsinki (UH) in Finland. Families for both the Australian and Finnish arms of the NAG were identified through smoking index cases by use of previously administered interview and/or questionnaire surveys of the community-based Australian and population-based Finnish registers of twins1,2. The Finnish arm of the NAG project (NAG-Fin) recruited twin pairs concordant for ever-smoking from the Finnish Twin Cohort, which consists of all Finnish twin pairs born between 1938 and 19573. Families chosen for the Australian arm of the NAG study (NAG-Aus) were identified from two cohorts of the Australian Twin Panel, which included spouses of the older of these two cohorts. The ancestry of the Australian samples is predominantly Anglo-Celtic or northern European (>90%). We also used data obtained from a third Australian Community-based family study, the Australian Big Sibship (BigSib). The BigSib sample comprises families ascertained through the Australian Twin Panel selected for five or more offspring sharing both biological parents. Families for the BigSib sample were recruited from the same Australian Twin Panel sources as were the NAG Australian families, and phenotypic information was obtained using the same assessment protocol as for the NAG. Clinical data for both Australian and Finnish subjects were collected using a computer-assisted telephone diagnostic interview (CATI), and adaptation of the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA)4,5 for telephone administration. The tobacco section of the CATI was derived from the Composite International Diagnostic Interview (CIDI)6 and incorporated standard FTND7 , DSM-III-R, and DSM-IV8 assessments of nicotine dependence. It also included a detailed history of cigarette and other tobacco use, including quantity and frequency of use for current, most recent, and heaviest period of use. The measure examined for the purposes of this study was the number of cigarettes smoked per day, during heaviest period of use. In addition age of onset of smoking was asked and educational level assessed on the basis of two questions on level of schooling and additional vocational training. All data-collection procedures were approved by institutional review boards at Washington University (WU), the QIMR, and the Ethics committee of the Hospital District of Helsinki and Uusimaa, including the use of appropriate and approved informed-consent procedures.

Smoking info from the study

For this meta-analysis, NAG/BigSib-Aus combined sample contributed information from a total of 1329 unrelated adult subjects (about 40% women; including 45% from the BigSib sample), 18-82 years of age (mean age: 44 years) at the time of assessment; including 592 who reported smoking 10 or fewer cigarettes, 489 subject who reported smoking 20 to 39, and 248 Australians who reported smoking 40 or more cigarettes during their heaviest period of smoking. Participants gave informed consent for an interview, for providing a blood sample for DNA extraction and cell lines, and for the sharing of their anonymous clinical and genotypic records with scientists outside of the NAG and/or BigSib research teams of investigators.

Analyzed as a separate sample, NAG-Fin contributed information from a total of 733 unrelated adult subjects (37% women); 36-66 years of age (mean age: 54.4 years) at the time of assessment. Participants were born between 1938 and 1965, and most of them (97%) between 1940-1959. Educational attainment information was from all participants; 23% had terminal degree of high school only or less while 77% had terminal degree greater than high school (these included all persons with vocational training). Participants with CPD information had smoked at least 100 cigarettes in their lifetime. Seventeen per cent of participants (n=121)reported smoking of ≤10 cigarettes, 34% (n=244) of >10 to 20, 32% (n=229) more than >20 to 30, and 18% (n=130) more than 30 CPD. Nine participants did not have CPD information. Age of onset of regular smoking information was from 709 participants ranging from 8 years to 45 years (mean18.1 years, SD 3.9). Thirty five percent of them had started regular smoking at 16 year old or younger. Participants gave informed consent for an interview, for providing a blood sample for DNA extraction, and for the sharing of their anonymous clinical and genotypic records with scientists outside of the NAG research teams of investigators.

How DNA was obtained

Participants gave blood samples at their local health center or laboratory, and blood samples were sent to the National Public Health Institute in Helsinki (currently National Institute for Health and Welfare), Finland. DNA was extracted from blood samples using standard procedures and genotyped using Illumina 670-Quad Custom chip at the Sanger Wellcome Trust Institute.

SNP info

The SNPs included in the analyses, rs1051730, rs1948, rs578776, rs6495306, and rs6265 all had genotyping success of >95% (0-3 missing genotypes per SNP) and Hardy-Weinberg equilibrium test p-values of 0.13, 0.23, 0.14, 0.77, and 0.12, respectively. The minor (MAF) and major alleles were T (37%) and C; T (30%) and C; T (29%) and C; G (36%) and A; and A (15%) and G, respectively.

References Cited:

1. Saccone SF, Pergadia ML, Loukola A, Broms U, Montgomery GW, Wang JC, Agrawal A, Dick DM, Heath AC, Todorov AA, Maunu H, Heikkila K, Morley KI, Rice JP, Todd RD, Kaprio J, Peltonen L, Martin NG, Goate AM, Madden PAF (2007) Genetic linkage to chromosome 22q12 for a heavy-smoking quantitative trait in two independent samples. Am J Hum Genet 80:856-866.
2. Loukola A, Broms U, Maunu H, Widén E, Heikkilä K, Siivola M, Salo A, Pergadia ML, Nyman E, Sammalisto S, Perola M, Agrawal A, Heath AC, Martin NG, Madden PAF, Peltonen L, Kaprio J (2008) Linkage of nicotine dependence and smoking behavior on 10q, 7q and 11p in twins with homogenous genetic background. The Pharmacogenomics Journal 8:209-219.
3. Kaprio J, Koskenvuo M. Genetic and environmental factors in complex diseases: the older Finnish Twin Cohort (2002) Twin Res 5:358-365.
4. Bucholz KK, Cadoret R, Cloninger CR, Dinwiddie SH, Hesselbrock VM, Nurnberger JI Jr, Reich T, Schmidt I, Schuckit MA (1994) A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA. J Stud Alcohol 55:149-158.
5. Hesselbrock M, Easton C, Bucholz KK, Schuckit M, Hesselbrock V (1999) A validity study of the SSAGA—a comparison with the SCAN. Addiction 94:1361-1370.
6. Cottler LB, Robins LN, Grant BF, Blaine J, Towle LH, Witthen HU, Sartorius N (1991) The CIDI-core substance abuse and dependence questions: cross-cultural and nosological issues: the WHO/ADAMHA field trial. Br J Psychiatry 159:653-658.
7. Heatherton, T.F., Kozlowski, L.T., Frecker, R.C., Fagerström, K.-O. (1991). The Fagerström Test for Nicotine Dependence: a revision of the Fagerström Tolerance Questionnaire. British Journal Addiction, 86, 1119–1127.
8. APA, American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders: DSM-IV, 4th ed. American Psychiatric Association, Washington DC, 1994.

Netherlands Study of Depression and Anxiety (NESDA):

The Netherlands Study of Depression and Anxiety (NESDA) (1) is a multi-centre study designed to examine the long-term course and consequences of depressive and anxiety disorders (http://www.nesda.nl). NESDA includes both individuals with depressive and/or anxiety disorders and controls without psychiatric conditions. Inclusion criteria were age 18-65 years and self-reported western European ancestry, exclusion criteria were not being fluent in Dutch and having a primary diagnosis of another psychiatric condition (psychotic disorder, obsessive compulsive disorder, bipolar disorder, or severe substance use disorder). For all participants DNA was isolated from the baseline blood sample (2) (collected between 2004-2007). The study protocol is approved by the Central Ethics Committee of the VU University Medical Center Amsterdam and all participating institutes. Through funding from the fNIH GAIN program (www.fnih.gov/gain), whole genome scan analysis was conducted for 1,859 NESDA (1,702 depressed cases and 157 controls) participants. (3) Perlegen Sciences (Mountain View, CA, USA) performed all genotyping according to strict standard operating procedures. Baseline data at time of DNA collection was used to determine smoking behavior.

References:

1. Penninx BW, Beekman AT, Smit JH, Zitman FG, Nolen WA, Spinhoven P, Cuijpers P, de Jong PJ, van Marwijk HWJ, Assendelft WJJ, van der Meer K, Verhaak P, Wensing M, de Graaf R, Hoogendijk WJ, Ormel J, van Dyck R. The Netherlands Study of Depression and Anxiety (NESDA): Rationale, objectives and methods. Int J Meth Psychiatr Res 2008;17:121-140.
2. Boomsma DI, Willemsen G, Sullivan PF, Heutink P, Meijer P, Sondervan D, Kluft C, Smit G, Nolen WA, Zitman FG, Smit JH, Hoogendijk WJ, van Dyck R, de Geus EJ, Penninx BW. Genome-wide association for Major Depression: Description of

samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. Eur J Hum Genet 2008;16:335-42.

3. Sullivan P, de Geus EJC, Willemsen G, James MR, Smit JH, Zandbelt T, Arolt V, Baune BT, Blackwood D, Cichon S, Coventry WL, Domschke K, Dumenil T, Farmer A, Fava M, Gordon SD, Heutink P, Holsboer F, Hoogendijk WJ, Hottenga JJ, Kohli M, Lin D, Lucae S, MacIntyre DJ, Maier W, McGhee KA, McGuffin P, Montgomery G, Muir WJ, Nolen W, Nöthen MM, Perlis R, Pirlo K, Posthuma D, Rietschel M, Schosser A, Smoller JW, AB Smit, Tzeng JY, van Dyck R, Zitman FG, Verhage M, Martin NG, Wray NR, Boomsma DI, Penninx BW. Genome-wide association for Major Depressive Disorder: a possible role for the protein PCLO. Mol Psychiatry 2009;14:359-75.

Netherlands Twin Registry (NTR1, NTR2):

The data come from a large-scale longitudinal study from the Netherlands Twin Registry (NTR). The NTR was established in 1987 and contains information about Dutch twins and their families voluntarily taking part in research[1]. The NTR study is approved by the Central Ethics Committee of the VU University Medical Center Amsterdam. Between 2004 and 2008 biological samples (including DNA) were collected[2]. Two subsamples with genotype data were available for the present study: NTR1[3] and NTR2[4]. Longitudinal survey data from 8 waves of data collection (1991-2010) were used to determine smoking behavior[5].

References:

1. Boomsma, D.I., et al., Netherlands Twin Register: from twins to twin families. Twin Research and Human Genetics, 2006. 9(6): p. 849-857.

2. Willemsen, G., et al., The Netherlands Twin Register Biobank:A Resource for Genetic Epidemiological Studies. Twin Res Hum Genet., 2010. 13(3): p. 231-245.

3. Boomsma, D.I., et al., Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects. European Journal of Human Genetics, 2008. 16: p. 335-342.

4. Ligthart, L. and e. al., Meta-analysis of genome-wide association for migraine in six population-based European cohorts. European Journal of Human Genetics, in press.

5. Vink, J.M., G. Willemsen, and D.I. Boomsma, Heritability of smoking initiation and nicotine dependence. Behavior Genetics, 2005. 35(4): p. 397-406.

Northern Finland Birth Cohort 1966 (NFBC1966)

All mothers expected to give birth in the two northernmost provinces of Finland, Oulu and Lapland, in 1966 were recruited to the study (n = 12,058 live births)[1]. At the age of 31 years, the cohort members were sent a postal questionnaire in which their smoking habits were inquired and N=8,767 (75%) returned it. Cohort members still living in the original target or capital area were also invited to a 31-year clinical examination and 6,033 (71%) participated. At this time, blood samples were collected of which DNA was extracted. Educational attainment was obtained from the Educational register of Statistics Finland in 1997. All participants included in this study gave written informed consent. The University of Oulu Ethics committee has approved the study.
DNA was extracted from blood using standard methods. Genome-wide genotyping was performed at the Broad Institute Biological Sample Repository in approximately 5500 participants with available DNA with the Illumina HumanCNV370DUO Analysis BeadChip[2]. For the current study, rs1051730 was directly genotyped. The NCBI build 35 genetic map and SNP positions were used.

Information on smoking behavior were collected at age 31 via postal questionnaire[4]. Participants were asked: 1. if they have ever smoked in their life, 2. if they have ever been smoking regularly (i.e. smoking at least one cigarette every day for at least one year), 3. how many cigarettes per day they have been smoking. Of the genotyped individuals 3,067 had smoked in their life and 1,896 had never smoked. Never smokers were excluded from the current analysis. The continuous smoking measured as cigarettes per day (CPD) was available for 3,025 individuals having mean CPD of 12.4 (7.9). The sex distribution in the NFBC1966 study sample was 49.7 % males (N=1,523) and 50.3 % females (N=1,544). Ever smokers were asked at what age they started smoking (age of initiation of smoking). This variable was available for 2,911 participants. Participants were also asked for how many years they have been smoking regularly. Considering that all individuals have been interviewed at 31 years, for current smokers, age at onset of regular smoking was indirectly computed by subtracting number of years they have been smoking regularly from current age (31 years). Individuals who were not current smokers but who have been regular smokers previously were asked at what age they quit smoking. For these individuals age at onset of regular smoking was indirectly computed by subtracting numbers of years they have been smoking regularly from the age at which they quit smoking. Age at onset of regular smoking was available for 2,245 participants.

References:

1. Rantakallio P. Groups at risk in low birth weight infants and perinatal mortality. Acta Paediatr Scand 1969;193:Suppl:1-71

2. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, et al. (2009): Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 41:35– 46.

3. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39: 906-913

4. Ducci F, Kaakinen M, Pouta A, Hartikainen AL, Veijola J, Isohanni M, CharoenP, Coin L, Hoggart C, Ekelund J, Peltonen L, Freimer N, Elliott P, Schumann G, Järvelin MR. TTC12-ANKK1-DRD2 and CHRNA5-CHRNA3-CHRNB4 Influence Different Pathways Leading to Smoking Behavior from Adolescence to Mid-Adulthood. Biol Psychiatry. 2010

Nurses' Health Study (NHS BrCa, NHS CHD, NHS T2D):

The three analysis samples that contribute to this meta-analysis were derived from case-control cohort studies nested within a multi-site U.S. cohort study, the Nurses' Health Study (NHS). The NHS cohort was initiated in 1976 across 11 states with 121,700 female registered nurses aged 30-55. In addition, blood samples were collected between 1989 and 1990 in the NHS.

Nested case-control cohort studies were conducted for Type 2 diabetes (T2D), cardiovascular disease (CHD), and breast cancer (BrCa). In the T2D study, controls were defined to be those free of diabetes at the time of diagnosis of the case, and were initially matched on year of birth, month of blood collection, and fasting status, with matched-pairs subsequently broken because not all subjects gave informed consent for the posting of their data on dbGaP[1]. In the CHD study, controls were randomly selected from participants who provided blood samples and did not experience CHD, with two controls for every case. Controls were matched on age, smoking, and month of blood draw. Finally, in the BrCa study, cases and controls were limited to post-menopausal women, who were not diagnosed with breast cancer during follow up. Controls were postmenopausal women matched with cases by age and post-menopausal hormone use at blood draw[2]. All studies from which the samples are derived obtained informed consent from participants and approval from the appropriate institutional review boards.

Current and former smokers were selected for analysis of the cigarettes-per day trait. The NHS samples contributed 1646 (NHS-T2D), 748 (NHS-CHD), and 1210 (NHS-BrCa) smokers. Cigarettes-per-day was measured using information from the baseline questionnaire and follow-up questionnaires through 2004, or the latest follow-up questionnaire available. The cigarettes-per-day trait reflects the average number of cigarettes per day over the period of observation, or pack-years/smoking duration. In the NHS samples, pack-years was calculated based on write-in values for cigarettes-per-day through 1982 and a categorical reporting of cigarettes-per-day after 1982[3].

In all samples, DNA was derived from white blood cells. Analysis samples were restricted to subjects of European ancestry, and genotyping used the Affymetrix 6.0 (T2D and CHD studies) and the Illumina 550 (BrCa study) platforms. Although exact protocols varied by sample, at a minimum DNA samples that did not meet a 90% completion threshold, and SNPs with low call rates (<90%), were dropped. Analyses based on principal components were conducted to assess self-reported race, and any self-reported "white" samples that had substantial similarity to non-European reference samples (either the HapMap YRI or CHB+JPT samples) were excluded.

References:

1. Qi L, Cornelis MC, Kraft P, Stanya KJ, Kao WH, Pankow JS, Dupuis J, Florez JC, Fox CS, Paré G, Sun Q, Girman CJ, Laurie CC, Mirel DB, Manolio TA, Chasman DI, Boerwinkle E, Ridker PM, Hunter DJ, Meigs JB, Lee CH; Meta-Analysis of Glucose and Insulin-related traits Consortium (MAGIC); Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, van Dam RM, Hu FB. (2010) Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Human Molecular Genetics*, in press.
2. Hunter DH, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ. (2007) A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics* 39: 870-874.

3. Caporaso N, Gu F, Chatterjee N, Sheng-Chih J, Yu K, Yeager M, Chen C, Jacobs K, Wheeler W, Landi MT, Ziegler RG, Hunter DJ, Chanock S, Hankinson S, Bergen AW, Kraft P (2009). Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS ONE* 4(2): e4653.

Acknowledgements: The Nurses Health Study (NHS) was supported by NCI (P01CA087969), NHGRI (5U01HG004399), NHLBI (5R01HL035464); genotyping for the NHS coronary heart disease study was supported by HL34594 and Merck Research Laboratories.

SardiNIA study:

Sardinia is the second largest island in the Mediterranean and constitutes a genetically isolated founder population. This population has aided in the identification of genes involved in several Mendelian disorders and is attractive for genetic studies due to its organization in long-established settlements that developed from an initial group of founder settlers (~1,000) thousands of years ago. The current study has recruited 6,148 Sardinians aged 14 and older, from a cluster of four towns in the Lanusei Valley in the Ogliastra region of the province of Nuoro. This sample corresponds to approximately 62% of the population eligible in the area for recruitment. Information collected during enrollment allowed the individuals to be organized into 711 complex pedigrees, each up to five generations deep, with an average kinship coefficient of 0.1628. All volunteers have been characterized for 98 quantitative traits. Traits include anthropomorphic measures, plasma and serum markers (including cholesterol and other markers of cardiovascular disease), and personality traits (using the five-factor model).

The main goal of the study is to examine phenotypic similarities between relatives that yield information on the overall contributions of genes to trait variability. Data given here provide p-values for 98 traits studied in 1,412 individuals, based on genotyping with the Affymetrix 500K chip and imputed markers using the HapMap population as a reference (N=2,259,179). Sharing this genome assessment data at high level of resolution for a variety of quantitative traits will be useful for other groups to validate newly observed associations and to investigate possible pleiotropic effects.

References:

1. Pilia, G. *et al*. Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet. 2, e132 (2006).

Acknowlegements: This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging.

The Study of Health in Pomerania (SHIP)

The Study of Health in Pomerania (SHIP) is a longitudinal, population-based survey from West Pomerania, Germany[1,2]. Data from the baseline cohort (n=4308) were used for this study. N=2458 Caucasian subjects were included into the present analyses (age $\geq$ 25, ever smoker). The study obtained informed consent from participants and approval from the appropriate institutional review boards.

Blood samples are obtained according to standardized procedures. Aliquots of blood samples are immediately placed on ice. The SHIP laboratories take part in the official German external quality proficiency testing programs. All assays are calibrated against the international reference preparations, whenever these are available. A bank of dummy samples allows the standardization of different laboratory methods. Serum, EDTA and citrate plasma, DNA and urine are stored at -80°C in a biobank.[2]

The SHIP samples were genotyped using the Affymetrix Human SNP Array 6.0. Hybridisation of genomic DNA was done in accordance with the manufacturer's standard recommendations. The genetic data analysis workflow was created using the Software InforSense. Genetic data were stored using the database Caché (InterSystems). Genotypes were determined using the Birdseed2 clustering algorithm.

For quality control purposes, several control samples where added. On the chip level, only subjects with a genotyping rate on QC probesets (QC callrate) of at least 86% were included. Finally, all arrays had a sample CallRate > 92%. Imputation of genotypes was performed with the software IMPUTE v0.5.0 based on HapMap II CEU reference panel.

This analysis included the following variables: gender, age, cigarettes per day ("How many cigarettes on average do you smoke per day")[3], age of onset ("How old were you started smoking regularly?"), educational attainment ("What is the highest educational degree or diploma you hold?").

References:

1. John, U. *et al*. Study of Health in Pomerania (SHIP): a health examination survey in an East German region: objectives and design. Soz. Praventivmed. 46, 186–194 (2001).

2. Völzke, H. *et al*. (2011). Study of Health in Pomerania (SHIP) – a community cohort and repeated survey approach to comprehensively assess main health determinants among the general adult population. International Journal of Epidemiology (in press)

3. Liu, J. *et al.* (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nature Genetics 42(5):436-40.

The Smoking in Families study (SMOFAM)

The Smoking in Families study (DA03706, Hyman Hops, PI, Oregon Research Institute) is a longitudinal, repeated measures age-sequential cohort study of environmental and psychosocial risk factors for adolescent and young adult substance use, including tobacco use and dependence, initiated in 1984. Subjects were recruited through advertisements in traditional media, and flyers distributed at middle and high schools in four mid-sized and small urban and rural Pacific Northwest cities with populations ranging from 30,000 to 175,000. The original SMOFAM study recruited 763 families, with at least one adolescent age 11 or older. Families with smoking parents and/or adolescents were of special interest since the focus was on adolescents at risk for tobacco and other substance use. Within each family one adolescent was designated as the proband if s/he had previously tried a substance. Each proband had to have at least one parent agree to participate. An attempt was made to encourage both parents and all sibs over the age of 11 to participate. The only other requirement was that all participants needed to be able to read basic level English. Repeated annual assessment of probands facilitated characterization of longitudinal phenotypes for tobacco use, including the acquisition and maintenance of smoking, as well as many potential psychosocial and environmental predictors of substance (Hops 2000). For an integrated research project study the environmental, genetic and metabolic determinants of tobacco use, probands and family members were recruited from those SMOFAM families in which the proband had completed at least seven of the first ten assessments on tobacco use and elected to provide a blood sample for DNA analysis. Probands and each first degree relative completed a family history of tobacco use including sex, age, relationship to proband (biological or nonbiological; full-, half-, or nonbiological sibling), vital status, lifetime “ever” smoking of 100 cigarettes, ever regular use of cigars, pipes, or smokeless tobacco, age at initiation of daily cigarette smoking, average number of cigarettes smoked per day when smoking, ever tried to quit, and success in permanent quitting (Swan, Hudmon et al. 2003).

For this meta-analysis, SMOFAM contributed a sample of 146 SMOFAM probands, European-Americans ever smokers (100 cigarettes lifetime). The study obtained informed consent from participants and approval from the appropriate institutional review boards. This analysis included the following variables: female gender, 59.6%; age, mean=28.3 years, SD=1.6; cigarettes per day ("Think about the year in your life when you smoked most. During that time, about how many cigarettes did you usually have per day?"), 49.3%, 40.4%, 6.9% and 3.4% for CPD=0, 1, 2, 3, respectively; age of onset ("How old were you the very first time you smoked cigarettes every day or nearly every day for a period of 2 months?"), mean=16.5 years, SD=3.6; educational attainment ("What is the highest educational degree or diploma you hold?"), 0=7.5%; 1=26.7%; 2=39.0%; 3=26.7% for no degree, basic degree, vocational or associates degree, and academic college or university degree, respectively. DNA was extracted from whole blood using standard procedures (Miller, Dykes et al. 1988). Genotyping of the DNA samples was carried out using Illumina GoldenGate technology (Conti, Lee et al. 2008). A multi-step genotype quality control procedure was performed (Conti, Lee et al. 2008; Bergen, Conti et al. 2009).

References:

1. Bergen, A. W., D. V. Conti, et al. (2009). "Dopamine genes and nicotine dependence in treatment-seeking and community smokers." Neuropsychopharmacology 34(10): 2252-2264.

2. Conti, D. V., W. Lee, et al. (2008). "Nicotinic acetylcholine receptor beta2 subunit gene implicated in a systems-based candidate gene study of smoking cessation." Hum Mol Genet 17(18): 2834-2848.

3. Hops, H. A., J.A.;Duncan, S.C.;Tildesley, E. (2000). Adolescent drug use development: a social interactional and contextual perspective. Handbook of Developmental Psychopathology. A. J. L. Sameroff, M.;Miller, S.M. New York, Kluwer Academic/Plenum Publishers: 589-605.

4. Miller, S. A., D. D. Dykes, et al. (1988). "A simple salting out procedure for extracting DNA from human nucleated cells." Nucleic Acids Res 16(3): 1215.

5. Swan, G. E., K. S. Hudmon, et al. (2003). "Environmental and genetic determinants of tobacco use: methodology for a multidisciplinary, longitudinal family-based investigation." Cancer Epidemiol Biomarkers Prev 12(10): 994-1005.

Acknowlegements: Grant support for the SRI sample is from NIH grants U01DA02830, R01 DA03706, and 7PT2000-2004 from the University of California Tobacco-Related Disease Research Program

The Utah Genetics of Addiction Project (Utah)

The Utah Genetics of Addiction Project contributed two cohorts (LHS and Utah) from a study of genetic risk markers for nicotine dependence and chronic obstructive pulmonary disease (COPD)[1141]. The UT cohort was made up of respondents to community advertising for persons who had smoked more than 100 cigarettes lifetime plus a subset of the Lung Health Study (LHS) participants originally recruited in Utah; these Utah LHS participants were excluded from the LHS cohort. UT participants were not drawn from a psychiatric treatment population, and no medical or behavioral treatments were offered as part of the study. UT volunteers were not excluded simply because they had a lifetime diagnosis of psychosis or Bipolar Disorder, but they were excluded if their current mental status made it impossible for them to complete the questionnaires or interviews. Pulmonary function testing determined 62% of the UT cohort had COPD. Of UT participants, 43% had not smoked for at least 2 years prior to participation in the study. All UT participants were of European descent, and all had smoked more than 100 cigarettes lifetime. Study procedures were approved by the University of Utah IRB.

DNA was isolated from peripheral blood lymphocytes collected in Salt Lake City, UT (UT cohort). SNP genotyping methods were previously described[1] and used either the SNPlex assay (Applied Biosystems) or TaqMan assay (Applied Biosystems). The call rate was 99.9%; and SNPs genotyped by both the TaqMan and SNPlex methods in 236 individuals had a concordance rate > 99.7%.

References:

1. Sanders AR, Duan J, Levinson DF, et al. No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics. Am J Psychiatry 2008;165(4):497-506.

2. Shi J, Levinson DF, Duan J, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 2009;460(7256):753-7.

3. Sanders AR, Levinson DF, Duan J, et al. The Internet-Based MGS2 Control Sample: Self Report of Mental Illness. Am J Psychiatry 2010.

4. Wittchen HU. Reliability and validity studies of the WHO--Composite International Diagnostic Interview (CIDI): a critical review. J Psychiatr Res 1994;28(1):57-84.

5. Kessler RC, Wittchen HU, Abelson JM, et al. Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey. Int J Methods Psychiatr Res 1998;7(1):33–55.

6. Kessler RC, Andrews G, Mroczek D, Ustun B, Wittchen HU. The World Health Organization Composite International Diagnostic Interview short-form (CIDI-SF). Int J Methods Psychiatr Res 1998;7(4):171-85.

7. Heatherton TF, Kozlowski LT, Frecker RC, Fagerstrom KO. The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. Br J Addict 1991;86(9):1119-27.

8. Eysenck SBG, Eysenck HJ, Barrett P. A revised version of the psychoticism scale. Pers Individ Dif 1985;6:21-9.

9. Nurnberger JI, Jr., Blehar MC, Kaufmann CA, et al. Diagnostic interview for genetic studies. Rationale, unique features, and training. NIMH Genetics Initiative. Arch Gen Psychiatry 1994;51(11):849-59.

10. Nelson CB, Kessler RC, Mroczek D. Scoring the World Health Organization's Composite International Diagnostic Interview Short Form (CIDI-SF; v1.0 NOV98 for all disorders except OCD which is from v1.1 MAR99) In; 2001.
11. Cloninger CR, Kaufmann CA, Faraone SV, et al. Genome-wide search for schizophrenia susceptibility loci: the NIMH Genetics Initiative and Millennium Consortium. Am J Med Genet 1998;81(4):275-81.

12. Suarez BK, Duan J, Sanders AR, et al. Genomewide Linkage Scan of 409 European-Ancestry and African American Families with Schizophrenia: Suggestive Evidence of Linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the Combined Sample. Am J Hum Genet 2006;78(2):315-33.

13. APA. Diagnostic and statistical manual of mental disorders : DSM-IV-TR. 4th ed., text revision ed. Washington, DC: American Psychiatric Association; 2000.

14. Weiss RB, Baker TB, Cannon DS, et al. A candidate gene approach identifies the CHRNA5-A3-B4 region as a risk factor for age-dependent nicotine addiction. PLoS Genet 2008;4(7):e1000125.

Virginia Adult Twin Study of Psychiatric and Substance Use Disorder (VA-twin)

The VA-Twins were selected from the Virginia Adult Twin Study of Psychiatric and Substance Use Disorder, which was a population-based epidemiology study. Tobacco smoking and nicotine dependence were assessed by the Fagerström Tolerance Questionnaire (FTQ) and/or Fagerström Test for Nicotine Dependence (FTND) during the time of heaviest lifetime nicotine use. In this study, only regular smokers (defined as those who used at some point in their lives an average of at least seven cigarettes per week for a minimum of four weeks) were included (N = 2388). One subject from each twin pair was selected, and all subjects were of Caucasian ancestry. The study obtained informed consent from participants and approval from the institutional review board of Virginia Commonwealth University. DNA was extracted from buccal brushes. Genotyping was performed with the TaqMan genotyping method. To ensure the quality of genotyping, negative control samples were included in each plate. Genotypes were scored using a semi-automated protocol.

References:

1. Chen X, Chen J, Williamson VS, An SS, Hettema JM, Aggen SH, Neale MC, Kendler KS (2009). Variants in nicotinic acetylcholine receptors alpha5 and alpha3 increase risks to nicotine dependence. *Am J Med Genet B Neuropsychiatr Genet*. 150B(7):926-33.

2. Kendler KS, Myers J, Prescott CA (2007) Specificity of genetic and environmental risk factors for symptoms of cannabis, cocaine, alcohol, caffeine, and nicotine dependence. *Arch.Gen.Psychiatry* 64(11): 1313-1320.

3. van den Oord EJ, Jiang Y, Riley BP, Kendler KS, Chen X (2003) FP-TDI SNP scoring by manual and statistical procedures: a study of error rates and types. *Biotechniques* 34: 610-6, 618-20, 622 passim.

Welcome Trust Case-Control Consortium (WTCCC) datasets: WTCCC-HT, WTCCC-CHD

CHD: WTCCC-CHD cases had a validated history of either myocardial infarction or coronary revascularization (coronary artery bypass surgery or percutaneous coronary angioplasty) before their 66th birthday.
Verification of the history of CAD was required either from hospital records or the primary care physician. Recruitment was carried out on a national basis in the UK through a direct approach to the public via (1) the media and (2) mailing all general practices (family physicians) with information about the study, as previously described [1]. In an initial pilot phase, potential participants were also identified and approached through local CAD databases in the two lead centres (Leeds and Leicester). Although the majority of subjects had at least one further sib also affected with premature CAD, only one subject from each family was included in the present study. Genotyping of these samples was carried out as part of the WTCCC1 genome-wide association studies [2].

As part of taking the medical history, a series of questions on smoking were asked which included the following: Do you smoke cigarettes now, if yes, when did you start smoking and how many per day? Did you smoke previously; if so how many cigarettes did you smoke previously; what age did you start and what age did you stop.

HT: The WTCCC-HT collection comprised severely hypertensive probands ascertained from families with multiple affected members in the UK as part of the MRC funded BRIGHT study [3]. Hypertensive cases had to have a diagnosis of hypertension prior to 50 years, and blood pressure recordings ≥ 150/100mmHg for a single reading or ≥ 145/95mmHg for 3 consecutive readings on a single visit. Exclusion criteria included BMI>35, presence of diabetes, secondary hypertension or a co-existing illness. As part of taking the medical history, a series of questions on smoking were asked which included the following: Do you smoke cigarettes now, if yes, what is the type and how many per day? Did you smoke previously, How many cigarettes did you smoke previously? What age did you start smoking.

References:

1. Samani, N. J. et al. A genomewide linkage study of 1,933 families affected by premature coronary artery disease: The British Heart Foundation (BHF) Family Heart Study. Am. J. Hum. Genet. 77, 1011–1020 (2005).

2. Wellcome Trust Case-Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007).

3. Caulfield, M., P. Munroe, J. Pembroke, N. Samani, A. Dominiczak, M. Brown, N. Benjamin, et al. (2003). "Genome-wide mapping of human loci for essential hypertension." Lancet 361(9375): 2118-23.

Yale Genetics of Cocaine Dependence, Genetics of Opioid Dependence and Genetics of Alcohol Dependence (YALE)

Subjects were recruited from substance abuse treatment centers and through advertisements at the University of Connecticut Health Center, Yale University School of Medicine, the Medical University of South Carolina, the University of Pennsylvania, and McLean Hospital (Harvard Medical School). The sample was recruited for substance abuse outcomes, including cocaine, opioid, alcohol, and nicotine dependence. Individuals were excluded if diagnosed with Axis I major psychotic illness (e.g., schizophrenia or schizoaffective disorder). The study protocol was approved by the institutional review board at each clinical site. After complete description of the study to the subjects, written informed content was obtained. Genetic studies of substance dependence disorders and related traits in a subset of this sample have been published[1-3].

Subjects were interviewed using an electronic version of the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA)[4,5] to derive diagnoses for lifetime nicotine, cocaine, opioid, and alcohol dependence according to DSM-IV criteria. The CPD phenotype was determined by the question "When you were smoking regularly, how many cigarettes did you usually smoke in a day?" The sample of unrelated individuals contains 912 individuals of European ancestry with genotype and phenotype data. All are smokers.

DNA was primarily extracted from immortalized cell lines or blood samples, with a small number from saliva. SNP genotyping was performed at Yale University using a closed-tube fluorescent TaqMan 5'-nuclease allelic discrimination assay ordered as "assays-on-demand" (Applied Biosystems Inc., Foster City, CA). Fluorescence plate reads and genotype calls were made using ABI 7900 Sequence Detection Systems. Two nanograms of genomic DNA were PCR amplified in 384-well plates using a 2-µl reaction volume. The insertion/deletion marker was genotyped by PCR amplification followed by agarose gel size fractionation.

At least two blank wells and two duplicate samples (for the purposes of cross-run confirmation of genotype assignment) were included in each 96-well plate. At least 8% of genotypes were repeated for quality control. The results were compared for verification. Results for 28 individuals for which the genotyping completely failed on the 3 SNPs were removed from analysis.

References:

1. Gelernter J, Panhuysen C, Weiss R, Brady K, Hesselbrock V, et al (2005) Genomewide linkage scan for cocaine dependence and related traits: Linkages for a cocaine-related trait and cocaine-induced paranoia. *Am J Med Genet Part B (Neuropsychiatric Genetics)* 136B:45–52.

2. Zhang H, Kranzler HR, Weiss RD, Luo X, Brady KT et al. (2009). Pro-opiomelanocortin gene variation related to alcohol or drug dependence: evidence and replications across family- and population-based studies. *Biol Psychiatry* 66:128-136.

3. Gelernter J, Yu Y, Weiss R, Brady K, Panhuysen C et al. (2006) Haplotype spanning TTC12 and ANKK1, flanked by the DRD2 and NCAM1 loci, is strongly associated to nicotine dependence in two distinct American populations. *Hum Mol Genet* 15:3498-3507.

4. Pierucci-Lagha A, Gelernter J, Chan G, Arias A, Cubells JF, et al. (2007). Reliability of DSM-IV diagnostic criteria using the Semi-structured Assessment for Drug Dependence and Alcoholism (SSADDA). *Drug Alcohol Depend* 91:85-90.
5. Pierucci-Lagha A, Gelernter J, Feinn R, Cubells JF, Pearson D, et al. (2005) Diagnostic reliability of the Semi-structured Assessment for Drug Dependence and Alcoholism (SSADDA). *Drug Alcohol Depend* 80:303-312.

Young Finns Study (YFS):

The Cardiovascular Risk in Young Finns (YFS) is a population-based follow up-study started in 1980 [1]. The main aim of the YFS is to determine the contribution made by childhood lifestyle, biological and psychological measures to the risk of cardiovascular diseases in adulthood. In 1980, over 3,500 children and adolescents all around Finland participated in the baseline study. The follow-up studies have been conducted mainly with 3-year intervals. The latest 27-year follow-up study was conducted in 2007 (ages 30-45 years) with 2,204 participants.

Age of onset of regular smoking and cigarettes per day at current moment (year 2007) or before quitting were assessed with a questionnaire. Smoking was considered to be regular when it had lasted at least one year. There were 694 smokers with genotype data available for analysis.

All subjects gave their written informed consent in 2007 and the study was approved by local ethics committees of the participating universities.

Genomic DNA was extracted from peripheral blood leukocytes using a commercially available kit and Qiagen BioRobot M48 Workstation according to the manufacturer's instructions (Qiagen, Hilden, Germany).

Genotyping was done for 2,556 samples using custom build Illumina Human 670k BeadChip at Welcome Trust Sanger Institute. Genotypes were called using Illuminus clustering algorithm [2]. Samples and SNPs with call rate < 0.95 and SNPs with MAF < 0.01 and HWE p-value < 1e-6 were filtered out. Quality control procedure has been described in detail elsewhere [3]. Genotype imputation was performed using MACH 1.0 and HapMap II CEU (release 22) samples as reference. After quality control and imputation there were 2,442 samples, 546,677 genotyped and 2,543,887 imputed SNPs available for analysis.

This analysis included the following variables: birth year, gender, cigarettes per day ("How many self rolled/factory made cigarettes do you currently smoke or have smoked before quitting?"), age of onset ("How old were you when you started smoking regularly?"), educational attainment ("What is the highest educational degree or diploma you hold?").

References:

1. Raitakari OT, Juonala M, Rönnemaa T, Keltikangas-Järvinen L, Räsänen L, Pietikäinen M, Hutri-Kähönen N, Taittonen L, Jokinen E, Marniemi J, Jula A, Telama R, Kähönen M, Lehtimäki T, Akerblom HK, Viikari JS. "Cohort profile: The cardiovascular risk in Young Finns Study." *Int J Epidemiol*. 2008 Dec;37(6):1220-6. "
2. Teo YY, et al. A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics. 2007;23:2741–2746
3. Smith EN, et al. Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa heart study. PLoS Genet. 2010 Sep 9;6(9)

Supplemental Results:

Here we present results from additional meta-analyses. In order to maximize the sample size for each analysis, and allow for a more direct comparison to previously published meta-analyses of smoking (Liu et al. 2010; TAG 2010; Thorgeirsson et al. 2010), the dependent variable used is cigarettes per day (CPD) coded 0 (CPD ≤ 10), 1 (10 < CPD ≤ 20), 2 (20 < CPD ≤ 30), and 3 (CPD > 30), and analyzed in a linear regression. Although the sample size is larger, the results are unchanged: early-onset smokers have a stronger genetic association with rs16969968 than late-onset smokers (Figure S1). There is no genetic interaction between this locus and birth cohort (Figure S2), gender (Figure S3), or educational attainment (Figure S4) on smoking quantity.

References

Liu, J. Z., F. Tozzi, D. M. Waterworth, S. G. Pillai, P. Muglia, L. Middleton, W. Berrettini, et al. (2010). "Meta-analysis and imputation refines the association of 15q25 with smoking quantity." Nat Genet **42**(5): 436-40.

TAG (2010). "Genome-wide meta-analyses identify multiple loci associated with smoking behavior." Nat Genet **42**(5): 441-7.

Thorgeirsson, T. E., D. F. Gudbjartsson, I. Surakka, J. M. Vink, N. Amin, F. Geller, P. Sulem, et al. (2010). "Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior." Nat Genet **42**(5): 448-53.

eTable 1: Measurement of CPD, Age of Onset of Regular Smoking, and Educational Attainment

Dataset	n	CPD definition	Age of Onset definition	Educational Attainment Definition
ACS_LCA	1,005	" How many cigarettes on average did or do you usually smoke per day?"	"How old did you begin smoking cigarettes?"	What is the highest education you have completed?
Addhealth	469	During the past 30 days, on the days you smoked, how many cigarettes did you smoke each day?	"How old were you the very first time you smoked cigarettes every day or nearly every day for a period of 2 months?"	What is the highest level of education you have achieved to date?
BOMA	1,050	Think about the year in your life when you smoked most. During that time, about how many cigarettes did you usually have per day?	How old were you when you started smoking regularly the first time in your life	What is the highest educational degree or diploma you hold?
COGA	1,704	Think about a period lasting a month or more when you were smoking the most. How many cigarettes did you usually smoke in a day?	How old were you the first time you smoked cigarettes regularly?	
COGEND	2,073	Think about the year in your life when you smoked most. During that time, about how many cigarettes did you usually have per day?	"How old were you the very first time you smoked cigarettes every day or nearly every day for a period of 2 months?"	“What is the highest educational degree or diploma you hold?”
Dental caries	639	"During the time in your life when you were smoking cigarettes most often, how many cigarettes did you have on a typical day?"	"How old were you the first time you smoked a cigarette when an adult was not around?"	What is the highest education you have completed?
DNBC	370	Women who smoked during pregnancy were asked, “How many cigarettes did you smoke on average?”		
GenMetS	648	How much do you daily smoke currently or did prior to quitting?		
GEOS	477	Maximum of "What was the average number of cigarettes per day you smoked 30 days prior to [REFERENCE DATE]?" and "When you smoke(d), what is (was) the average	How old were you when you first started to smoke cigarettes?	Including grade school, high school, and college, buisness, vocational, professional, and postgraduate schooling, what is the highest grade or year of
KORCULA	384	How many cigarettes on average do you smoke per day	How old were you the very first time you smoked cigarettes every day or nearly every day for a period of 2 months?	What is your highest vocational training
LHS	4,102	"On the average of the entire time you smoked, how many cigarettes did you smoke per day?"	At what age did you first become a daily cigarette smoker?	Years of Education
LHS-Utah	1,943	Think about the year in your life when you smoked most. During that time, about how many cigarettes did you usually have per day?	How old were you when you started smoking cigarettes daily?	
MGS	2,988	"Please think back to the time in your life when you smoked the most. How many cigarettes did you smoke on a typical day?"		What is the highest educational degree or diploma you received?
NAG FIN	733	During that same period in your life when you were smoking the most, how many cigarettes did you typically use on those days when smoked?	How old were you when you first smoked cigarette every day or nearly every day at least two months in a row?	What is the highest educational level that you have completed?

eTable 1 (cont.)

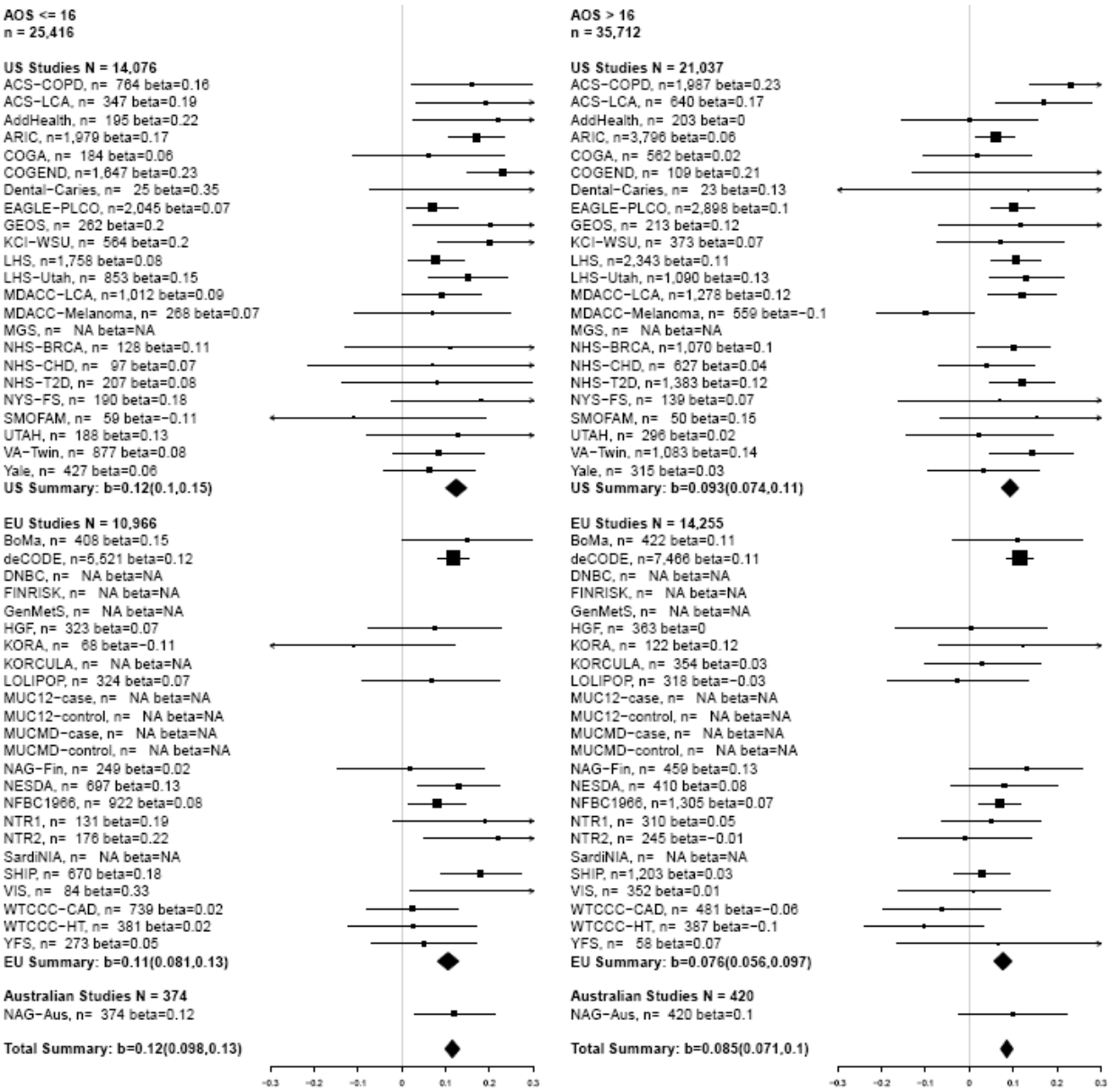
Dataset	n	CPD definition	Age of Onset definition	Educational Attainment Definition
NESDA	1,248	During your smoking periods how many cigarettes or self-rolled cigarettes did you smoke on average?	At what age did you start smoking?	What is the highest level of education that you completed (i.e. received a diploma)?
NFBC66	3,067	How many cigarettes per day have you been smoking?	How long have you been smoking regularly? (This value was subtracted from age at interview)	
NHS	1,198	Maximum of "Do you smoke cigarettes currently? How many cigs per day?" and "If not currently a smoker, did you smoke regularly in the past? How many cigs/day in	How old were you when you first started to smoke regularly?	Which of the follwing degrees have you received? Mark all that apply. RN,Bachelors, Masters, Doctorate
NTR	1,777	Smokers and ex-smoker: how many cigarettes/day?	At what age did you start smoking regularly?	What is the highest educational degree or diploma you hold?
NYSFS	751	"How many cigarettes did you smoke per day during the period when you were smoking most?"	How old were you when you began using tobacco on a regular basis, that is, at least once per month?	What is the highest educational degree or diploma you hold?
SardiNIA	531	“How many cigarettes a day?”	Smoke since when?	
SHIP	4,081	How many cigarettes on average do you smoke per day	How old were you the very first time you smoked cigarettes every day or nearly every day for a period of 2 months?	What is your highest degree (school, college, university)?
SMOFAM	146	Think about the year in your life when you smoked most. During that time, about how many cigarettes did you usually have per day?	How old were you the very first time you smoked cigarettes every day or nearly every day for a period of 2 months?	What is the highest educational degree or diploma you hold?
UTAH	484	Think about the year in your life when you smoked most. During that time, about how many cigarettes did you usually have per day?	How old were you when you started smoking cigarettes daily?	
WTCCC	1,221		How old were you the very first time you smoked cigarettes every day or nearly every day for a period of 2 months?	What is the highest educational degree or diploma you hold?
Yale	1,300	Average amount of cigarrettes per day during regular smoking	"How old were you the first time you smoked cigarettes at that rate?"	What is the highest grade in school you completed?
YFS	694	sum of questions: "Amount of factory-made cigarettes smoked per day (at present or before quitting)" and "Amount of self-rolled cigarettes smoked per day (at present or before quitting)"	How old were you when you started regular smoking? (mean of the three follow-up studies in 1980, 1983 and 1986)	What is the highest educational degree you hold?

eTable 2: Summary of Additional Meta-analyses to assess the Sensitivity of Results to coding of CPD and Age of Onset (AOS). To determine the sensitivity of the observed interaction between rs16969968 and early-onset smoking on smoking behavior, we permuted the coding of both age of onset and smoking behavior. Regardless of the coding of either variable, we see a stronger genetic effect in early-onset smokers. The discrepancy in sample sizes across permutations of the variables is because not all analyses were available for all datasets.

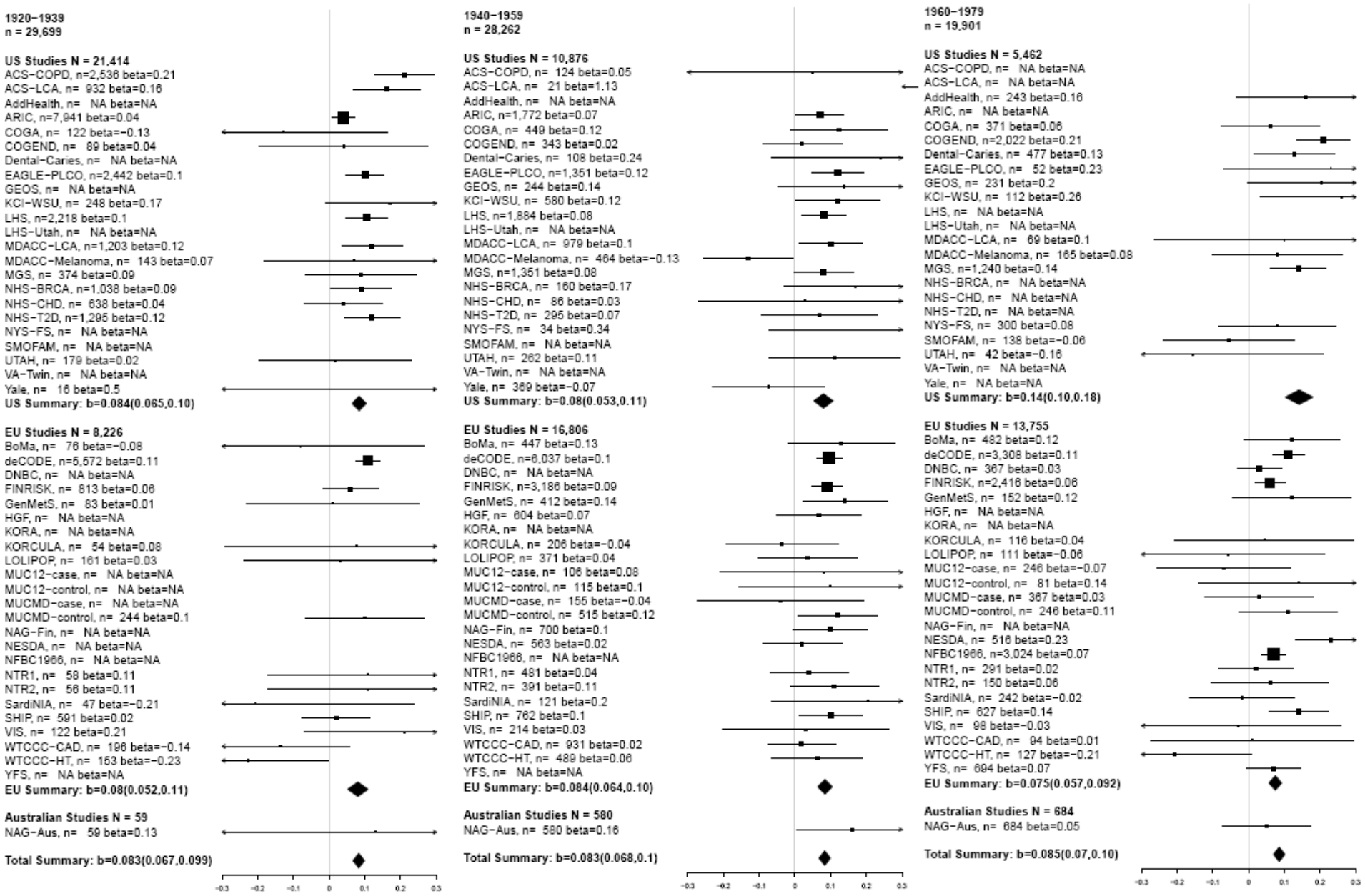
	Logistic Regression: CPD ≤ 10 vs. CPD > 20*				Linear Regression: CPD coded ≤ 10, 11-20, 21-30, >30		
	n	beta	sd		n	beta	sd
all AOS	41,760	0.30	0.02		51,128	0.08	0.005
AOS 2-level							
≤ 16	4,428	0.35	0.06		8,756	0.10	0.01
> 16	4,181	0.23	0.05		14,213	0.06	0.01
AOS 4-level							
≤ 15	6,651	0.35	0.05		9,003	0.11	0.01
16-17	4,992	0.37	0.05		6,339	0.13	0.02
18-19	4,867	0.23	0.05		5,446	0.09	0.02
≥ 20	6,233	0.23	0.04		6,675	0.10	0.02
AOS quartiles							
q1	4,337	0.29	0.06		6,773	0.11	0.02
q2	4,893	0.38	0.05		7,416	0.12	0.02
q3	6,213	0.24	0.04		9,551	0.09	0.01
q4	6,505	0.26	0.04		9,319	0.11	0.01

[a] AOS quartiles are computed within each individual study.
* smoking phenotype used in primary analyses

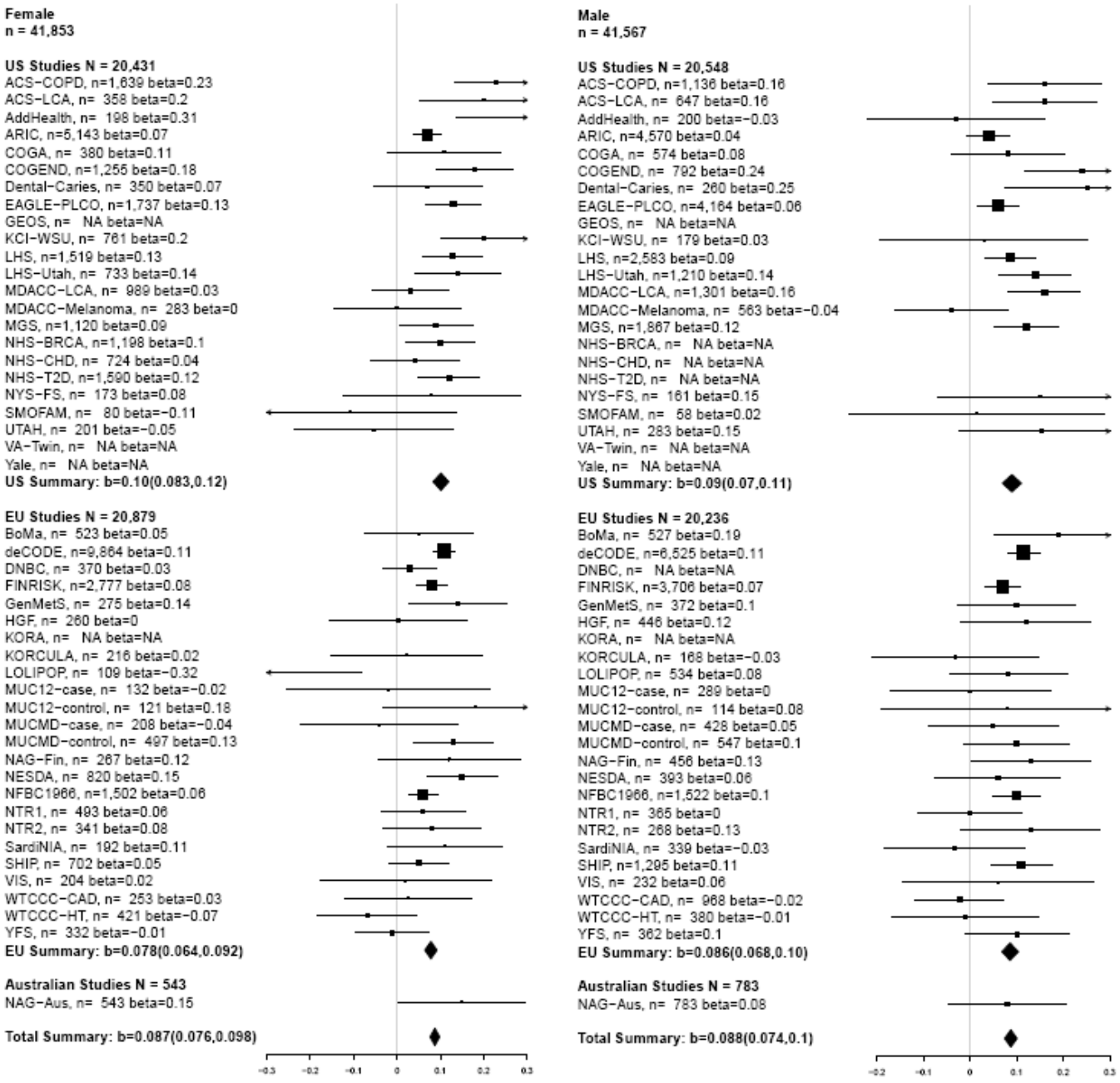
eFigure 1: Association between CPD and rs16969968 allele A across studies with subjects are stratified by age of onset of regular smoking (AOS). CPD is coded as an ordinal variable (0-3), and used as the dependent variable in a linear regression. The difference between the betas for the early and late onset subjects is 0.03 (95% CI 0.01-0.05, p=0.01), adjusted for gender and continent.



eFigure 2: Association between CPD and rs16969968 allele A across studies with subjects stratified by 20-year birth cohorts. CPD is coded as an ordinal variable (0-3), and used as the dependent variable in a linear regression.



eFigure 3: Association between CPD and rs16969968 allele A across studies with subjects stratified by gender. CPD is coded as an ordinal variable (0-3), and used as the dependent variable in a linear regression.



eFigure 4: Association between CPD and rs16969968 allele A across studies with subjects stratified by educational attainment (terminal degree of high school or less, versus terminal degree beyond high school). CPD is coded as an ordinal variable (0-3), and used as the dependent variable in a linear regression.