ORIGINAL ARTICLE

Linkage analysis of smoking initiation and quantity in Dutch sibling pairs

JM Vink[1]
AL Beem[1]
D Posthuma[1]
MC Neale[2]
G Willemsen[1]
KS Kendler[2]
PE Slagboom[3]
DI Boomsma[1]

[1]*Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands;* [2]*Departments of Psychiatry and Human Genetics, Virginia Institute for Psychiatry and Behavioral Genetics, Medical College of Virginia of Virginia Commonwealth University, Richmond, VI, USA;* [3]*Section Molecular Epidemiology, Sylvius Laboratory, Leiden Universitary Medical Centre, Leiden, The Netherlands*

Correspondence:
**Dr JM Vink, Dept of Biological Psychology, Vrije Universiteit, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands.
Tel: +31 20 4448787
Fax: +31 20 4448832
E-mail: jm.vink@psy.vu.nl**

ABSTRACT

The heritability of smoking initiation (SI) and number of cigarettes smoked (NC) was determined in 3657 Dutch twin pairs. For SI a heritability of 36% was found and for NC of 51%. Both SI and NC were also significantly influenced by environmental factors shared by family members. The etiological factors that influence these traits partly overlap. Linkage analyses were performed on data of 536 DZ twins and siblings from 192 families, forming 592 sibling pairs. Results suggested QTLs on chromosome 6 (LOD = 3.05) and chromosome 14 (LOD = 1.66) for SI and on chromosome 3 (LOD = 1.98) for NC. Strikingly, on chromosome 10 a peak was found in the same region for both SI (LOD = 1.92) and for NC (LOD = 2.29) which may partly explain the overlapping etiological factors for SI and NC.
The Pharmacogenomics Journal (2004) 4, 274–282. doi:10.1038/sj.tpj.6500255
Published online 1 June 2004

Keywords: QTL; linkage; sib pairs; heritability; smoking

INTRODUCTION

Large-scale population-based twin and family studies have shown that genetic factors contribute to individual differences in smoking behavior.[1–4] Several different, possibly correlated, dimensions of smoking behavior can be distinguished: smoking initiation (SI), number of cigarettes smoked per day and nicotine dependence (ND).[5] Koopmans *et al*[6] investigated the heritability of SI and the number of cigarettes smoked per day (NC) in adolescent Dutch twins by considering a single liability model, an independent liability model and a combined model. The combined model best described the data and showed that 39% of the variance in SI and 86% of the variance in NC was explained by genetic influences. Kendler *et al*[7] found that liabilities to SI and ND were substantially correlated but not identical, and that heritable factors played an important role in both SI and in ND.

The next step after obtaining evidence for significant heritability is to identify chromosomal regions involved in smoking behavior, either by linkage or association approaches.[8] Both human and animal studies have explored candidate genes for smoking behavior. Association studies point to dopamine receptor genes, dopamine transporter genes, cytochrome *P*450 and serotonergic genes.[9,10] Association studies have relatively high statistical power, and can detect quantitative trait loci (QTLs) with only small effects. A possible disadvantage of the candidate gene approach is that the focus is on known pathways, which may lead us to overlook genes that are etiologically important, because of our ignorance of other biological systems involved. In contrast, linkage analysis will identify chromosomal regions that harbor known and

unknown genes, although the statistical power to identify such regions is relatively low.

Linkage studies for smoking are, at present, sparse and those that were performed have used different definitions of smoking behavior. Using smoking data collected in the Collaborative Study on the Genetics of Alcoholism (COGA), the most promising linkage results were reported for chromosome 6, 9 and 14 using single point sibling pair analysis,[11] and for chromosome 3, 4, 5, 6, 9, 15 and 17 when applying a multipoint variance component method.[12] Smoking in these analyses was defined as ever having smoked daily for a month or having smoked more than 100 cigarettes during one's lifetime,[11] and as having smoked more than zero cigarettes/day for at least a year.[12] Using the same data set but focusing on heavy smoking (more than 20 cigarettes per day for at least 6 months), LOD scores greater than one were found on chromosomes 5, 9, 11 and 21.[13] Using data from two different populations, Straub et al[14] examined linkage for ND, defined as a score of 7 or over on the Fagerström Tolerance Questionnaire.[15] A genome scan was performed in a subsample of genotyped individuals from Christchurch (New Zealand), and confirmation of the results was sought by genotyping additional Christchurch subjects and linkage in an independent sample from Richmond (USA). For six of the most positive regions found in the first genome scan, located on chromosomes 2, 4, 10, 16, 17 and 18, replication was found. Thus, in the different studies, peaks have been found on most chromosomes. These results may reflect differences in populations but may also reflect the fact that different genes are involved in different aspects of smoking behavior.

In this paper we simultaneously examine SI and quantity smoked using longitudinal data from twins and siblings. Quantity smoked is often used as proxy measure for ND and both phenotypes are highly correlated. Phenotypic data were collected in a study on health related behavior of the Netherlands Twin Register.[16] First we fit a single liability, an independent liability and a combined model[17,18] to phenotypic data for SI and NC. From the model that best describes the data, heritability estimates for SI and NC are obtained. Next, we report the results from a complete genome scan on SI (ever/never smoked) and NC in a subsample of dizygotic (DZ) twin and sibling pairs.

RESULTS

Table 1 shows the distribution for smoking behavior in the genotyped sample and in the total sample. In the genotyped sample, approximately 57% of the subjects never smoked (regularly) while in the total sample 50% never smoked (regularly). The genotyped sample contained more heavy smokers than the total sample (Table 1). This is probably due to the fact that the average age (when reporting the maximum number of cigarettes per day) in the genotyped sample (DZ twins and siblings) was higher (28.3 years, SD 13.4) than in the total sample (24.7 years, SD 11.1).

Three models were fitted to the phenotypic data on SI and NC: single liability, independent liability model and a combined model.[6] The combined model gave the best description of the data. Under the combined model, several alternative explanations for familial resemblance in SI and NC were evaluated. Results are shown in Table 2a. For both SI and NC the most parsimonious model included genetic, shared environmental and unique environmental factors without sex differences. Table 2b depicts the parameter estimates. For SI, 36% of the variance in liability was

Table 1 Distribution of smoking behavior in genotyped sample ($n = 642$) and total sample ($n = 10\,623$)

	Genotyped sample		Total sample	
	N	%	N	%
Never smoked (regularly)	219	49.7	6006	56.6
Less than 1 cigarette per day	41	6.4	548	5.2
1–5 cigarettes per day	54	8.4	917	8.6
6–10 cigarettes per day	56	8.7	1038	9.8
11–20 cigarettes per day	113	17.6	1502	14.1
21–30 cigarettes per day	42	6.5	506	4.8
More than 30 cigarettes per day	17	2.6	106	1.0
Total	642	100	10\,623	100

Table 2a Model fitting results for a combined model with smoking initiation and maximum number of cigarettes smoked per day (best fitting model is given in boldface)

	Initiation	Nicotine dependence	χ^2	df	P	AIC
1	Full	Full	75.96	57	0.047	−38.04
2	ACE	Full	81.49	60	0.034	−38.51
3	AE	Full	101.45	61	0.000	−20.55
4	CE	Full	97.20	61	0.002	−24.79
5	Full	ACE	78.80	60	0.052	−41.20
6	Full	AE	84.96	61	0.023	−37.04
7	Full	CE	98.86	61	0.002	−23.14
8	**ACE**	**ACE**	**78.98**	**63**	**0.084**	**−47.01**

Full = full model with sex-dependent effects and a correlation between shared environmental factors in opposite sex twins (r_c) that is allowed to be less than 1; ACE = full model without sex differences; AE = additive genetic model; CE = shared environmental model; AIC = $\chi^2 - 2$df, this is a measure of the parsimony of the model, a lower value of AIC indicates a more parsimonious model.

Table 2b Proportion of the total variance in smoking initiation and maximum number of cigarettes smoked per day that is explained by additive genetic factors (h^2), shared environmental influences (c^2) and unique environmental influences (e^2) under the best fitting model

	h^2	c^2	e^2
Smoking initiation	0.36	0.56	0.07
Max n cigarettes	0.51	0.30	0.18
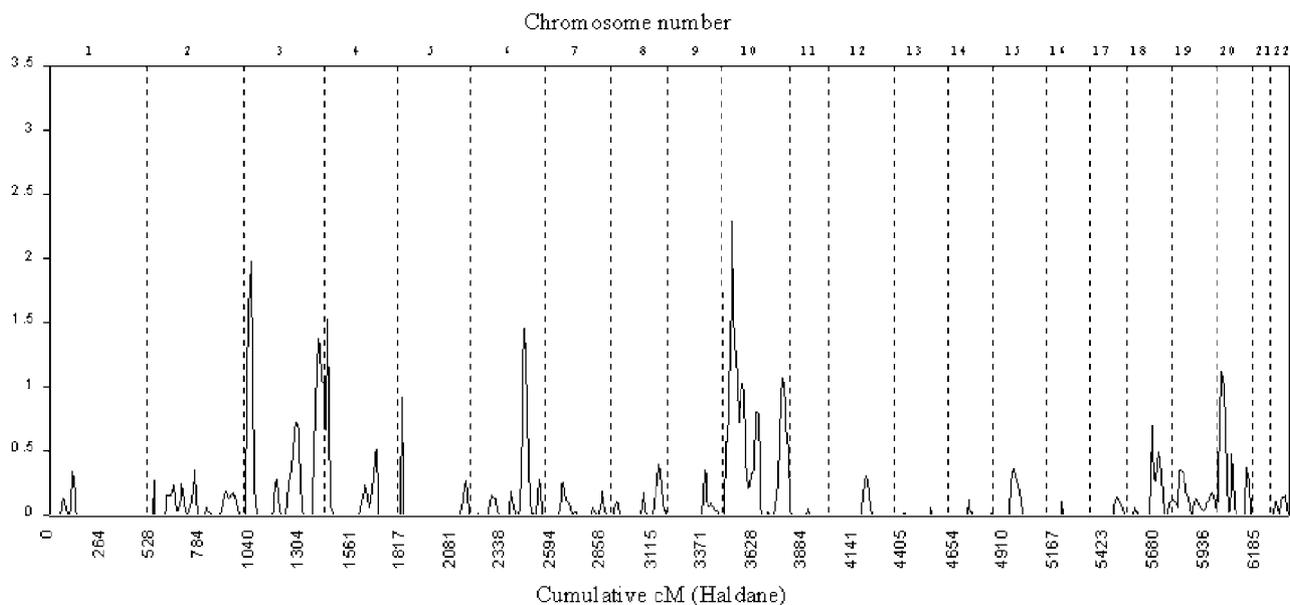
Figure 1 LOD scores across the genome for phenotype 'smoking initiation'. The cumulative Haldane centiMorgans are shown on the *x*-axis and LOD-score is shown on the *y*-axis. Chromosome number is shown at the top of the figure.
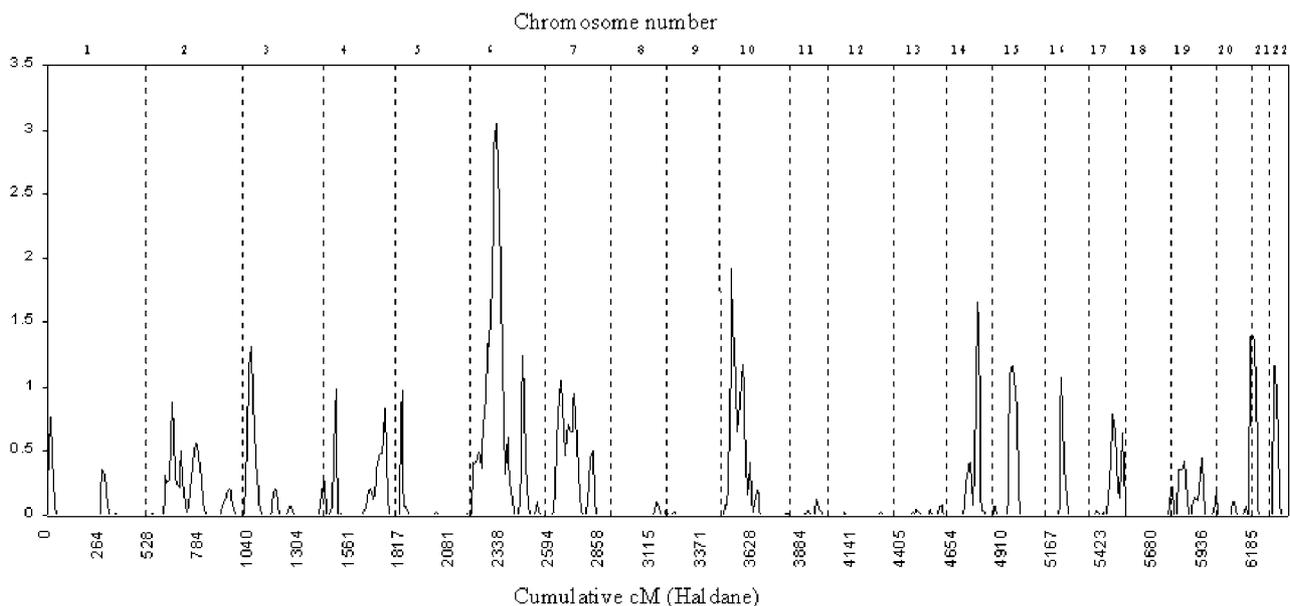
Figure 2 LOD scores across the genome for phenotype 'maximum number of cigarettes per day'. The cumulative Haldane centimorgans are shown on the *x*-axis and LOD-score is shown on the *y*-axis. Chromosome number is shown at the top of the figure.

explained by genetic factors and 56% by shared environmental factors. The remaining variance was explained by nonshared environmental factors (7%). For NC, 51% of the variance was explained by genetic, 30% by shared environmental and 18% by nonshared environmental factors.

As both SI and quantity were heritable traits, we explored both phenotypes in linkage analyses. For SI the highest LOD scores (>1.5) were found on chromosomes 6, 10 and 14 (Figure 1). For NC, the highest LOD scores (LOD >1.5) were found for chromosomes 3 and 10 (Figure 2).

Figures 3–6 show LOD score plots from the linkage analyses for SI and NC for chromosomes 3, 6, 10 and 14. For SI a peak was found on chromosome 6, in the region from approximately 98.1 to 143.3 cM (Haldane's map) with the highest peak (LOD = 3.05) at approximately 120.7 cM in the vicinity of markers D6S2410 and D6S1053. Another peak was found on chromosome 14 at approximately 143.3 cM, in the vicinity of markers Unk283 and D14S617.

For NC a peak LOD score (>1.5) was found on chromosome 3 in the region from approximately 7.5 to 15.1 cM
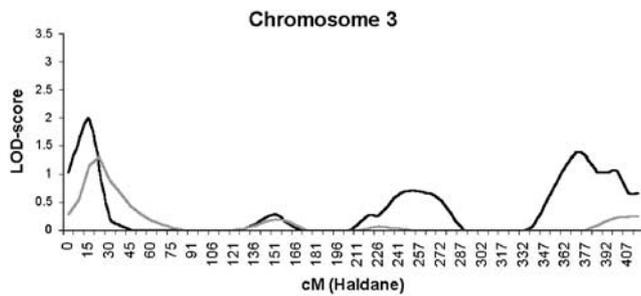
Figure 3   Linkage results for chromosome 3. Distance in Haldane cM is shown along the *x*-axis and the LOD-scores along the *y*-axis. The gray line represents the results for the phenotype 'smoking initiation' and the black line represents the results for the phenotype 'maximum number of cigarettes smoked per day'.
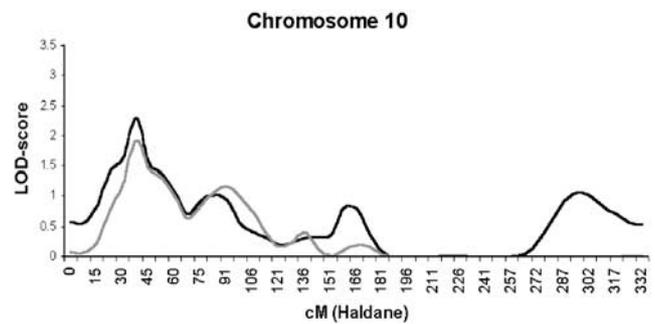
Figure 5   Linkage results for chromosome 10. Distance in Haldane cM is shown along the *x*-axis and the LOD-scores along the *y*-axis. The gray line represents the results for the phenotype 'smoking initiation' and the black line represents the results for the phenotype 'maximum number of cigarettes smoked per day'.
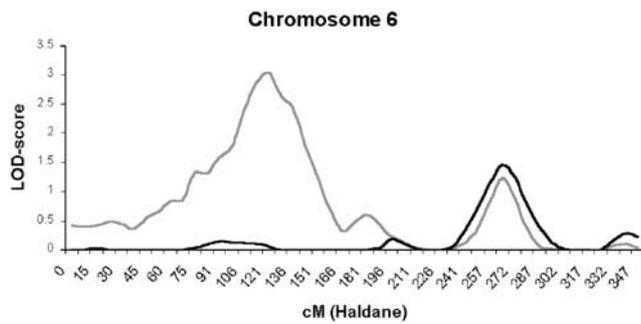
Figure 4   Linkage results for chromosome 6. Distance in Haldane cM is shown along the *x*-axis and the LOD-scores along the *y*-axis. The gray line represents the results for the phenotype 'smoking initiation' (SI) and the black line represents the results for the phenotype 'maximum number of cigarettes smoked per day' (NC).
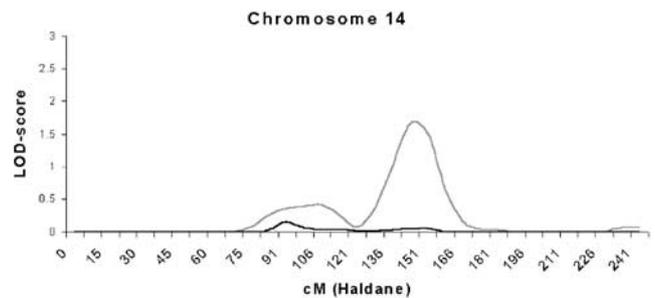
Figure 6   Linkage results for chromosome 14. Distance in Haldane cM is shown along the *x*-axis and the LOD-scores along the *y*-axis. The gray line represents the results for the phenotype 'smoking initiation' and the black line represents the results for the phenotype 'maximum number of cigarettes smoked per day'.

with the highest peak (LOD = 1.98) at approximately 15.1 cM in the vicinity of markers D3S3050 and D3S4545. No noteworthy peaks were seen for SI on this chromosome. Finally, on chromosome 10 there was a peak for both SI and NC in the same region (37.7–45.3 cM). The highest LOD scores for SI (LOD score = 1.92) and NC (LOD score = 2.29) were found at approximately 37.7 cM in the vicinity of markers D10S1412 and D10S1430.

DISCUSSION

Numerous twin studies of smoking behavior have been reported in the literature (reviewed by Heath and Madden,[1] Sullivan and Kendler[3] and Hopfer *et al*[2]). Most studies found evidence for both genetic and shared environmental contributions to familial resemblance in smoking behavior. We replicated this result in the combined genetic analyses of SI and NC and found the heritability for SI to be relatively low (36%). Based on 10 studies of SI, Sullivan and Kendler[3] reported a weighted mean heritability of 56% (range from 33 to 79%). Li *et al*[4] obtained heritability estimates of 37% for adult males and 55% for females in a meta-analysis of SI. These estimates are somewhat higher than the heritability

estimate in our study, which could be due to the fact that our sample is a relatively young one (on average 24.7 years). We observed a stronger genetic contribution to NC than to SI, as has consistently been reported others as well.[19] Furthermore, we found a significant contribution of environmental factors shared by family members to variation in SI as well as in the quantity dimension. Our sample was large, which facilitates detection of shared environmental influences. What these influences consist of remains largely unknown. They may include the effects of socioeconomic class,[20] religion,[21] social transmission, or the genetic effects of assortative mating.[22] There is significant nonrandom mating for smoking behavior[23] and the estimate of shared environment may reflect this assortment.[24]

After establishing the heritability for SI and NC, the next step was to localize chromosomal regions underlying these heritabilities by carrying out a genome scan. Linkage analyses showed peaks for SI on chromosomes 6 and 14 while for NC a peak on chromosome 3 was detected. For both SI and NC evidence for linkage was found on chromosome 10 at the same location. Those results suggest specific QTLs for SI on chromosome 6 and 14 and for NC on chromosome 3. Genetic factors common to both

phenotypes are found on chromosome 10. This is in line with the model fitting results, which suggested overlapping liabilities for SI and NC. The linkage results seem to argue against the suggestions made by Merikangas and Risch,[25] who questioned genetic studies of ND. Our genome scan suggests it is possible to find evidence for linkage for smoking behavior. This evidence consists of QTLs common to SI and NC, as well as of QTLs, which are unique to each phenotype. The unique QTLs that influence the quantity of cigarettes smoked become of importance only after an individual has crossed the threshold from nonsmoker to smoker.

There are relatively few other linkage studies of smoking behavior. All available linkage results are summarized in Table 3. Both our study in Dutch twin families and the study by Straub et al[14] on ND found positive results for chromosome 10. The multipoint analyses of Straub et al suggested a large peak at 125 cM (Kosambi's map, 216 cM on Haldane's map), while our results suggest a large peak at 38 cM (Haldane's map). Two studies found a peak on chromosome 3 (this study, Duggirala et al[12]). In a linkage study of substance dependence, Stallings et al[27] also found preliminary evidence for linkage to regions on chromosome 3 (LOD score 1.60) for the average number of dependence symptoms (ie the total symptom count across all classes of substances, including smoking, divided by the number of substances used more than five times).

However, there is no overlap between the peaks of Duggirala et al and Stallings et al and our linkage signal on chromosome 3.

The linkage result for SI on chromosome 6 was a replication of the LOD score of 1.10 reported by Duggirala et al[22] for the same region. Bergen et al[11] also reported positive findings for chromosome 6, though for another region. Both the present study and the study of Bergen et al[11] detected a peak on chromosome 14, and although the signals were not located at exactly the same position, they were in the same region. It should be noted that the location estimate of a linkage signal from a complex-trait may be many centiMorgans from the true disease locus.[28]

Thus, there is some overlap between our results using the phenotypes SI and NC and the previous genome scans looking at smoking behavior and substance dependence. The different phenotypes in the studies are probably related, for example the phenotype 'maximum number of cigarettes per day' correlates highly with the score on the Fagerström Test of ND ($r = 0.66–0.70$).[26] However, they are not the same and each phenotype is likely to be influenced by multiple genes leading to a different picture of the genetic architectures of substance use or smoking behavior. Based on a reanalysis of genome scans on alcohol dependence, drug abuse and ND, Uhl et al[29] report 15 regions that may harbor genes for substance abuse vulnerability, including regions on chromosome 10 and chromosome 3. These regions,

Table 3 Overview of positive results of linkage studies to smoking behavior

Reference	Sample	Phenotype	Chr	Position (Kosambi)	Position (Haldane)	LOD score
Straub et al[14]	Christchurch	Nicotine Dependence	2	130–180	226–325	1.50
Vink et al[26]	NETAD	Quantity smoked	3	7–13	7–15	2.42
Duggirala et al[12]	COGA	Ever smoked	3	105	176	1.71
Duggirala et al[12]	COGA	Ever smoked	4	65	99	2.17
Bierut et al[13]	COGA	Heavy smoking	5	119	204	1.12
Duggirala et al[12]	COGA	Ever smoked	5	217	399	3.20
Duggirala et al[12]	COGA	Ever smoked	6	63	95	1.10
Vink et al[26]	NETSAD	Smoking initiation	6	65–88	98–143	3.00
Bergen et al[11]	COGA	Ever smoked	6	134–165	234–295	3.00
Duggirala et al[12]	COGA	Ever smoked	9	0	0	1.14
Bierut et al[13]	COGA	Ever smoked	9	92 /116 /168	151/198/301	1.51
Bergen et al[11]	COGA	Ever smoked	9	165–170	295–305	3.00
Vink et al[26]	NETSAD	Smoking initiation	10	24–39	30–53	2.28
Vink et al[26]	NETSAD	Quantity smoked	10	24–39	30–53	2.65
Straub et al[14]	Christchurch	Nicotine Dependence	10	85–149	137–263	1.28
Bierut et al (2003)	COGA	Heavy smoking	11	87	141	1.64
Vink et al[26]	NETSAD	Smoking initiation	14	88	143	1.66
Bergen et al[11]	COGA	Ever smoked	14	95–110	156–186	3.00
Duggirala et al[12]	COGA	Ever smoked	15	135	236	1.97
Duggirala et al[12]	COGA	Ever smoked	17	20	24	2.88
Bierut et al (2003)	COGA	Ever smoked	21	0	0	1.51

Only the most positive results are shown. The z-scores and P-values in the study of Straub et al were transformed to a χ^2 distribution, which was divided by $2 \times \ln 10$ to calculate the LOD-scores. The P-values from the affected-sib-pair study of Bergen et al[11] were also transformed to a χ^2 distribution to calculate the LOD-score.
CADD = family, twin and adoption studies of Colorado Center on Antisocial Drug Dependence.
COGA = Collaborative Study on the Genetics of Alcholism (analysis are based on sample from wave 1).
NETSAD = Netherlands Twin family Study of Anxious Depression.
Christchurch = sample recruited in Christchurch, New Zealand. Inclusion criteria for a sibling pair included the presence of lifetime nicotine dependence.

however, do not overlap with the localization of the peaks in the present study.

To investigate the importance of chromosome 3, 6, 10 and 14 for SI and/or NC, the genome scan needs to be replicated in another sample using similar phenotypes. We have collected DNA samples in additional Dutch twin families selected when at least two siblings were nicotine dependent or when at least one sibling was nicotine dependent and one sibling was exposed to smoking but never smoked more than five cigarettes per day. We intend to carry out a genome scan in this selected sample.

If the positive results on chromosome 3, 6, 10 and 14 are replicated, candidate genes under the peaks can be considered for association analysis. Walton et al and Batra et al have summarized the current association studies of tobacco addiction. The strongest evidence linking particular alleles to nicotine addiction comes from studies on genetic variation in the dopaminergic system, in cytochrome P450 enzymes, the serotonin transporter gene and monoamine oxidase.[9,10] None of these genes is located on chromosomes 3, 6, 10 or 14. However, most genes that play a major role in tobacco addiction are not yet known. Localizing and identifying the genes responsible for the linkage results on chromosomes 3, 6, 10 and 14 will help to unravel another fraction of the molecular basis of tobacco addiction.

METHODS

Subjects

This study is part of an ongoing twin family study on health-related behavior in participants of the Netherlands Twin Register. Addresses of twin families were obtained from City Councils in 1991 and 1993. In later years, additional volunteer twin families also participated in the study. Surveys were mailed to twin families in 1991, 1993, 1995, 1997 and 2000.[16] A sixth wave of data collection is in progress. Twin pairs were invited to participate in all waves, and parents were invited in 1991, 1993 and 1995. Siblings of twins were included in the assessments since 1995 and spouses since 2000. Each survey collected data on smoking, other lifestyle factors, health, personality and psychopathology.

Data on smoking behavior were available for 8.039 twins and 2.529 siblings. Marker data were available for families selected for a linkage study of anxious depression.[30] Selection of extreme sibling pairs for anxious depression (according to EDAC design) was based on a composite score that included data on depression, neuroticism and anxiety. Selection of families took place at two occasions. The first selection used data from the first four surveys, the second selection from the 2000 survey. In this paper we use genome scan data from families from the first selection.[16]

If at least two offspring formed an extremely concordant or discordant sibling pair for anxious depression, the entire family, including parents and any additional siblings, were asked for a DNA sample for genotyping (MZ twin pairs were treated as a single offspring). As some families consisted of more than two siblings, this selection procedure resulted in

a (nonrandom) sample from the entire empirical distribution, not merely from its tails.

DNA Collection and Genotyping

The selected subjects were asked to provide a buccal swab for DNA isolation.[31] Of the subjects selected for the QTL study ($n = 2.724$), around 72% ($n = 1962$) returned a buccal swab and 917 subjects were genotyped over the entire genome. Selection of the first set of families for genotyping was based on family size (larger sibships) and on amount of DNA. Genotyping was conducted by the Marshfield Laboratory. For this scan the 10 cM spaced microsatellite screening set 10[32] was used with few alternative markers. On the autosomes 379 markers were measured. Pedigrees were checked for Mendelian errors with the program Unknown[33] and pedigree relationships in the entire sample with the GRR program.[34] Mendelian errors were removed by assigning missing values to the marker scores if the errors appeared incidental. One subject with an excessive error rate, two subjects with uncertain identities and two families for which apparent problems could not be resolved were removed from the analysis. This left a total of 896 subjects (606 siblings and 290 parents) from 215 families. A subset of 212 families contained two or more offspring, in which both parents were genotyped in 121 families, one parent in 43 families and no parent in 48 families (two families contained two parents and one offspring and one family contained one parent and one offspring).

Likelihoods for recombinations were checked using the program Merlin.[35] Excessive recombinations were observed for five markers indicating potential problems. Those markers were not included in the final analyses: two markers on chromosome 1 (D1S468-AFM280we5 and D1S1627-ATA25E0); two markers (D11S1985-GGAA5C04 and D11S2006-GATA46A12) in a group of five very closely or identically mapped markers on chromosome 11; and one marker on chromosome 20 (8; D20S159-UT1307). For all other recombination problems the data were cleaned using Merlin's default procedure. As a result of cleaning, 57 genotypings in 46 families were set to missing; for two subjects two marker scores were set to missing. For the linkage analyses, sibling pairs were selected for whom more than 50% of the markers were typed successfully. In total, successful genotyping data were available for 536 offspring and 278 parents from 192 families from which 592 sibling pairs were formed. Marker distances were assigned from the Decode map if available. For markers not mapped by Decode, the original distance provided on the Marshfield website[36] was transformed by linear interpolation from adjacent markers with known Decode map values.[37]

Phenotype

Data on smoking behavior were collected in every survey (1991, 1993, 1995, 1997, 2000) and most subjects participated more than once. The phenotypes were constructed by taking the answers to all surveys into account. The surveys contained several questions on smoking: 'Did you ever smoke a cigarette?', 'Did you smoke during the last

12 months?' and 'Did you smoke during the last 4 weeks?'. The answer categories were: no, a few times to try, yes. Furthermore was asked 'How many years did/do you smoke?' and another question was 'How often do you smoke now?' with the answer categories: I have never smoked regularly, I have quit smoking, I smoke less than once a week, I smoke several times a week but not every day, I smoke daily. Participants also reported the number of cigarettes they smoke per day or per week.

For the simultaneous model fitting of SI and NC, subjects were classified as never smokers (never smoked, or tried but never smoked regularly), 1–5 cigarettes per day, 6–10 cigarettes per day or more than 10 cigarettes per day (both smokers and ex-smokers).

In the linkage analyses of SI, subjects were classified as never smokers (never smoked, or tried but never smoked regularly) or ever smokers (including current smokers and ex-smokers). In the linkage analyses of NC, subjects (current smokers and ex-smokers) were classified as: never smoked regularly, <1 cigarettes per day, 1–5 cigarettes per day, 6–10 cigarettes per day, 11–20 cigarettes per day, 21–30 cigarettes per day and more than 30 cigarettes per day.

Genetic Model Fitting

Three different threshold models were fitted to the data: independent liability, single liability and a combined model (described in Koopmans et al[6]). The independent liability model assumes two independent liability dimensions for SI and NC. The single liability model postulates that the liability to smoking behavior is uni-dimensional and normally distributed with four categories (>10 cigarettes/day, 6–10 cigarettes/day, 1–5 cigarettes/day, nonsmoker). The independent liability model postulates two independent liability dimensions for initiation and quantity (>10 cigarettes/day, 6–10 cigarettes/day, 1–5 cigarettes/day, nonsmoker) that are each determined by completely separate genetic and environmental factors. The combined model includes features of both models. It consists of an initiation and a quantity dimension. Under the combined model there are two different routes to being a nonsmoker: an individual can be a nonsmoker due to genetic and/or environmental factors that influence the SI dimension or because the individual is low on the quantity dimension. The smoking behavior of the first twin was cross-classified with the smoking behavior of the second twin, resulting in 4×4 contingency tables for each zygosity group. Contingency tables were available for 595 monozygotic male (MZM), 476 dizygotic male (DZM), 1011 monozygotic female (MZF), 644 dizygotic female (DZF) and 931 dizygotic opposite sex twin pairs (DOS). Models were fitted to the contingency tables by maximum likelihood with Mx.[38]

Sources of variation that were considered in modeling the variation in liability to SI and NC were additive genetic variation (σ_a^2), shared environmental variation (σ_c^2) and unique environmental variation not shared by family members (σ_e^2). Sex-differences in variance components were tested by allowing the magnitude of the genetic and environmental effects to be different for males and females.

For all models, different thresholds were estimated for males and females, allowing for sex differences in the prevalence of smoking.

Genotyping and IBD Estimation

If a sibling pair receives the same chromosomal segment from a parent in a certain region of the genome, the pair is said to share the parent's alleles in that region identical by descent (IBD). Since offspring receive their alleles from two parents, a pair can share 0, 1 or 2 alleles IBD. IBD status is not always unambiguously known and has to be estimated using the specific allele pattern across chromosomes of two or more siblings and parents. The IBD status is usually estimated for a number of markers with (approximately) known location along the genome and is then used as the measure of genetic similarity. The estimate of the proportion of alleles shared identical by descent is referred to as $\hat{\pi}$, and is obtained as $\hat{\pi}_{ijk} = 0 \times p_{(IBD=0)_{ijk}} + 0.5 \times p_{(IBD=1)_{ijk}} + 1 \times p_{(IBD=2)_{ijk}}$ where $\hat{\pi}_{ijk}$ is the estimated proportion of alleles shared IBD between sib j and k for the ith family, and $p_{(IBD=0)_{ijk}}$, $p_{(IBD=1)_{ijk}}$ and, $p_{(IBD=2)_{ijk}}$ are the probabilities that sib j and k share 0, 1 or 2 alleles, respectively, conditional on the marker information. The probabilities of sharing zero, one or two alleles IBD at every 7.5 cM (Haldane map) over the genome were estimated with the program Merlin.[35]

Linkage Analyses

Linkage to a putative QTL was assessed by variance components analyses. We selected the sibling pairs for whom more than 50% of the markers were successfully measured (592 sib pairs). The average number of missing markers was 34 (SD = 58), which is 4.5% of the total number of markers measured.

A genome scan for SI was carried out in 592 siblings pairs (536 individuals) for whom both phenotypic and marker data were available. Different thresholds were estimated for males and females, allowing for sex differences in the prevalence of SI.

For the NC linkage analyses the never smokers were excluded. A genome scan for NC was carried out in 351 sibling pairs (424 individuals) for whom both phenotypic data and marker data were available and also included 763 MZ and 878 DZ twin pairs with only phenotypic data. Effects of sex and age were included.

Linkage analyses were performed with variance components analyses using Mx.[38] Estimates of the variance component associated with a putative QTL at or near a locus are commonly obtained from either of two approaches of modeling the contribution of the QTL to the covariance among sib pairs. The two approaches are the $\hat{\pi}$ approach and the mixture approach.[39] In the $\hat{\pi}$ approach, the covariance due to the marker or trait locus for a sib pair is modeled as a function of the $\hat{\pi}$ of the sib pair. In the mixture model, the likelihood for each sib pair is computed as the weighted sum of the likelihoods of the three models (for IBD = 0, 1 and 2) where the weights are the probabilities that the pair is IBD 0, 1 or 2. Apart from these variance components methods for linkage analyses, other statistical methods for conducting a

QTL linkage analysis have been proposed, most notably regression methods. The results presented in this paper are obtained with the $\hat{\pi}$ approach.[40]

For the dichotomous trait SI a threshold model with one threshold was used. The model assumes an underlying liability to SI that is a function of genetic and environmental factors. Subjects are affected if they cross a threshold.[41] Different thresholds were estimated for males and females, allowing for sex differences in prevalence of SI.

For NC the means were modeled according to the formula: $y_{ij} = \mu + \beta_1 \ Age_{ij} + \beta_2 \ Sex_{ij} + e_{ij}$, where y_{ij} is the observed phenotype for sibling j in the ith family, μ denotes the grand mean, β_1 represents the regression coefficient for age, β_2 represents the female deviation, age_{ij} and sex_{ij} represent the age and sex (male = 0 and female = 1) respectively of sib j from the ith family, and e_{ij} represents the residual term that is not explained by the fixed effects of age and sex.

For SI and NC, the variance in liability and the phenotypic variance of the residual term, respectively, were decomposed into additive genetic variance (σ_a^2), shared environmental variance (σ_c^2), variance due to nonshared environmental influences (σ_e^2), and variance due to the QTL (σ_q^2).[42] The variance–covariance matrix for pairs j, k of the ith family, Ω_{ijk} is given by $\sigma_a^2 + \sigma_c^2 + \sigma_q^2 + \sigma_e^2$ if $j = k$ and by $0.5\sigma_a^2 + \sigma_c^2 + \hat{\pi}_{ijk}\sigma_q^2$ if $j \neq k$. The analyses also included the phenotypic data from MZ and DZ twin pairs for whom no genotypic data were available to allow the distinction between background additive genetic and other familial effects (1596 MZ and 1943 DZ twin pairs for SI, 763 MZ and 878 DZ twin pairs for NC). For the twin pairs who were not genotyped, covariances were modeled as $\sigma_g^2 + \sigma_c^2$ for MZ pairs and $0.5\sigma_g^2 + \sigma_c^2$ for DZ pairs, where $\sigma_g^2 = \sigma_a^2 + \sigma_q^2$. Significance of genetic variation due to the QTL was evaluated by the likelihood ratio test, from which the LOD score can be calculated by dividing the test statistic χ^2 by 2ln 10 ($\sim$4.6).[43] In addition to the $\hat{\pi}$ approach for which the results are reported in this paper, linkage analyses were also carried out using a mixture distribution model and a regression approach in Merlin. The three methods yielded similar results.

DUALITY OF INTEREST

The authors certify there is no duality of interest that could be perceived to bias this work.

ACKNOWLEDGEMENTS

REFERENCES

1 Heath AC, Madden PAF. Genetic influences on smoking behavior. *Behavior Genetic Approaches in Behavioral Medicine*. Plenum Press: New York 1995; pp 45–66.
2 Hopfer CJ, Crowley TJ, Hewitt JK. Review of twin and adoption studies of adolescent substance use. *J Am Acad Child Adolesc Psychiatry* 2003; **42**: 710–719.
3 Sullivan PF, Kendler KS. The genetic epidemiology of smoking. *Nicotine Tobacco Res* 1999; **1**: S51–S57.
4 Li MD, Cheng R, Ma JZ, Swan GE. A meta-analysis of estimated and environmental effects on smoking behavior in male and female adult twins. *Addiction* 2003; **98**: 23–31.
5 Mayhew KP, Flay BR, Mott JA. Stages in the development of adolescent smoking. *Drug Alcohol Depend* 2000; **59**: S61–S81.
6 Koopmans JR, Slutske W, Heath AC, Neale MC, Boomsma DI. The genetics of smoking initiation and quantity smoked in Dutch adolescent and young adult twins. *Behav Genet* 1999; **29**: 383–393.
7 Kendler KS, Neale MC, Sullivan P, Corey LA, Gardner CO, Prescott CA. A population-based twin study in women of smoking initiation and nicotine dependence. *Psychol Med* 1999; **29**: 299–308.
8 Vink JM, Boomsma DI. Gene finding strategies. *Biol Pschol* 2002; **61**: 53–71.
9 Walton R, Johnstone E, Munafò M, Neville M, Griffiths S. Genetic clues to the molecular basis of tobacco addiction and progress towards personalized therapy. *Trends Mol Med* 2001; **7**: 70–76.
10 Batra V, Patkar AA, Berrettini WH, Weinstein SP, Leone FT. The genetic determinants of smoking. *CHEST* 2003; **123**: 1730–1739.
11 Bergen AW, Korczak JF, Wieissbecker KA, Goldstein AM. A genome-wide search for loci contributing to smoking and alcoholism. *Genet Epidemiol* 1999; **17**: S55–S60.
12 Duggirala R, Almasy L, Blangero J. Smoking behavior is under the influence of a major quantitative trait locus on human chromosome 5q. *Genet Epidemiol* 1999; **17**: S139–S144.
13 Bierut LJ, Rice JP, Goate A, Hinrichs AL, Saccone NL, Foroud T *et al*. A genomic scan for habitual smoking in families of alcholics: common and specific genetic factors in substance dependence. *Am J Med Genet* 2004; **124**: 19–27.
14 Straub RE, Sullivan PF, Ma Y, Myakishev MV, Harris-Kerr C, Wormley B *et al*. Susceptibility genes for nicotine dependence: a genome scan and followup in an independent sample suggest that regions on chromosomes 2, 4, 10, 16, 17 and 18 merit further study. *Mol Psychiatry* 1999; **4**: 129–144.
15 Fagerstrom KO. Measuring the degree of physical dependence to tobacco smoking with reference to individualization of treatment. *Addicit Behav* 1978; **3**: 235–241.
16 Boomsma DI, Vink JM, Beijsterveldt CEMv, Geus de EJC, Beem AL, Mulder EJCM *et al*. Netherlands Twin Register: a focus on longitudinal research. *Twin Res* 2002; **5**: 401–406.
17 Heath AC, Martin NC. Genetic models for the natural history of smoking: evidence for a genetic influence on smoking persistence. *Addict Behav* 1993; **18**: 19–34.
18 Heath AC, Martin N, Lynskey MT, Todorov AA, Madden PAF. Estimating two-stage models for genetic influences on alcohol, tobacco or drug use initiation and dependence vulnerability in twin and family data. *Twin Res* 2002; **5**: 113–124.
19 Madden PAF, Pedersen NL, Kaprio J, Koskenvuo MJ, Martin NG. The epidemiology and genetics of smoking initiation and persistence: crosscultural comparisons of twin study results. *Twin Res* 2004; **7**: 82–97.
20 Barbeau EM, Krieger N, Soobader MJ. Working class matters: socio-economic disadvantage, race/ethnicity, gender, and smoking in NHIS 2000. *Am J Public Health* 2004; **94**: 269–278.
21 Koopmans JR, Slutske WS, van Baal GC, Boomsma DI. The influence of religion on alcohol use initiation: evidence for genotype X environment interaction. *Behav Genet* 1999; **29**: 445–453.
22 Eaves LJ, Eysenck HJ, Martin NG. *Genes, Culture and Personality*. Academic Press: London 1989.
23 Boomsma DI, Koopmans JR, Doornen vLJP, Orleke JF. Genetic and social influences on starting to smoke: a study of Dutch adolescent twins and their parents. *Addiction* 1994; **89**: 219–226.
24 Willemsen G, Vink JM, Boomsma DI. Assortative mating may explain spouses' risk of same disease. *Br Med J* 2003; **326**: 396.

25 Merikangas K, Risch N. Genomic priorities and public health. *Science* 2003; **302**: 599–601.

26 Vink JM, Willemsen G, Beem AL, Boomsma DI. The fagerstrom test for nicotine dependence in a Dutch sample of daily smokers, ex-smokers. *Addict Behav*, in press.

27 Stallings MC, Corley RP, Hewitt JK, Krauter KS, Lessem JM, Mikulich SK *et al.* A genome-wide search for quantititative trait loci influencing substance dependence in vulnerability in adolescence. *Drug Alcohol Depend* 2003; **70**: 295–307.

28 Roberts SB, MacLean CJ, Neale MC, Eaves LJ, Kendler KS. Replication of linkage studies of complex traits: an examination of variation in location estimates. *Am J Hum Genet* 1999; **65**: 876–884.

29 Uhl GR, Liu QR, Naiman D. Substance abuse vulnerability loci: converging genome scanning data. *Trends Genet* 2002; **18**: 420–425.

30 Boomsma DI, Beem AL, Berg van den M, Dolan CV, Koopmans JR, Vink JM *et al.* Netherlands twin family study of anxious depression (NETSAD). *Twin Res.* 2000; **3**: 323–334.

31 Meulenbelt I, Droog S, Trommelen GJM, Boomsma DI, Slagboom PJ. High yield noninvasive human genomic DNA isolation method for genetic studies in geographically dispersed families and populations. *Am J Hum Genet* 1995; **57**: 1252–1254.

32 Yuan B, Vaske D, Weber JL, Beck J, Sheffield VC. Improved set of short-tandem-repeat polymorphisms for screening the human genome. *Am J Hum Genet* 1997; **60**: 459–460.

33 Schäffer AA. Faster linkage analysis computations for pedigrees with loops or unused alleles. *Hum Hered* 1996; **46**: 226–235.

34 Abecasis GR, Cherny SS, Cookson WO, Cardon LR. GRR: graphical representation of relationship errors. *Bioinformatics* 2001; **17**: 742–743.

35 Abecasis GR, Cherny SS, Cookson OC, Cardon L. Merlin-rapid analyses of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002; **30**: 97–101.

36 Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic map: individual and sex-specific variation in recombination. *Am J Hum Genet* 1998; **63**: 891–899.

37 Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardson B *et al.* A high-resolution recombination map of the human genome. *Nat Genet* 2002; **31**: 241–247.

38 Neale MC, Boker SM, Xie G, Maes HH. *Mx: Statistical Modeling*, Richmond, VA 23298, VCU Box 900126 1999.

39 Neale MC. QTL mapping with sib-pairs: the flexibility of Mx. In: Spector TD, Snieder H, MacGregor AJ (eds). *Advances in Twin and Sib-Pair Analysis*. Greenwich Medical Media Ltd: London 2000; pp 219–243.

40 Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* 1994; **54**: 535–543.

41 Falconer DS, Mackay TFC. Treshold Characters (Chapter 18). In *Quantitative Genetics*, editors Falconer DS, Mackay TFC. Longman Group Ltd: Essex 1996.

42 Fulker DW, Cherny SS. An improved multipoint sib-pair analysis of quantitative traits. *Behav Genet* 1996; **26**: 527–531.

43 Sham P. *Statistics in Human Genetics*. Oxford University Press: New York 1998.