

## Sequence variants at *CHRNA3–CHRNA6* and *CYP2A6* affect smoking behavior

Thorgeir E Thorgeirsson<sup>1,2\*</sup>, Daniel F Gudbjartsson<sup>1</sup>, Ida Surakka<sup>3,4</sup>, Jacqueline M Vink<sup>5</sup>, Najaf Amin<sup>6</sup>, Frank Geller<sup>1</sup>, Patrick Sulem<sup>1</sup>, Thorunn Rafnar<sup>1</sup>, Tõnu Esko<sup>7,8</sup>, Stefan Walter<sup>6</sup>, Christian Gieger<sup>9</sup>, Rajesh Rawal<sup>9</sup>, Massimo Mangino<sup>10</sup>, Inga Prokopenko<sup>11,12</sup>, Reedik Mägi<sup>11,12</sup>, Kaisu Keskitalo<sup>13</sup>, Iris H Gudjonsdottir<sup>1</sup>, Solveig Gretarsdottir<sup>1</sup>, Hreinn Stefansson<sup>1</sup>, John R Thompson<sup>14</sup>, Yurii S Aulchenko<sup>6</sup>, Mari Nelis<sup>7,8</sup>, Katja K Aben<sup>15,16</sup>, Martin den Heijer<sup>15,17</sup>, Asger Dirksen<sup>18</sup>, Haseem Ashraf<sup>18</sup>, Nicole Soranzo<sup>10,19</sup>, Ana M Valdes<sup>10</sup>, Claire Steves<sup>10</sup>, André G Uitterlinden<sup>6,20</sup>, Albert Hofman<sup>6</sup>, Anke Tönjes<sup>21,22</sup>, Peter Kovacs<sup>23</sup>, Jouke Jan Hottenga<sup>5</sup>, Gonneke Willemsen<sup>5</sup>, Nicole Vogelzang<sup>24</sup>, Angela Döring<sup>9</sup>, Norbert Dahmen<sup>25</sup>, Barbara Nitz<sup>9</sup>, Michele L Pergadia<sup>26</sup>, Berta Saez<sup>27</sup>, Veronica De Diego<sup>28</sup>, Victoria Lezcano<sup>28</sup>, Maria D Garcia-Prats<sup>29</sup>, Samuli Ripatti<sup>3,4</sup>, Markus Perola<sup>4</sup>, Johannes Kettunen<sup>19</sup>, Anna-Liisa Hartikainen<sup>30</sup>, Anneli Pouta<sup>31</sup>, Jaana Laitinen<sup>32</sup>, Matti Isohanni<sup>30</sup>, Shen Huei-Yi<sup>3,4</sup>, Maxine Allen<sup>11</sup>, Maria Krestyaninova<sup>33</sup>, Alistair S Hall<sup>34</sup>, Gregory T Jones<sup>35</sup>, Andre M van Rij<sup>35</sup>, Thomas Mueller<sup>36</sup>, Benjamin Dieplinger<sup>36</sup>, Meinhard Haltmayer<sup>36</sup>, Steinn Jonsson<sup>37</sup>, Stefan E Matthiasson<sup>38</sup>, Hogni Oskarsson<sup>39</sup>, Thorarinn Tyrfinngsson<sup>40</sup>, Lambertus A Kiemeny<sup>15,16,41</sup>, Jose I Mayordomo<sup>42</sup>, Jes S Lindholt<sup>43</sup>, Jesper Holst Pedersen<sup>44</sup>, Wilbur A Franklin<sup>45</sup>, Holly Wolf<sup>46</sup>, Grant W Montgomery<sup>47</sup>, Andrew C Heath<sup>26</sup>, Nicholas G Martin<sup>47</sup>, Pamela A F Madden<sup>26</sup>, Ina Giegling<sup>48</sup>, Dan Rujescu<sup>48</sup>, Marjo-Riitta Järvelin<sup>31,49–51</sup>, Veikko Salomaa<sup>4</sup>, Michael Stumvoll<sup>21</sup>, Tim D Spector<sup>10</sup>, H-Erich Wichmann<sup>9,52,53</sup>, Andres Metspalu<sup>7,8</sup>, Nilesh J Samani<sup>54</sup>, Brenda W Penninx<sup>24</sup>, Ben A Oostra<sup>55</sup>, Dorret I Boomsma<sup>5</sup>, Henning Tiemeier<sup>6</sup>, Cornelia M van Duijn<sup>6</sup>, Jaakko Kaprio<sup>3,13,56</sup>, Jeffrey R Gulcher<sup>1</sup>, The ENGAGE Consortium<sup>57</sup>, Mark I McCarthy<sup>11,12</sup>, Leena Peltonen<sup>3,19</sup>, Unnur Thorsteinsdottir<sup>1,38</sup> & Kari Stefansson<sup>1,38</sup>

Smoking is a common risk factor for many diseases<sup>1</sup>. We conducted genome-wide association meta-analyses for the number of cigarettes smoked per day (CPD) in smokers ( $n = 31,266$ ) and smoking initiation ( $n = 46,481$ ) using samples from the ENGAGE Consortium. In a second stage, we tested selected SNPs with *in silico* replication in the Tobacco and Genetics (TAG) and Glaxo Smith Kline (Ox-GSK) consortia cohorts ( $n = 45,691$  smokers) and assessed some of those in a third sample of European ancestry ( $n = 9,040$ ). Variants in three genomic regions associated with CPD ( $P < 5 \times 10^{-8}$ ), including previously identified SNPs at 15q25 represented by rs1051730[A] (effect size = 0.80 CPD,  $P = 2.4 \times 10^{-69}$ ), and SNPs at 19q13 and 8p11, represented by rs4105144[C] (effect size = 0.39 CPD,  $P = 2.2 \times 10^{-12}$ ) and rs6474412-T (effect size = 0.29 CPD,  $P = 1.4 \times 10^{-8}$ ), respectively. Among the genes at the two newly associated loci are genes encoding nicotine-metabolizing enzymes (*CYP2A6* and *CYP2B6*) and nicotinic acetylcholine receptor subunits (*CHRNA3* and *CHRNA6*), all of which have been highlighted in previous studies of smoking and nicotine dependence<sup>2–4</sup>. Nominal associations with lung cancer were

observed at both 8p11 (rs6474412[T], odds ratio (OR) = 1.09,  $P = 0.04$ ) and 19q13 (rs4105144[C], OR = 1.12,  $P = 0.0006$ ).

Smoking behavior and nicotine dependence are considered to be influenced by genetics<sup>5</sup>. Although environmental influences play a strong role in the initiation of smoking<sup>6</sup>, the heritability of smoking persistence, smoking quantity and nicotine dependence has been high in most twin studies<sup>6,7</sup>. Sequence variants within a cluster of genes on chromosome 15q25 that encode nicotinic acetylcholine receptors (nAChRs) have recently been shown to associate with CPD<sup>8,9</sup>, nicotine dependence<sup>3,8</sup> and smoking-related diseases such as lung cancer<sup>8,10,11</sup>, peripheral arterial disease (PAD)<sup>8</sup> and chronic obstructive pulmonary disease (COPD)<sup>12</sup>.

To search for additional common variants associated with smoking behavior, we performed meta-analyses of genome-wide association (GWA) studies, mainly using samples of European ancestry from the ENGAGE consortium (see URLs) and focusing on two smoking phenotypes: CPD and smoking initiation. The smoking initiation analysis was performed with a total of 30,431 ever-smokers and 16,050 never-smokers, using data from 12 GWA studies: Corogene, deCODE,

\*Complete list of author affiliations appears at the end of this paper.

**Table 1 Association of markers in four chromosomal regions with CPD**

SNP	Allele		Freq.	Chr.	Position	ENGAGE <sup>a</sup>		TAG and Ox-GSK <sup>b</sup>		ISL-AUS-DEN-GER-SPA <sup>c</sup>		Combined <sup>d</sup>				
	Effect	Other				Effect (s.e.m.)	P	Effect (s.e.m.)	P	Effect (s.e.m.)	P	n	Effect (s.e.m.)	P	P <sub>het</sub>	I <sup>2</sup>
rs1051730	A	G	0.339	15q25	76,681,394	0.84 (0.07)	2.1 × 10 <sup>-33</sup>	0.78 (0.06)	5.6 × 10 <sup>-38</sup>			76,972	0.80 (0.05)	2.4 × 10 <sup>-69</sup>	0.035	32
rs6474412	T	C	0.784	8p11	42,669,655	0.31 (0.08)	1.7 × 10 <sup>-4</sup>	0.30 (0.07)	2.6 × 10 <sup>-5</sup>	0.19 (0.18)	0.30	84,956	0.29 (0.05)	1.4 × 10 <sup>-8</sup>	0.24	13
rs13280604	A	G	0.784	8p11	42,678,743	0.31 (0.08)	1.2 × 10 <sup>-4</sup>	0.30 (0.07)	2.7 × 10 <sup>-5</sup>			76,670	0.31 (0.05)	1.3 × 10 <sup>-8</sup>	0.24	14
rs215614	G	A	0.356	7p14	32,313,860	0.38 (0.07)	2.4 × 10 <sup>-8</sup>	0.17 (0.06)	3.6 × 10 <sup>-3</sup>	-0.15 (0.16)	0.35	86,259	0.22 (0.04)	2.1 × 10 <sup>-7</sup>	0.018	34
rs215605	G	T	0.357	7p14	32,303,490	0.39 (0.07)	1.7 × 10 <sup>-8</sup>	0.17 (0.06)	3.5 × 10 <sup>-3</sup>			77,012	0.26 (0.04)	5.4 × 10 <sup>-9</sup>	0.12	22
rs7937	T	C	0.560	19q13	45,994,546	0.34 (0.07)	2.2 × 10 <sup>-7</sup>	0.19 (0.06)	1.1 × 10 <sup>-3</sup>	0.19 (0.14)	0.17	86,319	0.24 (0.04)	2.4 × 10 <sup>-9</sup>	0.45	1
rs1801272	A	T	0.961	19q13	46,046,373	1.08 (0.27)	7.0 × 10 <sup>-5</sup>	0.41 (0.24)	8.4 × 10 <sup>-2</sup>			66,380	0.68 (0.18)	1.1 × 10 <sup>-4</sup>	0.50	0
rs4105144	C	T	0.704	19q13	46,050,464	0.59 (0.10)	1.2 × 10 <sup>-9</sup>	0.31 (0.08)	5.8 × 10 <sup>-5</sup>	0.27 (0.15)	0.069	83,317	0.39 (0.06)	2.2 × 10 <sup>-12</sup>	0.51	0
rs7260329	G	A	0.687	19q13	46,213,478	0.43 (0.07)	1.1 × 10 <sup>-9</sup>	0.06 (0.06)	0.36	0.08 (0.16)	0.65	86,092	0.20 (0.04)	5.5 × 10 <sup>-6</sup>	0.12	21

Results are given for the ENGAGE analysis, the *in silico* replication obtained by combining results from TAG and Ox-GSK and the results of single-SNP assay replications in samples from Iceland, Australia, Denmark, Germany and Spain (ISL-AUS-DEN-GER-SPA). Samples that were both in ENGAGE and either TAG or Ox-GSK were removed before obtaining the combined *in silico* results. Shown are the number of smokers (n), the effect allele and the other allele, the allele frequencies (freq.), the chromosome number (chr.) and position, the estimated allelic effects on CPD and their standard errors in CPD (effect and s.e.m.), the P value for the test of association, the P value for the test for heterogeneity in effect size (P<sub>het</sub>) and an estimate of the proportion of total variation in study estimates that is due to heterogeneity (I<sup>2</sup>).

<sup>a</sup>Meta-analysis, n = 31,266. <sup>b</sup>*In silico* replication, n = 45,691. <sup>c</sup>Direct genotyping, n = 9,040. <sup>d</sup>n = 85,997.

The Estonian Genome Project of University of Tartu (EGPUT), the Erasmus Rucphen Family study (ERF), the Northern Finland Birth Cohorts (NFBC), the KORA study (Kooperative Gesundheitsforschung in der Region Augsburg), the Netherlands Twin Registry and Netherlands Study of Depression and Anxiety (NTR-NESDA), the Rotterdam study, the Sorbs study, the United Kingdom Twin Study (TwinUK) and the Wellcome Trust Case Controls Consortium Study of Coronary Heart Disease (WTCCC-CAD). For CPD, we combined data from these same 12 GWA studies, plus subjects from the Nijmegen Lung and Bladder Cancer sample (NL-BLC) study, for a total of 31,266 subjects. Information on the meta-analysis studies for CPD and smoking initiation is provided in **Supplementary Table 1**, the **Supplementary Note** and in the Online Methods. After genomic control correction of each component study, we combined association data for ~2,500,000 imputed and genotyped autosomal SNPs with a fixed-effects additive meta-analysis using the inverse-variance method for CPD and smoking initiation. Quantile-quantile plots for CPD, excluding markers in the 15q25 region, displayed only modest inflation of the  $\chi^2$ -test statistic (genomic control inflation factor ( $\lambda_{GC}$ ) = 1.02) (**Supplementary Fig. 1a**). In addition to the 15q25 locus, SNPs at two loci, 19q13 and 7p14, were genome-wide significant (GWS) for CPD ( $P < 5 \times 10^{-8}$ ) in the meta-analysis data. The quantile-quantile plot for smoking initiation displayed weak inflation of the  $\chi^2$ -test statistic ( $\lambda_{GC} = 1.03$ ) and no GWS associations (**Supplementary Fig. 1b**).

We selected 15 regions for smoking initiation totaling 277 SNPs and 14 regions for CPD totaling 443 SNPs for *in silico* replication in samples from the TAG and the Ox-GSK consortia (see accompanying papers published in this issue<sup>13,14</sup>) (**Supplementary Table 2**). For CPD, we included a region on chromosome 8p11 on the basis of (i) its large number of SNPs showing suggestive associations with CPD, (ii) the strong candidacy of genes in the region (*CHRNA6* and *CHRNAB3*, encoding nAChR subunits  $\alpha 6$  and  $\beta 3$ ) and (iii) previous suggestive evidence for association between SNPs within this region and nicotine dependence<sup>2,3</sup>.

In addition to the 15q25 locus, three new loci, 7p14, 8p11 and 19q13, were found to be GWS for CPD after combining the results from the ENGAGE meta-analysis set with those of TAG and Ox-GSK (**Fig. 1**, **Table 1** and **Supplementary Table 2**). No GWS associations for the selected smoking initiation regions were observed in the combined analysis of the meta-analysis and the *in silico* data (**Supplementary Table 2**).

For further confirmation of the CPD association signals at the 7p14, 8p11 and 19q13 loci, selected markers from these regions were genotyped in additional samples (n = 9,040) from Iceland, Australia, Denmark, Germany and Spain (**Table 1**). The markers at 8p11 and at the 19q13 loci had effects in the same direction but the marker on 7p14 did not (**Table 1**). After combining these data with ENGAGE results and the *in silico* replication, the 8p11 and the 19q13 loci remained GWS but the 7p14 locus did not (**Table 1**). The CPD association results for the SNPs in **Table 1** for each study are presented in **Supplementary Table 3** and **Supplementary Figure 2**.

Nominally significant heterogeneity in the strength of association with CPD was observed for rs1051730 at 15q25 ( $P = 0.035$ , fraction of variation due to heterogeneity ( $I^2$ ) = 32%) and rs215614 at 7p14 ( $P = 0.018$ ,  $I^2 = 34\%$ ). Given previous and current evidence, it is highly likely that there is a true association between 15q25 and CPD. Therefore, its heterogeneity must be due to differences between the study populations used, such as different CPD information ascertainment, different types of cigarettes being used, different phenotypic and demographic ascertainment strategies or different genetic structures. The heterogeneity observed at 7p14 could be caused by a false positive finding or, if it is indeed a true positive, some combination of the 'winner's curse', which inflates initial effect estimates, and the same sort of differences driving the heterogeneity at 15q25.

The strongest associations observed with CPD in the combined analysis were with SNPs within the previously identified region on chromosome 15q25 (rs1051730[A],  $P = 2.4 \times 10^{-69}$ , effect size = 0.80 CPD) (**Fig. 1** and **Table 1**). We searched for additional association signals in the 15q25 region which were not accounted for by rs1051730 by performing linear regression using the rs1051730 allele count as a covariate in a subset of the ENGAGE samples (n = 23,089). The residual signals were mostly tagged by two SNPs in relatively low linkage disequilibrium (LD), rs2869046[T] ( $P = 4.8 \times 10^{-5}$ , effect size = 0.5) and rs2036534[T] ( $P = 9.1 \times 10^{-5}$ , effect size = 0.50 CPD) ( $r^2 = 0.080$  and  $D' = 0.65$  in the HapMap CEU samples) (**Supplementary Fig. 3**). These two SNPs are also in fairly weak LD with rs1051730 ( $r^2 = 0.12$ ,  $D' = 0.49$  and  $r^2 = 0.18$ ,  $D' = 1.05$  in the HapMap CEU samples for rs2869046 and rs2036534, respectively). These data suggest that either the three variants (rs2869046[T], rs2036534[T] and rs1051730[A]) represent independent association signals or that a variant(s) captured by a combination of these three variants remains to be identified. A SNP, rs578776, in LD with rs2036534 ( $r^2 = 0.74$ ,  $D' = 0.95$  in the









