Institute for Molecular Medicine Finland, FIMM
University of Helsinki
Helsinki, Finland

NICOTINE DEPENDENCE – IDENTIFYING THE CONTRIBUTION OF SPECIFIC GENES

Jenni Hällfors

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Medicine of the University of Helsinki, for public examination in Seminar Room 3, Biomedicum Helsinki, on 2 June 2017, at 12 noon.

Helsinki 2017

FIMM

UNIVERSITY OF HELSINKI
DOCTORAL PROGRAMME
DP
BM
IN BIOMEDICINE
DOCTORAL SCHOOL IN HEALTH SCIENCES

UNIVERSITY OF HELSINKI

ISBN 978-951-51-3261-1 (pbk.)
ISBN 978-951-51-3262-8 (PDF)

Unigrafia Oy
Helsinki, Finland 2017

Supervisors

Professor Jaakko Kaprio
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Helsinki, Finland

Department of Public Health
University of Helsinki, Helsinki, Finland

Adjunct Professor Anu Loukola
Institute for Molecular Medicine Finland (FIMM),
University of Helsinki, Helsinki, Finland

Reviewers

Adjunct Professor Maija Wessman
Folkhälsan Institute of Genetics,
Helsinki, Finland

Institute for Molecular Medicine Finland (FIMM),
University of Helsinki, Helsinki, Finland

Assistant Professor, Dr. Rohan Palmer
Department of Psychology,
Emory University,
Atlanta, GA, USA

Opponent

Adjunct Professor Päivi Onkamo
Faculty of Biosciences,
University of Helsinki, Helsinki, Finland

For my grandfather

ABSTRACT

Cigarette smoking is a worldwide health burden and ranks among the greatest public health tragedies causing several diseases that have already reached epidemic levels. Chronic obstructive pulmonary disease, heart disease and several cancers remain the main avoidable causes for premature deaths. Among smokers, tobacco use is largely motivated by nicotine dependence (ND). High heritability estimates obtained from family, twin and adoption studies clearly state the unequivocal contribution of genetic factors in smoking behavior. However, smoking, as any other form of substance use, is influenced by the complex interplay between genetic vulnerability, environmental risk factors and individual characteristics. Despite the high heritability estimates of ND, the underlying genetic factors remain largely unknown. Recent genome-wide association studies (GWAS) have identified only a few genetic variants contributing to this genetic liability. This work aimed to identify the contribution of genes and specific genetic variants predisposing individuals to smoking behavior and ND, and to further assess their ability to explain inter-individual differences in other smoking-related comorbidities and traits, such as alcohol use.

Building upon previous linkage findings on chromosome 20, a Finnish twin family sample (*N*=759 subjects from 206 families) enriched for heavy smoking was used to map susceptibility genes for smoking-related phenotypes. Using a dense set of markers on chromosome 20, the previous linkage finding (Loukola et al., 2008) was replicated with an ND measure diagnosed by the Diagnostic and Statistical Manual of Mental Disorders (DSM), 4th edition (DSM-IV) criteria (maximum LOD score 3.8 on 20p11). The finding was further extended using a larger sample (N=1,302 subjects from 357 families). Sex-stratified analyses provided a clear distinction between genetic elements on chromosome 20 that influence ND in adult male and female smokers. This study highlights linkage between chromosome 20 and ND in Finnish adult smokers.

The same Finnish twin family sample was used in a GWAS for smoking behavior (*N*= 1,715). The data was imputed using a stochastic approach that combined the 1000 Genomes Phase I reference panel together with a whole genome sequence-based Finnish reference panel obtained from the Sequencing Initiative Suomi (SISu) project. The study yielded a genome-wide significant association (P=8.5x10^{-9}) on 16p12.3 with smoking quantity defined as self-reported cigarettes smoked per day (CPD). Several highly correlating single nucleotide polymorphisms (SNPs) within the association region map to an intergenic locus near gene *CLEC19A*. In addition, association (P=6.6x10^{-8}) was detected on 11p15.5 with nicotine withdrawal symptoms assessed by the DSM-IV criteria. The finding pinpointed genes *AP2A2* and *MUC6*. These findings on chromosomes 16 and 11 highlight the role of neurotrophin signaling pathway in smoking behavior.

The most robust smoking behavior locus to date has been mapped to chromosome 15q24-q25, which harbors a gene cluster (*CHRNA5-CHRNA3-CHRNB4*) encoding nicotinic acetylcholine receptor (nAChR) subunits α5, α3 and β4. Three of the most intensively studied SNPs on the locus were analyzed by combining two large and individual Finnish population-based samples, The National FINRISK Study and The Health 2000 Survey, into one sample (*N*=8,356). Using an association analyses approach, the study replicated an earlier finding reporting association between three distinct

loci on 15q25.1 and CPD (Saccone et al., 2010), and further confirmed that the alleles at this strongest smoking behavior locus explain approximately 1% of the variance in CPD.

Plausible pleiotropic effects of nAChRs were examined in two studies using SNPs tagging the robust smoking behavior locus on chromosome 15. First, we provided novel evidence of association between a genetic variant on 15q25.1 and alcohol use (OR=1.15, P=7.0x10^{-5}) in the National FINRISK Study and the Health 2000 Survey (*N*=31,812). Second, we participated in an international collaboration work aiming to estimate the causality between smoking and body mass index (BMI) with a Mendelian Randomization approach. The National FINRISK study was included in the Mendelian Randomization meta-analyses, which demonstrated that variants on 15q25.1 may contribute to BMI and the direction of the outcome is related to smoking status: never smokers are associated with higher BMI, whereas current smokers are associated with lower BMI.

Taken together, these results improve our knowledge of the genetic factors affecting smoking behavior and ND. They also provide further evidence on the pleiotropic effects of the nAChRs. However, more work is required, since the genetic architecture of variants influencing smoking remains scarce. In addition, bigger samples and large collaborations are needed to observe the many small effects of yet unidentified variants affecting complex traits – the ones that our genome currently keeps hidden.

TIIVISTELMÄ

Tupakoinnin terveyshaitat luovat yhden suurimmista uhkakuvista kansanterveydelle maailmanlaajuisesti. Keuhkoahtaumatauti, sydänsairaudet sekä useat syövät muodostavat mittavan osuuden tupakoinnin aiheuttamista ennaltaehkäistävistä kuolemista. Tupakointi ja sitä ylläpitävä nikotiiniriippuvuus ovat monitekijäisiä ilmiasuja, joihin vaikuttavat niin ympäristö, perimä kuin henkilön yksilölliset ominaisuudet. Perinnölliset tekijät selittävät noin puolet nikotiiniriippuvuuden normaalivaihtelusta, miksi onkin olennaista kartoittaa ja ymmärtää sen geneettinen tausta. Yllättäen, koko perimänlaajuisissa assosiaatiotutkimuksissa tähän mennessä identifioidut riskivariantit selittävät vain pienen osan nikotiiniriippuvuuden vaihtelusta. Tässä väitöstutkimuksessa kartoitettiin perimänlaajuisesti geenien sekä kirjallisuuden pohjalta valikoitujen yhden emäksen variaatioiden yhteyksiä tupakointikäyttäytymiseen ja nikotiiniriippuvuuteen. Tavoitteena oli selvittää geenivarianttien kyky selittää ihmistenvälisiä eroja niin nikotiiniriippuvuudessa kuin alkoholinkäytössä.

Aikaisempiin geneettisiin kytkentätutkimuksiin pohjautuen kromosomi 20 hienokartoitettiin mikrosatelliittigeenimerkkien avulla suomalaisessa kaksosperheaineistossa. Aikaisempi kytkös nikotiiniriippuvuuden ja kromosomin välillä vahvistettiin. Lisäksi sukupuolten välillä havaittiin selkeä ero nikotiiniriippuvuuteen vaikuttavissa geenimerkeissä.

Samassa suomalaisessa kaksosperheaineistossa selvitettiin koko perimänlaajuisella assosiaatiotutkimuksella geenivarianttien yhteyttä tupakointikäyttäytymiseen, nikotiiniripuvuuteen sekä nikotiinivieroitusoireisiin. Tutkimus osoitti selkeän yhteyden kromosomin 16 geenivarianttien ja päivittäisen tupakkamäärän välillä. Yhteys paikannettiin lähelle geeniä *CLEC19A*. Kromosomin 11 ja nikotiinivieroitusoireiden välillä havaittiin yhteys, joka paikannettiin geeneihin *AP2A2* ja *MUC6*. Nämä tulokset korostavat hermostollisen kehityksen ja selviytymisen kannalta oleellisen neurotrofiinien viestintäketjun merkitystä tupakointikäyttäytymisen taustalla.

Kromosomissa 15 sijaitsevien nikotiinireseptorigeenien *CHRNA5, CHRNA3* ja *CHRNB4* yhteyttä suomalaisten tupakointikäyttäytymiseen selvitettiin kolmen tunnetun geenivariantin avulla FINRISKI ja Terveys 2000 aineistoissa. Tutkimuksemme vahvisti aikaisemman havainnon kolmesta yhden emäksen variaatioista, joilla on itsenäinen yhteys päivättäiseen tupakkamäärään. Tutkimus varmisti vahvimman lokuksen selittävän noin yhden prosentin tupakkamäärässä havaitusta vaihtelusta.

Tupakointiin ja tupakoinnista aiheutuviin terveyshaittoihin vahvasti yhdistetyn geenialueen mahdollisia monivaikutteisia ominaisuuksia määritettiin kahdessa eri tutkimuksessa niin ikään kromosomin 15 tunnettujen geenivarianttien avulla. Ensimmäisessä tutkimuksessa havaittiin uusi yhteys nikotiinireseptoreiden ja alkoholin käytön välillä. Toisessa tutkimuksessa arvioitiin tupakoinnin vaikutusta painoindeksiin. Kyseinen kausaalitutkimus oli osa laajempaa kansainvälistä konsortiotutkimusta. Osatyön mukaan tupakoimattomilla henkilöillä geenivariantti oli yhteydessä korkeampaan painoindeksiin, kun taas tupakoivilla henkilöillä sama geenimuunnos oli yhteydessä alhaisempaan painoindeksiin.

Väitöstutkimuksessa havaittujen tulosten avulla olemme saaneet lisää tietoa tupakointikäyttäytymiseen ja nikotiiniriippuvuuteen vaikuttavasta geneettisestä muuntelusta. Havaitut geenimerkit kuitenkin lisäävät perimän osuutta nikotiiniriippuvuuden selittäjänä vain vähän. Lisäksi tulokset auttavat ymmärtämään geenien monivaikutteisia ominaisuuksia. Erityisesti nikotiinireseptoreiden genomitiedon kartoittaminen voisi antaa hyödyllistä lisätietoa eri riippuvuuksien hoidon suunnittelussa.

CONTENTS

ABSTRACT 6

TIIVISTELMÄ 8

LIST OF ORIGINAL PUBLICATIONS 12

ABBREVIATIONS 13

1 INTRODUCTION 15

2 REVIEW OF THE LITERATURE 16

2.1 Introduction to human genetics 16

2.1.1 Human genome 16

2.1.2 Determining the genetic component of a trait 20

2.1.3 Genetic mapping: methods for gene localization 23

2.1.4 Hiding heritability 27

2.1.5 Finnish population history and disease heritage 27

2.2 Smoking behavior and nicotine dependence 28

2.2.1 Life course of smoking behavior 28

2.2.2 Established biological pathways associated with smoking 30

2.2.3 Smoking quantity 35

2.2.4 Nicotine dependence 35

2.2.5 Nicotine withdrawal 37

2.3 The genetics of smoking behavior and nicotine dependence 39

2.3.1 Heritability estimates 40

2.3.2 Nicotinic acetylcholine receptors 40

2.3.3 Nicotine metabolism 43

2.3.4 Neuronal pathways 43

2.3.5 Epigenetic findings 44

2.4 Neuropsychiatric comorbidities 45

2.4.1 Alcohol and other substance use 45

2.4.2 Neuropsychiatric disorders 46

3 AIMS OF THE STUDY 47

4 MATERIALS AND METHODS 48

4.1 Participants 48

4.1.1 The Older Finnish Twin Cohort (Studies I and II) 48

4.1.2 The National FINRISK Study (Studies III and IV) 48

4.1.3 The Health 2000 Survey (Study III) 49

4.2 Phenotypes 50
4.2.1 Smoking quantity (Studies I, II, III and IV) 50
4.2.2 Nicotine dependence (Studies I and II) 51
4.2.3 Nicotine withdrawal (Study II) 51
4.2.4 Alcohol use (Study III) 51
4.2.5 Body mass index (BMI) (Study IV) 52
4.2.6 Confounders 53
4.3 Genotyping, imputation and quality control 53
4.3.1 Microsatellite markers (Study I) 53
4.3.2 Single nucleotide polymorphisms (Study II-IV) 55
4.4 Statistical methods 56
4.4.1 Linkage analysis (Study I) 56
4.4.2 Association analyses (Study II-IV) 57
4.4.3 Mendelian randomization meta-analysis (Study IV) 58
4.4.4 Infering the functional potential of the SNPs (Study II) 58
5 RESULTS AND DISCUSSION 60
5.1 Genetics of nicotine dependence 60
5.1.1. Chromosome 20 shows evidence of linkage with ND 60
5.1.2 Linkage results reveal sex differences 62
5.1.3 Genome-wide association study for smoking behavior highlights 16p12.3 and 11p15.5 (Study II) 64
5.1.4 The strongest genetic evidence for smoking behavior (Study III) 72
5.2. Pleiotropic effects of nAChR subunits 75
5.2.1 Association of the α5-α3-β4 subunit cluster with alcohol use 75
5.2.2 Causality between smoking and BMI (Study IV) 77
5.3 Strengths and Limitations 79
6 CONCLUSIONS AND FUTURE PROSPECTS 81
7 ACKNOWLEDGEMENTS 83
8 WEB RESOURCES 85
9 REFERENCES 86

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications that are referred to in the text by the Roman numerals I-IV:

I Keskitalo-Vuokko K*, Hällfors J*, Broms U, Pergadia ML, Saccone SF, Loukola A, Madden PA, Kaprio J (2012) Chromosome 20 shows linkage with DSM-IV nicotine dependence in Finnish adult smokers. *Nicotine Tob Res* **14**:153-60.

II Hällfors J, Palviainen T, Surakka I, Gupta R, Buchwald J, Raevuori A, Ripatti S, Korhonen T, Madden PA, Kaprio J, Loukola A. Genome-wide association study in Finnish twins highlights the connection between nicotine addiction and neurotrophin signaling pathway. Manuscript submitted to *Addict Biol*.

III Hällfors J, Loukola A, Pitkäniemi J, Broms U, Männistö S, Salomaa V, Heliövaara M, Lehtimäki T, Raitakari O, Madden PA, Heath AC, Montgomery GW, Martin NG, Korhonen T, Kaprio J (2013) Scrutiny of the CHRNA5-CHRNA3-CHRNB4 smoking behavior locus reveals a novel association with alcohol use in a Finnish population based study. *Int J Mol Epidemiol Genet* **4**:109-19.

IV Taylor AE, Morris RW, Fluharty ME, Bjorngaard JH, Åsvold BO, Gabrielsen ME, Campbell A, Marioni R, Kumari M, Hällfors J, Männistö S, Marques-Vidal P, Kaakinen M, Cavadino A, Postmus I, Husemoen LL, Skaaby T, Ahluwalia TS, Treur JL, Willemsen G, Dale C, Wannamethee SG, Lahti J, Palotie A, Räikkönen K, Kisialiou A, McConnachie A, Padmanabhan S, Wong A, Dalgård C, Paternoster L, Ben-Shlomo Y, Tyrrell J, Horwood J, Fergusson DM, Kennedy MA, Frayling T, Nohr EA, Christiansen L, Ohm Kyvik K, Kuh D, Watt G, Eriksson J, Whincup PH, Vink JM, Boomsma DI, Davey Smith G, Lawlor D, Linneberg A, Ford I, Jukema JW, Power C, Hyppönen E, Jarvelin MR, Preisig M, Borodulin K, Kaprio J, Kivimaki M, Smith BH, Hayward C, Romundstad PR, Sørensen TI, Munafò MR, Sattar N (2014) Stratification by smoking status reveals an association of CHRNA5-A3-B4 genotype with body mass index in never smokers. *PLoS Genet* **10**:e1004799.

* These authors contributed equally to this work.

The original articles are reproduced at the end of the thesis with permission of the respective copyright holders.

ABBREVIATIONS

A	adenine
ADHD	attention-deficit hyperactivity disorder
APA	American Psychiatric Association
BMI	body mass index
bp	base pair
C	cytosine
CEU	Utah residents with Northern and Western European ancestry
CI	confidence interval
cM	centimorgan
COPD	chronic obstructive pulmonary disease
CPD	cigarettes smoked per day
DA	dopamine
DNA	deoxyribonucleic acid
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, fourth edition
DZ	dizygotic
eQTL	expression quantitative trait loci
ExAC	Exome Aggregation Consortium
FDR	false discovery rate
FTND	Fagerström Test for Nicotine Dependence
G	guanine
g	grams
GABA	gamma-aminobutyric acid
GEMMA	Genome-wide Efficient Mixed Model Association
GTEx	Genotype-Tissue Expression
GWAS	genome-wide association study
h^2	heritability
HRC	Haplotype Reference Consortium

HWE	Hardy-Weinberg equilibrium
IBD	identity by descent
LD	linkage disequilibrium
LOD	logarithm-of-odds
MAF	minor allele frequency
MaxCigs24	maximum number of cigarettes smoked in a 24-hour period
meQTL	methylation quantitative trait loci
MLOD	maximum LOD score
MZ	monozygotic
nAChR	nicotinic acetylcholine receptor
ND	nicotine dependence
NF-κB	nuclear factor κ B
NMR	nicotine metabolite ratio
NW	nicotine withdrawal
OR	odds ratio
PCR	polymerase chain reaction
RNA	ribonucleic acid
SE	standard error
SISu	Sequencing Initiative Suomi
SNP	single nucleotide polymorphism
T	thymine
TAG	Tobacco and Genetics Consortium
TFBS	transcription factor binding site
TM	transmembrane protein domain
WES	whole exome sequence
WGS	whole genome sequence
WHO	World Health Organization

1 INTRODUCTION

Cigarette smoking remains the single major cause of avoidable global public health epidemics. More than five million yearly deaths can be attributed to tobacco use. It has been causally linked to multiple diseases and adverse health effects, including chronic obstructive pulmonary disease (COPD), cardiovascular disease and lung cancer, as well as several other cancers, which affect nearly all organs of the body (USDHHS, 2014).

Despite the well-documented risks, the World Health Organization (WHO) has estimated that in 2012, over 1.1 billion people smoked tobacco (WHO, 2015). The estimated worldwide prevalence of currently smoking adults has reached 22%, with smoking rates varying widely between countries (WHO, 2015). On average, smoking is still most prevalent in Europe, and least prevalent in Africa.

Although the worldwide ratios of women-to-men smoking prevalences vary across countries, currently, smoking among women is less prevalent than among men. However, the estimates of female smoking are projected to rise in the future, especially in many low- and middle-income countries (Hitchman and Fong, 2011). The rise of smoking among women can be attributed to several factors: changes in social acceptance, women's rising economic resources, and worryingly, to the tobacco industry's marketing of cigarettes to women as a symbol of emancipation (Amos and Haglund, 2000).

Marketing plans of the tobacco industry, although highly efficient, are not the sole perpetrators leading to the worldwide epidemic of smoking-related health consequences. Smoking is a highly efficient and fast form of drug use, and one force driving the behavior is nicotine dependence (ND), which is characterized by the persistent use of tobacco products and the emergence of withdrawal symptoms during abstinence (USDHHS 2014).

Ever since the identification of nicotine as the primary psychoactive drug in tobacco smoke, with an addiction potential comparable to heroine, a great amount of time and effort has been spent to unravel the neuropharmacological and behavioral aspects of its effects. Various neurotransmitter systems have been linked to mediate the psychoactive and addictive nature of nicotine (Laviolette and van der Kooy, 2004).

ND is a multifactorial and multidimensional phenomenon, with physiological, social, and psychological levels. Twin studies have established high heritability estimates, approximately 50% for persistent smoking (Tyndale, 2003; Hughes et al., 2006). These heritability estimates clearly state that genetic factors significantly contribute to the susceptibility for smoking behavior and ND. During the past decade, genetic studies have examined the human genome, aiming to identify genetic variants that would predispose us to smoking behavior. However, only a few genetic variants have been successfully identified, and they explain just a fraction of the variance in smoking behavior. This thesis focused on identifying specific genes predisposing individuals to smoking behavior and ND.

2 REVIEW OF THE LITERATURE

2.1 Introduction to human genetics

Inherited factors contribute in some way to nearly every human trait. Genetic factors, interacting with environmental and individual characteristics contribute significantly to the susceptibility of developing diseases which place a heavy burden on an individual as well as public health and the economy. Ever-evolving tools for identifying and characterizing the contribution of genetic variants have assisted medical research in shifting healthcare to personalized levels, since the inclusion of personal genomic portfolios in medical practice has become widely acknowledged as a pivotal part of a holistic approach in disease treatment. Since genetically inherited variation can affect how a person responds to drug treatments, adverse drug responses can be predicted or even prevented by applying the required information from personal genetic profiles of variants in genes encoding drug-metabolizing enzymes and drug transporter proteins (Su et al., 2014).

2.1.1 Human genome

The human genome consists of 46 chromosomes, 23 inherited from a mother and 23 from a father. There are 22 pairs of autosomes and 2 sex chromosomes (X and Y). Each nucleus in most cells, with a few exceptions, for example, red blood cells, contains the complete paired set of chromosomes, consisting of deoxyribonucleic acid (DNA) tightly wrapped around histone proteins, referred to as chromatin. Each of our cells also contains varying amounts of extracellular maternally-inherited mitochondrial DNA, which is stored in mitochondria.

DNA is a polymer compound made up of two complementary nucleotide strands, which are entwined in the shape of a double helix. Details of this double helix were first published in 1953 in the famous work by Watson and Crick (Watson and Crick, 1953). The building blocks of nucleotides are composed of deoxyribose sugar, a nitrogenous base and at least one phosphate group. Sugar and phosphate groups alternate to form the backbone structure for a DNA strand, and two strands are paired by complementary bases: adenine (A) pairs with thymine (T), and guanine (G) pairs with cytosine (C). Hydrogen bonds bind the complementary bases, and thus play a major role in stabilizing the DNA double helix shape. In 1961, less than a decade after the demonstration of DNA's structure, the genetic code was cracked and resolved, showing that each codon, a triplet of bases, translates into a different amino acid (Crick et al., 1961).

The human genome is made up of over 3 billion base pairs (Lander et al., 2001; Venter et al., 2001), and the order of them determines the meaning of our genes. Merely a small fraction, approximately 1.2%, of the entire DNA sequence codes for exons in genes, which further code for proteins, while other parts of the genome are non-coding. Some of the non-coding sequences contain regulatory regions (i.e., instruction sites for directing gene expression).

Currently, the estimated number of genes in the genome is around 23,000, considerably fewer than initially thought in the beginning of the era of genomic discoveries. These genes, however, can be transcribed into multiple different proteins through alternative splicing, and they can be adapted for use in different biological contexts. The majority of these genes are evolutionarily conserved, which is a strong indicator of function, and thus suggests that they correspond to genuine proteins (Pruitt et al., 2009). In addition, the genome harbors non-protein coding ribonucleic acid (RNA) genes, and pseudogenes (Naidoo et al., 2011), which challenges the definition of a gene. A specific feature of these "genes" is that they lack protein-coding abilities. Nevertheless, they may have important regulatory roles (Poliseno, 2012; Ulitsky, 2016).

How the human genome functions is only just beginning to be clarified. The term "functional element" is defined as a discrete region on the genome that encodes a defined product, like a protein, or a reproducible biochemical signature (e.g., a specific chromatin or transcription structure). Such signatures mark genomic sequences with important functions, including exons, sites of RNA processing (introns, sequences between exons that are cleaved from the original transcript, and 5' and 3' un-translated regions), and transcriptional regulatory elements, like promoters, enhancers, silencers, and insulators. Although certain biochemical signatures can be associated with specific functions, their ultimate biological roles and functions cannot be declared, but, rather, require further validation (ENCODE Project Consortium, 2011). While the research community desires to elucidate genomic function in human biology and disease, the means for defining these DNA elements require combining biochemical, evolutionary and genetic approaches (Kellis et al., 2014). Challenges in deciphering gene functions have recently spurred novel projects for cataloging gene annotations that will integrate genome exploration with gene expression, epigenomic and reference panel datasets (Mudge and Harrow, 2016). So far, the task remains challenging.

Genetic variation

Although human genomes between any two individuals are 99.9% identical, there are forms of genetic variation that distinguish us from each other. Sources leading to this variation are mutations and recombination events. Genetic variation consists of SNPs, insertions and deletions, repeated sequences and chromosomal abnormalities. Although these variable sites account for only a small portion of the human genome, they represent the mechanism for subtle differences that are observed among individuals, which can contribute to variation in the development of a trait or disease. In addition, these sites can be used for mapping the genetic variants affecting a phenotype of interest.

The human genome harbors several classes of repeated sequences, which can be further categorized based on their length. Modern genetic mapping techniques have mostly harnessed microsatellite markers. These are highly polymorphic repeats of 1-10 nucleotides, and the ones used in genetic mapping typically consist of 10-50 copies of di-, tri-, or tetranucleotide repeats. The length of repeat sequences ranges from tens to hundreds of base pairs (bps), and because they are flanked by unique DNA sequences, they are easily amplified by polymerase chain reaction (PCR). One advantage of microsatellites is the large number of alleles (variant forms of a gene), resulting in a high number of different genotypes (each consisting of two alleles, one inherited from a mother and the other from a father, at each genetic loci).

The most common type of genetic variation is a single nucleotide polymorphism (SNP), which is a variation in a single base. Because of the diploid nature of the human genome, a majority of the SNPs have two alleles. As an example, a SNP with alleles C and A can be present in a certain population with allele frequencies 15% and 85%, respectively. In this example, C is a minor allele at that particular locus, with a minor allele frequency (MAF) of 15%. The HapMap Project has estimated that, in the human population, about 10 million sites vary so that the alleles are observed in the population with at least 1% frequency (The International HapMap Consortium, 2003). The majority of these SNPs are common variants, meaning that their allele frequency in a population is at least 5%. The rest of the 10 million SNPs are low-frequency variants (MAF 1% < 5%). In addition, the human genome harbors large numbers of rare SNPs, with an MAF of less than 1%. The Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org/) has proposed the following evidence-based MAF categorization: common variants MAF > 10%, low-frequency variants 1-10%, rare variants 0.1-1.0%, and ultra-rare variants < 0.1% (Lek et al., 2016).

Depending on their locations, SNPs have varying effects on the phenotype. In the coding region, SNPs that alter the amino acid sequence, causing a missense mutation, may alter the final protein product to varying degrees. Intergenic regions (sequences between two genes) contain regulatory sites, where a SNP can have a regulatory effect on protein synthesis. SNPs are not equally distributed within the genome, but are more frequent within the non-coding, intergenic regions, compared to coding regions. In general, SNPs are located in sites where natural selection fixes the allele that contributes to the most favorable genetic adaptation at that particular moment in time. Other factors, like genetic recombination and mutation rates, can affect the SNP density as well. Genetic variation varies between populations, a SNP that is common in one population or geographical group can be rarer in another.

Linkage disequilibrium

Linkage disequilibrium (LD) refers to the occurrence of non-random association between two alleles at different loci. Usually, when a chromosome pair separates during meiosis, cross-over events, where two homologous chromosomes exchange genetic material, occur at various points resulting in recombinant daughter chromosomes that differ from the parental ones. Some regions are more prone to recombination than others, which is why they are referred to as recombination hotspots. In these regions, LD is smaller, whereas, the regions with little recombination have larger LD. Regions with LD patterns are usually described as haplotype blocks, which can be used further for SNP selection in gene mapping studies. Since a haplotype block harbors SNPs which are in high LD with each other, sometimes only one tagging SNP is adequate to capture the entire genetic variation within that region, thus lowering genotyping expenses and providing a means to map the genome more thoroughly. In addition, haplotype blocks provide an essential advantage for successive gene mapping in association analysis, as a majority of the observed genome-wide association study (GWAS) signals are functions of LD between the measured marker and the causal locus. For the same reason, however, mapping the correct causal gene variant remains challenging.

Some haplotypes clearly vary between populations, whereas some are amazingly similar (Rosenberg et al., 2002; Gabriel et al., 2002), which reflects both the immigration history of humans and geographic regions of different populations (Malaspinas et al., 2016; Mallick et al., 2016, Pagani et al., 2016). The greatest genetic variation has been observed in Africa, home of the Cradle of Humankind, and the diversity of haplotypes has been observed to decrease as distance from its origin increases (Conrad et al., 2006). This diversity needs to be acknowledged and thoroughly considered in all international gene mapping collaborations.

Reference of the human DNA sequence

The first draft of the human genome was published in 2000 (Lander et al., 2001, Venter et al., 2001), and the final draft in 2003 (International Human Genome Sequencing Consortium 2004). Building upon the known human genome, the International HapMap Project (http://hapmap.nci.nlm.nih.gov/) was launched in 2002, aiming to describe common patterns of human genetic variation that are involved in human health and disease (The International HapMap Consortium, 2003).

The publicly available human reference sequence, combined with high-throughput genotyping and sequencing technologies, have initiated several large-scale international projects to map the human genetic variation in diverse populations. One of them, the 1000 Genomes Project (http://www.internationalgenome.org/), set out to find most of the genetic variants in the genome, and after completion in 2015, has provided a benchmark for surveys of human genetic variation by enabling more accurate array designs and genotype imputation. The third phase of the project reconstructed the genomes of 2,504 individuals from 26 populations (1000 Genomes Project Consortium, 2015), and the reference panel includes > 99% of SNP variants with a frequency of > 1%. A more recent catalogue of human genetic diversity, ExAC, has provided the largest open access catalogue so far of variation in human protein-coding regions including 60,706 individuals in the study (Lek et al., 2016). The scale of this catalogue offers much anticipated insight into rare genetic variation across populations. In addition, another large-scale novel human reference panel has been constructed by the Haplotype Reference Consortium (HRC) (http://www.haplotype-reference-consortium.org/) including a total of 64,976 human haplotypes of purely European descent, and aims to increase the imputation accuracy of SNPs with MAFs as low as 0.1%, and helps to discover plausible causal loci for common diseases (McCarthy et al., 2016). In addition, an entirely population-based reference panel has been generated for a Finnish genome. As an advantage, it has been shown to increase imputation accuracy of rare and low-frequency variants in Finnish study cohorts (Surakka et al., 2016). Currently, the reference panel comprises 4,932 Finnish high-pass whole-exome sequences and 1,941 Finnish low-pass whole-genome sequences (Surakka et al., 2016).

In order to increase our understanding of the genome and how it is regulated and expressed in different species and between individuals, novel technologies have been developed. Using a novel gene editing tool based on bacteria Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) - associated protein-9 nuclease (Cas9), a recent study has provided insight into the origins of evolutionary adaptations (Nakamura et al., 2016). The overall function of the genome, which has remained mostly a mystery to scientists for over a decade, is finally starting to reveal its secrets.

All of these projects have contributed substantially to human genomics by more precisely and densely mapping the sequence and by giving insight into the functional complexity of the genome. These advances have directed the entire scientific quest toward an era of personal genomic sequencing and personalized medicine (Naidoo et al., 2011).

Imputation

The cost for large-scale whole-genome sequencing remains high, even though the prices have declined considerably during the past couple of years due to dramatic development in the sequencing technology (Metzker, 2010). The commercially available genotyping chips only contain some hundred thousand tagging SNPs, variants which are adequate to capture the entire genetic variation within one's specific haplotype block. As it is uncertain whether a SNP on a genotyping chip is a true causal variant, the missing genotypes are imputed to the data utilizing existing catalogues of human genetic variation. The catalogues offer constantly updated reference panels for use, such as the 1000 Genomes Phase I reference panel (The 1000 Genomes Project Consortium, 2012), the 1000 Genomes Phase III reference panel (The 1000 Genomes Project Consortium, 2015), the HRC reference panel (McCarthy et al., 2016), and the Finnish population-based reference panel provided by the Sequencing Initiative Suomi (SISu) project (www.sisuproject.fi).

In imputation, missing genotypes are predicted based on the observed genotypes on a reference panel. The race for developing more accurate imputation reference panels has already yielded impressive results because the accuracy has improved substantially, owing to the increased numbers of whole-genome and whole-exome sequenced genomes included in the panels. The greatest advantage of imputation is its ability to add the coverage of genetic data without increasing the costs. As an example, the 1000 Genomes Phase III reference panel increases the number of variants from ~500,000 SNPs detected by genotyping chips to up to ~36,000,000 genetic variants.

2.1.2 Determining the genetic component of a trait

Estimating heritability

It has been suggested that inherited factors contribute to nearly every human trait and condition. But how do we know that? The roots can be tracked to the time of the discovery of genetics. Although the word *genetics* was devised as late as 1905, the history of genetics reaches as far back as the work of the Augustinian friar Gregor Mendel, who perceived the basic laws of genetic inheritance in his studies with pea plants in the 1860s.

To determine whether a trait of interest is inherited from parents to offspring, family studies estimate the possibly elevated risk for siblings and relatives of an affected individual to manifest the same condition. These family studies assess whether the risk is higher than in the general population. Importantly, family studies do not reveal the mechanism behind the familial aggregation, which can be environmental or genetic factors, or both. Twin and adoption studies play a central role in elucidating the genetic component. In principle, the concordance rate of a trait is calculated between monozygotic (MZ) twins and further compared with the concordance rate calculated between dizygotic (DZ) twins. This produces an estimate of the proportion of the phenotype affected by genetic variation (Boomsma et al., 2002), in other words, it estimates the heritability of the trait. Narrow-sense heritability, h^2, refers to the portion of phenotypic variance in a population attributed by additive genetic factors (Hindorff et al., 2009).

Narrow-sense heritability can be estimated from genome-wide SNPs genotyped in unrelated individuals. Owing to the easy access and feasibility of the data, this method has several advantages over the traditional pedigree-based one. Using this approach, it has been estimated that approximately half of the heritability of human height can be attributed to ~300,000 common SNPs (Yang et al., 2010). However, merely 5-10% of the variance in height is explained by genome-wide significant common SNPs that exceed the significance level of $5x10^{-8}$ (Speed et al., 2012), and the method is sensitive to uneven LD between SNPs; high LD regions that harbor causal variants can overestimate the contribution of genetic variants to heritability, whereas regions of low LD may underestimate the explained contribution (Speed et al., 2012). A method of modified kinship matrix that weighs SNPs according to their local LD has been proposed to overcome the bias (Speed et al., 2012). Lately, a novel heritability estimation tool, linkage disequilibrium score regression has been developed (Bulik-Sullivan et al., 2015) to utilize summary statistics from GWAS and to estimate heritability and genetic correlations from several meta-analyses (Zheng et al., 2017; Anttila et al., 2016).

Heritability estimates can be affected by variations observed in the environment and allele frequencies. Hence, heritability estimates cannot be generalized to all populations, but they are rather population and sample specific. In addition, the heritability of a disorder or trait is not constant over time, since the environmental and genetic effects can change (Visscher et al., 2008).

Complex disease genetics

A small proportion of human traits and diseases are caused by mutations in a single gene, such as phenylketonuria and cystic fibrosis, and these are referred to as Mendelian disorders. They are mostly rare, whereas common traits and diseases, having the highest impact on public health, such as obesity and type II diabetes, are caused by several genetic variants together with an environment (Botstein and Risch, 2003; Gibson, 2012). Common disorders are also referred to as complex disorders. In general, these diseases are heritable in several dimensions, meaning that several genes together with an environment and lifestyle, which is strongly affected by the environment, all contribute to the disease onset. In these disorders, genetic variants individually increase the disease risk only a little, but together with other genetic risk variants and the environmental factors, they increase the chance of phenotypic heterogeneity and further the risk for inducing a disorder. For most disorders and traits, the majority of the risk factors, especially genetic risk variants, are unknown and yet to be discovered.

In addition, for some complex human traits and diseases the underlying genetic variants overlap. A phenomenon influencing that shared genetic predisposition is pleiotropy. In pleiotropy, one genetic variant influences two or more seemingly distinct and unrelated phenotypic traits simultaneously, and may induce a wide range of symptoms. A way of exploring pleiotropy is to calculate LD-score regression using summary statistics. As an example, a recent study utilizing this heritability-based method quantified the extent of shared genetic contributions across 23 brain disorders (N=842,820), and highlighted the importance of common genetic variation as a risk for brain disorders. Further, it demonstrated the potential of heritability-based methods for providing a more comprehensive view of the complex genetic architecture of brain phenotypes (Anttila et al., 2016).

In addition, epigenetics underlines the complexity of genetics. It refers to changes in the genome that affect gene expression and function. Currently, it has been acknowledged that genes or environmental exposures alone are hardly sufficient to completely explain a disease process, but instead they are components of a larger complex molecular system that, in combination, results in a phenotype. Epigenetics, being an integral component of this molecular system, may play a mechanistic role in the formulation of a trait or disease (Ladd-Acosta and Fallin, 2016). Incorporation of epigenetics into the field of epidemiology is likely to help distinguish biological mechanisms for disease.

Mendelian randomization

One of the fundamental public health improvement strategies is to identify environmentally modifiable factors that causally influence disease risk. In general, observational epidemiological studies are, however, plagued by their inability to reliably identify such causal association, since conventional genetic epidemiological association methods are affected by reverse causation and confounding (Smith, 2010). A way to overcome this would be to set up a randomized controlled trial, which, unfortunately, is limited in many cases by ethical permissions and practice (Latvala and Ollikainen, 2016). On the other hand, Mendelian Randomization, which has been created to model the random assortment of genes from parents to offspring during meiosis, provides one method for assessing the causal nature of environmental exposures (Davey Smith and Ebrahim, 2003).

In the method, association between a disease and a polymorphism that mimics the biological link between a proposed exposure and disease is not generally infringed upon by reverse causation or confounding, which can skew the interpretations of conventional observational studies (Davey Smith and Ebrahim, 2003). Therefore, Mendelian Randomization offers an alternative method for experimental studies, and utilizes genetic variants as proxy measures of exposure, and it can further contribute to strengthening causal inference (Smith, 2010). Still, not even this approach for determining the genetic component of a trait remains flawless, as factors such as LD, pleiotropy and population stratification may distort Mendelian Randomization studies (Gage et al., 2013).

Mendelian Randomization is based on a notion that, if X affects Y, factors affecting X must also have an effect on Y. A genetic variant Z that is known to associate with X is utilized as an instrument to investigate the causal nature of the association between X and Y (Latvala and Ollikainen, 2016). Causal inferences drawn from Mendelian Randomization studies are only valid if the rather strict assumptions of the method are met. According to these assumptions, the genetic instrument Z must be reliably associated with the exposure X, the genetic instrument Z must not be associated with the confounding factors U, and lastly, the genetic instrument Z should only be related to the outcome Y through its association with the exposure X (Lawlor et al., 2008). Figure 1 demonstrates the mechanism involved in the Mendelian Randomization method.

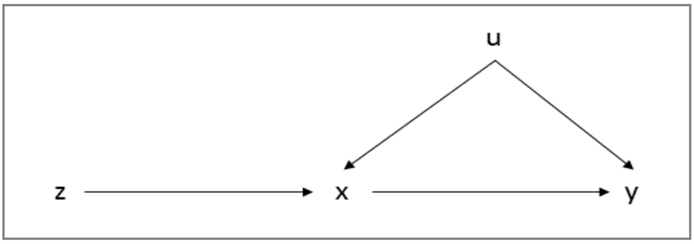


Figure 1. The basic instrumental variables model for Mendelian Randomization. Z, instrumental variable; X, modifiable variable; Y, outcome of interest; U, unmeasured confounders. (Adapted from Lawlor et al., 2008).

2.1.3 Genetic mapping: methods for gene localization

Genetic mapping aims to identify a gene or genes that increase the risk for a disease or trait of interest. The means for gene mapping largely involve estimations of correlation between a phenotype, a measurable feature one expresses, and genotypes assessed as genetic markers. The optimal gene mapping strategy depends on the underlying genetic architecture of the trait of interest.

Since rare variants are generally family-specific, they can be identified through linkage analysis, which investigates the co-segregation of chromosomal loci with the trait within families. Ideally, linkage analyses are conducted in large multi-generation families with multiple affected individuals, which provide higher statistical power to detect the signals. On the other hand, if the trait is thought to be caused by common variants, association analysis provides more power to identify genetic risk variants. Although association analysis can also be performed in families, the commonly adapted study design utilizes large case-control cohorts, which are easier to collect than large pedigrees, especially when the phenotype of interest manifests with a late age of onset, in which case the parents of index cases might not be available for the study. Both linkage and association analyses can be utilized to map the genetic contribution of a trait in two distinct settings: targeted candidate gene and genome-wide (i.e., hypothesis-free) approaches.

Hypothesis-free approaches do not require any previous biological knowledge of the modes of inheritance to detect genetic association. Whereas, candidate gene-based models require an *a priori* presumption of the gene of interest, derived either from previous studies highlighting genomic regions, or from an established function in a biological process or pathway connected to the phenotype of interest. Despite the differences between these two approaches, both have contributed to and increased our knowledge of the genetics of both Mendelian disorders and complex traits and diseases.

Linkage analysis

Linkage analysis is a powerful tool aimed to detect the chromosomal location of disease genes. Simply, it is based on the observation that genes that reside physically close on a chromosome remain linked during meiosis (Pulst, 1999). For several years during the late twentieth century, linkage analysis was the predominant statistical genetic mapping approach for Mendelian and complex traits with familial assembly. The aim of linkage analysis is to localize a chromosomal region that co-segregates with the disease within a pedigree. In other words, the analysis examines whether genetic markers within a distinct chromosomal region are inherited with a trait of interest more often than would be expected by chance.

During meiosis, recombination events generate a new combination of alleles in each generation, thus playing a significant role in genetic diversity. Genetic markers can be used to distinguish which chromosomal segments have been inherited from each parent. The probability for recombination to occur between two loci in the genome, the recombination fraction (θ), is related to their physical distance: θ=0 corresponds to the two loci co-segregating together, whereas θ=0.5 corresponds to independent segregation. The closer the two loci are located to each other, the more likely they co-segregate during meiosis. Linkage analysis tests whether θ between a genetic marker with a known position and a locus influencing the trait of interest differs significantly from 0.5. If the θ between two loci is less than 0.5, they are said to be linked. A measure for estimating the likelihood of linkage is expressed as a logarithm-of-odds (LOD) score, which is utilized to approximate the likelihood of co-segregation of a genetic marker with the trait of interest compared with the hypothesis of individual segregation (Morton, 1955). The LOD score (Z) is the logarithm base 10 of the ratio of two likelihoods, and expression of the likelihood as a logarithm allows for the summation of likelihoods observed in different pedigrees:

$$Z(\theta) = \log_{10} \frac{L(\theta)}{L(\theta=0.5)}$$

where L is the likelihood, θ the recombination fraction (estimated from the data), and θ = 0.5 represents the null hypothesis that the two loci are unlinked (Ott, 1974; Ott, 1989).

A LOD score higher than 3.0 is generally accepted as evidence for linkage (Pulst, 1999), and it corresponds to a point-wise significance of $p=1x10^{-4}$, which, on average, is regarded as significant. However, this assumption is applicable only to monogenic traits, whereas, the significance of a LOD score for a complex trait is more ambiguous (Lander and Kruglyak, 1995). To obtain a genome-wide significance level of 5%, corresponding to a point-wise significance of $p=1x10^{-5}$, the LOD score needs to be raised to 3.3 (Lander and Kruglyak, 1995). LOD scores can be calculated between two loci (one marker locus and the unknown disease locus) at a time (*single-point linkage analysis, also called two-point analysis*), and maximum likelihood estimates can be calculated for multiple loci at a time (*multi-point linkage analysis*).

Linkage analysis can be applied to at least two distinct study settings. In *parametric linkage analysis*, the genetic model for the disease or trait of interest must be specified, and the expected causative variant frequency, penetrance and phenocopy rate need to be estimated in advance. Unlike in monogenic diseases, these factors cannot usually be estimated precisely for complex traits. These prerequisites can be overcome by *non-parametric linkage analysis*, which does not make any assumptions about the disease model (Goring & Terwilliger, 2000). However, complex traits with ambiguous inheritance models can cause challenges in the positional mapping, since the location of the linkage signal is inclined to fluctuate (Altmuller et al., 2001, Roberts et al., 1999).

Linkage methods are particularly powerful for detection of variants with a large effect size, which almost always are rare in the population. The power to detect such loci using linkage methods can be enhanced by ascertaining families with aggregation of the trait of interest (i.e., “loading” the sample with the trait of interest). For example, the *BRCA1* and *BRCA2* genes underlying breast cancer have been mapped utilizing linkage studies of families specifically selected for strong aggregation of the disease (Bailey-Wilson and Wilson, 2011).

Association analysis

In association analysis, genetic effects of the trait of interest are examined by comparing allele frequencies. In general, association analysis calculates whether a marker allele is more frequent among affected individuals, who manifest the trait of interest, than would be expected by chance. The magnitude of the effect between cases and controls can be ascertained using a Chi-square test with one degree of freedom, or logistic regression. The latter allows the inclusion of several confounding factors as covariates, which need to be considered in complex trait research. Association analyses can also be used for estimating genetic effects for quantitative traits (such as smoking quantity and body mass index (BMI)) using linear regression. Association analysis is particularly suitable for detecting common variants in large samples.

Genome-wide approaches for gene mapping

Where candidate gene studies require an *a priori* knowledge of the targeted gene, genome-wide studies provide a hypothesis-free approach. Both linkage and association analyses can be utilized in a genome-wide approach. A genome-wide setting provides a suitable platform for scanning the entire genome before fine-mapping the highlighted regions and teasing out the possible causal locus in the following studies.

Within the past decade, GWAS have been widely used for mapping common genetic variants for complex traits. They test associations between common genetic variants and phenotypic variation in a trait of interest utilizing hundreds of thousands of genotyped, or millions of imputed, variants across the genome. The number of SNPs analyzed using GWAS has expanded since the first conducted GWASs utilizing SNP chips of approximately 100,000-250,000 variants (Klein et al., 2005; Maraganore et al., 2005) to millions due to denser arrays and evolving imputation methodology. While the quality and accuracy of imputation have improved due to more accurate reference panels, GWAS methods have been extended to include low-frequency and rare genetic variants as well. As of April 2017, the GWAS Catalogue contains 2,854 publications and 33,674 SNP-trait associations (http://www.ebi.ac.uk/gwas/).

Due to multiple testing, the likelihood for type I errors, increases. In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (i.e., detection of an effect that is not present: false positive), while a type II error is incorrectly retaining a false null hypothesis, and the analysis fails to detect an effect that is present (false negative). To account for multiple testing, a

Bonferroni correction with its stringent threshold for statistical significance, 5x10^{-8}, which corrects for approximately 1 million independent tests, has been applied as a standard rule for interpreting GWAS results. Other ways to account for multiple testing include the false discovery rate (FDR) correction (Benjamini et al., 2001) and permutation testing (Sham and Purcell, 2014). FDR specifically calls for controlling the expected proportion of falsely-rejected hypotheses, thus providing less stringent control on type I errors and greater power, at the cost of an increased rate of type I errors. The permutation procedure is a robust but computationally rather intensive alternative to calculate an empirically adjusted p-value.

In the category of common variants, association analyses have clearly excelled at linkage analyses. However, an emerging view states that rare variants could explain a substantial proportion of complex human traits and diseases (McClellan and King, 2010). Hence, the availability of whole-exome sequence (WES) and whole-genome sequence (WGS) data has once again raised the value of linkage analysis as a powerful method for detecting those rare genetic variants, which could play even more important roles in the disease etiology than initially estimated (Ott et al. 2015). The high cost of sequencing has been one of the key factors preventing large-scale *genome-wide linkage studies*. Along with the decreasing cost of sequencing, it will become plausible and even advantageous to have WGS data from large pedigrees of informative phenotype carriers, and to perform large-scale genome-wide linkage analysis (Ott et al., 2015). However, this approach requires bioinformatics tools and a computational environment which are efficient enough for such enormous data and capable of performing analyses in a reasonable timeframe.

Meta-analysis

Meta-analysis can be utilized to increase the power to detect SNP association or linkage, since it summarizes the results from several carefully harmonized studies (Evangelou and Ioannidis, 2013). Thus, the same imputation reference panel is required from all eligible GWAS cohorts. In meta-analysis, effect sizes from different data sets can be pooled with either fixed or random effects model. A fixed effects model assumes a common genetic effect through all data sets, and further explains any observed difference by chance alone. In a random effects model, the genetic effects are assumed to be different in all data sets (Ioannidis et al., 2009).

Exploring the functional potential of genetic variants

In order to infer the functional potential of the SNPs highlighted in a study, the analyses can be followed up by annotating the expression quantitative trait loci (eQTL) and methylation quantitative trait loci (meQTL) that are associated with the SNPs. eQTL refers to a SNP that has been associated with the expression of a local gene (*cis*-eQTL) or even a distant gene (*trans*-eQTL). Likewise, meQTL refers to a SNP that has been associated with methylation levels in *cis* or *trans.* eQTLs and meQTLs may provide a warranted means to explain the substantial inter-individual variance observed in traits that are not explained by the associated genetic variants.

2.1.4 Hiding heritability

The validity of the initial common disease–common variant model has been challenged by the existence of “missing heritability” (Manolio et al., 2009; Clarke and Cooper, 2010, Lander, 2011). It is evident that only a small portion of the heritability of a phenotype can be explained by common variants, individually or in combination. Several explanations for this phenomenon have been suggested. First, large numbers of variants with small effects are still to be highlighted, since the sample sizes in a majority of the GWASs are limited, and due to the stringent statistical thresholds imposed to ensure reproducibility (Lander, 2011). Second, limited sample sizes cause challenges in genotype calling, since rare variants as well as structural variants, which both could have large effects on the trait in question, are poorly detected (Manolio et al., 2009). Third, GWASs have limited power to detect gene–gene and gene–environment interactions (Manolio et al., 2009). Last, the current GWAS methodology has inadequate means for detecting shared environment effects among relatives (Manolio et al., 2009). On the other hand, as traits and diseases clearly run within families, the current methodology for mapping them on the genome might simply be inadequate. Thus, the missing heritability could merely be hiding.

At the same time, some of the hiding heritability may merely be an illusion. Since the estimates of heritability are inferred from additive genetic effects in epidemiological data, these estimates may be inflated by contributions of genetic interactions and gene-by-environment interactions, which are not accounted for in the estimations (Lander, 2011). For example, heritability estimates conducted via twin models do not account for shared environment, which might skew the numbers upward (Zaitlen et al., 2013). Also, the ExAC project has demonstrated that a surprisingly high portion of causal variants originally linked to rare Mendelian diseases are missclassifications in the literature and/or in databases (Lek et al., 2016), and this notion may provide novel insight into the estimation of hidden heritability.

2.1.5 Finnish population history and disease heritage

The Finnish population is one of the best-studied genetic isolates (de la Chapelle, 1993; Peltonen et al., 1999). Migration waves occurring some 4000 and 2000 years ago, arriving from east and south, respectively, have had a considerable effect on the Finnish gene pool. Both the paternally inherited Y chromosome and maternally inherited mitochondrial DNA show reduced genetic diversity compared to other populations (Sajantila et al., 1996). In detail, genetic evidence from phylogenetic analysis of Y chromosome haplotypes supports the theory of a dual origin of Finns, and reveals dramatic differences in Y haplotype variation between eastern and western Finland (Kittles et al., 1998). Evidence suggests that relatively few numbers of men and women have truly contributed to the genetic lineages that have been passed on and selected into the current generations of the Finnish population (Sajantila et al., 1996).

Reasons for this isolation can be described with geographical and geopolitical factors – being located between Sweden and Russia, two areas of distinct cultures, languages, and religion. In isolation, forces like *founder effects*, *genetic drift*, and *genetic bottlenecks* have shaped the genetic material in the Finnish gene pool, reducing the diversity. Ultimately, the Finnish population has expanded rapidly, mostly due to population growth, with relatively little external migration, creating regional sub-isolates as well (Peltonen et al., 1999).

Owing to this population history, Finland has its own unique diseases, referred to as Finnish Disease Heritage, which comprises altogether 36 monogenic disorders, that are more prevalent in Finland compared to the rest of the world (Norio, 2003a; Norio, 2003b, Norio, 2003c). While the frequency of some rare disorders has increased in the Finnish population due to reduced genetic diversity, for the same reason, some genetic variants predisposing one to certain diseases common in other populations (e.g., phenylketonuria and cystic fibrosis) are almost non-existent in Finland (Peltonen et al., 1999).

This genetic isolate has been advantageous in the process of mapping rare Mendelian genes (Peltonen et al., 2000). Using modern high-throughput exome sequencing technology, the allelic architecture of the Finnish population has also already revealed a significant enrichment of low-frequency (MAF 0.5-5%) loss-of-function variants in complex disorders (Lim et al., 2014). The SISu project, which combines nationwide health records with WES and WGS data, has also demonstrated a clear distinction between Finnish and non-Finnish European genome sequences (Lim et al., 2014), and indicated a clear need for a population-specific imputation reference panel for Finnish samples.

2.2 Smoking behavior and nicotine dependence

2.2.1 Life course of smoking behavior

Cigarette smoking is a complex behavior that includes several stages such as initiation, regular use, development of tolerance, dependence, cessation and relapse (Malaiyandi et al., 2005; Mayhew et al., 2000).

Several factors affect the probability of smoking *initiation*. A simplified list of these factors includes socioeconomic status, peer support and pressure, personality, cognitive factors, family background, gender, ethnicity, and previous substance use (Conrad et al., 1992; Buller et al., 2003; Mercken et al., 2009). The average age of onset of regular smoking is approximately 16 years (Filippidis et al., 2015). This age is a significant factor in predicting continued smoking behavior, and predicts difficulties in quitting smoking compared to those who have initiated smoking at a later age (Khuder et al., 1999). Consequently, an early onset of smoking is associated with substance use later in life (Siqueira and Brook, 2003), and the development of substance use disorders (Lewinsohn et al., 1999), and even earlier onset of depression and/or anxiety disorders (Jamal et al., 2011). In addition, early-onset smoking is a strong predictor associated with adverse smoking-related health consequences, such as the development of peripheral artery disease (Planas et al., 2002), and the risk of lung cancer (Hara et al., 2010).

Repeated exposure to nicotine causes neuroadaptation (i.e., *tolerance*) (Wang and Sun, 2005). As the tolerance develops, the number of nicotinic acetylcholine receptor (nAChR) binding sites in the brain increases, likely due to a response to nicotine-mediated receptor desensitization (Govind et al., 2009). Receptor desensitization refers to the ligand-induced closure and unresponsiveness of the receptor, and is suggested to mediate the development of tolerance and dependence. The symptoms of withdrawal and craving emerge in smokers during a period of abstinence, such as night-time sleep, when the receptors begin to respond again (Dani and Harris, 2005). Self-administered nicotine intake alleviates the unpleasant symptoms related to withdrawal and craving (described in more detail in subsection 2.2.5), because nicotine re-binds to the nAChRs. Subsequently, by sustaining sufficient levels of nicotine in the plasma to prevent withdrawal symptoms, smokers sense rewarding effects through the dopaminergic system (Balfour, 2004).

The development and maintenance of *ND* is mediated, at least in part, by neurobiological effects of nicotine and neuroadaptive changes with chronic nicotine exposure (Nestler, 2005). The α4β2 receptors are the most widely expressed nAChRs in the brain, with a high binding affinity for nicotine. In addition, these receptors are extensively connected to dopamine (DA) and gamma-aminobutyric acid (GABA) neurons in the brain's ventral tegmental area. Nicotine stimulates DA neurons directly via α4β2 nAChRs, and indirectly by activating excitatory glutamate neurons via α7 nAChRs (Nestler, 2005). These effects further enhance the DA-mediated learning processes that sustain nicotine self-administration and tighten the circle of dependence (Laviolette and van der Kooy, 2004).

Conditioning is a primitive form of learning, and it plays a central role in drug addictions. While the neurobiological processes, such as neuroadaptation to nicotine exposure, reinforce the route to ND, at the same time, the smoker begins to associate certain moods, places, situations and environments with the rewarding effect obtained from smoking (Benowitz, 2009). Animal studies have suggested that exposure to nicotine increases the behavioral control of conditioned stimuli, which could further contribute to the compulsive dimensions of smoking behavior (Olausson et al., 2004). Repeated exposure to smoking in the same situations causes them to become efficient cues for the urge to smoke. Overall, conditioning and neurobiological responses to nicotine in combination enhance the development of ND and maintain smoking behavior (Benowitz, 2009).

Smoking *cessation* causes the emergence of withdrawal symptoms: irritability, depressed mood, restlessness, and anxiety (Hughes and Hatsukami, 1986). Most smokers find difficulties in quitting and the majority of successful quitters relapse. Smoking behavior is maintained mainly by ND (Lerman et al., 2007), and the avoidance of withdrawal symptoms (Benowitz, 2010). During cessation, physiological changes in the central nervous system (i.e., neurobiological adaptation) causes withdrawal symptoms to emerge (Hughes, 2007). The symptoms are, however, temporary (West and Gossop, 1994), but a relatively small proportion of smokers attempting to quit manage to wait until the symptoms cease, and thus relapse.

2.2.2 Established biological pathways associated with smoking

Nicotine is the primary addictive compound in tobacco smoke. Cigarette smokers can control their nicotine levels by altering cigarette consumption, and by controlling the number, volume and depth of inhalation (Hukkanen et al., 2005). Evidently, nicotine intake, the effects it initiates in the central nervous system, and nicotine clearance from the body are regulated by biological pathways. Only a few of them are well characterized, while others lack current comprehension.

Neurobiological pathway

Nicotine is rapidly absorbed from the small airways and alveoli of the lungs, but also through the mucosal layers covering the nose and mouth, and even through the skin (Armitage et al., 1978; Benowitz et al., 1988). From the lungs, nicotine is rushed into the pulmonary venous circulation, reaching the brain within 10-20 seconds (Berridge et al., 2010; Benowitz, 2010).

In the brain, nicotine binds to nAChRs, which are ligand-activated neurotransmitter receptors, and consist of two major subtypes; the metabotropic muscarinic receptors and the ionotropic nicotinic receptors (Albuquerque et al., 2009). Both are activated by the endogenous neurotransmitter acetylcholine (ACh), and expressed by both neuronal and non-neuronal cells throughout the body (Albuquerque et al., 1995).

The ionotropic nicotinic receptors are fast cation channels, which are sensitive to activation by nicotine (Albuquerque et al., 2009). Each nicotinic receptor is built up from five subunits (Dajas-Bailador and Wonnacott, 2004). Each subunit contains a conserved extracellular NH2-terminal domain, four transmembrane protein domains (TM), a cytoplasmic loop of varying size and amino acid order, and a relatively short and variable extracellular COOH-terminal sequence (Albuquerque et al., 2009). The neuronal nicotinic receptors can be either homopentamers or heteropentamers. Figure 2 shows a simplified structure of nAChRs. To date, eight types of α-subunits (α_2, α_3, α_4, α_5, α_6, α_7, α_9, and α_{10}), and three β-subunits (from β_2 to β_4) have been cloned from mammalian neural tissues (Albuquerque et al., 2009). The α-subunits are characterized by a specific extracellular cysteine amino acid pair (Cys-Cys pair) near the entrance of TM1, and this Cys-Cys pair is further required for agonist binding (Karlin et al., 1986).

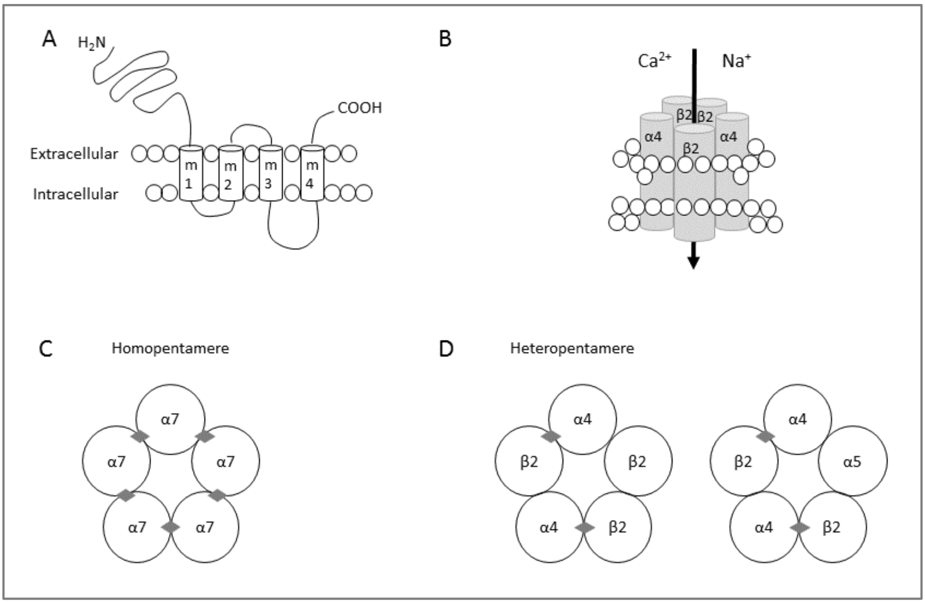


Figure 2. Nicotinic acetylcholine receptor (nAChR) structure. m1-m4 refer to the four transmembrane protein domains. (Adapted from Laviolette and van der Kooy, 2004).

Some neuronal nicotinic receptors are more abundant in the mammalian nervous system than others. For example, the $\alpha_4\beta_2$* (where * denotes the possible inclusion of additional nAChR subunits, complementing the number of required subunits to five) receptor is the most widely expressed receptor subtype in the mammalian brain, followed by expression of the sole homopentameric receptor α_7, and $\alpha_3\beta_4$* nAChRs (Dani and De Biasi, 2001). A recent study has provided an x-ray of the crystallographic structure of the human $\alpha_4\beta_2$ nicotinic receptor (Morales-Perez et al., 2016). The modeled structure provides further insight into the architecture of ligand recognition, heteromer assembly, ion permeability and desensitization (Morales-Perez et al., 2016). Nicotinic receptors are mainly located in the presynaptic terminals of axons, and have a role in the modulation rather than processing of fast postsynaptic transmission (Vizi and Lendvai, 1999).

The binding of nicotine at the receptor binding sites activates the release of several neurotransmitters in the brain (Benowitz, 2010; Dani and De Biasi, 2001; Kenny and Markou, 2001). Release of neurotransmitters, such as DA, serotonin, GABA, and glutamate (Dajas-Bailador and Wonnacott, 2004; Wonnacott, 1997), induces the complex behavioral effects of smoking, including pleasure, arousal, reduction of anxiety and tension, enhancement of memory, increased concentration, and suppression of appetite (Benowitz, 1999). Figure 3 illustrates the biology of ND. DA is responsible for the rewarding effects of nicotine, leading to behavioral reinforcement and addiction (Nestler, 2005; Balfour, 2004). Figure 4 shows a simplified image of neurotransmitter release by nicotine and the diverse psychological effects it initiates.

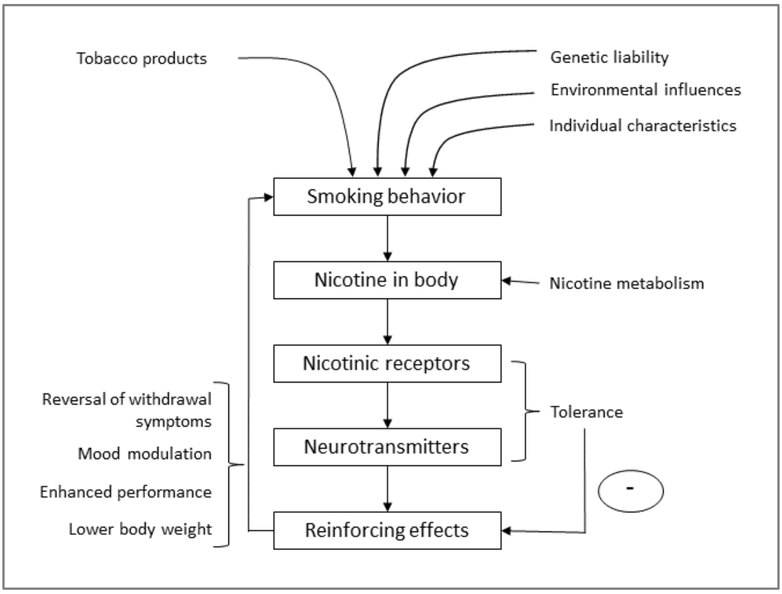


Figure 3. The biology of nicotine dependence. (Adapted from Benowitz, 1999).

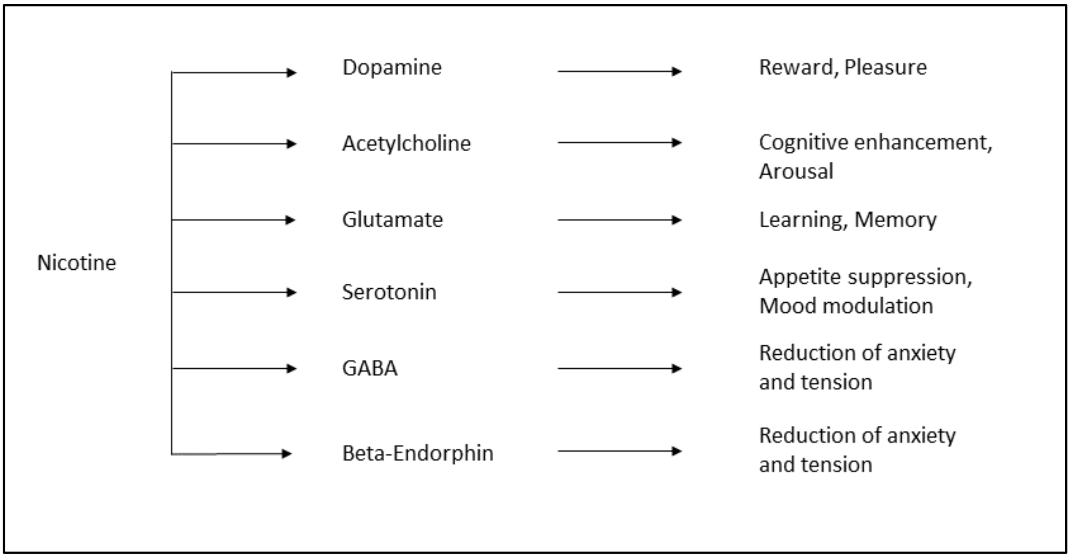

Nicotine
Dopamine
Reward, Pleasure
Acetylcholine
Cognitive enhancement, Arousal
Glutamate
Learning, Memory
Serotonin
Appetite suppression, Mood modulation
GABA
Reduction of anxiety and tension
Beta-Endorphin
Reduction of anxiety and tension

Figure 4. Neurochemical and subsequent psychological effects of nicotine. (Adaptated from Benowitz, 1999).

Nicotine seems to regulate signaling by modulating synaptic transmission. The process is mediated by the release of glutamate, which is believed to affect learning and memory by enhancing synaptic plasticity (Schilstrom et al., 2000). Nicotine also stimulates the release of GABA, which has been associated with reduced anxiety levels (Benowitz, 1999). Notably, both glutamate and GABA affect

DA release (Mansvelder et al., 2009). Long-term exposure to nicotine causes the desensitization of some receptors, which further generates upregulation. Thus, GABA-regulated inhibitory signals diminish, while glutamate-mediated excitatory signals persist (Mansvelder and McGehee, 2002). This relative deficiency of GABA over glutamate may lead to an enhanced DA release in the nucleus accumbens, and the mechanism may be crucial for perpetuating nicotine addiction (Mansvelder et al., 2009).

Nicotine increases the release of serotonin in the brain (Ribeiro et al., 1993), which is believed to be involved in mood regulation and aggression (Veenstra-VanderWeele et al., 2000). Low serotonin levels have been associated with behavioral traits, such as neuroticism and novelty seeking, which have been related to increased smoking behavior, increased ND, and difficulties in quitting (Lerman et al., 2000, Hu et al., 2000).

Long-term chronic exposure to nicotine leads to neuroadaptation (Wang and Sun, 2005) which, together with receptor desensitization, leads to the development of tolerance and dependence (Dani and Harris, 2005). Although the role of the neurobiological pathway in smoking behavior is well-established, the differences within it between any two smokers are less well understood. Evidence from electrophysiological experiments, as well as mouse studies, suggest that large differences in nicotine-induced behavioral effects may result from small differences in the sensitivity and connectivity of neurobiological pathways (Picciotto, 2003).

Nicotine metabolism

Nicotine is metabolized to several metabolites, primarily in the liver. The majority of nicotine, approximately 75%, is metabolized to cotinine by the CYP2A6 enzyme, a member of the cytochrome P450 system (Benowitz and Jacob, 1994). The rest of nicotine is metabolized to various compounds, such as nicotine-glucuronide, and nicotine *N*'-oxide, while around 10% of nicotine is excreted to urine unchanged (Hukkanen et al., 2005). Only 10-15% of cotinine is excreted to urine unchanged (Benowitz and Jacob, 1994), while the remaining is metabolized further, mainly to 3'-hydroxycotinine (Bowman and McKennis, 1962), by CYP2A6 enzyme exclusively (Hukkanen et al., 2005). The main path of nicotine metabolism is shown in Figure 5.

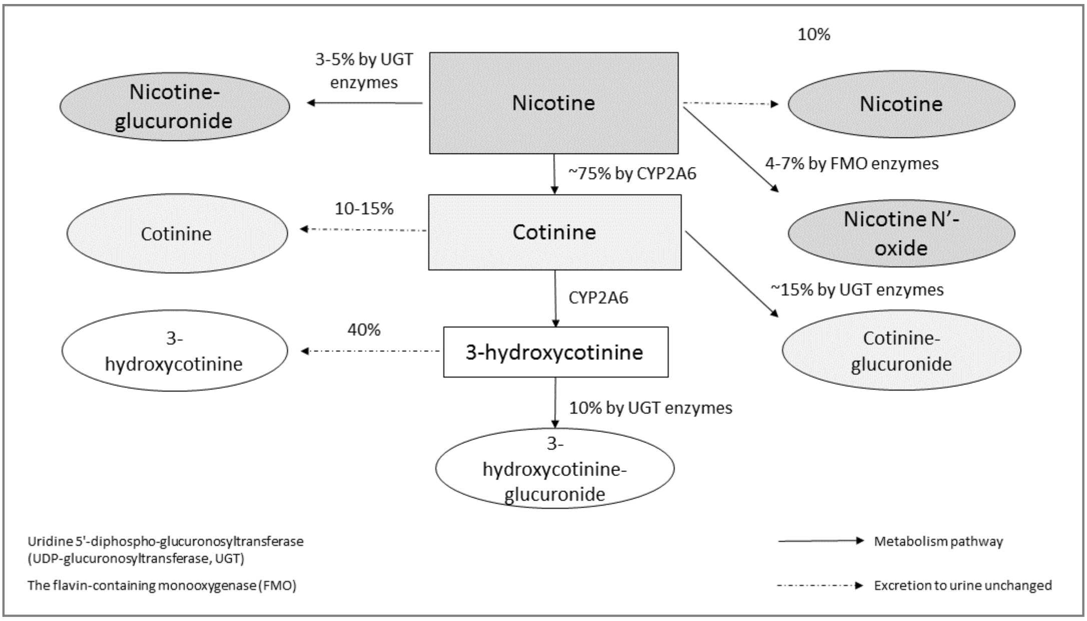


Figure 5. Main pathways of nicotine metabolism. (Adapted from Hukkanen et al., 2005).

The amount and duration of nicotine in the body is determined by the rate of nicotine metabolism. The average half-life of nicotine in the blood is approximately 2 hours, which is relatively fast compared to the half-life of cotinine (16 hours, on average) (Hukkanen et al., 2005). The rate of nicotine and cotinine clearance varies considerably between individuals. On top of genetic variations, which will be discussed in chapter 2.3.3, several factors may explain inter-individual variability in metabolism. These factors include (i) physiological factors, such as diet, age, gender, circadian rhythm, BMI and alcohol use, (ii) pathological conditions, (iii) medications, (iv) smoking and (v) ethnic- or population-specific differences (Hukkanen et al., 2005; Ross et al., 2016). Variations in the nicotine metabolism rate have roughly been categorized into slow (<50% of activity), intermediate (80% of activity) and normal (100% of activity). In general, fast metabolizers smoke more and are less likely to quit, resulting in greater vulnerability to ND (Ray et al., 2009).

2.2.3 Smoking quantity

The number of cigarettes smoked per day (CPD) is an essential phenotype describing daily smoking behavior, and it can be further used as a proxy phenotype for ND, since it builds on the reported evidence that heavier smoking results in heavier ND and causes greater difficulties in quitting (Kaprio and Koskenvuo, 1988; Etter et al., 1999; Senore et al., 1998). This reflects the mechanisms which regulate the development of tolerance, since higher levels of nicotine intake increase the number of nAChR binding sites in the brain (Govind et al., 2009). However, self-reported CPD is prone to reporting bias, since smokers may round the amount to the nearest, or even lower, quantity of ten. In addition, CPD does not measure the precise depth of inhalation from the cigarette in the way biomarkers would.

2.2.4 Nicotine dependence

ND is the main factor in maintaining smoking behavior. Highly dependent smokers find greater difficulty in quitting than non-dependent smokers (Xian et al., 2007; John et al., 2004). But not all smokers are nicotine dependent, only about half of them meet the criteria for ND (Hughes et al., 2006; MacKillop et al., 2010).

Different methods have been designed to assess ND, and the two most commonly utilized, traditional and global measures are the Diagnostic and Statistical Manual of Mental Disorders (DSM), 4th edition (DSM-IV) criteria (American Psychiatric Association (APA), 1994) and the Fagerström Test for Nicotine Dependence (FTND) (Heatherton et al., 1991). These measures are used both in clinics and research (Piper et al., 2008).

The difficulty in designing reliable ND measures lies in the multidimensional (Piper et al., 2008) nature of the phenotype. Both FTND and DSM-IV are unidimensional measures. These two measures also approach ND from different angles: the items included in the FTND were originally designed to measure the concept of physical dependence (Heatherton et al., 1991), while the DSM-IV definition includes concepts such as cognition, behavior and psychological symptoms (APA, 1994). However, despite the diffrences between the two ND measures, both show similar additive genetic influences according to common genetic variation (Bidwell et al., 2016).

The DSM-IV criteria are assessed by diagnostic interview and further structured into seven clusters, depicting loss of control regarding smoking behavior. The diagnosis of DSM-IV ND requires the presence of at least three out of seven criteria during a 12-month period. The DSM-IV ND criteria are listed in Table 1.

Table 1. The nicotine dependence criteria in the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (APA, 1994).

1	Tolerance (described as either of the following):
a.	Absence of nausea, dizziness, and other characteristic symptoms despite using substantial amounts of nicotine.
b.	Diminished effect observed with continued use of the same amount of nicotine-containing products.
2	Withdrawal (as manifested by either of the following):
a.	The characteristic withdrawal syndrome for nicotine (refers to Criteria A and B of the criteria sets for Withdrawal from the specific substances).
b.	Nicotine (or a closely related) substance is taken to relieve or avoid withdrawal symptoms.
3	Nicotine is often taken in larger amounts or over a longer period than was intended.
4	A persistent desire or unsuccessful efforts to cut down or control nicotine use.
5	A great deal of time is spent in activities necessary to obtain nicotine (e.g., driving long distances), use nicotine (e.g., chain-smoking), or to recover from its effects.
6	Important social, occupational, or recreational activities are given up or reduced because of nicotine.
7	Nicotine use is continued despite knowledge of having a persistent or recurrent physical or psychological problem that is likely to have been caused or exacerbated by nicotine.

DSM-IV ND is a psychiatric measure, whereas FTND is a symptom scale assessed by a questionnaire. It comprises six questions, the score ranging from 0 to 10 (Fagerström, 1978). FTND may also be used as a dichotomous measure by applying a cut-off threshold, typically individuals scoring ≥ 4 points being classified as nicotine dependent. The key component that best describes FTND is the difficulty in tolerating reduced nicotine levels (Moolchan et al., 2002). This component is depicted in two essential questions on the FTND questionnaire: “How soon after waking up do you smoke your first cigarette?” and “How many cigarettes do you smoke daily?”. Both questions have been weighted with high scores. The FTND questionnaire is described in Table 2.

Table 2. Fagerström Test for Nicotine Dependence questionnaire and score.

Questions		Score
How soon after waking up do you smoke your first cigarette?	5 minutes	3
	6-30 minutes	2
	31-60 minutes	1
	after 60 minutes	0
How many cigarettes do you smoke per day?	1-10 cigarettes	0
	11-20 cigarettes	1
	21-30 cigarettes	2
	31 cigarettes or more	3
Which cigarette would you hate most to give up?	first cigarette in the morning	1
	other	0
Do you smoke more frequently during the first hours after waking up than during the rest of the day?	yes	1
	no	0
Do you find it difficult to refrain from smoking in places where it is forbidden?	yes	**1**
	no	**0**
Do you smoke even when you are so ill that you are in bed most of the day?	yes	1
	no	0

In an effort to better understand the multidimensional nature underlying ND, which could be used to improve research and, ultimately, treatment possibilities, multidimensional measures have been developed, such as the Nicotine Dependence Syndrome Scale (Shiffman et al., 2004) and the Wisconsin Inventory of Smoking Dependence Motives (Piper et al., 2004). However, even these multidimensional measures have not fully succeeded in capturing ND, thus they have not been able to replace the unidimensional measures in research (Piper et al., 2008).

2.2.5 Nicotine withdrawal

Smoking cessation causes the appearance of nicotine withdrawal (NW) symptoms and these symptoms cause powerful stimuli to sustained smoking (Le Moal and Koob, 2007). Table 3 lists the eight DSM-IV NW criteria. Response to NW results in a cascade of events which involve, for example, increased expression and binding of corticotropin-releasing factor to corticotropin-releasing factor 1 receptors (George et al., 2007). This CRF-CRF1 receptor system has been shown to mediate the response to stress in rats (George et al., 2007). All in all, NW is a complex response to under-activation of the DA system and activation of the CRF-CRF1 receptor system, since they both contribute to NW symptoms, which often predict relapse. The positive reinforcement induced by the DA system, combined with avoidance of withdrawal symptoms, establishes the basis for the pharmacological and physiological aspects of ND.

Table 3. The nicotine withdrawal criteria in the Diagnostic and Statistical Manual of Mental Disorders, fourth edition (APA, 1994).

A	Daily use of nicotine for at least several weeks, and
B	Abrupt cessation of nicotine use, or reduction in the amount of nicotine used, followed within 24 hours by four or more of the following signs:
a.	Restlessness
b.	Anxiety
c.	Irritability, frustration, or anger
d.	Difficulty concentrating
e.	Dysphoric or depressed mood
f.	Insomnia
g.	Decreased heart rate
h.	Increased appetite or weight gain

2.3 The genetics of smoking behavior and nicotine dependence

Although environmental factors, such as the influence of peers and family members, as well as national culture and acceptance affect an individual's smoking behavior, genetic factors also have an established role in the process. A substantial variety of diseases and behavioral traits, including substance use and abuse, aggregate within families (Bierut et al., 1998). This may be due to shared environmental factors, but also due to shared genetic factors. A growing body of evidence from family, twin and adoption studies suggests that the initiation and maintenance of substance use is caused by the complex interplay between the environment, individual characteristics and genetic vulnerability (Kaprio et al., 1984; Carmelli et al., 1992; Heath and Martin, 1993; Madden et al., 1999; Lessov et al., 2004; Xian et al., 2007; Petersen and Sørensen, 2011) (Figure 6).

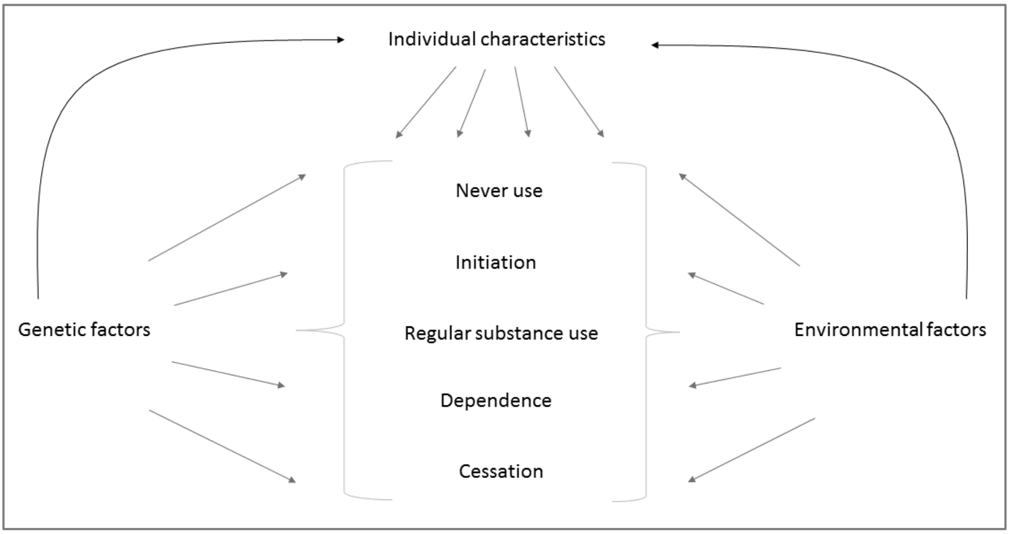


Figure 6. The development of dependence.

While the main picture of the biology influencing ND is known (Benowitz, 2010), the molecular genetic architecture of nicotine addiction is scarce and poorly understood. Over the past few decades, smoking behavior and ND have been extensively researched, with a full spectrum of methods, from positional cloning and linkage-based approaches to candidate-gene and whole-genome association studies, having scrutinized the genome. Still, a majority of the risk variants remain to be discovered.

2.3.1 Heritability estimates

Heritability estimates depict the percentage of phenotypic variation explained by additive genetic factors. Notably, the estimates are highly dependent on the sample sizes and target populations. Twin and adoption studies have demonstrated, via heritability estimates, that genetic factors play a major role in the initiation of smoking, the progression to ND, and the ability to quit smoking (Ho and Tyndale, 2007). ND is a highly heritable phenotype, with estimates varying from 31% to 75% (Vink et al., 2005; Rose et al., 2009; Lessov et al., 2004; Lessov-Schlaggar et al., 2008; True et al., 1999). The heritability estimates for smoking initiation are lower, varying from 35% to 55% (Li et al., 2003; Vink et al., 2005), and further varying between genders (Li et al., 2003). These lower estimates for smoking initiation emphasize the important roles of the shared and unique environments. The heritability estimates for NW vary between 26% and 50% (Broms et al., 2006; Xian et al., 2003; Pergadia et al., 2006; Pergadia et al., 2010), and the estimates are similar between adolescents and adults in both genders (Pergadia et al., 2010). Also, twin studies provide consistent support for the high heritability of nicotine metabolism (Swan et al., 2004; Swan et al., 2005; Loukola et al., 2015).

2.3.2 Nicotinic acetylcholine receptors

Since nicotine binds to nAChRs in the brain, it is likely that the genetic variation in the nAChR subunits have an impact on the inter-individual differences in smoking behavior. So far, members of the nAChR family have indeed provided the most solid genetic evidence for smoking behavior, although the effect sizes are small and explain merely a fraction of the proportion of phenotypic variance in smoking-related phenotypes.

Several candidate gene studies and GWASs studying smoking behaviors and ND have unequivocally identified SNP associations within the gene cluster *CHRNA5-CHRNA3-CHRNB4* on chromosome 15q25 encoding nicotinic receptor subunits α5, α3 and β4. In detail, the 15q25 locus has been implicated in an increased risk of ND (Saccone et al., 2007a; Thorgeirsson et al., 2008; Saccone et al., 2009; Shmulewitz et al., 2016; Olfson et al., 2016) and higher smoking quantity measured with self-reported CPD (Thorgeirsson et al., 2008; Thorgeirsson et al., 2010; The Tobacco and Genetics Consortium, 2010; Liu et al., 2010; Saccone et al., 2010; Li et al., 2010; Shmulewitz et al., 2016), smoking initiation (Li et al., 2010), earlier onset of regular smoking (Hartz et al., 2012), and craving (Shmulewitz et al., 2016). In addition, the locus has been associated with an increased risk of downstream diseases, such as lung cancer (Amos et al., 2008; Hung et al., 2008; Saccone et al., 2010) and COPD (Saccone et al., 2010).

Association signals have been reported on 15q25 with cotinine levels as well (Keskitalo et al., 2009; Munafo et al., 2012; Ware et al., 2016). These results have emphasized the increase in power that comes when using a highly informative phenotype, such as a biomarker: the alleles on 15q25 explain approximately 4-5% of the variance in cotinine levels (Keskitalo et al., 2009; Munafo et al., 2012; Ware et al., 2016), whereas the same alleles explain merely 1% of the variance in self-reported CPD (Thorgeirsson et al., 2008).

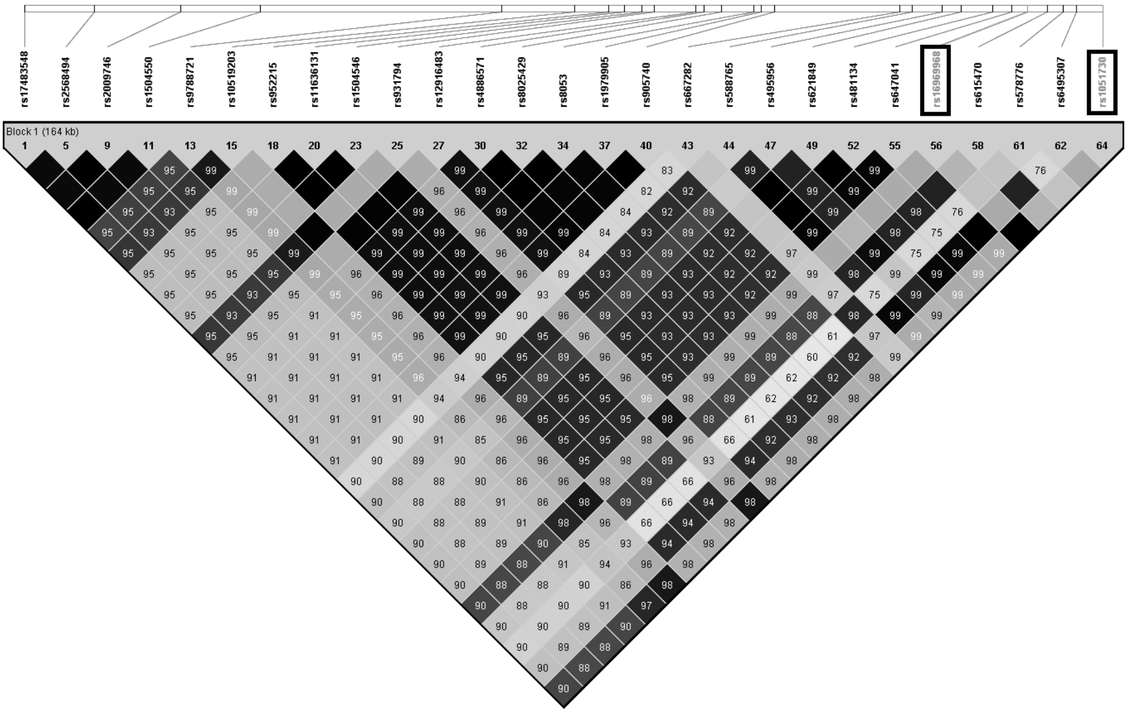


Figure 7. SNPs in high linkage disequilibrium (LD) on 15q25.1. LD pattern obtained from Finnish twin sample (N=2,063). Darker color refers to higher r2 value (correlation coefficient), describing the strength of pairwise LD. In black columns r2 is 1, meaning that two SNPs are fully correlated and provide identical information. Two rs-numbers, rs16969968 and rs1051730, are highlighted: they are fully correlated and perfect proxies for each other.

The associated 15q25 region harbors several highly-correlated SNPs (Figure 7), and subsequent investigations have identified at least two distinct loci effecting smoking behavior (Saccone et al., 2010). The most studied independent locus is tagged by a common missense variant, rs16969968 (D398N), located in the *CHRNA5* gene. This polymorphism changes a conserved amino acid from aspartic acid to asparagine (Bierut, 2011). The risk allele (398N) induces a more extensive receptor desensitization, thus decreasing calcium permeability and increasing nicotine intake (Fowler et al., 2011). Although the allele frequencies of the SNP vary significantly between populations, studies have consistently produced similar effects (Li et al., 2010; Saccone et al., 2009). The many associations of smoking-related phenotypes have been identified using the missense variant, rs16969968, or another polymorphism, rs1051730 located on *CHRNA3*, which is fully correlated with the functional variant (See Figure 7). Gene-by-environment interaction has been reported, since the variant has been associated with an increased risk of ND among men with childhood adversity (Xie et al., 2012) and among early-onset smokers (age at onset ≤16 years) (Hartz et al., 2012).

Despite the well-established role regarding smoking behavior, the *CHRNA5-CHRNA3-CHRNB4* gene cluster has also provided inconsistent results, particularly when examined in samples of African American origin. As an example, in one study, a variant, rs2036527, located at the 5’ enhancer region of the *CHRNA5* gene provided genome-wide significant evidence of association (David et al., 2012), whereas, another study did not detect genome-wide significant association on the gene cluster (Gelernter et al., 2015).

GWASs in general have mainly highlighted common variants contributing to smoking behavior and ND. However, the role of rare variation in the risk for ND in these nicotinic receptor genes has been less investigated. A study utilizing the sequencing approach has identified rare missense variants on *CHRNB4*, which are associated with a reduced risk of ND (Haller et al., 2012). In a recent study, targeted sequencing of *CHRNA5* identified novel rare and low-frequency variants contributing to the increased risk of ND (Olfson et al., 2016). Furthermore, their results indicate that common, low-frequency and rare *CHRNA5* variants are independently associated with the risk of ND (Olfson et al., 2016). However, the region has not been highlighted in linkage studies.

In addition to the gene cluster on chromosome 15q25, evidence from both GWASs and candidate gene-based association studies has highlighted the crucial role of another gene cluster of nAChR subunits, *CHRNB3–CHRNA6* on chromosome 8p11, in ND (Thorgeirsson et al., 2010; Bierut et al., 2007; Saccone et al., 2007a; Rice et al., 2012). These studies have identified two distinct loci within the region, both located upstream or within *CHRNB3* (Wen et al., 2016). Functional studies have suggested that the α6 subunit would have a role in nicotine-induced reward and withdrawal by mediating DA release (Grady et al., 2007).

CHRNA4 has an undisputed role in mediating the responses of nicotine as a receptor subunit of the most abundant nAChR (α4β2) in the brain. Thus, it has been under intense investigation in smoking-related research. Studies applying both linkage and association approaches have yielded evidence supporting the plausible role of *CHRNA4* in smoking behavior, however, the results have been inconsistent.

A meta-analysis of whole-genome linkage scans has reported genome-wide significant linkage with MaxCigs24 (maximum number of cigarettes smoked in a 24-hour period) on 20q13.12-q13.32 harboring *CHRNA4* (Han et al., 2010). Furthermore, a number of rare variants on *CHRNA4* have been identified in humans through sequencing and, notably, at least some of these variants appear to be underrepresented among smokers (Xie et al., 2011). In fact, a rare missense variant, P451L (rs55915440), showed an association with a decreased risk of ND (Xie et al., 2011). On the other hand, a functional study has reported that rare *CHRNA4* variants seem to alter the assembly of α4β2 nAChRs, making them more sensitive to nicotine exposure (McClure-Begley et al., 2014), and one variant, R336C (rs56175056), also included in this functional study, has been associated with an increased risk of ND and with downstream smoking-related diseases, such as lung cancer and COPD (Thorgeirsson et al., 2016).

Candidate gene studies have reported associations of common variants residing on *CHRNA4* with ND phenotypes in both European Americans (Han et al., 2011; Kamens et al., 2013) and African Americans (Han et al., 2011). Also, a common splice site acceptor variant (rs2273500) has provided genome-wide significant association with an increased risk of ND, and association with reduced *CHRNA4* expression levels in human brain and an increased risk of lung cancer (Hancock et al., 2015), although the elevated cancer risk might be mediated by the variant's effect on smoking.

As the research so far has demonstrated, no single variant appears to be responsible for the effects. Instead, the gene harbors several susceptibility variants, both rare and common, that create combinations, which affect inter-individual differences in smoking behavior, further predisposing one to ND.

Although several genetic variants of different nAChR subunits have been shown to affect smoking behavior and ND, a large proportion still remains uncovered. The reported effects are tiny, and the proportion of phenotypic variance explained by the identified genetic variants is modest. The sequencing approach is bound to bring some additional insight on the specific roles of the subunit genes, and possibly enable the discovery of novel rare variants. A denser picture of the susceptibility variants may explain the predisposition of downstream smoking-related diseases.

2.3.3 Nicotine metabolism

Nicotine metabolism, in principal, is mediated primarily by the cytochrome P450 2A6 (CYP2A6) enzyme (Hukkanen et al., 2005). The gene *CYP2A6* (*cytochrome P450, family 2, subfamily A, polypeptide 6*), residing on chromosome 19q13.2, is highly polymorphic, and has been linked with several smoking behavior phenotypes, such as CPD, ND and smoking cessation (Ray et al., 2009; Gold and Lerman, 2012). Individuals who carry either null or reduced activity alleles for CYP2A6 are less dependent on nicotine (Ho and Tyndale, 2007; Gold and Lerman, 2012).

A large-scale GWAS has identified associations between CPD and variants in *CYP2A6* and *CYP2B6* (*cytochrome P450, family 2, subfamily B, polypeptide 6*) (Thorgeirsson et al., 2010). A ratio of two metabolites, 3’ hydroxycotinine:cotinine, referred to as the nicotine metabolite ratio (NMR), is an alternative phenotypic measure of CYP2A6 enzyme activity. In a recent GWAS meta-analysis, the ratio has provided genome-wide significant association with *CYP2A6*, *CYP2B6* and *CYP2A7* (*cytochrome P450, family 2, subfamily A, polypeptide 7*), and notably, the results revealed three independent loci for *CYP2A6* (Loukola et al., 2015; Baurley et al., 2016). Another recent large-scale GWAS meta-analysis of levels of cotinine has identified a genome-wide significant association with a gene variant residing in close vicinity of *UGT2B10* (*UDP glucuronosyltransferase 2 family, polypeptide B10*) at 4q13.2. This gene plays a central role in nicotine metabolism by converting nicotine to nicotine-glucuronide (Hukkanen et al., 2005; Chen et al., 2010), and cotinine to cotinine-glucuronide (Chen et al., 2010). The minor allele of the top SNP, rs114612145, is associated with a reduction in *UGT2B10* function in both glucuronidation steps (Berg et al., 2010; Bloom et al., 2013).

2.3.4 Neuronal pathways

Neuregulin/ErbB signaling pathway

As imputation reference panels have developed, and thus enabled denser mapping of the entire human genome, novel pathways with intriguing connections to smoking behaviors and ND have started to emerge from GWASs. An example of one such novel finding is the association between ND and *ERBB4* (*Erb-B2 Receptor Tyrosine Kinase 4*), coding for a neuregulin receptor on 2q33 (Loukola et al., 2014). The same locus has yielded suggestive evidence of linkage with the regular smoker phenotype (Loukola et al., 2008). Also, a recent mouse study has detected an increase in *ErbB4* and *Nrg3* (*Neuregulin 3*) expression during chronic exposure and withdrawal periods (Turner et al., 2014). Association between *NRG3* and smoking cessation was also detected in a clinical trial (Turner et al., 2014). These genes are involved in the Neuregulin/ErbB signaling pathway. Furthermore, a meta-analysis of genome-wide linkage scans has yielded genome-wide significant evidence of linkage on chromosome 2q with schizophrenia (Ng et al., 2009). The precise mechanisms of this pathway in smoking behavior remain to be verified.

Neurotrophin signaling pathway

The neurotrophin signaling pathway is a vast and complex pathway modulating the effects of synaptic transmission (Bolaños and Nestler, 2004). It involves neurotrophins (Mattson and Meffert, 2006), a family of trophic factors, which have a role in the survival and differentiation of neural cells (Bibel and Barde, 2000). They regulate neurons in the nervous systems through vast signaling cascades, the neurotrophin signaling pathway (Reichardt 2006). Along with modulating synaptic transmission, this pathway also modulates the intricate processes of learning, memory, and even drug addiction (Bolaños and Nestler, 2004). Members of this pathway, *BDNF* (*brain-derived neurotrophic factor*) and *NTRK2* (*neurotrophic tyrosine kinase receptor 2*), have previously been associated with smoking behavior phenotypes – smoking initiation, progression and cessation (Lang et al, 2007; The Tobacco and Genetics Consortium, 2010; Wang and Li, 2010).

2.3.5 Epigenetic findings

The predisposing genetic variants identified to date explain merely a fraction of the phenotypic variation observed in smoking behavior traits and ND, hence, the research has shifted toward examining the epigenetics that mediates the regulation of environmental exposures on the genome. Changes in DNA methylation have been associated with cigarette smoking, and it is one possible mechanism by which tobacco exposure predisposes one to adverse downstream health outcomes. Using DNA methylation data, several genes have been associated with smoking, but some have been more highlighted than others. As an example, CpG sites on an *AHRR* (*Aryl-Hydrocarbon Receptor Repressor*) have consistently been associated with cigarette smoking (Joehanes et al., 2016; Baglietto et al., 2017; Zhang et al., 2016). In evaluating epigenetic studies, one must consider the time- and tissue specificity of epigenetic marks. In the studies of behavioral traits, blood and saliva are frequently used sources, instead of brain tissue, due to their easier accessibility, although it is widely acknowledged that brain tissue would be the correct target.

2.4 Neuropsychiatric comorbidities

Studies of genetics, neuroimaging, and nicotinic receptors support a neurobiological link between tobacco use and alcohol dependence, drug dependence, and several neuropsychiatric disorders, such as schizophrenia, depression, attention-deficit hyperactivity disorder (ADHD), and anxiety disorders (Williams and Ziedonis, 2004). Approximately 50–90% of individuals with mental illness or addiction are nicotine dependent, and the rates vary between diagnosed disorder and the study setting (Williams and Ziedonis, 2004). In the United States, roughly 44% of the cigarettes smoked are smoked by a persons with a psychiatric or substance-abuse disorder (Lasser et al., 2000). Evidence suggests that the rates of smoking appear to be the highest among individuals with a substance-use disorder or patients with a psychotic condition and slightly lower among patients with depression, anxiety, or personality disorder (Williams and Ziedonis, 2004).

The possible factors that increase the risk of co-occurrence of tobacco and other drugs or neuropsychiatric disorders are bound to be diverse. In the first plausible scenario, biological factors include an increased shared genetic vulnerability, a greater susceptibility to smoking behavior and further progression to dependence, owing to a greater subjective experience of reward or pleasure mediated by the dopaminergic system. According to the second scenario, tobacco itself, or a response to nicotine, or both in combination, alleviate some of the symptoms related to a behavioral disorder. Cigarette smoking could merely be an attempt to self-medicate hallucinations or symptoms of depression, anxiety, boredom, loneliness, and other feelings common in this population (Williams and Ziedonis, 2004).

2.4.1 Alcohol and other substance use

Alcohol dependence increases the likelihood of cigarette smoking. Nicotine may act as a conditioning cue for alcohol use. About 60–95% of patients in addiction treatment are tobacco dependent, and about 40–50% are heavy smokers, defined by smoking more than 20 cigarettes per day. Such high rates of tobacco dependence are not merely attributable to lower socioeconomic status (Hughes et al., 1986).

Substance use and abuse are highly heritable, and clearly aggregate within families (Bierut et al., 1998). The aggregation may be due to shared genetic factors, as well as shared environmental factors, or a combination of them. The heritability estimates increase with age, meaning that young individuals are more prone to their unique environments, but as they get older, genes begin to explain a greater portion of the phenotypic variance. This also applies to the co-occurrence of tobacco and alcohol use. Family and twin studies have reported that alcohol use, smoking and ND are transmitted within families (Dick et al., 2009; Rose et al., 2009). Studies clearly demonstrate that alcohol use and smoking tend to co-occur (Li et al., 2007), further suggesting a role of shared genetic predisposition. However, the molecular genetic mechanism underlying that shared predisposition is still largely unknown.

Tobacco use in adolescents is highly correlated with other substance use, and lowers the susceptibility threshold for the onset of other substance abuse and psychiatric illness. Adolescents with high vulnerability to mental illness or other addictions have an increased risk of progressing from experimenting to tobacco use, and further to dependence. Adolescents with adverse family environments with multiple stressors are more likely to experiment with illicit drugs, such as marijuana, and become regular users (Gil et al., 2002).

2.4.2 Neuropsychiatric disorders

The association between smoking and schizophrenia occurs across continents despite different environments and rates of smoking (de Leon et al., 2002). Thus, the increased risk of smoking in schizophrenia may not be merely attributable to environmental or medication effects and may instead be due to shared genetic vulnerability of the neurobiological components mediating the rewarding effects. Individuals with an underlying liability of developing schizophrenia are likely to have shared risk factors that increase their risk of initiating smoking behavior (de Leon and Diaz, 2005).

ADHD and tobacco smoking are among the most common and costly psychiatric and behavioral problems. ADHD has been estimated to occur in approximately 5–8% of children and 4–5% of adults (Visser et al., 2014; Kessler et al., 2006). ADHD is characterized by symptoms such as inattention, impulsivity, and hyperactivity (Kent et al., 2011; Sibley et al., 2012). Childhood ADHD increases the risk for alcohol and other drug use, abuse, and dependence (Charach et al., 2011; Lee et al., 2011). Also, children with ADHD are more likely to become regular cigarette smokers, initiate smoking earlier compared to children without the diagnosis (Kollins et al., 2005; Sibley et al., 2014), and are more likely to proceed as regular smokers in adolescence or adulthood (Molina et al., 2013; Sibley et al., 2014). The rates of co-occurrence of these two common problems are larger than expected by chance (McClernon and Kollins, 2008). The link between smoking and mental illness is well established in epidemiological studies, but its biological mechanism remains unclear, although it has been proposed that ADHD and smoking may involve the dysregulation of dopaminergic and nicotinic-acetylcholinergic circuits, and that these aberrations are partly influenced by genetic variations (McClernon and Kollins, 2008).

Furthermore, evidence shows a clear connection between individual characteristics and proneness to smoking behavior. In particular, impulsiveness-like aspects of novelty seeking were noted as a useful phenotypic marker for the increased risk of substance use initiation (Bidwell et al., 2015).

3 AIMS OF THE STUDY

Family and twin studies have shown the importance of genetic factors contributing to inter-individual differences in the susceptibility to nicotine dependence. Recent GWASs have identified only a few genetic variants contributing to this genetic liability, and the majority of the genetic variants explaining the variance in smoking behavior are still to be discovered. This thesis aimed to identify and characterize the contribution of genes on smoking behavior, ND and other smoking-related co-morbidities and traits, such as alcohol use.

1. To replicate and extend previous linkage findings of nicotine dependence on chromosome 20, and to reveal possible sex-related differences underlying them.
2. To reveal potential pleiotropic effects on 15q24-q25 by studying co-morbid smoking and alcohol use.
3. To investigate the causal relationship between smoking and body mass index.
4. To discover novel genes influencing smoking behavior in a Finnish twin family sample using detailed phenotype profiles.

4 MATERIALS AND METHODS

4.1 Participants

4.1.1 The Older Finnish Twin Cohort (Studies I and II)

The Older Finnish Twin Cohort has been established to examine the genetic, environmental, and psychosocial determinants affecting public health outcomes including several chronic diseases and health behaviors. Participants in the cohort were selected from the Central Population Registry by searching for twin pairs (i.e., individuals born on the same day) with the same last name at birth, of the same sex, and born in the same local municipality. The cohort included same-sex MZ and DZ twin pairs born in Finland before 1958, with both members still alive in 1975 (Kaprio and Koskenvuo, 2002). Questionnaires were distributed in 1975, 1981, and 1990. The 1981 questionnaire was delivered to all twins who were still alive, whereas in 1990, the questionnaire was restricted to pairs born between 1930-1957, with both co-twins resident in Finland in 1987, if they had replied to at least one of the previous surveys. The cohort was expanded in 1996 and 1997 to include opposite-sex twin pairs born in 1938-1949. Like same-sex twin pairs, the opposite-sex twin pairs were identified from the Central Population Registry. The entire Older Finnish Twin Cohort comprises 15,388 twin pairs.

Based on the health questionnaires conducted in 1975, 1981, 1990, and 1996-1997, twin pairs concordant for cigarette smoking were selected from the Older Finnish Twin Cohort and recruited along with their family members to form a *Family Study on Nicotine Dependence* (NAG-FIN) sample. The sample was established to examine genetic and environmental factors contributing particularly to smoking behavior and ND. Data collection, taking place in 2001-2005, included a computer-assisted telephone diagnostic interview based on an adaptation of the Semi-Structured Assessment for the Genetics of Alcoholism (Bucholz et al., 1994) and the Composite International Diagnostic Interview (Cottler et al., 1991), a blood sample for DNA extraction and a mailed self-report questionnaire. The tobacco section for the diagnostic interview was derived from the CIDI (Cottler et al., 1991), and included FTND (Heatherton et al., 1991) and DSM-IV (APA, 1994) assessment of ND. In addition, participants were asked about their lifetime smoking behavior, from smoking initiation to attempted cessation, from the quantity of use for current smokers or the most recent heaviest period of use to DSM-IV symptoms of NW. Thus, the sample consists of unique, detailed phenotypic profiles which aim to portray the complex landscape of smoking behavior. Data from this sample were used in Studies I (*N*=1,302) and II (*N*=1,715).

4.1.2 The National FINRISK Study (Studies III and IV)

The National FINRISK Studies have monitored risk factors in major chronic and non-communicable diseases from 1972 to 2012, in 5-year intervals (Borodulin et al, 2015). For each survey, an independent, random, and representative sample from four to six different parts of Finland, depending on the year of survey, was drawn from the national population register. In the early stages of the survey, in 1972 and 1977, data collection was restricted to North Karelia and Northern Savo as part of the North Karelia Project (Borodulin et al, 2015). In subsequent years of the survey, new areas were added to improve the representativeness of the survey (Borodulin et al, 2015); Turku and the Loimaa region in southwestern Finland in 1982, the capital area including two major cities, Helsinki and Vantaa, in 1992, and provinces of Northern Ostrobothnia and Kainuu in northwestern Finland in 1997. In addition, Lapland was included in the surveys in 2002 and 2007. The surveys included a self-administered questionnaire, with self-reported information about health-related habits, including tobacco and alcohol use, physical measurements and blood samples. Data from cohorts 1992, 1997, 2002, and 2007, including 26,800 genotyped participants (53% women) with a mean age of 48 years, were used in Studies III and IV. In 2007, no blood samples were collected in Lapland, restricting the participation of Lapland samples merely to year 2002. Table 4 describes the distribution and demographics of the National FINRISK Study.

Table 4. The distribution and demographics of the National FINRISK Study participants by year of the survey.

Year of Survey	Region A	B	C	D	E	F	Total
1992	1,316	1,513	1,447	1,284	0	0	5,560
1997	1,577	1,320	1,178	1,558	1,248	0	6,881
2002	1,623	1,190	1,227	1,427	1,244	1,644	8,355
2007	1,225	1,234	1,183	1,126	1,236	0	6,004
Total	5,741	5,257	5,035	5,395	3,728	1,644	26,800

Regions: A = North Karelia, B = Northern Savo, C = Turku and the Loimaa region in southwestern Finland, D = The capital area including two major cities, Helsinki and Vantaa, E = Provinces of Northern Ostrobothnia and Kainuu in northwestern Finland, and F = Lapland.

4.1.3 The Health 2000 Survey (Study III)

The study was aimed to provide an up-to-date comprehensive picture of health and functional ability in the working-aged and aged population in Finland. For the main survey, a random and nationally representative sample was drawn of the population aged 30 and over in order to investigate general health, major chronic conditions, determinants of health and disease, among other things. The survey was carried out in 2000-2001, and included interviews, self-administered questionnaires, physical measurements, determinations from blood samples and clinical examinations (Heistaro, 2008). Data of 5,792 genotyped participants (55% women), with a mean age of 54 years, were used in Study III. The data included self-reported information on tobacco and alcohol use.

Replication samples

Additional individual replication samples were included in Studies II and III. These samples are described in more detail in the original papers. Table 5 summarizes the replication samples and their references.

Table 5. Replication samples.

Sample	Abbreviation	*N*	Study	Reference
FinnTwin12		581	II	Kaprio 2006; Kaprio 2013
Australian twin family sample	NAG-OZALC	5,743	III	Heath et al., 2011
Young Finns Study	YFS	1,982	III	Raitakari et al., 2008

4.2 Phenotypes

4.2.1 Smoking quantity (Studies I, II, III and IV)

Smoking quantity was examined throughout the studies. In the NAG-FIN sample, the question of smoking quantity was addressed to participants who reported regular smoking (i.e., had smoked at least 100 cigarettes in their lifetime and at least once a week for at least two consecutive months). In the data, smoking quantity was defined as CPD and MaxCigs24. Originally, CPD was determined by the number of cigarettes smoked per day during the period of the heaviest smoking as a categorical measure comprising eight categories: 1-2, 3-5, 6-10, 11-15, 16-19, 20-25, 26-39, ≥40. Later, the variable was modified for quantitative use by taking class means from each category. CPD was analyzed as a continuous variable in Study II, whereas MaxCigs24 was examined in Studies I and II.

Study III focused on scrutinizing smoking quantity on a larger scale using two population-based samples, the National FINRISK Study and the Health 2000 Survey. In both samples, smoking quantity was derived from quantitative self-reported CPD. Equally, in both samples, the participants were asked, how many cigarettes, or equivalent tobacco products, they smoke daily or used to smoke before quitting. The answers from different tobacco products were summed as a single quantitative measure to be analyzed as a quantitative trait. CPD measures from the two different population samples were merged to be one large dataset. Additionally, the summed quantitative measure of CPD was analyzed as a binary variable with light smoking controls and heavy smoking cases. Table 6 shows the descriptive statistics of smoking quantity in Study III.

Table 6. Descriptive statistics of the smoking quantity variables utilized in Study III.

Smoking quantity in a Finnish population-based sample				
	% Males	**Mean**	**SD**	**Range**
Quantitative CPD, *N*=8,356	56	14.8	8.9	1-100
Heavy smokers (CPD>20), *N*=1,076	83	31.1	7.6	21-100
Light smokers (CPD≤10), *N*=3,624	40	6.9	3.1	1-10

CPD, cigarettes per day; SD, standard devision.

4.2.2 Nicotine dependence (Studies I and II)

In the NAG-FIN sample, ND has been determined with DSM-IV ND (APA, 1994) and FTND symptoms (Heatherton et al. 1991). They both measure ND, but with slightly differing aspects, and thus correlate only moderately (Broms et al, 2007; Hughes et al, 2004). In this thesis, we utilized these two ND measures. In Studies I and II, we examined DSM-IV ND as a binary diagnosis, requiring three symptoms or more out of seven to occur at the same time within a 12-month period. In addition, in Study II, the DSM-IV ND was analyzed as a continuous variable with the number of DSM-IV symptoms of ND ranging from 0 to 7. FTND was used in Study I as a continuous measure of scored points from the questionnaire, with a score ranging from 0 to 10, and as a binary trait with a cut-off of 4 points. In general, an individual scoring four points or more from the FTND questionnaire is considered to be a nicotine dependent smoker (Heatherton et al., 1991).

4.2.3 Nicotine withdrawal (Study II)

Regular smokers (see description in paragraph 4.2.1) in the NAG-FIN sample were asked about symptoms they might have had in the first 24 hours after they stopped or cut down on cigarettes. They were queried about each DSM-IV NW symptom (described in Table 3). Individuals were diagnosed with DSM-IV NW if they met the criteria of endorsement of at least four out of eight symptoms within 24 hours of quitting or reducing the use of cigarettes (APA, 1994).

4.2.4 Alcohol use (Study III)

Alcohol use was examined in Study III. In the National FINRISK Study, the self-reported measure of weekly alcohol consumption, defined as grams (g) per week, was derived from the question: "How many glasses (restaurant measures) or bottles of alcohol beverages did you drink during the last week?". In the Health 2000 Survey, on the other hand, the participants were asked: "How often and how much did you drink alcohol beverages during the past 12 months?". One drink (a 330-ml bottle of beer, a 12-cl glass of wine or a 4-cl portion of spirits) was estimated to include 12 g of pure alcohol. In both sample sets, an average consumption of alcohol per week (g/week) was calculated for each individual based on the knowledge of how many grams of ethanol the consumed alcohol beverages contain. In the National FINRISK Study, the week's drinks were summed in the g/week variable. In the Health 2000 Survey, alcohol drinks consumed during the past 12 months were summed and divided as an average alcohol use per week (g/week).

After excluding the possibility of sample stratification by first modeling independent samples and adjusting for study region (i.e., ensuring that the sample does not include subpopulations that would vary statistically from the overall population), the samples were merged in order to form a single large sample. Thus, the data included 31,812 individuals with information on alcohol use. The trait was analyzed as a quantitative, categorical, and binary trait. In the quantitative approach, g/week were logarithmically transformed, which provided the best fit for the sample. A total of 20,298 individuals were included in the analysis. Table 7 illustrates the descriptive statistics of alcohol use variables in Study III. For categorical and binary analyses, alcohol consumption was analyzed as standard drinks per week, one alcohol drink containing 12 grams of pure ethanol. For the categorical alcohol trait, individuals were categorized as abstainers and low-frequency drinkers (less than 1 drink per week), light drinkers (1-2 drinks per week), moderate drinkers (3-7 drinks per week for women, and 3-14 drinks per week for men), and heavy drinkers (over 7 drinks per week for women, and over 14 drinks per week for men). In the binary analyses, alcohol use was dichotomized into those drinking at least one drink per week (drinkers) versus those abstaining or drinking less than one drink per week (i.e., less than 12 g of alcohol per week) (abstainers and low-frequency drinkers). This low cut-off for drinking ensures that the sample has the power to detect associations. Also, as heavier alcohol use is frequently associated with smoking, the lower cut-off for drinking can assist in detecting an association which is not veiled by dual use of the two substances. In addition, evidence suggests that even one drink per week increases the likelihood of mortality, while two to six drinks per week increase mortality by 40% (Pyryla et al., 2016).

Table 7. Descriptive statistics of quantitative alcohol use analyzed in Study III.

Alcohol use (g/week) in a Finnish population-based sample				
N	**% Males**	**Mean**	**SD**	**Range**
20,298	53	97.6	124.6	4-2219

g, grams; SD, standard deviation.

4.2.5 Body mass index (BMI) (Study IV)

In Study IV, causal inference was assessed between self-reported smoking status and BMI. Participants were stratified into current, former and never cigarette smokers, and further stratified by sex. The variables height (m) and weight (kg) were directly measured, and BMI was calculated as weight/height^2.

4.2.6 Confounders

If a factor is statistically related to both the outcome and the exposure, it is a confounding variable. Such confounding can be controlled by adjusting the statistical model for that variable. The confounding factor may be commonly accepted as causing confounding for the exposure in question, and it should be accounted for in the analyses. Alternatively, its role might be uncertain, and it should be investigated if a potential confounding variable changes the statistical significance for the observed risk estimates each time in any study sample before conducting the final statistical model.

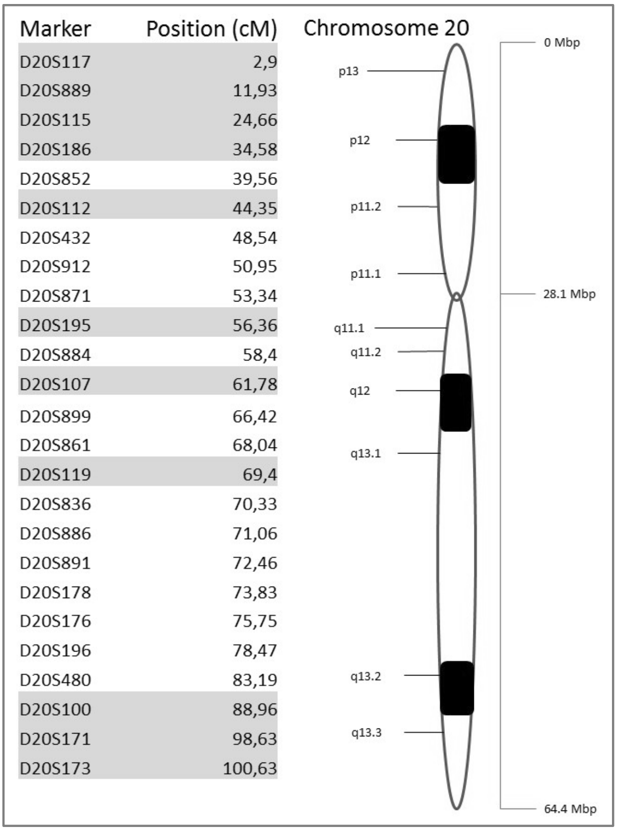
For smoking behavior and ND, sex and age are commonly accepted confounding factors, and thus they were included in the statistical models throughout the studies (Studies I-IV). In addition, Study III examined smoking and alcohol use, which are highly co-morbid traits. Hence, the role of confounding between them was tested. Smoking proved to be a significant confounding factor for alcohol use, and vice versa. Therefore, the analyses of alcohol use as an exposure were adjusted for age, sex and quantitative CPD, whereas the analyses of smoking quantity as an exposure were adjusted for age, sex and alcohol use (g/week; logarithmically transformed).

In Study II, population stratification and relatedness within the sample were accounted for by the covariance matrix. This was determined by calculating a relatedness matrix from the genome-wide genotype data, ultimately representing genetic similarity across individuals.

4.3 Genotyping, imputation and quality control

4.3.1 Microsatellite markers (Study I)

DNA samples from the NAG-FIN data were first genotyped in 2005 in a genome-wide scan which included 380 microsatellite markers (363 autosomal markers), 11 of which were located on chromosome 20 between 2.90 and 100.63 centimorgans (cM) (a distance along a chromosome), yielding an average distance of 9 cM between the markers. Two different platforms, ABI (Applied Biosystems, Foster City, CA, USA) and MegaBASE (Amersham Biosciences, Piscataway, NJ, USA) were used. Samples were genotyped at the former Finnish Genome Center, currently known as the Genotyping Unit of the Technology Center at the Institute for Molecular Medicine Finland (FIMM). The genetic location of each marker was defined using the published deCODE genetic map locations (Kong et al., 2002). For markers that were not included in the deCODE genetic map, linear interpolation was used for obtaining estimates of the genetic locations of these markers by using the physical and the genetic locations of the immediately flanking deCODE markers (Loukola et al., 2008). Figure 8 shows the position of the markers on chromosome 20.



Marker	Position (cM)
D20S117	2,9
D20S889	11,93
D20S115	24,66
D20S186	34,58
D20S852	39,56
D20S112	44,35
D20S432	48,54
D20S912	50,95
D20S871	53,34
D20S195	56,36
D20S884	58,4
D20S107	61,78
D20S899	66,42
D20S861	68,04
D20S119	69,4
D20S836	70,33
D20S886	71,06
D20S891	72,46
D20S178	73,83
D20S176	75,75
D20S196	78,47
D20S480	83,19
D20S100	88,96
D20S171	98,63
D20S173	100,63

Figure 8. Microsatellite markers and their positions on chromosome 20. Markers with gray highlight are the ones that were genotyped in the earlier genome-wide scan, white markers highlight fine-mapped regions.

Based on an earlier whole-genome scan showing the strongest evidence of linkage between MaxCigs24 and chromosome 20 in the NAG-FIN sample (Saccone et al., 2007b), the top four markers were selected and genotyped in an additional sample (*N*=759 participants from 206 families) in 2009. In addition to these four markers, 14 additional markers were genotyped for fine-mapping purposes in both the original whole-genome scan sample and the additional sample. The genotyping was performed using the ABI (Applied Biosystems) platform, and the genotyping was done at the Finnish Genome Center. After fine-mapping, chromosome 20 harbored 25 microsatellite markers in total, with an average distance between markers of 4 cM. Within the fine-mapped region including the 18 markers, the average distance between markers was 2.4 cM.

A genotyping success rate was calculated for each sample and marker, after which Mendelian inconsistencies were checked with the programme PedCheck (O'Connel and Weeks, 1998), which resulted in the removal of eight families yielding more than three Mendelian inconsistencies. No further errors in pedigrees were detected. For locating the unlikely but Mendelian consistent genotypes, we utilized the programme MERLIN (Abecasis et al., 2002), which identifies the unlikely genotypes by an error-detection algorithm and erases them from the data with the command Pedwipe. A genotyping success rate threshold of > 85% was applied to the final data.

Study I was performed in two phases using two data sets: replication data (Study I A) and combined data (Study I B) (reported as Study 1 and Study 2, respectively, in the original publication). In the analyses, data for Study I A consisted of 759 participants (genotyped in 2009) with 18 genotyped markers (4 whole-genome scan markers and 14 fine-mapping markers) on chromosome 20. For Study I B, two data sets, the existing whole-genome linkage scan (genotyped in 2005) with 508 participants (Loukola et a., 2008; Saccone et al., 2007b), and the data included in Study I A (759 participants) were merged. Thus, Study I B consisted of 1,302 participants, of whom (a) 485 were genotyped with all 25 markers (11 genome-scan markers and 14 additional markers for fine-mapping), (b) 794 were genotyped with 18 markers (4 genome-scan markers and 14 fine-mapping markers), and (c) 23 were genotyped with only 11 genotyped genome-scan markers (i.e., sample was included in the whole-genome linkage scan, but the fine-mapping was unsuccessful).

4.3.2 Single nucleotide polymorphisms (Study II-IV)

Genome-wide genotyping for the NAG-FIN sample was performed using two high-density SNP arrays, the Human670-QuadCustom Illumina BeadChip (Illumina, Inc., San Diego, CA, USA) (*N*=1,162) and the Illumina Human Core Exome BeadChip (*N*=901), at the Welcome Trust Sanger Institute and at the Broad Institute of MIT and Harvard. In the pre-imputation quality control protocol, SNPs with a MAF < 1%, a SNP and sample call rate < 95% (< 99% for SNPs with MAF < 5%), and a Hardy–Weinberg equilibrium (HWE) test P < 1x10^{-6} were excluded from the data generated with the Human670-QuadCustom Illumina BeadChip. Similarly, SNPs with a minor allele count < 2, sample call rate < 98%, SNP call rate < 95% (< 99% for SNPs with MAF < 5%), and HWE test P < 1x10^{-6} were excluded from the data generated with the Illumina Human Core Exome BeadChip. Samples were also excluded if they contained non-Finnish ancestry, excess heterozygosity, duplicates or were of ambiguous sex. Pre-phasing of the data was performed with SHAPEIT2 (Delaneau et al. 2013).

The data was imputed with IMPUTE version 2.3.1 (Marchini et al., 2007; Howie et al., 2009; Howie et al., 2012) utilizing a combined reference panel consisting of the 1000 Genomes September 2013 release and an all-Finnish low-pass whole genome sequence (*N*=1,941) from the SISu project, both of which have been conducted using the human reference sequence GRCh37/hg19 assembly. Imputation was followed by post-imputation exclusion criteria, which were applied for each SNP. The criteria included an effect allele frequency < 1% and > 99%, a SNP call rate < 95%, HWE P < 1x10^{-6}, and imputation info < 0.8. Both quality controls and imputation for the GWAS data were done at FIMM, University of Helsinki, Helsinki, Finland.

For Study III, three SNPs were selected for genotyping based on a previous meta-analysis of binary CPD contrasting heavy smoking cases vs. light smoking controls (Saccone et al., 2010) in both the National FINRISK Study and the Health 2000 Survey samples, at the same time. The SNPs included rs16969968 (*CHRNA5*; locus 1), rs578776 (*CHRNA3*; locus 2), and rs588765 (*CHRNA5*; locus 3). DNA was derived from whole blood samples, from which the DNA was extracted at the laboratory of molecular genetics at the National Institute of Health and Welfare (previously called the National Public Health Institute). Genotyping of the DNA samples was carried out using standard protocols of iPLEX Gold technology on the MassARRAY System (Sequenom, San Diego, CA, USA). For all SNPs, the success rate was > 99%, and they were in HWE (P > 0.05). MAFs were 32%, 32%, and 38% for rs16969968, rs578776, and rs588765, respectively.

4.4 Statistical methods

4.4.1 Linkage analysis (Study I)

Linkage analysis was performed in Study I using selected smoking behavior traits and microsatellite markers on chromosome 20. Both single- and multipoint linkage analyses were performed using MERLIN (Abecasis et al., 2002). In singlepoint analysis, linkage was calculated for each marker separately, while multipoint analysis took into account the adjacent markers as well. The binary DSM-IV diagnosis for ND and FTND were analyzed using the nonparametric linkage analysis by Whittmore and Halpern's NPL statistics (Whittmore and Halpern, 1994), which tests for allele sharing among affected individuals. The analysis was performed using two types of allele sharing statistics: pairs and all. The pairs option counts the number of identity-by-descent (IBD) (referring to segments of DNA shared by individuals that has been inherited from a common ancestor without any intervening recombination events) shared alleles for each pair of affected relatives, whereas the all option creates arbitrary groups of individuals to count the IBD alleles shared by all. Continuous traits, FTND and MaxCigs24, were analyzed with variance components linkage analysis using MERLIN (Abecasis et al., 2002).

In cases where significant linkage was observed, the signal was followed up by sex-stratified analyses. Additionally, in Study I B, regular smokers were divided into groups of current and former smokers, and analyzed separately. The results of linkage analysis are presented as LOD scores, and for any trait reaching statistically significant linkage, the maximum LOD (MLOD) score, showing the most significant results of the trait, is presented.

Simulation studies provide a means to determine the empirical significance for detected linkage findings. In Study I, simulations were performed using MERLIN (Abecasis et al., 2002), which simulated 1,000 genome-wide linkage scans of comparable structure, and analyzed each simulated scan identically to the original data analysis. In MERLIN, simulation is created by a gene-dropping pattern. While maintaining the genetic map, phenotype data, pedigree structure, and missing genotype data patterns, it creates comparable data with random marker genotypes. The data are simulated under the hypothesis of no linkage, and thus, any detected linkage is due to chance alone. This allows for the evaluation of the false-positive rate of the data set. The empirical p-value for a certain LOD score was defined as the proportion of simulated genomes where the LOD score in question was reached or exceeded.

4.4.2 Association analyses (Study II-IV)

GWAS

In Study II, we performed a GWAS with selected smoking behavior phenotypes. The analysis was performed using GEMMA (Genome-wide Efficient Mixed Model Association) (Zhou and Stephens, 2012). Allelic dosage data were used to model the genotypes. The genetic associations were acquired with a linear mixed model in which the phenotypeS of interest (smoking quantity defined as CPD and MaxCigs24), DSM-IV NW (diagnosis and symptoms), and DSM-IV ND (diagnosis and symptoms) were the dependent variables and the coded allele dose (posterior mean genotypes) the independent variable. The obtained effect sizes of the binary traits, DSM-IV NW diagnosis and DSM-IV ND diagnosis, were transformed to the odds-scale for more meaningful interpretation using the formula suggested by Pirinen and colleagues (2013). The transformation has been shown to provide accurate effect sizes when the data fulfills the following criteria: the genetic effects are small, the case-control ratio is balanced and the MAF is above 5% (Pirinen et al., 2013). P-values below $5x10^{-8}$ were considered to be genome-wide significant.

In Study II, genomic loci showing genome-wide significant association were further targeted with conditional analyses to determine whether the signal represents an independent effect, or whether there is more than one signal affecting the smoking behavior. Thus, we performed additional association analyses for loci of interest, adjusting the analysis for the SNP with the lowest p-value. The next strongest signal was identified from the results and included in the subsequent round of conditional analyses in an iterative fashion until no residual genome-wide significant signal ($P < 5x10^{-8}$) remained.

Candidate gene analysis

In Study III, associations were tested between three SNPs of interest, rs16969968 (locus 1), rs578776 (locus 2), and rs588765 (locus 3), and smoking quantity, using both quantitative self-reported CPD and dichotomous CPD (light versus heavy smokers). For quantitative CPD, the associations were modeled using negative binomial regression, which provided the best fit for the trait distribution in the large, merged population-based sample. Associations between the loci and dichotomous CPD were evaluated using logistic regression. Each SNP was tested separately in a single-SNP model. In addition, to test whether additional loci contribute to dichotomous smoking quantity over and above the effect of locus 1, joint models for locus 1 and each of the other two loci were analyzed. Throughout the analyses, genotypes for all three SNPs were coded additively, as the number of copies of the minor allele.

Alcohol use was examined as either a quantitative, categorical, or dichotomous trait. Associations between log-transformed alcohol use (g/week) and the SNPs were modeled using linear regression, 4-level categorical alcohol use (abstainers and low-frequency drinkers, light, moderate, and heavy drinkers) was tested using multinomial logistic regression, and logistic regression was performed for dichotomous alcohol use (abstainers and low-frequency drinkers versus drinkers).

A goodness-of-fit test was used in order to capture the best fit for the locus-specific dominance model (recessive, dominant, additive, and co-dominant). The additive model provided the best fit for the traits and loci, with the exception of dichotomous alcohol use and locus 3, for which the co-dominant model was most suitable.

Acknowledging the fact that large population-based cohorts might include former alcoholics as participants, and that they might end up in the abstainers group when categorizing the group of abstainers and low-frequency drinkers, the analysis of dichotomous alcohol use was repeated among never smokers. This model inhibits the confounding effect of smoking.

All statistical analyses in Study III were performed with Stata 11.1 (StataCorp. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP, 2009).

4.4.3 Mendelian randomization meta-analysis (Study IV)

In Study IV, the National FINRISK Study sample was included in a large-scale meta-analysis conducted by the Causal Analysis Research in Tobacco and Alcohol (CARTA) consortium (http://www.bris.ac.uk/expsych/research/brain/targ/research/collaborations/carta/). The meta-analysis was performed using data on individuals (≥ 16 years) of European ancestry from a total of 29 studies. For each cohort, on-site Mendelian Randomization analyses between a SNP, rs16969968 (*CHRNA5*), and BMI were performed using linear regression, stratified by smoking status and sex. Hence, placed into the model of Mendelian Randomization (see chapter 2.1.2), rs16969968 refers to factor Z, which has a robust effect on smoking (X), but does not influence BMI (Y). BMI was analyzed as a log-transformed variable assuming an additive genetic model. Hence, each effect size could be exponentiated to represent the percentage increase in BMI per minor (risk) allele. Analyses were conducted using Stata 13 (StataCorp. Stata Statistical Software: Release 13 College Station, TX: StataCorp LP, 2013).

The results from individual studies were pooled for fixed effects meta-analyses using the “metan” command in Stata (version 13). The “metareg” command was used to examine whether SNP effects varied by sex. Since no evidence for effect modification by sex was observed, the estimates of effect sizes from different data sets were pooled together.

4.4.4 Infering the functional potential of the SNPs (Study II)

In Study II, the functional potential of the genome-wide significant SNPs was assessed utilizing the publicly available database Ensembl Variant Effect Predictor (McLaren et al., 2016). To examine whether the SNPs have an effect on expression levels of nearby genes, eQTL analyses were performed using blood- and brain tissue-derived data from the Genotype-Tissue Expression (GTEx) database (GTEx Consortium, 2015) and brain tissue-derived data from the Brain eQTL Almanac (BRAINEAC) database (Ramasamy et al., 2014). To assess whether the SNPs affect the methylation within adjacent genes, blood-derived meQTL data from the mQTLdb (mqtldb.org) (Gaunt et al.,

2016) and BIOSqtl (genenetwork.nl/biosqtlbrowser) (Bonder et al., 2016) databases was examined, as well as meQTL data derived from fetal brain (epigenetics.essex.ac.uk/mQTL) (Hannon et al., 2016).

5 RESULTS AND DISCUSSION

5.1 Genetics of nicotine dependence

5.1.1. Chromosome 20 shows evidence of linkage with ND

Chromosome 20 is a highly relevant candidate chromosome for genetic mapping studies for smoking-related phenotypes, since it harbors the nAChR subunit α4 at 20q13.2–13.33. The α4β2 receptor is the most widely expressed receptor subtype in the mammalian brain (Dani and De Biasi, 2001), and thus plays a major role in mediating and modulating the release of various neurotransmitters (Benowitz, 2010). Previous studies utilizing a subset of 505 individuals from the NAG-FIN sample have yielded linkage on 20p13 with DSM-IV ND (Loukola et al., 2008) and on 20q13 with MaxCigs24 (Saccone et al., 2007b; Han et al., 2010). Thus, in Study I, chromosome 20 was specifically targeted with linkage analyses using the NAG-FIN sample, in order to replicate and extend previous findings of ND. The analyses were performed in two phases, using a replication sample (Study I A), and a combined sample (Study I B).

Table 8. Linkage results from single- and multipoint linkage analyses in Study I A and I B.

Chromosome 20 linkage results for DSM-IV ND diagnosis						
			Study I A (N=759)		**Study I B (N=1,302)**	
			Singlepoint LOD	Multipoint LOD	Singlepoint LOD	Multipoint LOD
Marker	cM	Location	Pairs	Pairs	Pairs	Pairs
D20S852	39.56	20p12.1	1.122	1.937	0.403	0.793
D20S112	44.35	20p12.1	2.101	2.138	1.743	1.074
D20S432	48.54	20p11.23	0.89	2.767	0.894	1.483
D20S912	50.95	20p11.23	1.581	**3.412**	1.462	2.348
D20S871	53.34	20p11.21	**3.806**	**3.221**	1.675	1.819
D20S195	56.35	20q11.21	1.297	2.894	1.332	1.65
D20S884	58.40	20q11.23	1.83	2.363	1.442	1.182
D20S107	61.78	20q12	1.133	1.761	0.978	0.824
D20S899	66.42	20q13.11	1.924	1.005	1.071	0.383

Bolded values refer to statistifically significant LOD scores.

In Study I A, a sample of 759 individuals yielded significant linkage with the DSM-IV ND diagnosis on 20p11.21 (D20S871, MLOD score 3.8 (singlepoint)) (Table 8). Only pairs-results are presented in the table, since pairs and all statistics provided similar results. This region harbors several genes, such as *NXT1* (*NTF2-like export factor 1*), *GZF1* (*GDNF-inducible zinc finger protein 1*), *NAPB* (*N-ethylmaleimide-sensitive factor attachment protein, beta*), and a cystatin locus, which contains the majority of the type-2 cystatin genes and pseudogenes. These type-2 cystatin proteins are a class of cysteine proteinase inhibitors found in a variety of human fluids and secretions. The cystatin locus within this region is of interest, since cystatins have been shown to resemble the α subunits of nAChRs (Steinlein et al., 1996). Both cystatins and α subunits contain four cysteine residues, which form two disulfide bonds serving as agonist binding sites in nAChRs (Steinlein et al., 1996). Other genes within the linkage locus do not provide as clear a connection as the cystatins.

Increasing the sample size from 759 to 1,302 in Study I B shifted the linkage peak from 20p11.21 (53.34 cM) to 20p11.23 (50.95 cM), showing linkage with DSM-IV ND diagnosis (D20S912, MLOD score 2.3 (multipoint)) (Table 8). In the previous genome-wide linkage scan utilizing a subset of 505 individuals from the NAG-FIN sample (Loukola et al., 2008), linkage was observed at 20p13. Evidently, an increase in sample size has shifted the linkage region from the distal part of the short arm toward the centromere. In our study, linkage was observed at 20p12.1 (D20S112, MLOD score 1.7 (singlepoint)). This marker is located in an intron of *PCSK2* (*proprotein convertase subtilisin/kexin type 2*); the gene has three transcript variants, and the numerical order of the intron is related to the transcript variant. PCSK2 serves as a catalyst, releasing protein hormones and neuropeptides from their precursors. *PCSK2* is predominantly expressed in thyroid, and only at moderate levels in other organs and tissues, including the human brain (genome.ucsc.edu). Interestingly, *PCSK2*, along with other adjacent genes within that region, has been reported to associate with smoking cessation in a GWAS (Uhl et al., 2010). Stratification by smoking status did not yield any significant linkage signals.

No evidence of linkage was observed with the binary FTND trait, and the continuous FTND yielded merely moderate linkage at 20p12.1 (D20S112, MLOD score 1.5 (singlepoint); data not shown). However, both binary DSM-IV ND diagnosis and continuous FTND highlight the same marker (D20S112) in singlepoint analyses. Further scrutiny showed significant correlation between the two ND measures, with a tetrachoric correlation of 0.69 ($P < 0.001$, standard error (SE) 0.03). Although the two ND measures demonstrate a similar direction in the results, it has been proposed that they measure ND from slightly different points of view (Hughes et al., 2004, Broms et al., 2007). Still, the results could have been improved using a measure of shared variance across the two traits (Bidwell et al., 2016), which, however, was not tested in Study I. Clinical trials support the distinction between the two measures, as DSM-IV ND and FTND rarely yield consistent results in them (Moolchan et al., 2002; Piper et al., 2006; DiFranza et al., 2010).

MaxCigs24 did not yield significant linkage on chromosome 20 (D20S119, MLOD score 1.3 (singlepoint); data not shown). Thus, the linkage detected in Study I B is modest compared with the two previous studies reporting linkage on chromosome 20 with MaxCigs24 (Saccone et al., 2007b; Han et al., 2010). Both previous reports included a subset of the sample used in Study I B.

Whenever significant linkage was observed in Study I B, regular smokers were divided into groups of current and former smokers, and analyzed separately, which did not provide any evidence of linkage.

5.1.2 Linkage results reveal sex differences

Men and women differ in their smoking behaviors. Women are less likely to attempt quitting and may be more vulnerable to relapse compared to men (Becker and Hu, 2008). However, it is unclear which factors influence the difference. Women tend to respond more to environmental smoking cues than men, which may be influenced by the hormonal changes during the menstrual cycle (Lynch, 2009). Clear pharmacological sex differences in smoking and ND suggest that genetic factors may play a role in the sex differences of ND (Benowitz and Hatsukami 1998).

To address the importance of identifying genetic factors that explain sex differences in ND, sex-stratified linkage analyses for DSM-IV ND were performed on chromosome 20 (Study I). A better understanding of the factors that influence the neurobiology underlying sex differences in smoking behavior, including genetic factors, will aid in designing individualized and optimized smoking cessation therapies, and further lead the way to sophisticated precision medicine.

In the replication sample of Study I A, a clear distinction between linkage loci was observed. Males (*N*=374) provided suggestive evidence for linkage at 20p11.21 (D20S871, MLOD score 2.6 (singlepoint)), whereas females (*N*=385) yielded significant linkage at 20q11.23 (D20S884, MLOD 3.4 (singlepoint)). Table 9 shows the most significant linkage results from the sex-stratified analyses. Only results from all statistics are presented in the table since pairs and all statistics provided similar results.

Table 9. Sex-stratified linkage results.

Sex-stratified linkage results on chromosome 20 for DSM-IV ND diagnoses									
		Study I A (N=759)				**Study I B (N=1,302)**			
		Males N=188		**Females N=149**		**Males N=313**		**Females N=235**	
Marker	Location	LOD [a]	LOD [b]	LOD [a]	LOD [b]	LOD [a]	LOD [b]	LOD [a]	LOD [b]
D20S852	20p12.1	0.332	0.901	1.741	2.430	0.001	0.113	1.827	2.325
D20S112	20p12.1	0.810	0.938	2.591	2.944	0.245	0.126	2.232	2.620
D20S432	20p11.23	0.597	0.992	0.933	2.520	0.188	0.147	0.653	1.969
D20S912	20p11.23	0.794	1.266	1.177	2.742	0.580	0.498	1.377	2.023
D20S871	20p11.21	2.654	1.410	2.155	2.224	0.879	0.720	1.727	1.376
D20S195	20q11.21	0.457	1.212	1.390	2.357	0.633	0.775	0.910	1.244
D20S884	20q11.23	0.441	0.953	**3.441**	2.319	0.183	0.461	**3.259**	1.269
D20S107	20q12	0.701	0.691	1.196	1.931	1.224	0.377	0.715	0.913
D20S899	20q13.11	1.235	0.459	1.243	1.078	1.258	0.254	1.370	0.519

[a] LOD score from a singlepoint linkage analysis using all statistics; [b] LOD score from a multipoint linkage analysis using all statistics. Bolded values refer to statistifically significant LOD scores.

Male-specific linkage signal peaks were seen from the same region as the signal for the pooled analyses in Study I A (See chapter 5.1.1). The analysis of the female group highlights a region at 20q11.23, which harbors several genes in close proximity to marker D20S884. These genes include *SRC (v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian))*, *BLCAP* (*bladder cancer associated protein*), and *NNAT* (*neuronatin*).

In Study I B, an increase in the sample size resulted in findings similar to the Study I A results in females (*N*=650), because the evidence for linkage remained invariable at 20q11.23 (D20S884, MLOD 3.3 (singlepoint)) (Table 8). Figure 9 shows the single- and multipoint linkage results for females (A) and males (B). In males (*N*=652), singlepoint analysis yielded a modest MLOD score of 1.3 at 20q13 with marker D20S899 (Table 8), and thus the suggestive evidence of linkage observed with marker D20S871 in a smaller sample in Study I A diminished. Nevertheless, these results clearly suggest that chromosome 20 harbors genetic elements which may influence sex differences in ND.

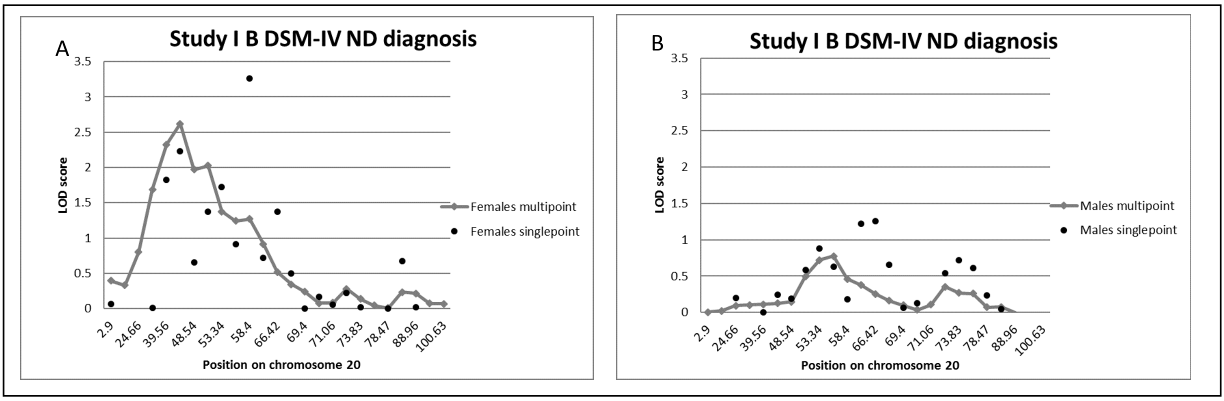


Figure 9. Study I B sex-stratified linkage results on chromosome 20, all statistics, in females (N=650) (A), and in males (N=652) (B). X-axis of the figures describes the marker location (cM) along chromosome 20; Y-axis presents the LOD score for each marker obtained from the linkage analyses.

In the simulation analyses, the highest LOD score (single-point 3.3 in females) produced an empirical p-value of <0.001 in the permutation analyses based on 1,000 permutations. A LOD score this high (3.3) was not reached even once (highest LOD score obtained in the simulations for females was 3.2).

None of the analyses yielded linkage near 20q13.2–q13.33, which harbors *CHRNA4* encoding the nAChR subunit α4. However, the rather wide multipoint linkage signal peak detected approximately 26 Mb upstream of *CHRNA4* may indicate genetic elements in close proximity to *CHRNA4*. Hence, the role of *CHRNA4* cannot be ruled out. In fact, a recent GWAS meta-analysis detected an association between a *CHRNA4* splice site variant and ND measured by FTND (Hancock et al., 2015). The NAG-FIN sample was included in that study as a replication cohort. Since GWAS identify mainly common variants, it is not surprising that the linkage approach in Study I did not yield evidence of linkage close to *CHRNA4*. Whether the genes residing at 20q11.23 have an impact on sex differences in ND remains unsolved in this study, and further investigation with larger samples is required. In both Studies I A and I B the sample sizes of the sex groups are small, which may result in false positive signals. However, the linkage signal on the 20q11.23 region detected in the female group remains significant in Study I B, which strongly argues against a false positive finding.

It is worth emphasizing that the linkage result on chromosome 20, predominantly driven by females, is captured by DSM-IV ND. The binary FTND measure did not yield any linkage in Studies I A or I B in either of the sex groups. FTND provides more accurate measures of pharmacological and physical dependence, while DSM-IV ND focuses more on the behavioral and cognitive layers of ND (DiFranza et al., 2010). Based on these arguments, and since females may be more prone to environmental smoking cues than males (Lynch 2009; Franklin et al., 2015), our sex-stratified results are consistent with the assumption of the differences between the two ND measures.

5.1.3 Genome-wide association study for smoking behavior highlights 16p12.3 and 11p15.5 (Study II)

Chromosome 16p12.3 and a region near gene CLEC19A

Owing to the scarce underlying genetic contribution for smoking behavior and ND, Study II aimed to identify novel genetic variants influencing smoking behavior in a Finnish twin family sample utilizing genotype data imputed using a 1000 Genomes Phase I reference panel (1000 Genomes Project Consortium, 2012) and Finnish population-specific reference panel from SISu project. In Study II, we conducted a GWAS to examine the impact of common and low-frequency variants on smoking behavior traits (*N*=1,715).

Smoking quantity (both CPD and MaxCigs24) yielded genome-wide significant association on 16p12.3, and thus strengthened our previously reported GWAS findings with an overlapping sample (a subset of 1,114 individuals from the current sample) (Loukola et al., 2014). The strongest association signal for CPD was detected with rs4300632 (P=8.5x10^{-9}) and for MaxCigs24 with rs2353663 (P=7.0x10^{-9}). Table 10 shows the top 20 association results for CPD. Beta coefficients are reported for minor alleles, creating a more meaningful approach for reporting the results. The association signal peaks from region chr16:19,326,020 – 19,357,771, and harbors altogether 23 genome-wide significant SNPs. Also, according to conditional analysis, the association region harbors only one independent locus. Within the strongest association region on chr16:19,326,020 – 19,357,771, the SNPs are in LD with each other (Figure 10).

Table 10. Top 20 genetic variant results for cigarettes per day.

Chr	Position[a]	SNP	Allele 1	Allele 2	*p*-value	β	SE	MAF
16	19350508	rs4300632	T	A	8.5E-09	4.7	0.82	0.046
16	19351749	rs11074386	T	C	9.8E-09	4.7	0.82	0.046
16	19355577	rs11074388	A	G	1.1E-08	4.4	0.76	0.055
16	19353064	rs916078	T	C	1.2E-08	4.7	0.82	0.046
16	19351837	rs11074387	G	A	1.6E-08	4.4	0.77	0.055
16	19352221	rs4782229	T	C	1.6E-08	4.3	0.77	0.055
16	19352463	rs916077	A	G	1.8E-08	4.3	0.77	0.055
16	19352499	rs2353661	A	C	1.8E-08	4.3	0.77	0.055
16	19353107	rs139902244	ATTATTACCCC	A	1.9E-08	4.3	0.77	0.055
16	19353159	rs762761	A	G	1.9E-08	4.3	0.77	0.055
16	19353246	rs762762	C	T	2.0E-08	4.3	0.77	0.055
16	19353328	rs4238673	C	A	2.0E-08	4.3	0.77	0.055
16	19353409	rs4782232	G	A	2.0E-08	4.3	0.77	0.055
16	19353429	rs4782233	C	T	2.0E-08	4.3	0.76	0.056
16	19353430	rs4782234	T	C	2.0E-08	4.3	0.76	0.056
16	19353940	rs2213280	A	G	2.1E-08	4.3	0.77	0.055
16	19352688	rs2353664	G	A	2.3E-08	4.4	0.78	0.055
16	19354089	rs2227005	T	C	2.3E-08	4.3	0.77	0.055
16	19352685	rs2353663	T	G	2.4E-08	4.4	0.78	0.055
16	19352684	rs2353662	C	T	2.6E-08	4.4	0.78	0.055

[a] Base pair position according to human reference sequence Build37/hg19 assembly

Abbreviations: Chr, chromosome; SNP, single nucleotide polymorphism; Allele 1, effect allele (minor allele); Allele 2, alternative allele (major allele); β, effect size provided by beta-value; SE, standard error; MAF, minor allele frequency.

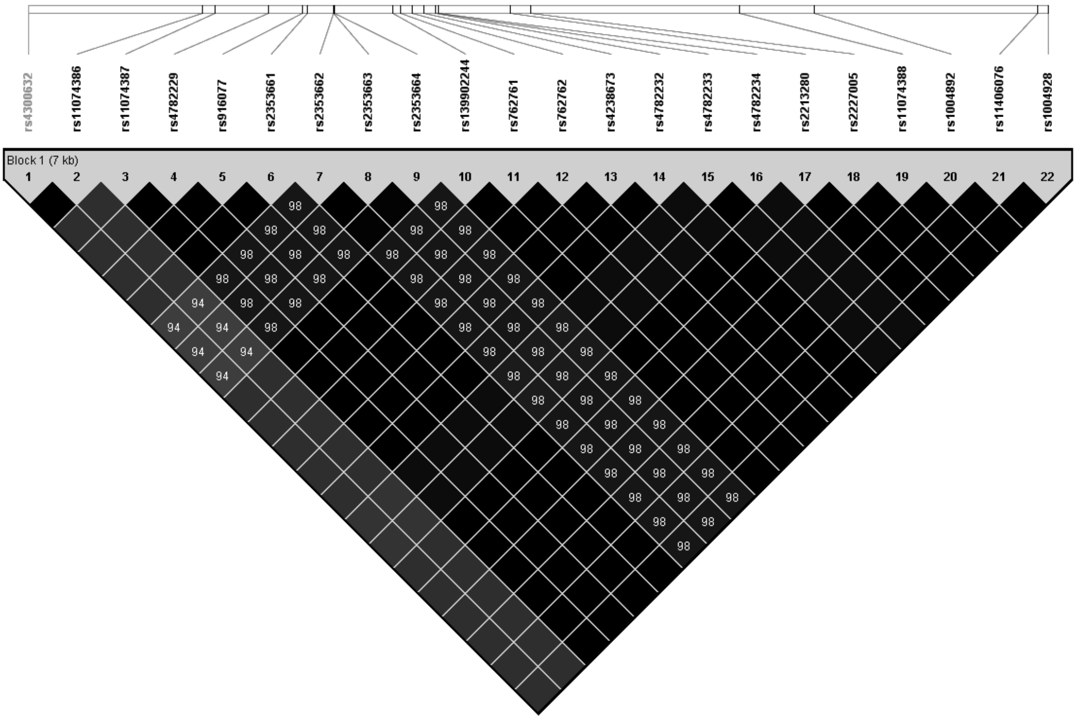


Figure 10. The linkage disequilibrium structure for genome-wide significant gene variants located in 16p12.3. Darker color indicates higher r2 values. In black columns, two SNPs are fully correlated. The green rs-number flags the strongest association signal for the self-reported cigarettes per day.

The signal is located in an intergenic region, 28 kb from *CLEC19A* (*C-type lectin domain family 19, member A*)*;* the function of which still lacks a more detailed description and has low levels of expression throughout the various tissues (gtexportal.org). It is possible that *CLEC19A* SNPs tag causal variants or regulatory motifs within this region, including neighboring genes *SYT17* (*synaptotagmin XVII*) and *TMC5* (*transmembrane channel-like 5*). In fact, no eQTLs or meQTLs were seen for *CLEC19A,* since the gene was insufficiently expressed in the public databases. However, a majority of the 23 genome-wide significant SNPs were identified as eQTLs for genes *SYT17* and *TMC5* in brain tissue databases GTEx and BRAINEAC. Figure 11 summarizes the main results for the eQTL and meQTL search. It is no wonder that eQTLs are seen, since the association region is located on a DNAse I hypersensitivity site. These sites span approximately 42% of the human genome (Maurano et al., 2012). In these specific regions of the genome, chromatin has lost its condensed structure, exposing the DNA and making it accessible for DNA degradation enzymes, but also for increased transcriptional activity.

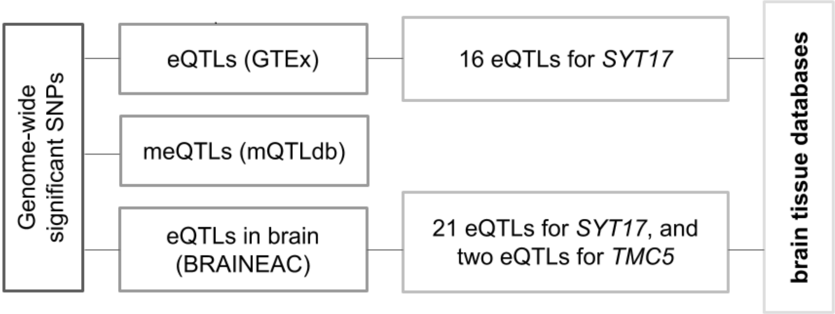

Genome-wide significant SNPs
eQTLs (GTEx)
16 eQTLs for SYT17
meQTLs (mQTLdb)
eQTLs in brain (BRAINEAC)
21 eQTLs for SYT17, and two eQTLs for TMC5
brain tissue databases

Figure 11. A summary of the main findings from the expression quantitative trait loci (eQTL) and methylation quantitative trait loci (meQTL) analyses.

The 16p12 locus has previously shown linkage with ADHD (Romanos et al, 2008), which is one of the most common neurodevelopmental disorders in childhood. ADHD has been associated with smoking in adulthood (Brook et al, 2008). Moreover, the locus 16p12.3 is not associated with cotinine (Ware et al., 2016), the primary metabolite of nicotine, suggesting that the signal is most likely driven by co-morbidity factors rather than genetic variants directly influencing inter-individual differences in smoking amount. Hence, the suggested co-morbidity effect cannot be ruled out.

The 16p12.3 region has yielded evidence of linkage with MaxCigs24 (Han et al., 2010), whereas large-scale GWAS meta-analyses have not highlighted this locus (Saccone et al, 2010; Liu et al, 2010; Thorgeirsson et al, 2010; Tobacco and Genetics Consortium, 2010). The associating variants on the 16p12.3 locus are low-frequency variants (MAF 0.04-0.06) that produce remarkably large effects (Table 9), corresponding to an increase in smoking quantity of approximately five cigarettes per day, with respect to each minor allele at the locus. When these effect sizes are compared with the ones produced by the most well-established smoking quantity locus at 15q25.1 (tagged by rs16969968), which corresponds to an increase of only one CPD, the effects captured by 16p12.3 locus variants are large. However, the MAFs between the variants on these two loci vary significantly, the MAF for rs16969968 is 0.32 (in the large Finnish population sample utilized in Study III), while the genetic variants on the 16p12.3 association region are low-frequency variants. Hence, the impact of the polymorphisms in the 16p12.3 region is less notable at the population level. However, in the sample used in Study II, the minor allele of variant rs4300632 was well-represented, with altogether 157 copies of the minor allele being found among the 1,714 smokers included in the study. A boxplot in Figure 12 shows the large effect of variant rs4300632 on CPD.

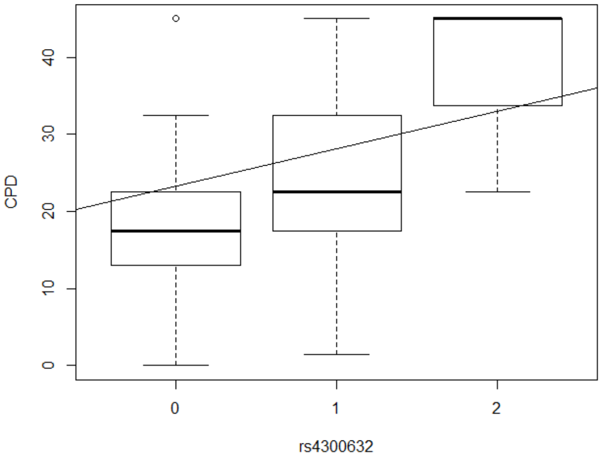


Figure 12. A boxplot of the rs4300632 (on 16p12.3) genotype distribution among smokers. The number of individuals carrying 0, 1 and 2 copies of the minor allele were 1,560, 151 and 3, respectively.

The data in Study II was imputed using the 1000 Genomes Project reference panel Phase 1 (1000 Genomes Project Consortium, 2012) together with a low-pass whole genome sequence-based Finnish panel from the Sequencing Initiative Suomi project, whereas the data used in the previous work (Loukola et al., 2014) was imputed using the HapMap2 reference panel. In order to determine whether the gain of strength in the association signal is due to a mere increase in the sample size or a result of a higher imputation quality, the association region of 16p12.3 was scrutinized in parallel using three data sets from the same sample with different imputation reference panels: HapMap2, 1000 Genomes Phase 1, and 1000 Genomes Phase 1 + SISu. The comparison showed an increase in the number of SNPs included in the association region (HapMap2: 526 SNPs, 1000 Genomes Phase 1: 1,282 SNPs, and 1000 Genomes + SISu: 1,304 SNPs). Despite differences in numbers of SNPs, and their imputation quality scores, the association signals in the different data sets map to exactly the same region. This result suggests that the increase in the strength of the association is due to merely an increase in the power, enabled by the larger study sample.

Neurotrophin signaling pathway

Although the function of *CLEC19A* remains undetermined, an intriguing and established transcription factor binding site (TFBS) is located at the position chr16:19,328,414 – 19,328,427, which stands within the association locus (chr16:19,326,020 – 19,357,771). This TFBS serves as a binding site for nuclear factor κ B (NF-κB) transcription factors (genome.ucsc.edu). NF-κBs are a conserved family of pleiotropic transcription factors operating, for example, in the immune system through the regulation of immune and inflammatory responses triggered via Toll-like receptors (Medzhitov

2001). Additionally, they are involved in complex pathways that regulate synaptic plasticity (Mattson and Meffert, 2006). One of these pathways, the neurotrophin signaling pathway, include neurotrophins (Mattson and Meffert, 2006), which have important roles in neural development (Bibel and Barde, 2000). Thus, it is a promising addition to the scarce list of candidates affecting the inter-individual differences in smoking behavior.

In Study II, a GWAS was performed in order to confirm and discover novel target genes and factors predisposing smokers to sustain the smoking behavior. This aim was achieved by strengthening a previously reported finding of association near *CLEC19A* with smoking quantity, and further linking the finding with the neurotrophin signaling pathway through the TFBS for NF-κB transcription factors. This pathway is not a novel discovery in smoking-related research. In fact, previous studies have reported association between the neurotrophin signaling pathway and smoking-related traits: smoking initiation and progression (Lang et al, 2007; Tobacco and Genetics Consortium, 2010; Wang and Li, 2010), and smoking cessation (Wang and Li, 2010). However, these findings targeted and highlighted only a couple of members of the pathway*, BDNF* and *NTRK2*. In Study II, no association was detected with either of these genes. Thus, further studies are required in order to verify the result and the plausible link between smoking behavior and the neurotrophin signaling pathway.

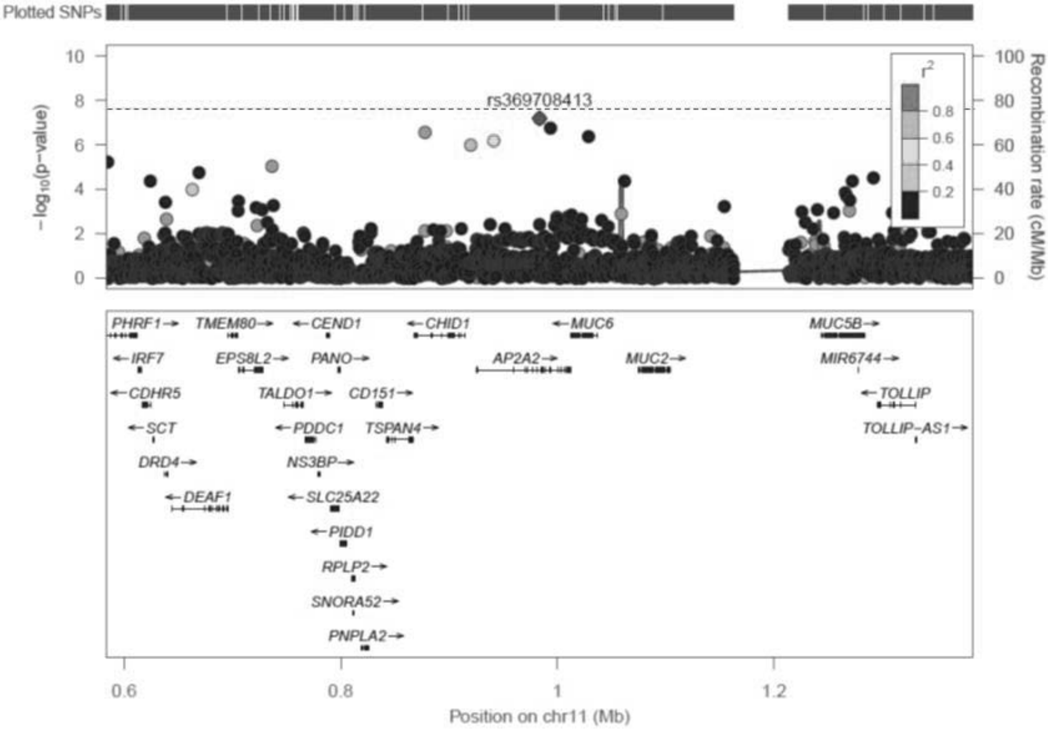


Figure 13. A regional plot of the genome-wide association study for DSM-IV NW symptom count showing association on 11p15.5.

In addition, an intriguing association signal supporting the role of the neurotrophin signaling pathway was observed with the DSM-IV NW symptom count, with locus 11p15.5 showing association approaching genome-wide significance (P=6.58x10^{-8}) for rs369708413. This locus harbors several genes, which are illustrated in Figure 13. The signal peaks at *AP2A2* (*adaptor-related protein complex 2, alpha 2 subunit*) and *MUC6* (*mucin 6*). Adaptor-protein 2 (AP-2) is an important mediator in endocytic vesicle formation, together with clathrin (Smythe, 2002), and is involved in the neurotrophin signaling pathway (Beattie et al., 2000). The other gene, *MUC6*, is up-regulated by NFκB1 (Sakai et al., 2005). In addition, the top *MUC6* variant, rs201137338 (P=4.2x10^{-7}), is a missense variant (H524Q), which changes histidine to glutamine, consequently changing the electrically-charged sidechain into a polar uncharged sidechain. Thus, this polymorphism is a functional variant, with a potential capability of altering the end product through an impaired folding of the polypeptide. The association results for DSM-IV NW highlighting locus 11p15.5 are presented in Table 11.

Table 11. Association results for DSM-IV NW on locus 11p15.5.

Chr	Position[a]	SNP	Allele 1	Allele 2	*p*-value	β	SE	MAF
11	920049	rs561547189	A	G	9.9E-07	1.34	0.27	0.018
11	941391	rs529797424	C	T	6.3E-07	1.37	0.27	0.018
11	983584	rs369708413	C	T	6.6E-08	1.61	0.30	0.017
11	993507	rs560149619	A	G	1.7E-07	1.50	0.28	0.017
11	1028665	rs201137338	G	C	4.2E-07	1.40	0.27	0.018

[a] Base pair position according to human reference sequence Build37/hg19 assembly

Abbreviations: Chr, chromosome; SNP, single nucleotide polymorphism; Allele 1, effect allele (minor allele); Allele 2, alternative allele (major allele); β, effect size provided by beta-value; SE, standard error; MAF, minor allele frequency.

Another interesting gene, *DRD4* (*dopamine receptor D4*), is located approximately 343 kilobases upstream from the association signal. The DA system has been under close investigation in smoking-related research (Moran et al, 2013; George and O'Malley 2004; Picciotto 2003), making *DRD4* an important candidate gene in smoking behavior. It encodes the D4 subtype of the DA receptor, and thus plays an important role in processes in which the highly rewarding effect of DA is pivotal, such as learning and memory (Wise, 2004). Both are essential in the development of smoking behavior and ND.

Previous linkage studies have highlighted the connection between locus 11p15 and smoking-related traits. Pergadia and colleagues (2009) yielded genome-wide significant linkage of the locus with DSM-IV NW using a subset of 505 individuals from the Study II sample. Also, linkage has been detected with DSM-IV ND, FTND, and a co-morbid phenotype of ND and alcohol use (Loukola et al, 2008). Additionally, the same locus has shown genome-wide significant association with chronic bronchitis, which is one of the classic phenotypes of COPD (Lee et al, 2014). These previous reports, combined with the results observed in Study II, clearly suggest that locus 11p15 harbors an element or possibly several elements that influence the occurrence of smoking-related traits. *DRD4* plays an undisputable role in smoking behavior, at least in theory, though solid evidence of association is still lacking using the GWAS approach. Study II, like so many other GWASs, detected no association with variants located in the close vicinity of *DRD4*, but instead highlighted other genes (*AP2A2* and *MUC6*) within the region, which have links to the neurotrophin signaling pathway.

The variants in the 11p15.5 region are more frequent in Finland than other European populations, which is likely due to the Finnish population history of founding bottlenecks. Thus, the genetic background of the Finnish population offers a clear advantage for studying low-frequency and rare variants (Surakka et al. 2016). It is worth emphasizing that this study highlights genetic variants (in close proximity to *CLEC19A*, and within *AP2A2* and *MUC6*) that have low MAFs. The associating variants and their allele frequencies in the study sample and in other populations are compared in Table 12.

Table 12. Genetic variants highlighting the neurotrophin signaling pathway - a global comparison of minor allele frequencies.

					MAF			
SNP	**Position**[a]	**Chr**	**Allele 1**	**Allele 2**	**Sample**	**CEU**[b]	**YRI**[b]	**JPT**[b]
rs369708413	983584	11	C	T	0.017	0.005	0.009	0
rs560149619	993507	11	A	G	0.017	0	0	0
rs201137338	1028665	11	G	C	0.018	0	0	0
rs4300632	19350508	16	T	A	0.046	0.04	0	0.337
rs11074386	19351749	16	T	C	0.046	0.04	0	0.337
rs11074388	19355577	16	A	G	0.055	0.04	0.352	0.438

[a] Base pair position according to human reference sequence Build37/hg19 assembly; [b] Allele frequencies have been obtained using publicly available database Ensembl (ensembl.org)

Abbreviations: SNP, single nucleotide polymorphism; Chr, chromosome; Allele 1, effect allele (minor allele); Allele 2, alternative allele (major allele); MAF, minor allele frequency; CEU, Utah residents with Northern and Western European ancestry; YRI, Yoruba in Ibadan, Nigeria; JPT, Japanese in Tokyo, Japan.

The results in Study II highlight factors which are involved in the neurotrophin signaling pathway. This suggests that the pathway might regulate neuronal modifications that have a plausible impact on the sustained use of tobacco products.

No genome-wide significant association for DSM-IV ND

The GWAS performed in Study II included DSM-IV ND and aimed to discover novel genetic variants predisposing individuals to ND, and also to confirm previous suggestive findings of association reported in an earlier GWAS (Loukola et al., 2014). Despite the well-known reputation of DSM-IV to provide efficient tools for classifying and diagnosing currently recognized mental disorders, including substance-related disorders, DSM-IV ND is remarkably underutilized in research studies. The NAG-FIN sample used in Studies I and II remains one of the few to include DSM-IV-classified disorders and traits (ND, NW, and alcohol dependence). The inclusion of distinct measures and tools for capturing smoking as a behavior has assisted in creating a sample consisting of unique and highly detailed phenotype profiles.

Table 13. Top 10 GWAS results for DSM-IV ND.

Chr	Position[a]	SNP	Allele 1	Allele 2	*p*-value	β	SE	MAF
5	5786702	rs187632	A	G	7.2E-07	-0.09	0.02	0.462
11	21472503	rs148586747	A	G	7.4E-07	-0.27	0.05	0.027
7	19566286	rs6461441	C	A	8.5E-07	-0.12	0.02	0.156
7	19554431	rs2192487	G	A	8.7E-07	-0.12	0.02	0.152
7	19563837	rs2192483	G	A	1.1E-06	-0.12	0.02	0.151
7	19573641	rs17141561	C	T	1.3E-06	-0.12	0.02	0.155
14	99275537	rs188595723	G	A	1.9E-06	0.31	0.06	0.020
10	99790398	rs61873668	G	C	2.4E-06	-0.11	0.02	0.167
7	19569253	rs7796093	C	A	2.6E-06	-0.12	0.02	0.147
4	15542197	rs16892135	G	A	2.9E-06	0.10	0.02	0.218

[a] Base pair position according to human reference sequence Build37/hg19 assembly

Abbreviations: Chr, chromosome; SNP, single nucleotide polymorphism; Allele 1, effect allele (minor allele); Allele 2, alternative allele (major allele); β, effect size provided by beta-value; SE, standard error; MAF, minor allele frequency.

The study did not yield genome-wide significant association with DSM-IV ND. The results of DSM-IV ND diagnosis are presented in Table 13. In the previous study (using a subset of 1,114 individuals from the current sample), DSM-IV ND diagnosis showed a suggestive association (min P=1.68x10^{-6}) on 2q33, harboring a neuregulin receptor *ERBB4* (Loukola et al, 2014). Furthermore, evidence of linkage has previously been reported on 2q33 with a regular smoker phenotype, using a subset of 505 individuals from the current sample (Loukola et al, 2008). Although the sample size in the current study was increased compared to the previous GWAS, no statistically significant association was detected for *ERBB4* (P=1.0x10^{-4} for ND diagnosis; P=9.5x10^{-5} for ND symptom count). There are a few plausible explanations for this. First, a false positive finding in the previous GWAS cannot be ruled out. However, Turner and colleagues (2014) highlighted the connection between ERBB4/NRG3 signaling and ND and NW using both behavioral mouse models and association analyses in a clinical trial sample, providing further evidence for the involvement of *ERBB4* in smoking behavior. Second, although sample size was increased, the current sample was diluted with respect to the prevalence of ND. In the previous GWAS, 592 out of the 1,114 subjects (54%) fulfilled the DSM-IV ND criteria and scored, on average, 3.0 on the symptom scale, whereas among the 601 subjects added to the current study, 281 (47%) fulfilled the criteria and scored, on average, 2.6 on the symptom scale (t-test p-value < 0.0001). Similarly, smoking quantity differs between the previous sample (average CPD of 19.7) and the additional subjects (average CPD of 17.4). Third, the previous GWAS analyses were done with the Plink QFAM algorithm, which differs from the linear mixed model used in the current study.

5.1.4 The strongest genetic evidence for smoking behavior (Study III)

While the broad picture of smoking behavior and ND is known, the basis for underlying molecular genetics remains in its early steps. Only one locus, 15q25.1, has proved solid evidence for association with smoking-related traits and diseases in several independent studies, large-scale GWASs, GWAS meta-analyses, and replication studies (Saccone et al., 2010; Thorgeirsson et al., 2010; Liu et al., 2010; Tobacco and Genetics Consortium (TAG) 2010; Amos et al., 2008; Hung et al., 2008). Chromosome 15q25.1 harbors the *CHRNA5-CHRNA3-CHRNB4* gene cluster, encoding the nAChR subunits α5, α3, and β4. Study III aimed to replicate the reported finding between three distinct loci at 15q25.1 and smoking quantity (Saccone et al., 2010) in a large Finnish population-based sample. In addition, Study III provided precise estimates of the strength of the association in the Finnish population.

We utilized the National FINRISK study and the Health 2000 Survey samples, and grouped smokers into heavy smoking "cases" (CPD>20) and light smoking "controls" (CPD≤10), thus generating a dichotomous CPD variable. The associations between dichotomous CPD and locus 1 (rs16969968, prevalence odds ratio (OR)=1.39, P=1.43 x 10^{-9}) and locus 2 (rs578776, OR=0.80, P=0.00005) were successfully replicated, but the analyses for the dichotomous trait did not yield statistical significance for locus 3 (rs588765, OR=0.91, P=0.091). However, the association between smoking quantity and locus 3 was detected with quantitative CPD (β=0.023, P=0.036), but only after adjusting for rs16969968 (locus 1). This result is in line with a previous finding reported by Saccone and colleagues (2010). The results for both dichotomous and quantitative CPD are presented in Table 14. These results likely reflect the differences in power; 8,356 current smokers were included in the quantitative CPD variable, while the dichotomous CPD variable consisted of 3,624 controls and 1,076 cases.

Table 14. Single-SNP association results of dichotomous and quantitative CPD analyses.

	Dichotomous CPD (*N*=4,700) Logistic regression			Quantitative CPD (*N*=8,356) Negative binomial regression		
Single-SNP analysis	**OR**	**95% CI**	***p*-value**	**β**	**95% CI**	***p*-value**
rs16969968	1.39	1.25,1.55	1.43 x 10^{-9}	0.067	0.048,0.086	1.11 x 10^{-12}
rs578776	0.80	0.72,0.90	0.00005	-0.042	-0.061,-0.023	6.50 x 10^{-6}
rs588765	0.91	0.81,1.01	0.091	-0.020	-0.038,-0.002	0.028
Joint model analysis	**OR**	**95% CI**	***p*-value**	**β**	**95% CI**	***p*-value**
rs16969968	1.34	1.18,1.52	2.56 x 10^{-6}	0.062	0.040,0.083	7.14 x 10^{-9}
rs578776	0.93	0.81,1.06	0.262	-0.012	-0.034,0.009	0.266
rs16969968	1.49	1.30,1.70	1.43 x 10^{-9}	0.081	0.058,0.104	8.95 x 10^{-13}
rs588765	1.13	0.99,1.29	0.066	0.023	0.001,0.045	0.036

CPD, cigarettes per day; OR, Odds Ratio; CI, confidence interval; β, effect size provided by beta-value.

Interestingly, three other meta-analyses conducted by large consortia have been unsuccessful in detecting confirmatory evidence of association for rs588765 (locus 3) and smoking quantity (Thorgeirsson et al., 2010; Liu et al., 2010; TAG 2010). Methodological differences could provide a plausible explanation for this discrepancy. Saccone and colleagues (2010) used logistic regression in their analyses to test for association with dichotomous CPD, while the three other meta-analyses used linear regression and analyzed either quantitative or categorical CPD (Thorgeirsson et al., 2010; Liu et al., 2010; TAG 2010). In Study III, the associations between quantitative CPD and the three loci, including locus 3 (rs588765), were detected using negative binomial regression.

Table 15. Estimated number of cigarettes smoked per day, stratified by genotypes of three single nucleotide polymorphisms.

Locus 15q25.1	***N* of minor alleles per genotype**					
	0		**1**		**2**	
SNP	mean CPD	95% CI	mean CPD	95% CI	mean CPD	95% CI
rs16969968	13.9	13.6, 14.2	14.8	14.5, 15.1	16.0	15.4, 16.6
rs578776	14.9	14.7, 15.2	14.2	13.9, 14.5	13.9	13.9, 14.5
rs588765	14.8	14.5, 15.1	14.3	14.0, 14.6	14.4	13.9, 14.9

CPD, cigarettes per day; CI, confidence interval.

The most well-established locus within the 15q25.1 region is tagged by the functional SNP rs16969968. The variant has been suggested to contribute to increased nicotine consumption by impairing the capacity of (α4β2)α5 nAChRs to produce a normal inhibitory motivational signal intended to limit nicotine intake (Fowler et al., 2011). In Study III, precise estimates of the strength of association were provided for each locus tagged by SNPs rs16969968, rs578776 and rs588765. Notably, the effect size of rs16969968 was approximately one CPD for each minor allele (Table 15), in agreement with the original GWAS report by Thorgeirsson and colleagues (2008).

Smoking and alcohol use frequently co-occur, resulting in high correlation between the two traits (Dani and Harris, 2005). Hence, in order to determine the specificity of the three loci, the CPD analyses were adjusted for log-transformed alcohol consumption in an additional model. The results showed remarkable similarity with and without the adjustment for alcohol use. The lack of difference, specifically in locus 1, suggests that the association signal from rs16969968 is specific to smoking. A previous study supports this finding (Wang et al., 2009), since no evidence of association was observed between alcohol dependence and rs16969968, while the effect of the SNP's specificity to smoking was confirmed.

On chromosome 15q25.1, two distinct loci have been consistently highlighted in smoking-related research: locus 1, tagged by rs16969968 or rs1051730 (a highly correlating SNP located on *CHRNA3*), and locus 2, tagged by rs578776. Study III, along with several other association studies, suggests that locus 1 is a risk factor for smoking behavior with the minor allele predisposing an individual to smoking behavior. Although the effect of locus 1 is undeniably robust throughout different populations, the effect of the minor allele corresponds to merely a small increment in the daily smoking quantity. Locus 2, on the other hand, seems to have a protective effect on smoking, with an OR of 0.8 being reported in Study III. This result is consistent with previous reports (Saccone et al., 2010). Lastly, the role of an additional distinct locus 3 (tagged by rs588765) in smoking behavior remains unconfirmed, since an association signal was detected with quantitative CPD only after adjusting for rs16969968, but the dichotomous trait failed to achieve statistical significance.

5.2. Pleiotropic effects of nAChR subunits

5.2.1 Association of the α5-α3-β4 subunit cluster with alcohol use

The *CHRNA5-CHRNA3-CHRNB4* gene cluster on chromosome 15q25.1 has been thoroughly scrutinized in research with smoking-related traits and diseases. Altough smoking often co-occurs with alcohol use, and growing evidence from animal and human studies suggest that different drugs of abuse utilize a common circuitry in the brain's limbic system (Nestler, 2001; Di Chiara et al., 2004; Pierce and Kumaresan, 2006; Ross and Peselow, 2009; White and White, 2016), only a few studies have explored the 15q25.1 locus with alcohol-related traits. Clearly, further studies are required to define the role of genetic variants in the nAChR genes contributing to inter-individual differences in alcohol use. In Study III, the goal was to examine the role of genetic variants on this robust smoking behavior locus in alcohol use, and to investigate whether such a role is independent of smoking.

In Study III, association was tested between three SNPs, rs16969968, rs578776, and rs588765, and alcohol use. The trait was analyzed as quantitative, categorized and dichotomized alcohol use. Out of all the distinct association analyses, only dichotomous alcohol use, defined as 'abstainers and low-frequency drinkers' (*N*=12,246) vs. 'drinkers' (*N*=19,566) yielded a statistically significant association with rs588765 (locus 3, OR=1.15, P=0.00007). Table 16 shows the logistic regression association results for the binary alcohol use variable. Individuals with a homozygous genotype for the minor allele were more likely to be drinkers. No evidence of association was detected with the quantitative or categorical alcohol use traits.

Table 16. Results for logistic regression analyses of binary alcohol use (abstainers and low-frequency drinkers vs. drinkers).

SNP	N of minor alleles	OR	95% CI	*p*-value
rs16969968	0	REF		
	1	0.93	0.89,0.98	0.007
	2	0.97	0.89,1.05	0.430
rs578776	0	REF		
	1	0.98	0.94,1.03	0.532
	2	0.98	0.90,1.06	0.569
rs588765	0	REF		
	1	1.00	0.95,1.05	0.945
	2	1.15	1.07,1.24	0.00007

SNP, single nucleotide polymorphism; OR, odds ratio; CI, confidence interval; REF, reference.

Evidence of association has been reported between locus 1 (rs16969968) and symptoms of alcohol abuse or dependence (Chen et al., 2009). Also, association signals have been detected from locus 2 (rs578776) and locus 3 (rs615470) with DSM-IV alcohol dependence in an African-American sample, while in the same study, samples of European decent provided no evidence of association (Sherva et al., 2010). In another study, Wang and colleagues (2009), reported association between

DSM-IV alcohol dependence and locus 3 (rs588765), whereas their study, in unison with Study III, detected no association with either locus 1 rs1696968 or locus 2 rs578776. On top of the works focusing on alcohol dependence, association has been reported between alcohol use initiation in young Americans of European descent and two SNPs (rs1948 and rs11634351) located on *CHRNB4* (Schlaepfer et al., 2008). The accumulating amount of contradicting evidence in study results may partly reflect population heterogeneity and different LD structures in different study samples.

It is worth noting that abstainers and low-frequency drinkers form a truly heterogeneous group of subjects. The group potentially includes some participants who are former alcoholics, but are abstaining at the time of the data collection. These former alcoholics bring a possible confounder aspect to the study, which was addressed by repeating the analysis among never smokers. Due to the established co-occurrence between alcoholism and heavy smoking, it is unlikely that many former alcoholics are included in the analysis restricted to never smokers. In addition, this approach excludes the possibility of confounding in the association signal emerging from smoking. Intriguingly, limiting the analysis to never smokers, thus narrowing the total sample size to 16,786 individuals, had no notable impact on the effect of locus 3 (OR=1.18, P=0.001) (Table 17).

Table 17. Results for logistic regression analyses of binary alcohol use among never smokers (abstainers and low-frequency drinkers vs. drinkers).

SNP	N of minor alleles	OR	95% CI	*p*-value
rs16969968	0	REF		
	1	0.96	0.90,1.03	0.230
	2	0.95	0.85,1.06	0.368
rs578776	0	REF		
	1	0.96	0.90,1.03	0.271
	2	0.93	0.84,1.04	0.217
rs588765	0	REF		
	1	1.04	0.97,1.12	0.231
	2	1.18	0.97,1.12	0.001

SNP, single nucleotide polymorphism; OR, odds ratio; CI, confidence interval; REF, reference genotype.

This strongly suggests that the association signal detected with variant rs588765 is specifically produced by regular alcohol use. Although the effect of locus 3 is somewhat modest, this result confidently provides new direction for the entire body of research of locus 3 in the *CHRNA5-CHRNA3-CHRNB4* gene cluster suggesting that the effects of alcohol use may, at least to some extent, be mediated through nAChRs and their downstream transmitters. In fact, the most studied transmitter molecule system is the mesolimbic DA pathway, which consists of dopaminergic neurons and their targets (Nestler, 2005). This pathway links several brain regions, and some of them are involved in the brain's natural memory system. Notably, powerful emotional memories have further been hypothesized as an important aspect resulting in addictive behaviors (Wise, 2004; Everitt et al., 2003). The emotional effects of natural rewards, such as food and social interactions, may be mediated by the DA system (Kelley and Berridge, 2002; Tobler et al., 2005). nAChRs are closely linked to this system, and thus they may mediate responses to other substances and natural rewards. Emerging evidence from rodent studies indicate that ethanol interacts with nAChRs in the dopaminergic reward circuitry to affect brain reward systems (Nestler, 2005; Hendrickson et al., 2013).

These results from Study III provide highly welcomed fragments of evidence explaining the molecular genetic basis underlying the shared genetic predisposition between smoking and alcohol use, which has largely been unknown. Thus, these results suggest that nAChR genes contribute, not only to inter-individual differences in smoking, but in alcohol use as well.

5.2.2 Causality between smoking and BMI (Study IV)

Cigarette smoking is a modifiable environmental factor among many others that has a strong lifestyle influence on BMI. Consistent results from epidemiological studies show that smokers have a lower BMI than non-smokers (Sneve and Jorde, 2008; Munafo et al., 2009). Questions of whether this association is causal (i.e., exposure to smoking causes lower BMI) can be addressed and investigated using the Mendelian Randomization approach. Indeed, stratification by smoking status has been reported to modify the association between rs1051730 (a genetic variant on 15q25 in complete LD with rs16969968) and BMI (Freathy et al., 2011). This result strengthens the evidence of causality between smoking and lower BMI. Study IV aimed to extend these findings by increasing the sample size from 24,198 (Freathy et al., 2011) to 148,730 (Taylor et al., 2014 (Study IV)), thus ensuring increased power for the analyses. Hence, a large-scale international meta-analysis was conducted by the Causal Analysis Research in Tobacco and Alcohol (CARTA) consortium (http://www.bris.ac.uk/expsych/research/brain/targ/research/collaborations/carta/), and the National FINSIRK study sample (*N*=20,368), including 5,120 current smokers, 5,493 formers smokers, and 9,755 never smokers, was included in the analyses.

Table 18. Association results of rs16969968[a] with body mass index, stratified by smoking status: a Mendelian randomization analysis.

Smoking status	***N***	**Effect (% change in BMI)**	**95% CI**	***p*-value**
Current smokers	38,912	-0.74	-0.97, -0.51	$2.00\text{x}10^{-10}$
Former smokers	43,009	-0.14	-0.34, 0.07	0.19
Never smokers	66,809	0.35	0.18, 0.52	$6.38\text{x}10^{-5}$
Total	148,730	-0.07	-0.19, 0.04	0.22

[a] or rs1051730, which is in full LD with rs16969968; BMI; body mass index; CI, confidence interval.

The total sample size of the Mendelian Randomization analysis comprised altogether 148,730 never, former and current smokers. Among *never smokers* (*N*=66,809), a positive association was observed between the genetic variant (rs16969968 or rs1051730 depending on the cohort) and BMI, and the results are presented in Table 18. The result indicates a 0.35% increase in BMI per minor allele (95% CI 0.18, 0.52; P=$6.38\text{x}10^{-5}$). In addition, this study confirmed the expected inverse association between the genetic variant and BMI in *current smokers* (*N*=38,912) (Table 18), indicating a 0.74% decrease in BMI per minor allele (95% CI -0.97, -0.51; P=2.00×10^{-10}). The result is consistent with a causal, anorexic effect of smoking on BMI. *Former smokers* (*N*=43,009) provided no evidence of association (percentage change -0.14, 95% CI -0.34, 0.07; P=0.19). Furthermore, an interaction of the estimates was tested, and they differed significantly from each other (P=4.95x10^{-13}). In addition, the effects observed between smoking status and BMI were not modified by sex. Heterogeneity between the 29 cohorts included in the meta-analysis was low. It is critical to note that when the data was examined without stratification by smoking status no clear association was detected between rs16969968 and BMI (P=0.22), suggesting that a conventional GWAS using BMI as the phenotype of interest would have failed in detecting this signal. However, smoking status was not adjusted for in this additional test.

The observed 0.35% increase per minor allele in BMI among never smokers represents a change of approximately 0.09 kg/m^2. This change is rather small when compared to the effect of *FTO* (*fat mass and obesity associated*) gene variants (~0.4 kg/m^2) (Frayling et al., 2007). However, the proportion of variance is comparable to the other variants identified for BMI (Speliotes et al., 2010).

In Study IV, evidence of association for smoking cessation was observed (current vs. former smokers: OR per minor allele 1.08, 95% CI 1.06, 1.10; P=1.44x10^{-12}), whereas, the data yielded no association for smoking initiation (ever vs. never smokers: OR per minor allele 1.01, 95% CI 0.99, 1.03; P=0.50). These results are consistent with previous studies which have reported that the rs16969968/rs1051730 variant does not influence the ultimate risk of being a smoker (TAG, 2010), but it has been associated with short-term smoking cessation in treatment-seeking smokers (Munafo et al., 2011), suggesting a pleiotropic effect for the variant.

These results indicate that rs16969968 may predispose never smokers to higher BMI in a manner similar to how it increases the risk for smoking heaviness among current smokers. At present, the exact mechanism through which the variant may induce a positive effect on BMI among never smokers remains uncertain, and can merely be speculated. In addition, the 15q25 region has provided evidence of pleiotropy (i.e., also being associated with other diseases or traits), since GWASs have established the association between the *CHRNA5-CHRNA3-CHRNB4* gene cluster and smoking heaviness, as well as downstream health consequences, such as lung cancer and peripheral artery disease (TAG, 2010; Li et al., 2010, Thorgeirsson et al., 2010; Amos et al., 2008). On top of these results, candidate gene studies have suggested an association with cocaine use (Grucza et al., 2008) and alcohol use (Study III, see 5.2.1), as well as alcohol dependence (Sherva et al., 2010). Therefore, it can be hypothesized that nAChRs play a role in mechanisms that mediate responses to rewarding stimuli in general. Another possibility is that nAChR subunits play a direct role in mediating food intake, since evidence from animal models indicates that activation of hypothalamic α3β4 nAChRs initiates the activation of pro-opiomelanocortin neurons and triggers further, subsequent activation of melanocortin 4 receptors, which have proved to be critical for nicotine-induced decreases in food intake (Mineur et al., 2011). However, the exact mechanism for mediating food intake through nAChRs remains yet unspecified. It is worth noting that the effects observed in Study IV among current and never smokers may operate via other genes (namely *CHRNA3* and *CHRNB4*) and their variants in LD with rs16969968. Clearly, further work, with possibly more detailed body composition measures, such as percent body fat and its distribution, is required to explore the causality and mechanism.

The association observed in never smokers was detected merely due to stratification by smoking status. Hence, if confirmed, these results will provide an important implication for future GWAS and their design, since additional variants may be identified using the GWAS approach when the known environmental exposures with pronounced effects on the phenotype of interest are stratified.

To conclude, the pleiotropic effect of rs16969968 (or other variants in full LD) influences BMI. The results of Study IV demonstrate how the stratification of well-characterized environmental factors with a known impact on health outcomes may reveal novel genetic association with health outcomes. Accordingly, these associations may operate through genetic influences on the environmental factors themselves, or via other pathways which are veiled by the environmental factors.

5.3 Strengths and Limitations

The sample utilized in Studies I and II is a unique combination of highly detailed profiles describing broad substance use behaviors in a family setting. At the same time, the sample is limited by the low number of participating parents, leading to incomplete family structures, and thus, decreased power in family-based analysis. When the data was collected, the mean age of the twins was 57 years, which is why the majority of parents were unavailable. In addition, biochemical verification of the smoking status was not assessed with cotinine or exhaled carbon monoxide measurements. However, in our analyses only self-reported smokers were included; it is unlikely that many non-smokers would have claimed to be smokers in the extensive interview.

While comparing Study II GWAS results with a previous GWAS using a subsample of the NAG-FIN sample (Loukola et al., 2014), a clear difference in the two sample sets was observed, but the direction of the difference was inconsistent between study phenotypes. Increasing the sample size with mainly siblings strengthened the CPD association findings, while the previously detected effects observed with DSM-IV ND diminished, which may have resulted in diluted variable variance in the sample. This is in line with the study design of the NAG-FIN sample: twin pairs concordant for heavy smoking were initially identified and used in our previous GWAS; in the current study, all available family members were also included. Family members included heavy smokers and light smokers, as well as non-smokers (which were excluded from the GWAS). It is plausible that some of the less exposed subjects may not have developed ND, but would do so given additional exposure; yet, since they are first degree relatives of the twins used in the original GWAS, there is a 50% probability that they carry the same risk alleles. Having risk allele carriers who have not yet developed ND reduces the power to detect the association.

In Study I, both single-point and multipoint linkage analyses were performed. Unlike single-point analyses, multipoint analysis uses haplotype information from several markers to infer the IBD relationships (Halpern and Whittemore, 1999), and thus should increase power. However, cautious interpretation of the results is recommended since multipoint analysis has been suggested to be sensitive to power loss due to misspecification of intermarker order and distances, especially with closely spaced markers (Halpern and Whittemore, 1999).

In Study III, the likely absence or underrepresentation of severely alcohol and nicotine addicted individuals needs to be acknowledged. Although the sample is adequately sized and genetically homogenous, these highly addicted individuals are unlikely to participate in health-related surveys in the first place. However, their absence is unlikely to explain the novel association results with alcohol use.

The outcome of smoking quantity in Study III differs from the one utilized in Studies I and II. This is due to phenotyping techniques in different samples. The NAG-FIN sample has been enriched for ND, and thus the criteria for qualifying as an ever smoker has been set high, 100+ cigarettes in the individual's lifetime. However, the Finnish population-based samples used in Studies III and IV are large-scale health surveys, without any detailed emphasis on smoking. Thus, any participant reporting information on smoking quantity were qualified for the analyses. Hence, the sample size was ensured to be as large as possible to gain power. Therefore, the difference in measuring CPD must be acknowldeged when interpreting the results.

The study design in Study IV, using Mendelian Randomization to detect causality between smoking and BMI, did not attempt to stratify smoking status according to smoking quantity or cotinine values, which could have possibly been a more informative and reliable measure of smoking status. Instead, the stratification relied solely on self-report.

6 CONCLUSIONS AND FUTURE PROSPECTS

This thesis aimed to identify specific genes that would predispose individuals to smoking behavior and increased vulnerability to ND. Specifically, this study assessed genetic variants and their ability to explain the differences in smoking behavior between individuals. Considering the high heritability estimates of ND, the rationale to expect a genetic contribution was clear.

Study I provided further evidence for chromosome 20 harboring genetic elements that do affect ND. In addition, Study I highlighted a significant sex-difference in the genomic regions linked to ND. The exact gene, be it a single genetic variant or several, however, remaines unidentified after this study, and requires additional research focusing, for example, on whole-genome sequence analyses. Also, whether the genes residing at the linkage loci have an impact on sex differences in ND remains unsolved. Follow-up studies with an epigenetic approach would assist in describing the detailed epigenetic profiles that may be involved in causing sex differences in ND, since the possibility of gene-environment interaction cannot be ruled out. In addition, ND variables in future studies would benefit from inter-item correlation tests in search of a common factor for dependence. Modeling the covariance across symptoms would ensure greater power for genetic analysis than the mere arbitrary summing of items allows.

Low-frequency variants were predominantly highlighted in the GWAS (Study II) with smoking quantity and NW phenotypes. Owing to the Finnish population history and subsequent enrichment of rare genetic variation in the Finnish gene pool, many population-specific rare and low-frequency variants predisposing individuals to complex diseases and traits are likely to be found in Finns.

Study III confirmed that a locus, 15q24-q25, harboring a gene cluster coding for nAChR subunits α5-α3-β4 is a strong candidate to describe interindividual differences in smoking behavior that can be identified with the current gene mapping and association analyses methods. However, the effects of the variants on that locus are small and require large samples in order to emerge in statistical analyses as significant findings.

It was also confirmed that nAChRs have pleiotropic effects. First, a novel association with alcohol use was detected with a variant residing on the *CHRNA5-CHRNA3-CHRNB4* gene cluster on 15q25.1 (Study III). The finding suggests that the effects of alcohol use may be mediated by nAChRs as one mechanism along with several others. Second, it was demonstrated that genetic variants on 15q25.1 may contribute to higher BMI in never smokers and to lower BMI in current smokers (Study IV). The overall effect on BMI would have remained undiscovered with a traditional GWAS approach. Instead, this causality was revealed in a Mendelian Randomization study, which shows how stratification on established environmental factors with a known impact on traits and diseases may provide a means to unveil novel genetic factors that increase the risk of an outcome. However, Mendelian Randomization requires notably large samples to be capable of detecting causality.

Despite all the global effort and time invested in several attempts to identify and characterize the secrets of the human genome, we are barely on the verge of starting to understand its many whimsical functions. The same applies to every complex trait and disorder, including smoking behavior and ND – the genetic architecture remains scarce and requires innovative leaps in technical development, a higher capacity in progressing and analyzing gigantic data sets, along with better storage solutions for the rest of the genetic factors to become visible.

7 ACKNOWLEDGEMENTS

This work was carried out at the Institute for Molecular Medicine Finland (FIMM), University of Helsinki. I thank the director of FIMM, Professor Jaakko Kaprio, for providing excellent research facilities and an inspiring scientific working environment.

My supervisors Professor Jaakko Kaprio and Adjunct Professor Anu Loukola have guided me through this enormous project with an impeccable degree of knowledge, skill and kindness. Jaakko, thank you for giving me the opportunity to do research and learn. I am grateful for your steady guidance and reliable support through all these years, which ended up being a little longer than initially agreed on. Your friendly and kind dedication towards your students combined with the tight schedule of a globally acknowledged and esteemed top researcher, as you are, is exceptionally respectable. I most warmly thank you for all the inspiring conversations and patience you showed in teaching me the many things about genetic epidemiology and smoking-related phenotypes. It has been a privilege to work in your group. Anu, thank you for your unconditional encouragement, inspiration, support, and most of all friendship throughout all the unexpected turns I have had along this journey. Your skills in guiding and teaching a student while creating high-quality research are amazing. I have been very fortunate to have met you. Thank you for standing by my side and believing in me.

I would also like to express my gratitude to the official reviewers of this thesis, Docent Maija Wessman and Assistant Professor Rohan Palmer. Thank you for your excellent work in improving my thesis. Your words of wisdom and clarifying insight on this work were highly appreciated. I would like to thank Adjunct Professor Päivi Onkamo for accepting the position as the official opponent at the public examination of this thesis.

Besides those already mentioned, several other inspiring and talented colleagues have been crucial for this PhD project. I have been so fortunate to be a part of an exceptionally supportive, talented and caring working community. I want to thank all of my current and former colleagues and friends at the Department of Public Health and later at FIMM for your constant support and guidance: Anja Häppölä, Kauko Heikkilä, Ulla Broms, Kaisu Keskitalo-Vuokko, Pia Ruokolinna, Pirkko Särkijärvi, Janne Pitkäniemi, Tellervo Korhonen, Antti Latvala, Miina Ollikainen, Maarit Piirtola, Eero Vuoksimaa, Samuli Ripatti, Mari Siltala, Eila Voipio, Katja Sohlo, Mia Urjansson, and many others.

The collaboration and friendship among the doctoral students and post-docs at FIMM have been more important to me than I can ever describe. I want to express my deepest gratitude to all the current and former colleagues who have made this long journey worth every step: Aileen, Alyce, Anu R., Arvydas, Beenish, Diana, Eeva, Emmi, Ida, Irene, Jade, Khadeeja, Leonie, Mahes, Marjis, Pauliina, Regis, Richa, Sailalitha, Sara, Sari, Sini, Teemu, Tiia, Tuuli, and many others not mentioned here. This journey would not have been the same without you.

This work is dedicated to my late grandfather, Vaari, who took me by the hand and guided me to the trail of learning and education. We were kindred spirits, and your never-ending love and support made me who I am. After a long journey, it finally seems that I have found my way where I am supposed to be. It was a path we set out to walk together almost 24 years ago. *Can you see me?*

The love and support of my family and friends has been invaluable. All my love and gratitude goes to my parents, my brother and my dear grandmother, Mumi, who have been standing by my side and supporting me all the way. Finally, I thank my dear husband Janne for all the happiness, joy and love that we have, and for being my harbor and shelter through all the ups and downs. Right beside you along with our lovely children, Max and Alex, is the best place for me. The three of you mean everything to me, and I love you so much.

8 WEB RESOURCES

The URLs for data presented within are as follows:

1000 Genomes Project, http://www.internationalgenome.org/

BIOSqtl, http://www.genenetwork.nl/biosqtlbrowser

Causal Analysis Research in Tobacco and Alcohol (CARTA) consortium, http://www.bris.ac.uk/expsych/research/brain/targ/research/collaborations/carta/

Ensembl, http://www.ensembl.org/index.html

Exome Aggregation Consortium (ExAC), http://exac.broadinstitute.org/

GTEx Portal, https://www.gtexportal.org/home/

Haplotype Reference Consortium, http://www.haplotype-reference-consortium.org/

International HapMap Project, http://hapmap.nci.nlm.nih.gov/

mQTLdb, http://www.mqtldb.org/

meQTL data, http://www.epigenetics.essex.ac.uk/mQTL

Sequencing Initiative Suomi (SISu) project, www.sisuproject.fi

USCS Genome Browser, https://genome.ucsc.edu/

9 REFERENCES

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**:56-65.

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. (2015) A global reference for human genetic variation. *Nature* **526**:68-74.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**:97-101.

Albuquerque EX, Pereira EF, Alkondon M, Rogers SW. (2009) Mammalian nicotinic acetylcholine receptors: from structure to function. *Physiol Rev* **89**:73-120.

Albuquerque EX, Pereira EF, Castro NG, Alkondon M, Reinhardt S, Schroder H, et al. (1995) Nicotinic receptor function in the mammalian central nervous system. *Ann N Y Acad Sci* **757**:48-72.

Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M. (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* **69**:936-950.

Ambrose JA, Barua RS. (2004) The pathophysiology of cigarette smoking and cardiovascular disease: an update. *J Am Coll Cardiol* **43**:1731-1737.

American Psychiatric Association. (1994) Diagnostic and statistical manual of mental disorders: DSM-IV (4th ed.). Washington, DC: American Psychiatric Association.

Amos A, Haglund M. (2000) From social taboo to "torch of freedom": the marketing of cigarettes to women. *Tob Control* **9**:3-8.

Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, 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.

Anttila V, Bulik-Sullivan B, Finucane H, Bras J, Duncan L, Escott-Price V, et al. (2016) Analysis of shared heritability in common disorders of the brain. *bioRxiv* preprint; doi: http://dx.doi.org/10.1101/048991.

Armitage A, Dollery C, Houseman T, Kohner E, Lewis PJ, Turner D. (1978) Absorption of nicotine from small cigars. *Clin Pharmacol Ther* **23**:143-151.

Ashare RL, Lerman C, Cao W, Falcone M, Bernardo L, Ruparel K, et al. (2016) Nicotine withdrawal alters neural responses to psychosocial stress. *Psychopharmacology* **233**:2459-2467.

Ashare RL, Wileyto EP, Perkins KA, Schnoll RA. (2013) The first 7 days of a quit attempt predicts relapse: validation of a measure for screening medications for nicotine dependence. *J Addict Med* **7**:249-254.

Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH, et al. (2017) DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *Int J Cancer* **140**:50–61.

Bailey-Wilson JE, Wilson AF. (2011) Linkage analysis in the next-generation sequencing era. *Hum Hered* **72**:228-236.

Balfour DJ. (2004) The neurobiology of tobacco dependence: a preclinical perspective on the role of the dopamine projections to the nucleus accumbens. *Nicotine Tob Res* **6**:899-912.

Baurley JW, Edlund CK, Pardamean CI, Conti DV, Krasnow R, Javitz HS, et al. (2016) Genome-Wide Association of the Laboratory-Based Nicotine Metabolite Ratio in Three Ancestries. *Nicotine Tob Res* **18**:1837-1844.

Beattie EC, Howe CL, Wilde A, Brodsky FM, Mobley WC. (2000) NGF signals through TrkA to increase clathrin at the plasma membrane and enhance clathrin-mediated membrane trafficking. *J Neurosci* **20**:7325-7333.

Becker JB, Hu M. (2008) Sex differences in drug abuse. *Front Neuroendocrinol* **29**:36-47.

Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. (2001) Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **125**:279-284.

Benowitz NL. (2010) Nicotine addiction. *N Engl J Med* **362**:2295-2303.

Benowitz NL. (2009) Pharmacology of nicotine: addiction, smoking-induced disease, and therapeutics. *Annu Rev Pharmacol Toxicol* **49**:57-71.

Benowitz NL. (1999) The biology of nicotine dependence: from the 1988 Surgeon General's Report to the present and into the future. *Nicotine Tob Res* **1** Suppl 2:S159-63.

Benowitz NL, Hatsukami D. (1998) Gender differences in the pharmacology of nicotine addiction. *Addict Biol* **3**:383-404.

Benowitz NL, Jacob P,3rd. (1994) Metabolism of nicotine to cotinine studied by a dual stable isotope method. *Clin Pharmacol Ther* **56**:483-493.

Benowitz NL, Porchet H, Sheiner L, Jacob P,3rd. (1988) Nicotine absorption and cardiovascular effects with smokeless tobacco use: comparison with cigarettes and nicotine gum. *Clin Pharmacol Ther* **44**:23-28.

Berg JZ, von Weymarn LB, Thompson EA, Wickham KM, Weisensel NA, Hatsukami DK, et al. (2010) UGT2B10 genotype influences nicotine glucuronidation, oxidation, and consumption. *Cancer Epidemiol Biomarkers Prev* **19**:1423-1431.

Berridge MS, Apana SM, Nagano KK, Berridge CE, Leisure GP, Boswell MV. (2010) Smoking produces rapid rise of [11C]nicotine in human brain. *Psychopharmacology* **209**:383-394.

Bibel M, Barde YA. (2000) Neurotrophins: key regulators of cell fate and cell shape in the vertebrate nervous system. *Genes Dev* **14**:2919-2937.

Bidwell LC, Knopik VS, Audrain-McGovern J, Glynn TR, Spillane NS, Ray LA, et al. (2015) Novelty Seeking as a Phenotypic Marker of Adolescent Substance Use. *Subst Abuse* **17:**(Suppl 1):1-10.

Bidwell LC, Palmer RH, Brick L, McGeary JE, Knopik VS. (2016) Genome-wide single nucleotide polymorphism heritability of nicotine dependence as a multidimensional phenotype. *Psychol Med* **46**:2059-2069.

Bierut LJ. (2011) Genetic vulnerability and susceptibility to substance dependence. *Neuron* **69**:618-627.

Bierut LJ, Dinwiddie SH, Begleiter H, Crowe RR, Hesselbrock V, Nurnberger JI,Jr, et al. (1998) Familial transmission of substance dependence: alcohol, marijuana, cocaine, and habitual smoking: a report from the Collaborative Study on the Genetics of Alcoholism. *Arch Gen Psychiatry* **55**:982-988.

Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, et al. (2007) Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet* **16**:24-35.

Bloom AJ, von Weymarn LB, Martinez M, Bierut LJ, Goate A, Murphy SE. (2013) The contribution of common UGT2B10 and CYP2A6 alleles to variation in nicotine glucuronidation among European Americans. *Pharmacogenet Genomics* **23**:706-716.

Boden JM, Fergusson DM, Horwood LJ. (2010) Cigarette smoking and depression: tests of causal linkages using a longitudinal birth cohort. *Br J Psychiatry* **196**:440-446.

Bolanos CA, Nestler EJ. (2004) Neurotrophic mechanisms in drug addiction. *Neuromolecular Med* **5**:69-83.

Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. (2016) Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet* **49**: 131-138.

Boomsma D, Busjahn A, Peltonen L. (2002) Classical twin studies and beyond. *Nat Rev Genet* **3**:872-882.

Borodulin K, Vartiainen E, Peltonen M, Jousilahti P, Juolevi A, Laatikainen T, et al. (2015) Forty-year trends in cardiovascular risk factors in Finland. *Eur J Public Health* **25**:539-546.

Botstein D, Risch N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33** Suppl:228-237.

Bowman ER, Mckennis H,Jr. (1962) Studies on the metabolism of (-)-cotinine in the human. *J Pharmacol Exp Ther* **135**:306-311.

Broms U, Madden PA, Heath AC, Pergadia ML, Shiffman S, Kaprio J. (2007) The Nicotine Dependence Syndrome Scale in Finnish smokers. *Drug Alcohol Depend* **89**:42-51.

Broms U, Silventoinen K, Madden PA, Heath AC, Kaprio J. (2006) Genetic architecture of smoking behavior: a study of Finnish adult twins. *Twin Res Hum Genet* **9**:64-72.

Brook JS, Duan T, Zhang C, Cohen PR, Brook DW. (2008) The association between attention deficit hyperactivity disorder in adolescence and smoking in adulthood. *Am J Addict* **17**:54-59.

Bucholz KK, Cadoret R, Cloninger CR, Dinwiddie SH, Hesselbrock VM, Nurnberger JI,Jr, et al. (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.

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**:291-295.

Buller DB, Borland R, Woodall WG, Hall JR, Burris-Woodall P, Voeks JH. (2003) Understanding factors that influence smoking uptake. *Tob Control* **12** Suppl 4:IV16-25.

Burden S, Yarden Y. (1997) Neuregulins and their receptors: a versatile signaling module in organogenesis and oncogenesis. *Neuron* **18**:847-855.

Burns DM. (2003) Epidemiology of smoking-induced cardiovascular disease. *Prog Cardiovasc Dis* **46**:11-29.

Carmelli D, Swan GE, Robinette D, Fabsitz R. (1992) Genetic influence on smoking--a study of male twins. *N Engl J Med* **327**:829-833.

Charach A, Yeung E, Climans T, Lillie E. (2011) Childhood attention-deficit/hyperactivity disorder and future substance use disorders: comparative meta-analyses. *J Am Acad Child Adolesc Psychiatry* **50**:9-21.

Chen X, Chen J, Williamson VS, An SS, Hettema JM, Aggen SH, et al. (2009) Variants in nicotinic acetylcholine receptors alpha5 and alpha3 increase risks to nicotine dependence. *Am J Med Genet B Neuropsychiatr Genet* **150B**:926-933.

Chen G, Giambrone NE,Jr, Dluzen DF, Muscat JE, Berg A, Gallagher CJ, et al. (2010) Glucuronidation genotypes and nicotine metabolic phenotypes: importance of functional UGT2B10 and UGT2B17 polymorphisms. *Cancer Res* **70**:7543-7552.

Clarke AJ, Cooper DN. (2010) GWAS: heritability missing in action? *Eur J Hum Genet* **18**:859-861.

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**:1251-1260.

Conrad KM, Flay BR, Hill D. (1992) Why children start smoking cigarettes: predictors of onset. *Br J Addict* **87**:1711-1724.

Cottler LB, Robins LN, Grant BF, Blaine J, Towle LH, Wittchen HU, et al. (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.

Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. (1961) General nature of the genetic code for proteins. *Nature* **192**:1227-1232.

Dahl JP, Jepson C, Levenson R, Wileyto EP, Patterson F, Berrettini WH, et al. (2006) Interaction between variation in the D2 dopamine receptor (DRD2) and the neuronal calcium sensor-1 (FREQ) genes in predicting response to nicotine replacement therapy for tobacco dependence. *Pharmacogenomics J* **6**:194-199.

Dajas-Bailador F, Wonnacott S. (2004) Nicotinic acetylcholine receptors and the regulation of neuronal signalling. *Trends Pharmacol Sci* **25**:317-324.

Dani JA, De Biasi M. (2001) Cellular mechanisms of nicotine addiction. *Pharmacol Biochem Behav* **70**:439-446.

Dani JA, Harris RA. (2005) Nicotine addiction and comorbidity with alcohol abuse and mental illness. *Nat Neurosci* **8**:1465-1470.

Davey Smith G, Ebrahim S. (2003) Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**:1-22. doi: 10.1093/ije/dyg070.

David SP, Hamidovic A, Chen GK, Bergen AW, Wessel J, Kasberger JL, et al. (2012) Genome-wide meta-analyses of smoking behaviors in African Americans. *Transl Psychiatry* **2**:e119.

de Leon J, Becona E, Gurpegui M, Gonzalez-Pinto A, Diaz FJ. (2002) The association between high nicotine dependence and severe mental illness may be consistent across countries. *J Clin Psychiatry* **63**:812-816.

de Leon J, Diaz FJ. (2005) A meta-analysis of worldwide studies demonstrates an association between schizophrenia and tobacco smoking behaviors. *Schizophr Res* **76**:135-157.

de la Chapelle A. (1993) Disease gene mapping in isolated human populations: the example of Finland. *J Med Genet* **30**:857-865.

Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. (2013) Haplotype estimation using sequencing reads. *Am J Hum Genet* **93**:687-696.

Di Chiara G, Bassareo V, Fenu S, De Luca MA, Spina L, Cadoni C, et al. (2004) Dopamine and drug addiction: the nucleus accumbens shell connection. *Neuropharmacology* **47** Suppl 1:227-241.

Dick DM, Bernard M, Aliev F, Viken R, Pulkkinen L, Kaprio J, et al. (2009) The role of socioregional factors in moderating genetic influences on early adolescent behavior problems and alcohol use. *Alcohol Clin Exp Res* **33**:1739-1748.

DiFranza J, Ursprung WW, Lauzon B, Bancej C, Wellman RJ, Ziedonis D, et al. (2010) A systematic review of the Diagnostic and Statistical Manual diagnostic criteria for nicotine dependence. *Addict Behav* **35**:373-382.

Doll R, Peto R, Wheatley K, Gray R, Sutherland I. (1994) Mortality in relation to smoking: 40 years' observations on male British doctors. *BMJ* **309**:901-911.

ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**:e1001046.

Etter JF, Duc TV, Perneger TV. (1999) Validity of the Fagerstrom test for nicotine dependence and of the Heaviness of Smoking Index among relatively light smokers. *Addiction* **94**:269-281.

Evangelou E, Ioannidis JP. (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**:379-389.

Everitt BJ, Cardinal RN, Parkinson JA, Robbins TW. (2003) Appetitive behavior: impact of amygdala-dependent mechanisms of emotional learning. *Ann N Y Acad Sci* **985**:233-250.

Fagerstrom KO. (1978) Measuring degree of physical dependence to tobacco smoking with reference to individualization of treatment. *Addict Behav* **3**:235-241.

Filippidis FT, Agaku IT, Vardavas CI. (2015) The association between peer, parental influence and tobacco product features and earlier age of onset of regular smoking among adults in 27 European countries. *Eur J Public Health* **25**:814-818.

Fowler CD, Lu Q, Johnson PM, Marks MJ, Kenny PJ. (2011) Habenular alpha5 nicotinic receptor subunit signalling controls nicotine intake. *Nature* **471:**597-601.

Franklin TR, Jagannathan K, Wetherill RR, Johnson B, Kelly S, Langguth J, et al. (2015) Influence of menstrual cycle phase on neural and craving responses to appetitive smoking cues in naturally cycling females. *Nicotine Tob Res* **17**:390-397.

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**:889-894.

Freathy RM, Kazeem GR, Morris RW, Johnson PC, Paternoster L, Ebrahim S, et al. (2011) Genetic variation at CHRNA5-CHRNA3-CHRNB4 interacts with smoking status to influence body mass index. *Int J Epidemiol* **40**:1617-1628.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. (2002) The structure of haplotype blocks in the human genome. *Science* **296**:2225-2229.

Gage SH, Smith GD, Zammit S, Hickman M, Munafo MR. (2013) Using Mendelian randomisation to infer causality in depression and anxiety research. *Depress Anxiety* **30**:1185-1193.

Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. (2016) Systematic identification of genetic influences on methylation across the human life course. *Genome Biol* **17**:61-016-0926-z.

Gelernter J, Kranzler HR, Sherva R, Almasy L, Herman AI, Koesterer R, et al. (2015) Genome-wide association study of nicotine dependence in American populations: identification of novel risk loci in both African-Americans and European-Americans. *Biol Psychiatry* **77**:493-503.

George O, Ghozland S, Azar MR, Cottone P, Zorrilla EP, Parsons LH, et al. (2007) CRF-CRF1 system activation mediates withdrawal-induced increases in nicotine self-administration in nicotine-dependent rats. *Proc Natl Acad Sci U S A* **104**:17198-17203.

George TP, O'Malley SS. (2004) Current pharmacological treatments for nicotine dependence. *Trends Pharmacol Sci* **25**:42-48.

Gibson G. (2012) Rare and common variants: twenty arguments. *Nat Rev Genet* **13**:135-145.

Gil AG, Vega WA, Turner RJ. (2002) Early and mid-adolescence risk factors for later substance abuse by African Americans and European Americans. *Public Health Rep* **117** Suppl 1:S15-29.

Gold AB, Lerman C. (2012) Pharmacogenetics of smoking cessation: role of nicotine target and metabolism genes. *Hum Genet* **131**:857-876.

Goring HH, Terwilliger JD. (2000) Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet* **66**:1310-1327.

Govind AP, Vezina P, Green WN. (2009) Nicotine-induced upregulation of nicotinic receptors: underlying mechanisms and relevance to nicotine addiction. *Biochem Pharmacol* **78**:756-765.

Grady SR, Salminen O, Laverty DC, Whiteaker P, McIntosh JM, Collins AC, et al. (2007) The subtypes of nicotinic acetylcholine receptors on dopaminergic terminals of mouse striatum. *Biochem Pharmacol* **74**:1235-1246.

Grucza RA, Wang JC, Stitzel JA, Hinrichs AL, Saccone SF, Saccone NL, et al. (2008) A risk allele for nicotine dependence in CHRNA5 is a protective allele for cocaine dependence. *Biol Psychiatry* **64**:922-929.

GTEx Consortium (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**:648-660.

Haller G, Druley T, Vallania FL, Mitra RD, Li P, Akk G, et al. (2012) Rare missense variants in CHRNB4 are associated with reduced risk of nicotine dependence. *Hum Mol Genet* **21**:647-655.

Halpern J, Whittemore AS. (1999) Multipoint linkage analysis. A cautionary note. *Hum Hered* **49**:194-196.

Han S, Gelernter J, Luo X, Yang BZ. (2010) Meta-analysis of 15 genome-wide linkage scans of smoking behavior. *Biol Psychiatry* **67**:12-19.

Han S, Yang BZ, Kranzler HR, Oslin D, Anton R, Gelernter J. (2011) Association of CHRNA4 polymorphisms with smoking behavior in two populations. *Am J Med Genet B Neuropsychiatr Genet* **156B**:421-429.

Hancock DB, Reginsson GW, Gaddis NC, Chen X, Saccone NL, Lutz SM, et al. (2015) Genome-wide meta-analysis reveals common splice site acceptor variant in CHRNA4 associated with nicotine dependence. *Transl Psychiatry* **5**:e651.

Hannon E, Spiers H, Viana J, Pidsley R, Burrage J, Murphy TM, et al. (2016) Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci* **19**:48-54.

Hara M, Inoue M, Shimazu T, Yamamoto S, Tsugane S, Japan Public Health Center-based Prospective Study Group. (2010) The association between cancer risk and age at onset of smoking in Japanese. *J Epidemiol* **20**:128-135.

Hartz SM, Short SE, Saccone NL, Culverhouse R, Chen L, Schwantes-An TH, et al. (2012) Increased genetic vulnerability to smoking at CHRNA5 in early-onset smokers. *Arch Gen Psychiatry* **69**:854-860.

Heath AC, Martin NG. (1993) Genetic models for the natural history of smoking: evidence for a genetic influence on smoking persistence. *Addict Behav* **18**:19-34.

Heath AC, Whitfield JB, Martin NG, Pergadia ML, Goate AM, Lind PA, et al. (2011) A quantitative-trait genome-wide association study of alcoholism risk in the community: findings and implications. *Biol Psychiatry* **70**:513-518.

Heatherton TF, Kozlowski LT, Frecker RC, Fagerstrom KO. (1991) The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *Br J Addict* **86**:1119-1127.

Heistaro S. (2008) Methodology report Health 2000 Survey. Edited by Heistaro S. Publications of the National Public Health Institute B 26/2008.

Hendrickson LM, Guildford MJ, Tapper AR. (2013) Neuronal nicotinic acetylcholine receptors: common molecular substrates of nicotine and alcohol dependence. *Front Psychiatry* **4**:29.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**:9362-9367.

Hitchman SC, Fong GT. (2011) Gender empowerment and female-to-male smoking prevalence ratios. *Bull World Health Organ* **89**:195-202.

Ho MK, Tyndale RF.(2007) Overview of the pharmacogenomics of cigarette smoking. *Pharmacogenomics J* **7**:81-98.

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**:955-959.

Howie BN, Donnelly P, Marchini J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**:e1000529.

Hu S, Brody CL, Fisher C, Gunzerath L, Nelson ML, Sabol SZ, et al. (2000) Interaction between the serotonin transporter gene and neuroticism in cigarette smoking behavior. *Mol Psychiatry* **5**:181-188.

Hughes JR. (2007) Effects of abstinence from tobacco: valid symptoms and time course. *Nicotine Tob Res* **9**:315-327.

Hughes JR, Hatsukami D. (1986) Signs and symptoms of tobacco withdrawal. *Arch Gen Psychiatry* **43**:289-294.

Hughes JR, Hatsukami DK, Mitchell JE, Dahlgren LA. (1986) Prevalence of smoking among psychiatric outpatients. *Am J Psychiatry* **143**:993-997.

Hughes JR, Helzer JE, Lindberg SA. (2006) Prevalence of DSM/ICD-defined nicotine dependence. *Drug Alcohol Depend* **85**:91-102.

Hughes JR, Oliveto AH, Riggs R, Kenny M, Liguori A, Pillitteri JL, et al. (2004) Concordance of different measures of nicotine dependence: two pilot studies. *Addict Behav* **29**:1527-1539.

Hukkanen J, Jacob P,3rd, Benowitz NL. (2005) Metabolism and disposition kinetics of nicotine. *Pharmacol Rev* **57**:79-115.

Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. (2008) A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**:633-637.

Huxley RR, Woodward M. (2011) Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies. *Lancet* **378**:1297-1305.

International HapMap Consortium. (2003) The International HapMap Project. *Nature* **426**:789-796.

International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**:931-945.

Ioannidis JP, Thomas G, Daly MJ. (2009) Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet* **10**:318-329.

Jackson KJ, Marks MJ, Vann RE, Chen X, Gamage TF, Warner JA, et al. (2010) Role of alpha5 nicotinic acetylcholine receptors in pharmacological and behavioral effects of nicotine in mice. *J Pharmacol Exp Ther* **334:**137-146.

Jamal M, Does AJ, Penninx BW, Cuijpers P. (2011) Age at smoking onset and the onset of depression and anxiety disorders. *Nicotine Tob Res* **13**:809-819.

Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. (2016) Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet* **9**:436-447.

John U, Meyer C, Hapke U, Rumpf HJ, Schumann A. (2004) Nicotine dependence, quit attempts, and quitting among smokers in a regional population sample from a country with a high prevalence of tobacco smoking. *Prev Med* **38**:350-358.

Kamens HM, Corley RP, McQueen MB, Stallings MC, Hopfer CJ, Crowley TJ, et al. (2013) Nominal association with CHRNA4 variants and nicotine dependence. *Genes Brain Behav* **12**:297-304.

Kaprio J. (2006) Twin studies in Finland 2006. *Twin Res Hum Genet* **9**:772-777.

Kaprio J. (2013) The Finnish Twin Cohort Study: an update. *Twin Res Hum Genet* **16**:157-162.

Kaprio J, Koskenvuo M. (1988) A prospective study of psychological and socioeconomic characteristics, health behavior and morbidity in cigarette smokers prior to quitting compared to persistent smokers and non-smokers. *J Clin Epidemiol* **41**:139-150.

Kaprio J, Koskenvuo M. (2002) Genetic and environmental factors in complex diseases: the older Finnish Twin Cohort. *Twin Res* **5**:358-365.

Kaprio J, Koskenvuo M, Langinvainio H. (1984) Finnish twins reared apart. IV: Smoking and drinking habits. A preliminary analysis of the effect of heredity and environment. *Acta Genet Med Gemellol* **33**:425-433.

Karlin A, Cox RN, Dipaola M, Holtzman E, Kao PN, Lobel P, et al. (1986) Functional domains of the nicotinic acetylcholine receptor. *Ann N Y Acad Sci* **463**:53-69.

Kelley AE, Berridge KC. (2002) The neuroscience of natural rewards: relevance to addictive drugs. *J Neurosci* **22**:3306-3311.

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. (2014) Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**:6131-6138.

Kenfield SA, Stampfer MJ, Rosner BA, Colditz GA. (2008) Smoking and smoking cessation in relation to mortality in women. *JAMA* **299**:2037-2047.

Kenny PJ, Markou A. (2001) Neurobiology of the nicotine withdrawal syndrome. *Pharmacol Biochem Behav* **70**:531-549.

Kent KM, Pelham WE,Jr, Molina BS, Sibley MH, Waschbusch DA, Yu J, et al. (2011) The academic experience of male high school students with ADHD. *J Abnorm Child Psychol* **39**:451-462.

Keskitalo K, Broms U, Heliovaara M, Ripatti S, Surakka I, Perola M, et al. (2009) Association of serum cotinine level with a cluster of three nicotinic acetylcholine receptor genes (CHRNA3/CHRNA5/CHRNB4) on chromosome 15. *Hum Mol Genet* **18**:4007-4012.

Kessler RC, Adler L, Barkley R, Biederman J, Conners CK, Demler O, et al. (2006) The prevalence and correlates of adult ADHD in the United States: results from the National Comorbidity Survey Replication. *Am J Psychiatry* **163**:716-723.

Khuder SA, Dayal HH, Mutgi AB. (1999) Age at smoking onset and its effect on smoking cessation. *Addict Behav* **24**:673-677.

Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, et al. (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. *Am J Hum Genet* **62**:1171-1179.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308**:385-389.

Kollins SH, McClernon FJ, Fuemmeler BF. (2005) Association between smoking and attention-deficit/hyperactivity disorder symptoms in a population-based sample of young adults. *Arch Gen Psychiatry* **62**:1142-1147.

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

Kuper H, Adami HO, Boffetta P. (2002) Tobacco use, cancer causation and public health impact. *J Intern Med* **251**:455-466.

Ladd-Acosta C, Fallin MD. (2016) The role of epigenetics in genetic and environmental epidemiology. *Epigenomics* **8**:271-283.

Lakier JB. (1992) Smoking and cardiovascular disease. *Am J Med* **93**:8S-12S.

Lander E, Kruglyak L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* **11**:241-247.

Lander ES. (2011) Initial impact of the sequencing of the human genome. *Nature* **470**:187-197.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**:860-921.

Lang UE, Sander T, Lohoff FW, Hellweg R, Bajbouj M, Winterer G, et al. (2007) Association of the met66 allele of brain-derived neurotrophic factor (BDNF) with smoking. *Psychopharmacology* **190**:433-439.

Lasser K, Boyd JW, Woolhandler S, Himmelstein DU, McCormick D, Bor DH. (2000) Smoking and mental illness: A population-based prevalence study. *JAMA* **284**:2606-2610.

Latvala A, Ollikainen M. (2016) Mendelian randomization in (epi)genetic epidemiology: an effective tool to be handled with care. *Genome Biol* **17**:156-016-1018-9.

Laviolette SR, van der Kooy D. (2004) The neurobiology of nicotine addiction: bridging the gap from molecules to behaviour. *Nat Rev Neurosci* **5**:55-65.

Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. (2008) Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* **27**:1133-1163.

Le Moal M, Koob GF. (2007) Drug addiction: pathways to the disease and pathophysiological perspectives. *Eur Neuropsychopharmacol* **17**:377-393.

Lee JH, Cho MH, Hersh CP, McDonald ML, Crapo JD, Bakke PS, et al. (2014) Genetic susceptibility for chronic bronchitis in chronic obstructive pulmonary disease. *Respir Res* **15**:113-014-0113-2.

Lee SS, Humphreys KL, Flory K, Liu R, Glass K. (2011) Prospective association of childhood attention-deficit/hyperactivity disorder (ADHD) and substance use and abuse/dependence: a meta-analytic review. *Clin Psychol Rev* **31:**328-341.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**:285-291.

Lerman C, Caporaso NE, Audrain J, Main D, Boyd NR, Shields PG. (2000) Interacting effects of the serotonin transporter gene and neuroticism in smoking practices and nicotine dependence. *Mol Psychiatry* **5**:189-192.

Lerman C, LeSage MG, Perkins KA, O'Malley SS, Siegel SJ, Benowitz NL, et al. (2007) Translational research in medication development for nicotine dependence. *Nat Rev Drug Discov* **6**:746-762.

Lessov CN, Martin NG, Statham DJ, Todorov AA, Slutske WS, Bucholz KK, et al. (2004) Defining nicotine dependence for genetic research: evidence from Australian twins. *Psychol Med* **34**:865-879.

Lessov-Schlaggar CN, Pergadia ML, Khroyan TV, Swan GE. (2008) Genetics of nicotine dependence and pharmacotherapy. *Biochem Pharmacol* **75**:178-195.

Leventhal AM, David SP, Brightman M, Strong D, McGeary JE, Brown RA, et al. (2012) Dopamine D4 receptor gene variation moderates the efficacy of bupropion for smoking cessation. *Pharmacogenomics J* **12**:86-92.

Leventhal AM, Lee W, Bergen AW, Swan GE, Tyndale RF, Lerman C, et al. (2014) Nicotine dependence as a moderator of genetic influences on smoking cessation treatment outcome. *Drug Alcohol Depend* **138**:109-117.

Lewinsohn PM, Rohde P, Brown RA. (1999) Level of current and past adolescent cigarette smoking as predictors of future substance use disorders in young adulthood. *Addiction* **94**:913-921.

Li MD, Cheng R, Ma JZ, Swan GE. (2003) A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction* **98**:23-31.

Li MD, Yoon D, Lee JY, Han BG, Niu T, Payne TJ, et al. (2010) Associations of variants in CHRNA5/A3/B4 gene cluster with smoking behaviors in a Korean population. *PLoS One* **5**:e12183.

Li TK, Volkow ND, Baler RD, Egli M. (2007) The biological bases of nicotine and alcohol co-addiction. *Biol Psychiatry* **61**:1-3.

Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnstrom K, et al. (2014) Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* **10**:e1004494.

Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* **42**:436-440.

Loukola A, Broms U, Maunu H, Widen E, Heikkila K, Siivola M, et al. (2008) Linkage of nicotine dependence and smoking behavior on 10q, 7q and 11p in twins with homogeneous genetic background. *Pharmacogenomics J* **8**:209-219.

Loukola A, Buchwald J, Gupta R, Palviainen T, Hallfors J, Tikkanen E, et al. (2015) A Genome-Wide Association Study of a Biomarker of Nicotine Metabolism. *PLoS Genet* **11**:e1005498.

Loukola A, Wedenoja J, Keskitalo-Vuokko K, Broms U, Korhonen T, Ripatti S, et al. (2014) Genome-wide association study on detailed profiles of smoking behavior and nicotine dependence in a twin sample. *Mol Psychiatry* **19**:615-624.

Lutz SM, Cho MH, Young K, Hersh CP, Castaldi PJ, McDonald ML, et al. (2015) A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet* **16**:138-015-0299-4.

Lynch WJ. (2009) Sex and ovarian hormones influence vulnerability and motivation for nicotine during adolescence in rats. *Pharmacol Biochem Behav* **94**:43-50.

Mackillop J, Obasi E, Amlung MT, McGeary JE, Knopik VS. (2010) The Role of Genetics in Nicotine Dependence: Mapping the Pathways from Genome to Syndrome. *Curr Cardiovasc Risk Rep* **4**:446-453.

Madden PA, Heath AC, Pedersen NL, Kaprio J, Koskenvuo MJ, Martin NG. (1999) The genetics of smoking persistence in men and women: a multicultural study. *Behav Genet* **29**:423-431.

Malaiyandi V, Sellers EM, Tyndale RF. (2005) Implications of CYP2A6 genetic variation for smoking behaviors and nicotine dependence. *Clin Pharmacol Ther* **77**:145-158.

Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. (2016) A genomic history of Aboriginal Australia. *Nature* **538**:207-214.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**:201-206.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. (2009) Finding the missing heritability of complex diseases. *Nature* **461**:747-753.

Mansvelder HD, McGehee DS. (2002) Cellular and synaptic mechanisms of nicotine addiction. *J Neurobiol* **53:**606-617.

Mansvelder HD, Mertz M, Role LW. (2009) Nicotinic modulation of synaptic transmission and plasticity in cortico-limbic circuits. *Semin Cell Dev Biol* **20**:432-440.

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, et al. (2005) High-resolution whole-genome association study of Parkinson disease. *Am J Hum Genet* **77**:685-693.

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.

Mathers CD, Loncar D. (2006) Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med* **3**:e442.

Mattson MP, Meffert MK. (2006) Roles for NF-kappaB in nerve cell survival, plasticity, and disease. *Cell Death Differ* **13**:852-860.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**:1190-1195.

Mayhew KP, Flay BR, Mott JA. (2000) Stages in the development of adolescent smoking. *Drug Alcohol Depend* **59** Suppl 1:S61-81.

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**:1279-1283.

McClellan J, King MC. (2010) Genetic heterogeneity in human disease. *Cell* **141:**210-217.

McClernon FJ, Kollins SH. (2008) ADHD and smoking: from genes to brain to behavior. *Ann N Y Acad Sci* **1141**:131-147.

McClure-Begley TD, Papke RL, Stone KL, Stokes C, Levy AD, Gelernter J, et al. (2014) Rare human nicotinic acetylcholine receptor alpha4 subunit (CHRNA4) variants affect expression and function of high-affinity nicotinic acetylcholine receptors. *J Pharmacol Exp Ther* **348**:410-420.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. (2016) The Ensembl Variant Effect Predictor. *Genome Biol* **17**:122-016-0974-4.

Medzhitov R. (2001) Toll-like receptors and innate immunity. *Nat Rev Immunol* **1**:135-145.

Mercken L, Candel M, Willems P, de Vries H. (2009) Social influence and selection effects in the context of smoking behavior: changes during early and mid adolescence. *Health Psychol* **28**:73-82.

Metzker ML. (2010) Sequencing technologies - the next generation. *Nat Rev Genet* **11**:31-46.

Mineur YS, Abizaid A, Rao Y, Salas R, DiLeone RJ, Gundisch D, et al. (2011) Nicotine decreases food intake through activation of POMC neurons. *Science* **332**:1330-1332.

Molina BS, Hinshaw SP, Eugene Arnold L, Swanson JM, Pelham WE, Hechtman L, et al. (2013) Adolescent substance use in the multimodal treatment study of attention-deficit/hyperactivity disorder (ADHD) (MTA) as a function of childhood ADHD, random assignment to childhood treatments, and subsequent medication. *J Am Acad Child Adolesc Psychiatry* **52**:250-263.

Moolchan ET, Radzius A, Epstein DH, Uhl G, Gorelick DA, Cadet JL, et al. (2002) The Fagerstrom Test for Nicotine Dependence and the Diagnostic Interview Schedule: do they diagnose the same smokers? *Addict Behav* **27**:101-113.

Morales-Perez CL, Noviello CM, Hibbs RE. (2016) X-ray structure of the human alpha4beta2 nicotinic receptor. *Nature* **538**:411-415.

Moran LV, Sampath H, Kochunov P, Hong LE. (2013) Brain circuits that link schizophrenia to high risk of cigarette smoking. *Schizophr Bull* **39**:1373-1381.

Morton NE. (1955) Sequential tests for the detection of linkage. *Am J Hum Genet* **7**:277-318.

Mucha L, Stephenson J, Morandi N, Dirani R. (2006) Meta-analysis of disease risk associated with smoking, by gender and intensity of smoking. *Gend Med* **3**:279-291.

Mudge JM, Harrow J. (2016) The state of play in higher eukaryote gene annotation. *Nat Rev Genet* **17**:758-772.

Munafo MR, Johnstone EC, Walther D, Uhl GR, Murphy MF, Aveyard P. (2011) CHRNA3 rs1051730 genotype and short-term smoking cessation. *Nicotine Tob Res* **13**:982-988.

Munafo MR, Tilling K, Ben-Shlomo Y. (2009) Smoking status and body mass index: a longitudinal study. *Nicotine Tob Res* **11**:765-771.

Munafo MR, Timofeeva MN, Morris RW, Prieto-Merino D, Sattar N, Brennan P, et al. (2012) Association between genetic variants on chromosome 15q25 locus and objective measures of tobacco exposure. *J Natl Cancer Inst* **104**:740-748.

Mwenifumbo JC, Tyndale RF. (2007) Genetic variability in CYP2A6 and the pharmacokinetics of nicotine. *Pharmacogenomics* **8**:1385-1402.

Naidoo N, Pawitan Y, Soong R, Cooper DN, Ku CS. (2011) Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Hum Genomics* **5**:577-622.

Nakamura T, Gehrke AR, Lemberg J, Szymaszek J, Shubin NH. (2016) Digits and fin rays share common developmental histories. *Nature* **537**:225-228.

Nestler EJ. (2005) Is there a common molecular pathway for addiction? *Nat Neurosci* **8**:1445-1449.

Nestler EJ. (2001) Molecular basis of long-term plasticity underlying addiction. *Nat Rev Neurosci* **2**:119-128.

Ng MY, Levinson DF, Faraone SV, Suarez BK, DeLisi LE, Arinami T, et al. (2009) Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol Psychiatry* **14**:774-785.

Norio R. (2003a) Finnish Disease Heritage I: characteristics, causes, background. *Hum Genet* **112**:441-456.

Norio R. (2003b) Finnish Disease Heritage II: population prehistory and genetic roots of Finns. *Hum Genet* **112**:457-469.

Norio R. (2003c) The Finnish Disease Heritage III: the individual diseases. *Hum Genet* **112**:470-526.

O'Connell JR, Weeks DE. (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* **63**:259-266.

Olausson P, Jentsch JD, Taylor JR. (2004) Repeated nicotine exposure enhances responding with conditioned reinforcement. *Psychopharmacology* **173**:98-104.

Olfson E, Saccone NL, Johnson EO, Chen LS, Culverhouse R, Doheny K, et al. (2016) Rare, low frequency and common coding variants in CHRNA5 and their contribution to nicotine dependence in European and African Americans. *Mol Psychiatry* **21**:601-607.

Ott J. (1989) Computer-simulation methods in human linkage analysis. *Proc Natl Acad Sci U S A* **86**:4175-4178.

Ott J. (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet* **26**:588-597.

Ott J, Wang J, Leal SM. (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* **16**:275-284.

Pagani L, Lawson DJ, Jagoda E, Morseburg A, Eriksson A, Mitt M, et al. (2016) Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**:238-242.

Pauly JR. (2008) Gender differences in tobacco smoking dynamics and the neuropharmacological actions of nicotine. *Front Biosci* **13**:505-516.

Peltonen L, Jalanko A, Varilo T. (1999) Molecular genetics of the Finnish disease heritage. *Hum Mol Genet* **8**:1913-1923.

Peltonen L, Palotie A, Lange K. (2000) Use of population isolates for mapping complex traits. *Nat Rev Genet* **1**:182-190.

Pergadia ML, Agrawal A, Heath AC, Martin NG, Bucholz KK, Madden PA. (2010) Nicotine withdrawal symptoms in adolescent and adult twins. *Twin Res Hum Genet* **13**:359-369.

Pergadia ML, Agrawal A, Loukola A, Montgomery GW, Broms U, Saccone SF, et al. (2009) Genetic linkage findings for DSM-IV nicotine withdrawal in two populations. *Am J Med Genet B Neuropsychiatr Genet* **150B**:950-959.

Pergadia ML, Heath AC, Martin NG, Madden PA. (2006) Genetic analyses of DSM-IV nicotine withdrawal in adult twins. *Psychol Med* **36**:963-972.

Petersen L, Sørensen TI. (2011) Studies based on the Danish Adoption Register: schizophrenia, BMI, smoking, and mortality in perspective. *Scand J Public Health* **39**:191-195.

Picciotto MR. (2003) Nicotine as a modulator of behavior: beyond the inverted U. *Trends Pharmacol Sc*i **24**:493-499.

Pierce RC, Kumaresan V. (2006) The mesolimbic dopamine system: the final common pathway for the reinforcing effect of drugs of abuse? *Neurosci Biobehav Rev* **30**:215-238.

Piper ME, McCarthy DE, Baker TB. (2006) Assessing tobacco dependence: a guide to measure evaluation and selection. *Nicotine Tob Res* **8**:339-351.

Piper ME, McCarthy DE, Bolt DM, Smith SS, Lerman C, Benowitz N, et al. (2008) Assessing dimensions of nicotine dependence: an evaluation of the Nicotine Dependence Syndrome Scale (NDSS) and the Wisconsin Inventory of Smoking Dependence Motives (WISDM). *Nicotine Tob Res* **10**:1009-1020.

Piper ME, Piasecki TM, Federman EB, Bolt DM, Smith SS, Fiore MC, et al. (2004) A multiple motives approach to tobacco dependence: the Wisconsin Inventory of Smoking Dependence Motives (WISDM-68). *J Consult Clin Psychol* **72**:139-154.

Pirinen M, Donnelly P, Spencer CC. (2013) Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics* **7**:369-390.

Planas A, Clara A, Marrugat J, Pou JM, Gasol A, de Moner A, et al. (2002) Age at onset of smoking is an independent risk factor in peripheral artery disease development. *J Vasc Surg* **35**:506-509.

Pogun S, Yararbas G. (2009) Sex differences in nicotine action. Handb Exp Pharmacol **192**:261-91.

Poliseno L. (2012) Pseudogenes: newly discovered players in human cancer. *Sci Signal* **5**:re5.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19**:1316-1323.

Pulst SM. (1999) Genetic linkage analysis. *Arch Neurol* **56**:667-672.

Raitakari OT, Juonala M, Ronnemaa T, Keltikangas-Jarvinen L, Rasanen L, Pietikainen M, et al. (2008) Cohort profile: the cardiovascular risk in Young Finns Study. *Int J Epidemiol* **37**:1220-1226.

Ramasamy A, Trabzuni D, Guelfi S, Varghese V, Smith C, Walker R, et al. (2014) Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat Neurosci* **17**:1418-1428.

Ray R, Tyndale RF, Lerman C. (2009) Nicotine dependence pharmacogenetics: role of genetic variation in nicotine-metabolizing enzymes. *J Neurogenet* **23**:252-261.

Reichardt LF. (2006) Neurotrophin-regulated signalling pathways. *Philos Trans R Soc Lond B Biol Sci* **361**:1545-1564.

Ribeiro EB, Bettiker RL, Bogdanov M, Wurtman RJ. (1993) Effects of systemic nicotine on serotonin release in rat brain. *Brain Res* **621**:311-318.

Rice JP, Hartz SM, Agrawal A, Almasy L, Bennett S, Breslau N, et al. (2012) CHRNB3 is more strongly associated with Fagerstrom test for cigarette dependence-based nicotine dependence than cigarettes per day: phenotype definition changes genome-wide association studies results. *Addiction* **107**:2019-2028.

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

Romanos M, Freitag C, Jacob C, Craig DW, Dempfle A, Nguyen TT, et al. (2008) Genome-wide linkage analysis of ADHD using high-density SNP arrays: novel loci at 5q13.1 and 14q12. *Mol Psychiatry* **13**:522-530.

Rose RJ, Broms U, Korhonen T, Dick D, Kaprio J. (2009) Genetics of smoking behavior. In Handbook of behavior genetics. Y. K. Kim (eds). New York: Springer. pp 411–432.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. (2002) Genetic structure of human populations. *Science* **298**:2381-2385.

Ross KC, Gubner NR, Tyndale RF, Hawk LW,Jr, Lerman C, George TP, et al. (2016) Racial differences in the relationship between rate of nicotine metabolism and nicotine intake from cigarette smoking. *Pharmacol Biochem Behav* **148**:1-7.

Ross S, Peselow E. (2009) The neurobiology of addictive disorders. *Clin Neuropharmacol* **32**:269-276.

Saccone NL, Culverhouse RC, Schwantes-An TH, Cannon DS, Chen X, Cichon S, et al. (2010) Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet* **6**:10.1371/journal.pgen.1001053.

Saccone NL, Saccone SF, Hinrichs AL, Stitzel JA, Duan W, Pergadia ML, et al. (2009) Multiple distinct risk loci for nicotine dependence identified by dense coverage of the complete family of nicotinic receptor subunit (CHRN) genes. *Am J Med Genet B Neuropsychiatr Genet* **150B:**453-466.

Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA, et al. (2007a) Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet* **16**:36-49.

Saccone SF, Pergadia ML, Loukola A, Broms U, Montgomery GW, Wang JC, et al. (2007b) Genetic linkage to chromosome 22q12 for a heavy-smoking quantitative trait in two independent samples. *Am J Hum Genet* **80**:856-866.

Sakai H, Jinawath A, Yamaoka S, Yuasa Y. (2005) Upregulation of MUC6 mucin gene expression by NFkappaB and Sp factors. *Biochem Biophys Res Commun* **333**:1254-1260.

Sajantila A, Salem AH, Savolainen P, Bauer K, Gierig C, Paabo S. (1996) Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc Natl Acad Sci U S A* **93**:12035-12039.

Schilstrom B, Fagerquist MV, Zhang X, Hertel P, Panagis G, Nomikos GG, et al. (2000) Putative role of presynaptic alpha7* nicotinic receptors in nicotine stimulated increases of extracellular levels of glutamate and aspartate in the ventral tegmental area. *Synapse* **38**:375-383.

Schlaepfer IR, Hoft NR, Collins AC, Corley RP, Hewitt JK, Hopfer CJ, et al. (2008) The CHRNA5/A3/B4 gene cluster variability as an important determinant of early alcohol and tobacco initiation in young adults. *Biol Psychiatry* **63**:1039-1046.

Schnoll RA, Johnson TA, Lerman C. (2007) Genetics and smoking behavior. *Curr Psychiatry Rep* **9**:349-357.

Schnoll RA, Patterson F. (2009) Sex heterogeneity in pharmacogenetic smoking cessation clinical trials. *Drug Alcohol Depend* **104** Suppl 1:S94-9.

Senore C, Battista RN, Shapiro SH, Segnan N, Ponti A, Rosso S, et al. (1998) Predictors of smoking cessation following physicians' counseling. *Prev Med* **27**:412-421.

Sham PC, Purcell SM. (2014) Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* **15**:335-346.

Sherva R, Kranzler HR, Yu Y, Logue MW, Poling J, Arias AJ, et al. (2010) Variation in nicotinic acetylcholine receptor genes is associated with multiple substance dependence phenotypes. *Neuropsychopharmacology* **35**:1921-1931.

Shiffman S, Waters A, Hickcox M. (2004) The nicotine dependence syndrome scale: a multidimensional measure of nicotine dependence. *Nicotine Tob Res* **6**:327-348.

Shmulewitz D, Meyers JL, Wall MM, Aharonovich E, Frisch A, Spivak B, et al. (2016) CHRNA5/A3/B4 Variant rs3743078 and Nicotine-Related Phenotypes: Indirect Effects Through Nicotine Craving. *J Stud Alcohol Drugs* **77**:227-237.

Sibley MH, Pelham WE, Molina BS, Coxe S, Kipp H, Gnagy EM, et al. (2014) The role of early childhood ADHD and subsequent CD in the initiation and escalation of adolescent cigarette, alcohol, and marijuana use. *J Abnorm Psychol* **123**:362-374.

Sibley MH, Pelham WE, Molina BS, Gnagy EM, Waxmonsky JG, Waschbusch DA, et al. (2012) When diagnosing ADHD in young adults emphasize informant reports, DSM items, and impairment. *J Consult Clin Psychol* **80**:1052-1061.

Sipila P, Rose RJ, Kaprio J. (2016) Drinking and mortality: long-term follow-up of drinking-discordant twin pairs. *Addiction* **111**:245-254.

Siqueira LM, Brook JS. (2003) Tobacco use as a predictor of illicit drug use and drug-related problems in Colombian youth. *J Adolesc Health* **32**:50-57.

Smith GD. (2010) Mendelian Randomization for Strengthening Causal Inference in Observational Studies: Application to Gene x Environment Interactions. *Perspect Psychol Sci* **5**:527-545.

Smythe E. (2002) Regulating the clathrin-coated vesicle cycle by AP2 subunit phosphorylation. *Trends Cell Biol* **12**:352-354.

Sneve M, Jorde R. (2008) Cross-sectional study on the relationship between body mass index and smoking, and longitudinal changes in body mass index in relation to change in smoking status: the Tromso Study. *Scand J Public Health* **36**:397-407.

Speed D, Hemani G, Johnson MR, Balding DJ. (2012) Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* **91**:1011-1021.

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42**:937-948.

Steinlein O, Weiland S, Stoodt J, Propping P. (1996) Exon-intron structure of the human neuronal nicotinic acetylcholine receptor alpha 4 subunit (CHRNA4). *Genomics* **32**:289-294.

Su SC, Chung WH, Hung SI. (2014) Digging up the human genome: current progress in deciphering adverse drug reactions. *Biomed Res Int* **2014**:824343.

Surakka I, Sarin AP, Ruotsalainen SE, Durbin R, Salomaa V, Daly MJ, et al. (2016) The rate of false polymorphisms introduced when imputing genotypes from global imputation panels. *bioRxiv* preprint; doi: http://dx.doi.org/10.1101/080770.

Swan GE, Benowitz NL, Jacob P,3rd, Lessov CN, Tyndale RF, Wilhelmsen K, et al. (2004) Pharmacogenetics of nicotine metabolism in twins: methods and procedures. *Twin Res* **7**:435-448.

Swan GE, Benowitz NL, Lessov CN, Jacob P,3rd, Tyndale RF, Wilhelmsen K. (2005) Nicotine metabolism: the impact of CYP2A6 on estimates of additive genetic influence. *Pharmacogenet Genomics* **15**:115-125.

Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. (2008) A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**:638-642.

Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, et al. (2010) Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* **42**:448-453.

Thorgeirsson TE, Steinberg S, Reginsson GW, Bjornsdottir G, Rafnar T, Jonsdottir I, et al. (2016) A rare missense mutation in CHRNA4 associates with smoking behavior and its consequences. *Mol Psychiatry* **21**:594-600.

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

Tobler PN, Fiorillo CD, Schultz W. (2005) Adaptive coding of reward value by dopamine neurons. *Science* **307**:1642-1645.

True WR, Xian H, Scherrer JF, Madden PA, Bucholz KK, Heath AC, et al. (1999) Common genetic vulnerability for nicotine and alcohol dependence in men. *Arch Gen Psychiatry* **56**:655-661.

Turner JR, Ray R, Lee B, Everett L, Xiang J, Jepson C, et al. (2014) Evidence from mouse and man for a role of neuregulin 3 in nicotine dependence. *Mol Psychiatry* **19**:801-810.

Tyndale RF. (2003) Genetics of alcohol and tobacco use in humans. *Ann Med* **35**:94-121.

Uhl GR, Drgon T, Johnson C, Walther D, David SP, Aveyard P, et al. (2010) Genome-wide association for smoking cessation success: participants in the Patch in Practice trial of nicotine replacement. *Pharmacogenomics* **11**:357-367.

Ulitsky I. (2016) Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet* **17**:601-614.

U.S. Department of Health and Human Services. The Health Consequences of Smoking – 50 Years of Progress. A Report of the Surgeon General. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, (2014). (https://www.surgeongeneral.gov/library/reports/50-years-of-progress/full-report.pdf).

Veenstra-VanderWeele J, Anderson GM, Cook EH,Jr. (2000) Pharmacogenetics and the serotonin system: initial studies and future directions. *Eur J Pharmacol* **410**:165-181.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. (2001) The sequence of the human genome. *Science* **291**:1304-1351.

Vink JM, Willemsen G, Boomsma DI. (2005) Heritability of smoking initiation and nicotine dependence. *Behav Genet* **35**:397-406.

Visscher PM, Hill WG, Wray NR. (2008) Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* **9**:255-266.

Visser SN, Danielson ML, Bitsko RH, Holbrook JR, Kogan MD, Ghandour RM, et al. (2014) Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United States, 2003-2011. *J Am Acad Child Adolesc Psychiatry* **53**:34-46.e2.

Vizi ES, Lendvai B. (1999) Modulatory role of presynaptic nicotinic receptors in synaptic and non-synaptic chemical communication in the central nervous system. *Brain Res Brain Res Rev* **30**:219-235.

Wang H, Sun X. (2005) Desensitized nicotinic receptors in brain. *Brain Res Brain Res Rev* **48**:420-437.

Wang J, Li MD. (2010) Common and unique biological pathways associated with smoking initiation/progression, nicotine dependence, and smoking cessation. *Neuropsychopharmacology* **35**:702-719.

Wang JC, Grucza R, Cruchaga C, Hinrichs AL, Bertelsen S, Budde JP, et al. (2009) Genetic variation in the CHRNA5 gene affects mRNA levels and is associated with risk for alcohol dependence. *Mol Psychiatry* **14**:501-510.

Ware JJ, Chen X, Vink J, Loukola A, Minica C, Pool R, et al. (2016) Genome-Wide Meta-Analysis of Cotinine Levels in Cigarette Smokers Identifies Locus at 4q13.2. *Sci Rep* **6**:20092.

Warren CW, Jones NR, Eriksen MP, Asma S, Global Tobacco Surveillance System (GTSS) collaborative group. (2006) Patterns of global tobacco use in young people and implications for future chronic disease burden in adults. *Lancet* **367**:749-753.

Watson JD, Crick FH. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**:737-738.

Wen L, Yang Z, Cui W, Li MD. (2016) Crucial roles of the CHRNB3-CHRNA6 gene cluster on chromosome 8 in nicotine dependence: update and subjects for future research. *Transl Psychiatry* **6**:e843.

West R, Gossop M. (1994) Overview: a comparison of withdrawal symptoms from different drug classes. *Addiction* **89**:1483-1489.

White W, White IM. (2016) Amphetamine and morphine may produce acute-withdrawal related hypoactivity by initially activating a common dopamine pathway. *Physiol Behav* **165**:187-194.

Whittmore AS, Halpern J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics* **50**:118–127.

WHO global report on trends in prevalence of tobacco smoking. Geneva: World Health Organization (2015). (http://apps.who.int/iris/bitstream/10665/156262/1/9789241564922_eng.pdf).

Williams JM, Ziedonis D. (2004) Addressing tobacco among individuals with a mental illness or an addiction. *Addict Behav* **29**:1067-1083.

Wise RA. (2004) Dopamine, learning and motivation. *Nat Rev Neurosci* **5**:483-494.

Wonnacott S. (1997) Presynaptic nicotinic ACh receptors. *Trends Neurosci* **20**:92-98.

Xian H, Scherrer JF, Eisen SA, Lyons MJ, Tsuang M, True WR, et al. (2007) Nicotine dependence subtypes: association with smoking history, diagnostic criteria and psychiatric disorders in 5440 regular smokers from the Vietnam Era Twin Registry. *Addict Behav* **32**:137-147.

Xian H, Scherrer JF, Madden PA, Lyons MJ, Tsuang M, True WR, et al. (2003) The heritability of failed smoking cessation and nicotine withdrawal in twins who smoked and attempted to quit. *Nicotine Tob Res* **5**:245-254.

Xie P, Kranzler HR, Krauthammer M, Cosgrove KP, Oslin D, Anton RF, et al. (2011) Rare nonsynonymous variants in alpha-4 nicotinic acetylcholine receptor gene protect against nicotine dependence. *Biol Psychiatry* **70**:528-536.

Xie P, Kranzler HR, Zhang H, Oslin D, Anton RF, Farrer LA, et al. (2012) Childhood adversity increases risk for nicotine dependence and interacts with alpha5 nicotinic acetylcholine receptor genotype specifically in males. *Neuropsychopharmacology* **37**:669-676.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**:565-569.

Zaitlen N, Kraft P, Patterson N, Pasaniuc B, Bhatia G, Pollack S, et al. (2013) Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet* **9**:e1003520.

Zhang Y, Florath I, Saum KU, Brenner H. (2016) Self-reported smoking, serum cotinine, and blood DNA methylation. *Environ Res* **146**:395-403.

Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, et al. (2017) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**:272-279.

Zhou X, Stephens M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**:821-824.