Maiken Elvestad Gabrielsen

Genetic Risk Factors for Lung Cancer: Relationship to Smoking Habits and Nicotine Addiction

The Nord-Trøndelag (HUNT) and Tromsø Health Studies

Thesis for the degree of Philosophiae Doctor

Trondheim, March 2013

Norwegian University of Science and Technology
Faculty of Medicine
Department of Cancer Research and Molecular Medicine

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Medicine
Department of Cancer Research and Molecular Medicine

ISBN 978-82-471-4274-5 (printed ver.)
ISBN 978-82-471-4275-2 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2013:87

Printed by NTNU-trykk

NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET

DET MEDISINSKE FAKULTET

Maiken Elvestad Gabrielsen

Genetiske risikofaktorer for lunge kreft: sammenheng med røykevaner og nikotin avhengighet; en studie basert på helseundersøkelsene i Nord-Trøndelag (HUNT) og Tromsø

Lungekreft en den kreftformen som tar flest liv årlig på verdensbasis og hvert år dør omkring 1,1 millioner mennesker av sykdommen. Det er allment kjent at tobakksrøyking er den viktigste årsaken til lungekreft. Fra å være en relativt sjelden sykdom rundt begynnelsen på 1900-tallet har antallet tilfeller økt jevnt i takt med tobakksforbruket. Tall fra kreftregisteret viser at det i Norge tilkommer rundt 2600 nye tilfeller av lungekreft og 2000 dødsfall som følge av lungekreft hvert år. Også andre lunge-sykdommer er forårsaket av tobakksrøyking. Kronisk obstruktiv lungesykdom (KOLS) har en klar sammenheng med røyking, og er en progressiv kronisk betennelse i lungevev som resulterer i en gradvis irreversibel reduksjon av lungekapasiteten. I tillegg til tobakksrøyking øker risikoen for både lungekreft og KOLS ved andre miljømessige eksponeringer. Epidemiologiske studier viser også en økt risiko for lungekreft og KOLS relatert til variasjoner i arvematerialet, DNA.

Genetisk variasjon er et begrep som benyttes for å beskrive forskjeller i DNA mellom ulike individ. Selv om to ubeslektede individ deler omtrent 99,9 % av arvematerialet, så utgjør forskjellene ca. 3 millioner ulikheter på nukleotidnivå bare pga. den enorme størrelsen på genomet. Den vanligste formen for genetisk variasjon kalles singel-nukleotid-polymorfisme (SNP, uttales *snipp*). Dette er i realiteten en «staveforskjell» i DNA'et hvor man i samme posisjon har to alternative skrivemåter. Dette kan medføre en endring i betydningen av «ordet» (dvs. endring i funksjon), men trenger ikke alltid gjøre det. Forekomsten av de to alternative variantene kan variere mellom ulike populasjoner, og i enkelte tilfeller er det forbundet en økt eller redusert sykdomsrisiko med den ene varianten.

Etter at sekvensen av det humane genomet ble ferdigstilt i 2001/2003 har man sett en enorm økning i antall studier som undersøker betydningen av naturlig forekommende genetisk variasjon og sammenhengen med risiko for en rekke vanlige sykdommer og egenskaper. Den teknologiske utviklingen har gjort at det nå er mulig å studere et stort antall (hundretusener til flere millioner) SNP'er per individ. Ved å sammenligne en gruppe syke personer med en frisk kontrollgruppe kan man undersøke hvorvidt noen av disse variantene opptrer oftere i sykdomsgruppen enn i kontrollgruppen.

I denne studien ble sammenhengen mellom vanlig forekommende SNP'er og risiko for lungekreft, KOLS og nikotinavhengighet undersøkt. Det ble benyttet DNA og data fra Helseundersøkelsen i Nord Trøndelag (HUNT) og Tromsøundersøkelsen. Gjennom deltagelse i en stor internasjonal helgenomsstudie klarte man å identifisere to kromosomale regioner assosiert med økt risiko for lungekreft. Den ene av disse regionene, på kromosom 15q25, ligger i et område hvor man finner gener som koder for subenheter av nikotin acetylcholine reseptor (nAChR). Disse genene har over lengere tid vært studert i forhold til nikotinavhengighet da

nAChR er del av systemet for frigjøring av dopamin. Vår oppfølgingsstudie for en av de relevante variantene (rs16969968) basert på hele HUNT 2 populasjonen konkluderer i denne avhandlingen med at i sær denne varianten gir økt risiko for nikotinavhengighet og dermed en indirekte effekt på både lungekreft og KOLS. Dette kommer tydelig fram da varianten også er assosiert med snusforbruk.

Det faktum at frekvensen av ulike genetiske varianter varierer mellom populasjoner har ført til en utvikling av studier som fokuserer på genetiske populasjonsstrukturer. Dette er viktig da forskjeller i genetisk variasjon mellom populasjoner kan resultere i utilsiktede misvisninger (bias) i helgenomsstudier. I denne studien ble forskjeller i genetisk variasjon mellom de to helseundersøkelsene HUNT og Tromsø kartlagt. Det ble funnet betydelige forskjeller i genetisk variasjon mellom disse to regionene, og at disse forskjellene vil kunne føre til bias i helgenomsstudier dersom utvalget i sykdomsgruppe og i kontrollgruppe ikke er balansert mellom regionene. I tillegg ble det funnet klare forskjeller i genetisk variasjon innad i HUNT-gruppen. Arbeidet i denne avhandlingen utgjør en pilotstudie for videre undersøkelse av den genetiske variasjonen i Norge og danner basis for en grundig kartlegging av genetiske strukturer innad og mellom norske helseundersøkelser for framtidige genetiske studer.

Kandidat: Maiken Elvestad Gabrielsen
Institutt: Institutt for Kreftforskning og Molekylær Medisin
Veiledere: Professor Hans E. Krokan og Professor Frank Skorpen
Finansieringskilde: Stipend fra NTNU
Finansiell støtte: Svanhild og Arne Must's fond for medisinsk forskning og Den Norske Kreftforening

Overnevnte avhandling er funnet verdig til å forsvares offentlig for graden PhD i Molekylæmedisin.
Disputas finner sted i Auditoriet MTA, Medisinsk teknisk forskningssenter, Torsdag 21. Mars 2013 kl. 12:15.

ACKNOWLEDGEMENTS

The work presented in this thesis has been carried out at the Department of Cancer Research and Molecular Medicine, Faculty of Medicine at NTNU. I am grateful for the fellowship granted by NTNU which has allowed me to explore the world of SNPs and GWAS.

I've wanted to make this "My thesis". To tell my own story about the journey I've had through the years involved in this work. I honestly believe I had no idea what I was getting myself into when Hans and Frank presented me with this really interesting study on SNPs and lung cancer using the HUNT population. I was thrilled to be offered such an exciting and at the time forward looking project. Throughout these years I´ve had the opportunity to work independently and develop my skills as a researcher while still being under your guidance. I would like to take this opportunity to thank you both for the opportunities you have given me. Frank, you truly deserve a special acknowledgement in this thesis. I am deeply grateful for the help you have given me. Your door is always open and you are ready to answer any question.

I am very grateful for the contribution from co-authors. I appreciate all the help and good ideas from Arnulf Langhammer and Pål Romundstad in the jungle of statistics, smoking and lung diseases. A great thank you to Einar Ryeng and Arnar Flatberg for always trying to help me through me feeble attempts at understanding R and making some sense out of our population structures data. Oddgeir L. Holmen, your ideas and visions have taken our "odd one out project" to new heights. In this last year you have given me many interesting discussions and opened my eyes to a bigger picture. I am also grateful for the opportunity to collaborate with the genetic epidemiology group at IARC. This has given us the opportunity to be part of a large international collaboration contributing to the epidemiology of lung cancer.

Dear colleagues at IKM you make my working days joyful. Whether discussing challenges related to our corridor's ever-growing number of children or scientific issues, questions or frustrations over a coffee you all contribute to a nice working environment. A special thank you to Berit and Linda in the office, you always try to answer any question. Linda, I am forever grateful for all your help and support. I'm sure you'll make an excellent supervisor in the years to come. Mona, your words of wisdom have helped med through the days of writing. Siv Anita, a comrade in writing, it is a great comfort to have someone in the same situation.

I acknowledge all the participants in the HUNT and Tromsø health studies for their contribution to enabling this research.

Doing a PhD is not just a job, it is part of you. I could not have done this without the support from family and friends. Whatever I do, I appreciate your constant support and I know you are always there. My family and in-laws have stepped up the baby-sitting allowing me to keep a steady focus for which I am truly grateful. Martine you deserve a special thank you. Coming home to your dinners and a clean house has made my everyday a great deal easier, thank you for a wonderful job.

To my beloved parents and sister, thank you for your constant love and support. You never really understand what I do at work on but you always try, and keep asking questions until you think you do. Mum and Dad, your support especially in the last few months have meant the world to me.

Finally, my dearest Christian, your love and support means everything to be. Sitting side by side at the dinner table late at night working together has helped me through to the finishing line and your weekend getaways with the kids have given me the extra time to complete the writing. I could not have done this without you.

En spesiell takk til min to fantastiske barn, Ingrid og Astrid, som sørger for at man alltid har føttene godt plantet i den virkelige verden og som har hjulpet meg med å holde målet friskt i minne, middag i Tyholttårnet etter innlevering av oppgaven.

Trondheim, October 2012

TABLE OF CONTENTS

LIST OF PAPERS

PAPER I

Hung, R. J. McKay, J. D. Gaborieau, V. Boffetta, P. Hashibe, M. Zaridze, D. Mukeria, A. Szeszenia-Dabrowska, N. Lissowska, J. Rudnai, P. Fabianova, E. Mates, D. Bencko, V. Foretova, L. Janout, V. Chen, C. Goodman, G. Field, J. K. Liloglou, T. Xinarianos, G. Cassidy, A. McLaughlin, J. Liu, G. Narod, S. Krokan, H.E. Skorpen, F. Elvestad, M.B. Hveem, K. Vatten, L. Linseisen, J. Clavel-Chapelon, F. Vineis, P. Bueno-de-Mesquita, H.B. Lund, E. Martinez, C. Bingham, S. Rasmuson, T. Hainaut, P. Riboli, E. Ahrens, W. Benhamou, S. Lagiou, P. Trichopoulos, D. Holcatova, I. Merletti, F. Kjaerheim, K. Agudo, A. Macfarlane, G. Talamini, R. Simonato, L. Lowry, R. Conway, D.I. Znaor, A. Healy, C. Zelenika, D. Boland, A. Delepine, M. Foglio, M. Lechner, D. Matsuda, F. Blanche, H. Gut, I. Heath, S. Lathrop, M. Brennan, P (2008). **A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25.** Nature, Apr 3; 452(7187): 633-7

PAPER II

J. D. McKay, R. J. Hung, V. Gaborieau, P. Boffetta, A. Chabrier, G. Byrnes, D. Zaridze, A. Mukeria, N. Szeszenia-Dabrowska, J. Lissowska, P. Rudnai, E. Fabianova, D. Mates, V. Bencko, L. Foretova, V. Janout, J. McLaughlin, F. Shepherd, A. Montpetit, S. Narod, H. E. Krokan, F. Skorpen, M. B. Elvestad, L. Vatten, I. Njolstad, T. Axelsson, C. Chen, G. Goodman, M. Barnett, M. M. Loomis, J. Lubinski, J. Matyjasik, M. Lener, D. Oszutowska, J. Field, T. Liloglou, G. Xinarianos, A. Cassidy, P. Vineis, F. Clavel-Chapelon, D. Palli, R. Tumino, V. Krogh, S. Panico, C. A. Gonzalez, J. Ramon Quiros, C. Martinez, C. Navarro, E. Ardanaz, N. Larranaga, K. T. Kham, T. Key, H. B. Bueno-de-Mesquita, P. H. Peeters, A. Trichopoulou, J. Linseisen, H. Boeing, G. Hallmans, K. Overvad, A. Tjonneland, M. Kumle, E. Riboli, D. Zelenika, A. Boland, M. Delepine, M. Foglio, D. Lechner, F. Matsuda, H. Blanche, I. Gut, S. Heath, M. Lathrop and P. Brennan (2008). **Lung cancer susceptibility locus at 5p15.33.** Nat Genet 40(12): 1404-1406.

PAPER III

Maiken E. Gabrielsen, Pål Romundstad, Arnulf Langhammer, Hans E. Krokan, Frank Skorpen (2012) **Association between 15q25 gene variants, nicotine related habits, lung cancer and COPD in the HUNT study, Norway.** :Manuscript submitted to Eur. J. Hum. Genet

PAPER IV

Maiken E. Gabrielsen, Oddgeir Lingaas Holmen, Arnar Flatberg, Einar Ryeng, Kristian Hveem, Frank Skorpen, Hans E. Krokan (2012)
The genetic structures of stable populations – the HUNT and Tromsø cohorts in Norway. Manuscript

OTHER WORK, NOT INCLUDED IN THIS THESIS:

1. E. H. Lips, V. Gaborieau, J. D. McKay, A. Chabrier, R. J. Hung, P. Boffetta, M. Hashibe, D. Zaridze, N. Szeszenia-Dabrowska, J. Lissowska, P. Rudnai, E. Fabianova, D. Mates, V. Bencko, L. Foretova, V. Janout, J. K. Field, T. Liloglou, G. Xinarianos, J. McLaughlin, G. Liu, F. Skorpen, M. B. Elvestad, K. Hveem, L. Vatten, E. Study, S. Benhamou, P. Lagiou, I. Holcatova, F. Merletti, K. Kjaerheim, A. Agudo, X. Castellsague, T. V. Macfarlane, L. Barzan, C. Canova, R. Lowry, D. I. Conway, A. Znaor, C. Healy, M. P. Curado, S. Koifman, J. Eluf-Neto, E. Matos, A. Menezes, L. Fernandez, A. Metspalu, S. Heath, M. Lathrop and P. Brennan (2010). **"Association between a 15q25 gene variant, smoking quantity and tobacco-related cancers among 17 000 individuals."** Int J Epidemiol 39(2): 563-577.

2. S. C. Heath, I. G. Gut, P. Brennan, J. D. McKay, V. Bencko, E. Fabianova, L. Foretova, M. Georges, V. Janout, M. Kabesch, H. E. Krokan, M. B. Elvestad, J. Lissowska, D. Mates, P. Rudnai, F. Skorpen, S. Schreiber, J. M. Soria, A. C. Syvanen, P. Meneton, S. Hercberg, P. Galan, N. Szeszenia-Dabrowska, D. Zaridze, E. Genin, L. R. Cardon and M. Lathrop, (2008). **"Investigation of the fine structure of European populations with applications to disease association studies."** Eur J Hum Genet 16(12): 1413-1429.

3. R. Kazma, M. C. Babron, V. Gaborieau, E. Genin, P. Brennan, R. J. Hung, J. R. McLaughlin, H. E. Krokan, M. B. Elvestad, F. Skorpen, E. Anderssen, T. Vooder, K. Valk, A. Metspalu, J. K. Field, M. Lathrop, A. Sarasin and S. Benhamou (2012). **"Lung Cancer and DNA Repair Genes: Multilevel Association Analysis from the International Lung Cancer Consortium."** Carcinogenesis **33**(5): 1059-1064

4. Timofeeva MN, Hung RJ, Rafnar T, Christiani DC, Field JK, Bickeboller H, Risch A, McKay JD, Wang Y, Dai J, Gaborieau V, McLaughlin J, Brenner D, Narod S, Caporaso NE, Albanes D, Thun M, Eisen T, Wichmann HE, Rosenberger A, Han Y, Chen W, Zhu D, Spitz M, Wu X, Pande M, Zhao Y, Zaridze D, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Krokan HE, Gabrielsen ME, Skorpen F, Vatten L, Njølstad I, Chen C, Goodman G, Lathrop M, Benhamou S, Vooder T, Välk K, Nelis M, Metspalu A, Raji O, Chen Y, Gosney J, Liloglou T, Muley T, Dienemann H, Thorleifsson G, Shen H, Stefansson K, Brennan P, Amos CI, Houlston R, Landi MT; for TRICL Research Team. (2012). **"Influence of Common Genetic Variation on Lung Cancer Risk: Meta-Analysis of 14,900 Cases and 29,485 Controls."** Hum Mol Genet. [Epub ahead of print]

ABBREVIATIONS

α_1 AT	α_1 Antitrypsin
AKT	v-akt murine thymoma viral oncogene homolog 1
AMD	Age-related macular degeneration
CDCV	Common disease – Common variant
CDKN2A (p16(INK4))	Cyclin-dependent kinase inhibitor 2A
CDRV	Common disease – Rare variant
CEU	Utah residents with Northern and Western European ancestry from the CEPH (Centre d'Etude du Polymorphisme Humain) collection
CHRNA3	Cholinergic receptor, nicotinic, alpha 3 (neuronal)
CHRNA5	Cholinergic receptor, nicotinic, alpha 5 (neuronal)
CHRNB4	Cholinergic receptor, nicotinic, beta 4 (neuronal)
CLPTM1L	Cisplatin resistance-related protein 9/ Cleft lip and palate transmembrane protein 1-like protein
CNG	Centre National de Génotypage
CNV	Copy number variant
COPD	Chronic obstructive pulmonary disease
CPD	Cigarettes per day
DNA	Deoxyribonucleic acid
EGFR	Epidermal growth factor receptor
FEV_1	Forced expiratory volume at 1 s
FHIT	Fragile histidine triad
FVC	Forced vital capacity
FTND	Fagerström Test for Nicotine Dependence
GOLD	Global Initiative on Obstructive Lung Disease
GWA	Genome-wide associaton
GWAS	Genome-wide association study
HGP	Human Genome Project
HHIP	Hedgehog interacting protein
HUNT	The Nord-Trøndelag Health study
HR	Hazard ratio
HWE	Hardy-Weinberg equilibrium
IARC	International agency for research on cancer
IBD	Identity by descent
IBS	Identity by state
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
LD	Linkage disequilibrium
LDU	Linkage disequilibrium unit
LOH	Loss of heterozygosity
MAF	Minor allele frequency
MAPK	Mitogen-activated protein kinase
MDS	Multiple dimensional scaling
mRNA	Messenger ribonucleic acid
NAcc	Nucleus accumbens
nAChr	Nicotinic acetylcholine receptor

ND	Nicotine dependence
NF-κB	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1
NHGRI	National Human Genome Research Institute
NNN	N-nitrosonornicotine
NNK	4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone
NPR	Norwegian patient register
NSCLC	Non-small cell lung carcinoma
OR	Odds Ratio
PAH	Polycyclic aromatic hydrocarbon
PCA	Principal component analysis
PCR	Polymerase chain reaction
PI3K	Phosphatidylinositol-4,5-bisphosphate 3-kinase
QC	Quality control
RNA	Ribonucleic acid
ROH	Runs of homozygosity
SCLC	Small cell lung carcinoma
SERPINA1	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1
SNP	Single nucleotide polymorphism
TERT	Telomerase reverse transcriptase
TNF-α	Tumour necrosis factor
TP53	Tumour protein p53
VNTR	Variable number tandem repeat
WISDM	Wisconsin Inventory of Smoking Dependence Motives
WTCCC	Welcome Trust Case Control Consortium

GENETIC TERMS GLOSSARY

Allele	Alternate forms of a gene or a specific variant/base at a particular locus in the genome that differ in DNA sequence.
Association analysis	Analysis of the relationship between a phenotype and a genotype. The genotype and phenotype is said to be associated if the genotype-phenotype combination occurs more frequently than would be expected from their separate frequencies.
Candidate gene	A gene believed to be involved in a complex trait or disease based on known biological and/or physiological properties of its products, or its location near a region of association or linkage.
Complex traits	A trait that is influenced by multiple genes, environmental factors and the interaction between them
Copy number variant (CNV)	A form of structural variation of the DNA where stretches of genomic sequence (1kb-3Mb in size) are deleted or duplicated in varying numbers.
Deoxyribonucleic acid (DNA)	A double helix molecule consisting of 4 bases; Adenine (A), Thymine (T), Guanine (G) and Cytosine (C), together, forming the molecular basis of the genome.
Gene	Traditionally, the basic physical unit of heredity; a sequence of DNA that gives the coding instructions for the synthesis of RNA. The human genome contains approximately 25,000 genes distributed on 23 pairs of chromosomes. New research from the ENCODE project show that about 75% of the genome is transcribed at some point in some cells, and that genes are highly interlaced with overlapping transcripts that are synthesized from both DNA strands [1]
Genetic code	The set of rules by which information encoded in genetic material (DNA or mRNA sequences) is translated into amino acid sequences. A specific sequence of three nucleotides, a codon, determines the amino acid.
Genetic variation	Variation in alleles of genes, both within and among populations. Provides the “raw material” for natural selection.
Genome	The total of an individual organism’s entire genetic material.
Genome-wide association studies	The study of genetic variation across the entire genome aimed at identifying genetic variation associated with a complex disease or trait.
Genomics	Genomics is a discipline in genetics concerning the study of the genomes of organisms. Traditionally genomics concerns everything that has to do with DNA. A broader definition is used by the United States Environmental Protection Agency, to also include mRNA and proteins.

Genotype	The combination of alleles on corresponding loci in the two copies of the chromosomes. When two sequence alternatives exist at a given locus, e.g. A and G 3 different genotypes are possible, AA and GG when the allele is identical on each chromosome and AG when the allele differs.
Haplotype	A combination of alleles at adjacent loci on the chromosome that are transmitted together.
HapMap	A genome-wide database of patterns of common human genetic sequence variation among multiple ancestral population samples.
Hardy-Weinberg equilibrium	The population distribution of 2 alleles (with frequencies p and q) such that the distribution is stable from generation to generation. Genotypes occur at frequencies of p^2, 2pq and q^2 for the major allele homozygote, heterozygote and minor allele homozygote.
Heritability	The proportion of observable differences between individuals that is due to genetic differences.
Linkage disequilibrium (LD)	The non-random association of allele at two or more loci. Occurs when two or more loci on a chromosome have reduced recombination between them because of their physical proximity to each other. LD describes the extent to which a variant at one locus predicts the variant at another locus.
Locus	Any given specific site in a genome. Often used to describe a particular site where sequence or functional alternatives exist.
Mendelian disease	Disease or trait caused by a single major gene with an inheritance pattern such that the disease is only manifested in 1 (recessive) or 2 (dominant) of the 3 possible genotype groups.
Minor allele	The allele with the lowest frequency of a biallelic polymorphisms.
Minor allele frequency	The frequency of the least common of 2 alleles in a population.
Mutation	A change in the genomic sequence of DNA as a result of DNA damage, replication error, incomplete repair or other intrinsic events.
Phenotype	A phenotype is the composite of an organism´s observable characteristics or traits and result from the expression of the organism's genes as well as the influence of environmental factors and the interactions between the two.
Single Nucleotide Polymorphism (SNP)	A type of genetic variation where, at a specific locus in the genome two sequence alternatives exists and where the least common alternative is found in minimum 1% of the population in question.
Tag-SNP	A SNP measured in a genotyping array in strong LD with multiple other SNP. Serves as a proxy for these SNPs on large scale genotyping platforms.

"The capacity to blunder slightly is the real marvel of DNA. Without this special attribute, we would still be anaerobic bacteria and there would be no music"

Lewis Thomas (1913-1993)

1 INTRODUCTION

The last decade has seen an enormous upsurge in large scale genetic analyses of a wide range of phenotypes. Scientific and technological advances and reduction in prices have opened the doors to a new dimension of molecular epidemiology; genome-wide association studies (GWASs).

Lung cancer is a complex and heterogeneous disease dependent on many genes, environmental- and lifestyle factors. It is also the number one killer of all cancers with approximately 1.1 million deaths per year worldwide [2]. The work described in this thesis includes participation in an international GWAS that aimed to uncover genetic predispositions for lung cancer and a follow-up study in a large homogenous cohort, the Nord-Trøndelag Health study (HUNT). In the follow-up the phenotypic outcomes were extended to include chronic obstructive pulmonary disease (COPD), smoking habits, and the use of smokeless tobacco (snus). Lastly we have utilised genome-wide SNP-data to uncover population structures of particular interest for future large scale genomic studies using Norwegian samples.

The work involved in this thesis has taken part in the GWAS revolution, surfed on its waves of enthusiasm and humbly accepted its limitations.

1.1 The Book of life

The study of genes

The general concept of a unit of inheritance was first coined by Gregor Medel in 1865. His experiments with *Pisum sativum* [3] was the beginning of the understanding of heredity. Since then, several historic events have shaped genetic research into the highly advanced science we know today. One of the most fundamentally important of these events, and what has been called the dawn of the molecular revolution, is the discovery of the molecular structure of DNA by Watson and Crick in 1953 [4, 5]. The discovery was based on the X-ray diffraction image, referred to as “Photo 51”, by Rosalind Franklin and Raymond Gosling [6] and solved one of the great mysteries of biology, how information is passed on from one generation to the next [7]

"*It struck us with a tremendous impact just how beautiful and exciting it was, because there before us was the answer to one of the fundamental problems in biology; how do genes replicate? And it was very simple and you couldn't miss it."*[1].

Seventy-five years before Watson and Crick uncovered the molecular structure of DNA, a Swiss medical doctor by the name of Fredrich Miescher discovered what he then named *nuclein* [8, 9]. It may seem however, that Miescher was ahead of his time. It was not until the 1940´s and 50´s when DNA was suggested as the hereditary material of bacteria [10] [11], together with Watson and Crick's discoveries of the DNA structure, that molecular biology gathered serious headway. It sparked a mad rush to understand the complex functions of this relatively simple molecule. In the 1960`s several researchers worked to unravel the genetic code (reviewed in [12]) and in 1968 Khorana, Nirenberg and Holley received the Nobel Prize in Physiology or Medicine for their work showing how the specific sequence of three nucleotides codes for different amino acids ("The Nobel Prize in Physiology or Medicine 1968". Nobelprize.org. 25 Oct 2012 http://www.nobelprize.org/nobel_prizes/medicine/laureates/1968/ accessed 30.10.2012). The early perception of the DNA molecule was of a highly stable molecule [13]. According to Errol C. Friedberg this delayed efforts into the understanding of mutations and repair [13]. Even Francis Crick admitted to missing the role of DNA repair "*We totally missed the possible role of enzymes in repair*" [14].

Eventually, the scientific advances finally culminated in the jewel of crown in modern molecular biology, the complete sequence of the human genome. The Human Genome Project (HGP) started in 1990, though the idea was conceived already in the 1980's. It aimed to identify all protein coding genes and determine the sequence of the approximate 3 billion base pairs in the human DNA. A draft sequence was published in 2001 [15, 16] and a more complete sequence in 2003 [17]. This was the result of a race between two groups, one public, The Human Genome Sequencing Consortium and one private, the Celera group. A statement from The White House so eloquently expresses the hope that this achievement would "lead to a new era of molecular medicine, an era that will bring new ways to prevent, diagnose, treat and cure disease." [18]. The Human genome project gave unfathomed, and at the time surprising, knowledge into composition of the human genome [19]. The book of life had finally been unravelled.

[1] Francis Crick in a television interview, (http://www.youtube.com/watch?v=UxJ-NrHw2B4&feature=related)

"The most surprising discovery about the human genome was that the majority of the functional sequence does not encode protein."[2]
Eric S. Lander 2011.

1.2 Genomics

"The genome revolution is only just beginning" [3]
Craig Venter 2010

The completion of a reference sequence for the human genome opened the doors to large scale genomic research. One of the hallmarks of genomics is "comprehensiveness", meaning genomics is concerned with creating large scale, complete data sets [20]. Genomics is also driven by the development of new technology. The gathering and analysis of large scale data sets require a reduction in costs and increase in data storage and analysis capabilities. It is an area of science developing at an enormous speed. Since the year 2000, more than 3,800 organisms have had their genomes sequenced (Figure 1). Craig Venter said in an Opinion for the ten year anniversary of the human genome: "*Nearly ten years after Francis Collins and I stood at the White House with President Bill Clinton to announce the first two drafts of the human genome, the technology for DNA sequencing has progressed more dramatically than any of us could have predicted.*" [21].

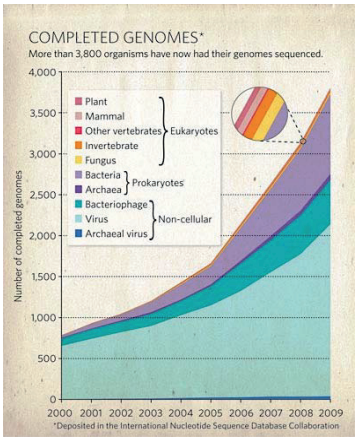


Figure 1. The number of completed genomes from the year 2000-2009 registered in the International Nucleotide Sequence Database Collection. Reprinted by permission from Macmillan Publishers Ltd: Nature [21], © (2010)

[2] Lander 2011, "The initial impact of the sequencing of the human genome", Nature; 470:187-197
[3] Craig Venter 2010, "Multiple personal genomes await", Nature; 464: 676-677

1.3 Genetic variation

"The more biologists look, the more complexity there seems to be" [4]
Erika Check Hayden 2010

Two unrelated individual share on average 99.9% of their genome at the nucleotide level. The sheer size of the human genome means that this amounts to approximately 3 million single nucleotide differences between two genomes. Variations in DNA can arise from a number of sources. Most common variants are old, and ancient polymorphisms account for about 90% of our variation (reviewed in [22]). It is likely that these variations developed parallel to the evolution of our species and have followed the first people out of Africa [23, 24]. Based on research on the Y-chromosome, the mutation rate in germ line cells is approximately $3.0x10^{-8}$ mutations/nucleotide/generation, meaning that 100-200 new mutations are accumulated in the entire genome from generation to generation [25]. Another study [26] found that approximately 175 new alleles arise per generation. Mutations can arise as a result of a number of processes, such as replication errors, DNA damage and erroneous bypass of the lesion, or incomplete and incorrect DNA repair. A large range of mechanisms has evolved to keep the mutation rate at a minimum and multiple highly efficient DNA repair pathways, including nucleotide excision repair, base excision repair, mismatch repair and recombinational repair, act to correct damage to the DNA molecules (reviewed in [27]. DNA damage escaping repair may give rise to mutations, which may then be passed on to the next generation. Such mutations are left in the hands of evolution in the form of natural selection and random genetic drift, which determines their frequency in the population [28]. If the frequency of a mutation is found in >1% of the chromosomes in the population, it has traditionally been referred to as a polymorphism [29].

The identification of the ABO blood groups in 1919 by Hirszfeld and Hirszfeld [30] was the first demonstration of molecular genetic variation in humans. Since then a wealth of different genetic variations has been described. They can be divided into two main categories, single nucleotide variants and structural variants (Figure 2) [31, 32]. Single nucleotide polymorphisms (SNPs) are the most studied genetic variation and the focus of this thesis. It was known already

[4] Hayden 2010, "Life is Complicated", Nature; 464:664-667

in the early 1980s that heterozygous sites were found approximately every 1,300 bases (reviewed in [19]. They will be described in more detail in the following chapter.

Structural variants embrace a range of genetic variations that are not single nucleotide variants (Figure 2). These include copy number variants (CNVs), insertion-deletion variants, block substitutions and inversion variants (reviewed in [32]). Studies by Kidd *et al.* [33] suggested that structural variants account for at least 20% of all genetic variation and, because of their size, approximately 70% of all variant bases. In 2006 Redon *et al.* [34] published a map of CNVs in the human genome, describing a considerable source of genetic variation affecting the risk of complex diseases [35, 36]. A CNV is a segment of DNA, 1 kb or larger which is present at variable copy numbers [34]. Structural variants have been linked to a number of diseases such as schizophrenia [37, 38], autism [39, 40] and Crohn´s disease [41]. It has also been shown that not only specific variants, but also the total load of structural variants in a person's genome could influence the risk of schizophrenia [37, 42].

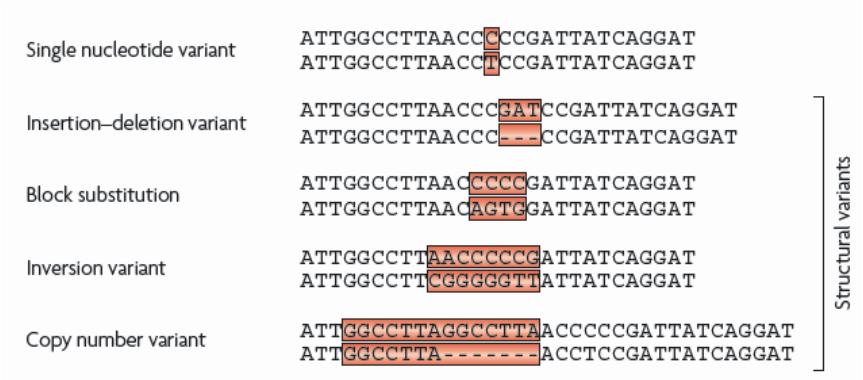


Figure 2. Genetic variations found in the human genome. Single nucleotide variants are single base-pair changes found at regular intervals in the sequence. Insertion–deletion variants are one or more base-pairs which are present or absent in one genome and not the other, described in Levy *et al.* 2007[43]. Block substitutions occurs when a set of adjacent nucleotides are substituted (from one individual to the other). Inversion variants describe the case where a DNA sequence is inverted, that is the base-pairs are reversed in a defined section. A copy number variant is a stretch of genomic sequence (1kb-3Mb in size) that is deleted or duplicated in varying numbers between individuals. [34] Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics [32] © (2009).

1.3.1 Single Nucleotide Polymorphisms

SNPs are the most common form of genetic variation (Figure 3). Approximately 38 million SNPs are currently (based on build 173, June 26th 2012) known and validated in the human genome (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi accessed 30.10.2012). They have varying effect depending on location (regulatory, coding or non-coding region) and the type of SNP. A non-synonymous SNP (also called missense) changes the codon in such a way that a different amino acid is inserted, while a synonymous SNP leaves the amino acid sequence unchanged. A type of SNP with a much larger potential to affect the phenotype, but found at a lower frequency than non-synonymous SNPs, is a nonsense SNP [44]. This SNP introduces a premature stop-codon resulting in a truncated gene product.

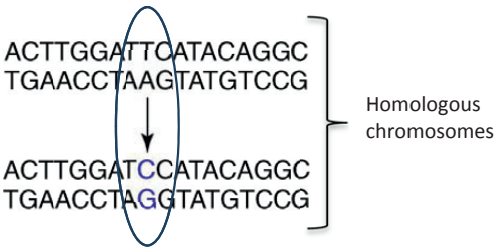


Figure 3. Visualisation of a SNP. A SNP is a specific position in the genome at which different sequence alternatives (alleles) exist in normal population(s) wherein the least frequent allele has an abundance of 1% or greater; here two different alternatives are seen. Here a heterozygous individual displaying a TA base pair and a CG base pair at the same position on homologous chromosomes.

Systematic research and cataloguing of SNPs began in the late 1990s (reviewed in [19]). Upon the completion of the HGP, the International SNP Map Working Group, consisting of the SNP Consortium and The International Human genome Sequencing Consortium, published a map of 1.42 million SNPs [45]. This spurred the research into the role of these variations in disease aetiology.

HapMap

The HapMap project was officially launched in 2002 [29] with the goal to "*determine the common patterns of DNA sequence variations in the human genome and to make this information freely available in the public domain*" [29]. They aimed to genotype SNPs in three different populations, European, African and Asian, and describe the pattern at which SNPs were inherited. Linkage disequilibrium (LD) is the non-random inheritance of genetic markers. The LD between two SNPs is measured as r^2 or D´ and their value decreases with increasing physical distance between them. The term LD was first used in 1960 [46] and initially applied in population genetics. SNPs inherited together form a haplotype block [47]. This means that by genotyping one SNP one can obtain information about other SNPs in LD with the genotyped SNP. Haplotype structures based on LD were described in a number of papers in the early 2000s [47-54]. It enabled the use of SNPs to "tag" nearby variation [55]. Instead of having to genotype all known variants, a subset of informative SNPs can be chosen which will cover a large percentage of all genetic variants. This opened the door to cost-efficient assessment of common genetic variants, GWASs [56, 57].

1.4 Complex Traits and Genetic Approaches

1.4.1 Complex Traits

The search for genes responsible for Mendelian diseases was of great impact for medical genetics during the 1980s [58]. Mendelian diseases are recognised by their often predictable mode of inheritance and are often caused by mutation in a single gene [59]. The hunt for disease genes proved fruitful and by the mid-1990s more than 400 diseases had been genetically mapped [60]. Today we know the molecular basis of over 4,000 Mendelian disorders [61]. The term complex trait refers to any phenotype that does not follow the classical Mendelian order of dominant or recessive inheritance [58], such as cardiovascular disease, Crohn´s disease and type 2 diabetes. Complex diseases or traits are caused by many genes, gene-gene and gene-environment interactions (reviewed in [32]). Therefore, the genetic architecture of complex diseases has proven more difficult to unravel. The linkage and candidate gene studies came short in identifying genes associated with common complex diseases. *"Has the genetic study of complex disorders reached its limit?"* Risch and Merikangas

asked in a *Science* paper from 1996 [62]. They suggested GWASs to be the future for uncovering the genetic basis of diseases or traits.

1.4.2 Genome-Wide Association Studies

The "GWAS idea" was discussed by several researchers in the second half of the 1990´s [62-64]. Wang *et al*. [65] showed in 1998 using a prototype genotyping chip that it could be feasible. The completion of the sequence of the human genome helped open the doors to this new era in genetics. Efficient genotyping technologies developed at an astonishing rate allowing for large GWASs to emerge. The original goal of a GWAS is to link common genetic variants to common diseases or traits [32]. In the years following the first successful GWAS published in 2005 for Age-related macular degeneration (AMD), [66] the number of studies published have sky-rocketed (www.genome.gov/gwastudies/) [67] (Figure 4). GWAS is a powerful and efficient approach for the identification of genetic variants associated with common and complex diseases or traits. GWASs are hypothesis-generating studies investigating a large number of genetic variants (minimum > 100,000, however today generally between 500,000 and millions) across the entire genome (reviewed in [57]). The goal is the identification of novel genes/genomic loci related to the disease under investigation, to increase the understanding of the molecular mechanisms involved, or to predict the risk of the disease.

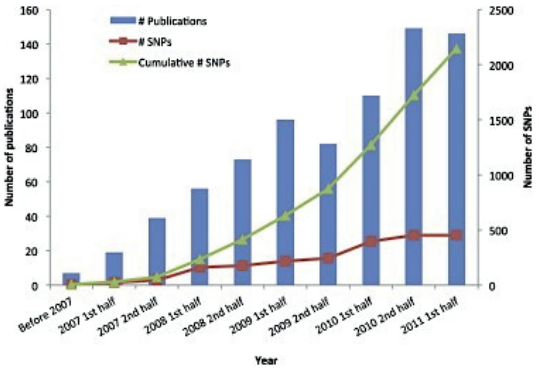
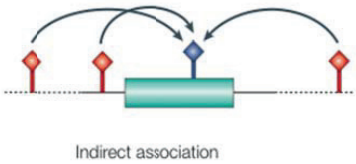
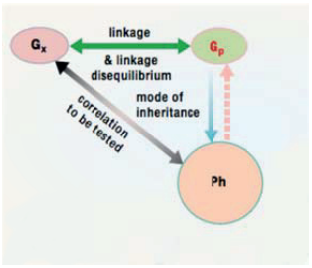


Figure 4. Overview of the number of GWASs published from before 2007 and until 2011. Reprinted from The American Journal of Human Genetics 90, 7-24, Vissher *et al*, Five Years of GWAS Discovery, © (2012), with permission from Elsevier.

SNPs have proven useful as markers for complex diseases and have been linked to a variety of diseases through GWASs. However a SNP associated with a disease through a GWAS is not necessarily the predisposing allele [68]. Although a SNP may sometimes be causative, it more often serves as a marker for a locus at which disease association can be found (Figure 5a). What is tested is really the correlation between a specific genotyped marker and a phenotype, and this is dependent on the correlation between the genotyped marker and the allele(s) that influence the phenotype (Figure 5b) [68].



a



b

Figure 5. Shows an example of an indirect association. a) The blue marker is the causal variant and is not genotyped. The red variants are the genotyped variants and are in LD with the causal variant. Adapted by permission from Macmillan Publishers Ltd: Nature Review Genetics [57], © (2005). b) The correlation tested in a GWAS is the correlation between the genotyped SNP G_x and the given phenotype (Ph). The strength of this correlation is dependent on the linkage disequilibrium with the causal SNP (G_p) and its influence on the phenotype (Ph). Adapted by permission from Macmillan Publishers Ltd: Nature Genetics [68], © (2000)

Common Disease Common Variant Hypothesis

Human genetic variation can be divided into common and rare variants. The upsurge of GWASs was built on the common disease - common variant (CDCV) hypothesis [62-64, 69, 70], which states that common diseases or traits may be caused by a limited number of common variants (frequency >1%) with low penetrance, each contributing to the disease risk or trait. Figure 6 shows the relationship between allele frequencies and penetrance. As results from GWASs started to mount it became clear that most common variants also have low effect size (mean OR around 1.3[71]) and explain little of the heritability of a trait [72-74]. For example, in type 2 diabetes, despite a very large sample size (> 10,000 individuals in the discovery set and around 50,000 in replication) the 18 common variants found only explained about 6% of the increased risk [74, 75].

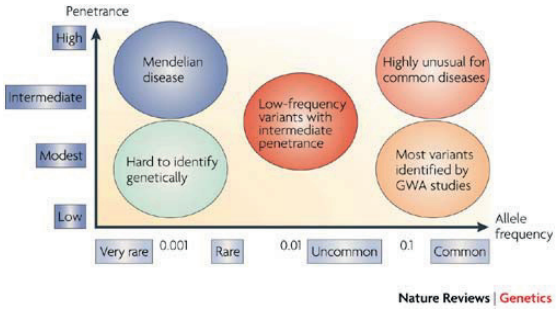


Figure 6. Shows the relationship between allele frequency and penetrance for Mendelian disease, rare and common variants. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics [76], © (2008)

The opposing hypothesis, the rare variant hypothesis (common disease rare variants, CDRV) states that summations of rare variants with higher penetrance and larger effect size are the genetic cause of common complex diseases or traits [71, 77]. Evidence exists for both rare [78, 79] and common [66, 80] variants influencing common diseases and they can perfectly well co-exist [81]. David Altshuler was quoted in a Nature Genomics technology editorial saying: *"right now no one actually knows which one is going to apply to which disease"* [82].

GWASs and especially the CDCV hypothesis have been vigorously discussed even before the first large GWASs were published. Followers and sceptics have written numerous scientific papers, reviews, commentaries and editorials discussing all aspects of GWASs [67, 68, 83-91]. Some of the aspects concerning methodological consideration will be discussed further in Chapter 5.

This thesis stretches from single SNP analysis in paper III based on the initial results from the GWASs in paper I and II, to investigating population structures based on available whole-genome SNP data and evaluating the effect of potential bias in GWASs in paper IV. Aspects central to these papers, including population structures and the phenotypes studied in paper I-III will be discussed in the following sections.

1.5 Population Structures

In 1999 Cargill *et al.* [92] studied the distribution of 560 SNPs found in 106 genes among Europeans, African Americans, African and Asian samples and found an excess of SNPs that were only seen in one of the ethnic subgroups. Their findings were in concordance with previous observations [93, 94] and they raised the issue of the need for a comprehensive SNP data-base which described genetic variation in different populations. Such data sets are highly valuable addressing the genetic structure of populations. Today, biological anthropology has reached new heights with the emergence of large scale genetic studies making whole genome SNP data sets available for the scientific community. Two aspects are central in understanding population structures. One is population genetics, understanding and uncovering the demographic history of populations. The second is genetic association studies of complex diseases or traits and understanding the potential bias in case control studies introduced by non-random distribution of SNPs in the population.

Large studies have investigated the population structures of Europe [95-102], as well as of our neighbouring Nordic countries [103-107]. Interestingly the pattern of genetic variation reflects the geographic map of Europe in the plotted individuals (Figure 7). The Nordic Centre of Excellence in Disease Genetics has created a database collection of genome-wide SNP data for Nordic samples (http://www.nordicdb.org/database/Home.html accessed 30.10.2012). They have investigated the difference in population structures in these samples and find the similar mirror of geographic map of Sweden, Finland and Denmark [105].

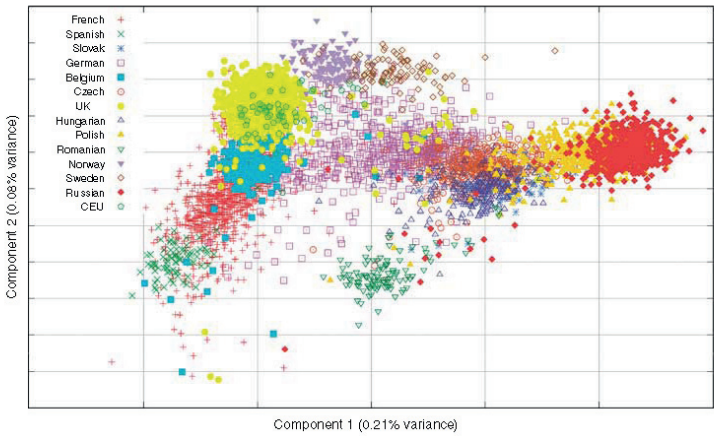


Figure 7. Investigation into population structures in the European population. The plot which mirrors the geography of Europe shows the first two principal components in a principal component analysis (PCA) of the European population. Reprint from Heath *et al.* 2008[95].

Differences in allele frequencies underlie population structures and can be detected using a principal component analysis (PCA) [108]. It is a statistical method for investigating data sets with a large number of measurements and reducing the large number of observations to principal components which explain the variance within the sample. PCA have three main applications; 1) detecting population structures, 2) correcting for this stratification in case control studies and 3) making inference about human history [109].

Identity by state (IBS) and identity by decent (IBD) are commonly applied in describing differences and similarities in populations. Two individuals share an allele IBD if it is inherited from a common ancestor. An IBD analysis requires genome-wide SNP coverage and in general, the analysis uncovers individuals who look more similar to each other than expected by chance [110]. The aim of an IBD analysis is to identify unknown family relations, siblings or parent-child pairs that are expected to share approximately half of their alleles IBD. Alleles IBS on the other hand are identical alleles not inherited from a common ancestor. An IBS analysis aims to identify individuals who look more different to each other than would be expected in a homogenous sample [110].

Another commonly investigated feature of population structures is runs of homozygosity (ROH) which has been characterised in a number of European populations [99, 111, 112]. This structure, seen as a stretch of homozygous alleles, represents elevated levels of background parental relatedness [113]. The frequencies of these ROH, and the total length of the genome found in ROH, vary between populations. These aspects are also found to have a positive correlation with consanguinity [114-116]. In that respect, ROH have been utilised in the identification of recessive disease genes [117-122]. Meiosis and recombination have the potential to break up these structures and reduce the size of ROH through the courses of generations in outbred populations [115, 121]. Even so, ROH >1Mb have been found to be widespread in all populations [113, 115, 116, 123, 124]. LD can also be a contributing factor to ROH. Patterns of LD differ between different populations and have been investigated in detail in the HapMap project [29, 56] and others [47, 125, 126]. In population studies LD is often characterised using LD-unit (LDU) maps. A LDU is the product of the physical distance between SNPs and a parameter that reflects the decline in the probability of association between markers according to physical distance [126].

1.5.1 Population Structures as a Bias in GWASs

It is well known that differences in population structures, where allele frequencies differ systematically between cases and controls, can cause bias in the form of greater number of type I errors (false positives) and spurious associations in genetic association studies [127-134]. This is due to the fact that in GWASs we are looking for alleles which differ significantly in frequency between cases and controls. This difference in allele frequencies between cases and controls will be sensitive to inflations or deflations in allele frequencies caused by individuals with admixed ancestry or familial relations where allele frequencies naturally differ or have a higher degree of sharing. [135, 136]. Careful considerations must be made when selecting cases and controls for large scale genetic studies [137, 138].

1.6 Lung Cancer

Lung cancer is the leading cause of cancer death in the western world [2]. There is no doubt that tobacco consumption, more specifically cigarette smoking, is the major cause of this disease [139]. From being a rather rare disease until the beginning of the 20th century, incidence rates of lung cancer have risen with the increasing tobacco consumption to become the most common cancer in men in most countries [140] with an incidence rate of >60/100,000 in Central and Eastern Europe [141, 142] and the second most common cause of cancer amongst men in Norway (http://kreftregisteret.no/ accesses 23.10.12) (numbers for the Norwegian population can be found in table 2). Several aspects of cigarette smoking, of which smoking duration is paramount, play a role in lung cancer risk: smoking quantity, duration of smoking, time since quitting, age at start, type of tobacco product consumed and inhalation pattern [139]. The cumulative risk of lung cancer for continuous smokers is approximately 15% at age 75 compared to <1% for never-smokers [143-145]. Other environmental and occupational factors known to increase the risk of lung cancer are exposure to polycyclic aromatic hydrocarbon (PAH), asbestos and radon [146, 147] (for a complete list of occupational agents and exposure circumstances classified by IARC as carcinogenic to humans with the lung as target organ see table 1.01 in [147]).

Lung cancer is divided into two main histological categories; Non-small-cell lung carcinoma (NSCLC) derived from bronchial epithelial cells and Small-cell lung carcinoma (SCLC) derived from neuroendocrine cells. NSCLC is further divided into three main subtypes namely squamous-cell carcinoma, adenocarcinoma and large-cell carcinoma [140, 147]

Table 2. Lung cancer statistics for Norway 2009 (total number of inhabitants 4,842,676), based on numbers from the Cancer registry of Norway [148]

	Men	Women	Total
Incidence (2009)	1519	1129	2648
Prevalence (per 31.12.2009)	-	-	4987
Accumulative risk by age 75	4.4	3.1	-
Survival (5 yrs. relative survival, %)	11.5	15.1	-
Number of deaths (2009)	1230	830	2060

1.6.1 Genetics of Lung Cancer

A large number of genetic changes are implicated in the initiation and development of lung cancer. These include chromosomal aberrations, point mutations and epigenetic alterations [149-152]. The development of lung cancer is referred to as a multistep carcinogenesis, which is a stepwise malignant progression of the cancer cells (reviewed in[153]). Genetic changes tend to vary between the different histological subtypes of lung cancer, however three genetic changes are common, *TP53* mutations, inactivation of the retinoblastoma pathway and loss of heterozygosity (LOH) at chromosome 3p (reviewed in[153]), the most common being *TP53* mutations. *TP53* is a tumour suppressor gene encoding the protein p53 that has an important role in response to genotoxic stress [154, 155]. Smokers have a higher frequency of *TP53* mutations and the most common of which is G to T transversions [156-159] at positions where DNA adducts are formed after exposure to PAH [158]. Other mutations commonly found amongst smokers are G to T transversions in the proto-oncogene *KRAS*. Mutations in this gene are often early events and are associated with poorer survival [160, 161]. In the cases of non-smokers, mutations and overexpression of the epidermal growth factor receptor (EGFR) are more common [162].

Epigenetic changes commonly observed in lung cancer patients are promoter hypermethylation [163] leading to gene silencing. The tumour suppressor gene *CDKN2A* is frequently inactivated by DNA hypermethylation [163-166]. Inactivation of *CDKN2A* leads to loss of G1 arrest control, and hence deregulation of cell proliferation [165].

LOH at chromosome 3p is the third most common event occurring in all types of lung cancer (reviewed in [153]). This region is particularly prone to deletions due to damage caused by carcinogens in cigarette smoke [167], and contains several tumour suppressor genes.

Telomerase activity is associated with various cancers and telomerase is expressed at high levels in over 90% of human malignancies (reviewed in [168]). Telomerase is an important enzyme in the maintenance of chromosome ends and is normally inactivated in somatic cells (reviewed in [169]. Gain of the chromosomal arm 5p containing the reverse transcriptase telomerase, *TERT*, is one of the most common chromosomal gains and a study by Weir *et al*. [170] found copy number gain at 5p in 60% of their total samples. In relation to lung cancer, amplification of the *TERT* locus on 5p15.33 is the most common event in early stage lesion of NSCLC [171].

Other genetic changes found in various types of lung cancer involve activation of intracellular signals such as PI3K, AKT, MAPK, NF-κB and TNF-α (reviewed in [149]). Tobacco-specific nitrosamines such as 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanon (nicotine-derived nitrosamine ketone, NNK) have been shown to activate the AKT-pathway, leading to an increase in cellular proliferation [172].

Recent years have seen the emergence of deep sequencing of various lung cancer cell lines, investigating both variants at the nucleotide level and transcriptional level [173-178]. Pleasance *et al*. [177] identified 22,910 somatic substitutions and 344 copy number segments in a SCLC cell line. Weir *et al*. [170] found copy number gains and losses comprising about half the human genome in lung adenocarcinoma tumours. These studies highlight the vast amounts of genetic alterations in lung cancers, and represent important steps toward characterisation and understanding of the molecular background for the disease.

Genetic Susceptibility to Lung Cancer

Familial aggregation of lung cancer seen in epidemiological studies strongly suggests a genetic component to the susceptibility of lung cancer [179, 180]. In particular, three large cohort studies have contributed to this knowledge; the Utah, Swedish and Icelandic cancer registries [140]. The increased risk of lung cancer for an individual with a family history of lung cancer is approximately 2.5 [181, 182].

DNA polymorphisms have been studied to elucidate the genetic susceptibility to lung cancer. Pre-GWASs studies focused on genes involved in the Phase I/II of xenobiotic metabolism and DNA repair (Summarised in [183]). Many of the studies have lacked sufficient power and the results have been inconsistent (reviewed in [183]). Paper I in this thesis was one of three GWASs on the risk of lung cancer published simultaneously [184-186]. These independent studies identified a region on chromosome 15q25 containing the nicotinic acetylcholine receptors (nAChR)-subunits *CHRNA5/A3/B4.* In an extension to our first study (paper II), a second locus was identified on chromosome 5p15.33. This locus contains two genes, *TERT* and *CLPTM1L* [187]. In addition to this, association with lung cancer has been found on chromosomal region 6p21 in successive studies [187-190]. However, despite the efforts to

elucidate the genetic architecture of lung cancer, only 10% of the familial risk of lung cancer can be explained by the three susceptibility loci 15q25, 5p15 and 6p21 [140].

1.7 Chronic Obstructive Pulmonary Disease

COPD is currently the 4th leading cause of death world-wide [191]. In 2002 it killed a total of 2.75 million people and accounted for 4.8% of all deaths [192]. In fact, the death rate from COPD in the United States has doubled since 1970 [193]. As for lung cancer, the largest risk factor for developing COPD is cigarette smoking (reviewed in [194]. Other environmental risk factors include exposure to air pollutant, in particular indoor from burning biomass fuels for cooking purposes, and occupational exposure to fumes [194, 195]. In addition, childhood asthma and respiratory infections and tuberculosis have been associated with chronic respiratory symptoms [194, 196]. Prevalence data have previously been difficult to compare due to differences in the diagnostic criteria. The current criteria recommended by The Global Initiative on Obstructive Lung Disease´s (GOLD) is based on post-bronchodilator spirometry ratio of forced expiratory volume in one second (FEV_1), and forced vital capacity (FVC) being less than 0.7. Furthermore, severity of COPD has been categorised according to FEV_1 in per cent of predicted; mild $\geq$ 80 %, moderate 50-79%, severe 30-49 % and very severe < 30%. Many studies, have defined COPD according to pre-bronchodilator spirometry FEV1/FVC< 0.7. Post-bronchodilator values gives about 30% lower prevalence [197].

COPD is a progressive and chronic inflammatory disease [198]. The GOLD definition states that COPD is “a disease state characterized by airflow limitation that is not fully reversible. The airflow limitation is usually progressive and associated with an abnormal inflammatory response of the lungs to noxious particles and gases” [199]. A narrowing of the small airways is caused by a non-specific inflammatory response [200], mucosal hyperplasia and disturbance in tissue repair [201]. The immunological mechanisms leading to COPD can be seen in a step-wise manner from the initial response by the innate immune system, to T-cell activation and proliferation, and the adaptive immune reactions [202, 203]. Several key inflammatory cells including macrophages, T-lymphocytes, B-lymphocytes and neutrophils have been found associated with COPD (figure 8) [204].

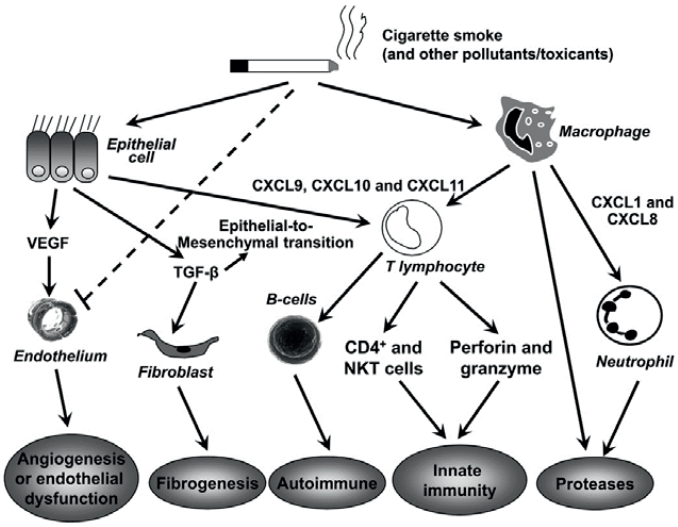


Figure 8. A schematic view of cells and mediators involved in the pathogenicity of COPD. Macropages and epithelial cells release chemokines attracting inflammatory and immune cells (T-cells, B-cells and Neutrophils) to the lungs. This leads to an increase in the release of proteases, perforin and granzyme resulting in alveolar wall destruction and mucus hypersecretion. Reprinted from Curr Opin Pharmacol, 9(4): 375-383, Yao, H. and I. Rahman, Current concepts on the role of inflammation in COPD and lung cancer, © (2009), with permission from Elsevier

1.7.1 Genetics and COPD

The only well-established genetic cause of COPD is deficiency in the protease inhibitor α_1 antitrypsin (α_1 AT). This deficiency was first reported in association with emphysema in 1963 [205], and approximately 1-2% of all individuals with COPD display this defect [206]. α_1 AT is a glycoprotein, coded for by the *SERPINA1* gene, which main function is the inhibition of neutrophil elastase [207]. Deficiency in this protein predisposes to early onset emphysema.

Other evidence of a genetic component is based on familial aggregation of COPD, candidate gene studies and GWASs, and this has been vigorously reviewed [208-214]. However, results have been varying and often inconsistent in replication studies. Several large GWASs and meta-analyses have been conducted over the last five years [215-219]. Of the genes most commonly found associated with COPD are the *CHRNA5/A3/B4* gene cluster, also found associated with lung cancer and nicotine addiction, and the *HHIP*-gene located on

chromosome 4. The meta-analysis by Hancock *et al.* [215] identified eight different loci with moderate impact on pulmonary function. With the exception of α_1 AT deficiency, smoking is indisputably the largest risk factor.

1.8 Smoking and Nicotine Addiction

Should we all blame Christopher Columbus? Well, probably not, even though it was he who first brought tobacco to Europe around 1492 [220]. More than 500 years later, where has it lead us? To lung cancer, cardiovascular disease and COPD, all amongst the top ten leading causes of death [192, 193] and all strongly associated with smoking [2, 194, 221]. Figure 9 shows a schematic representation of smoking related cancers and chronic diseases.

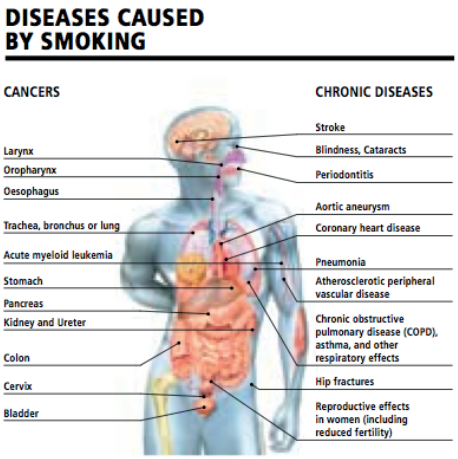


Figure 9. Schematic representation of the cancers and chronic diseases caused by cigarette smoking. Source: U.S. Department of Health and Human Services. The health consequences of smoking: a report of the Surgeon General. Atlanta, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Centre for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2004 (http://www.cdc.gov/tobacco/data_statistics/sgr/2004/index.htm accessed 30.10.2012.

It is not surprising that cigarette smoke has such a devastating effect on the human body; tobacco smoke contains around 4,800 compounds where at least 60 are known carcinogens [222-224]. Of the most potent are the tobacco specific nitrosamines [225]. There are several different tobacco specific nitrosamines but the most carcinogenic and most widely discussed is nicotine-derived nitrosamine ketone (NNK), and *N*-nitrosonornicotine (NNN) [225, 226]. The metabolic pathway of NNK, initiated by cytochrome P450 [227], leads to the formation of compounds able to bind to DNA and form potentially mutagenic DNA adducts, such as 7-methylguanine or O^6-methylguanine [228]. The process is referred to as metabolic activation [227]. The other important class of carcinogenic compounds, PAHs, also have the ability to form DNA adducts leading to DNA damage. Failure to repair theses damages may lead to cytotoxic effects due to block of DNA replication and/or transcription, or mutation if bypassed incorrectly (Reviewed in [27]). It has been shown that DNA adducts from both NNK and PAH can lead to mutations in genes such as *KRAS* and *TP53*, which are central in lung cancer (figure 10) [229-231].

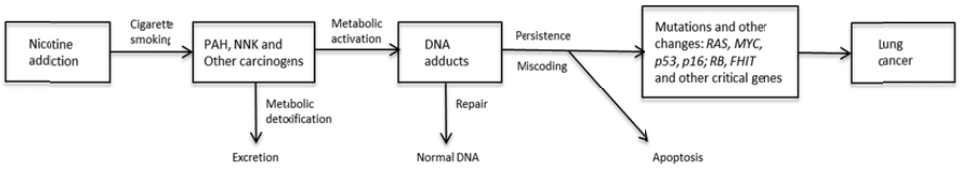


Figure 10. A simplified model for tobacco carcinogenesis via adduct formation by PAH and tobacco specific nitrosamines. Hecht, S.S. Tobacco Smoke Carcinogens and Lung Cancer, Journal of the National Cancer Institute, 1999, Vol. 91, No. 14, 1194-1210, by permission of Oxford University Press

In addition to cigarettes, another tobacco containing product is highly popular in Norway and other Nordic countries, namely snus, often referred to as Swedish snus. Snus is a smokeless tobacco product placed under the upper lip and kept there [232]. Snus contains many of the same harmful substances as cigarettes though the levels of tobacco specific nitrosamines and PAH are shown to be lower [233]. The nicotine exposure is believed to be similar to that of cigarette smoking. [234]. Measured blood concentrations of nicotine show a more gradual increase compared to that of smoking a cigarette, and peaks around 30-40 minutes after

placement [232]. Hazardous health effects from the use of snus compared to cigarettes have been highly debated. An extensive review [235] concludes that the use of snus is clearly less hazardous than smoking.

Most people are aware of the dangers represented by cigarette smoking. So why do they continue? Nicotine is a highly addictive substance leading to nicotine dependence (ND), a dependence almost as strong as to that of cocaine or heroin (reviewed in [220]). Nicotine triggers the release of dopamine in the nucleus accumbens (NAcc) and this elevated level of dopamine reinforces the abuse [236-238]. Nicotine and some tobacco specific nitrosamines can act as agonists and have the ability to activate nAChR [239, 240]. nAChRs are ligand-gated ion channels consisting of a combination of five subunits (α1-10, β1-4, δ, γ, ε) and can be either homomeric (α7, α8 or α9) or heteromeric (combination of α2-α6 or α10 with β2-β4 or α1 with β1, γ, δ or ε) [241]. nAChR are divided into two main categories, neuronal and muscular, based on their original identification in the nervous system and at the junction between nerve endings and muscles [242, 243]. However, today it is know that they are found in a wide range of tissues and both types have been found in cancer cells [244]. The different subunits of nAChRs are encoded by separate genes (*CHRNA1-10*, *CHRNB1-4, CHRND, CHRNG, CHRNE*) spread across eight different chromosomes (http://www.ncbi.nlm.nih.gov/gene/ accessed 30.10.2012). Activation of nAChRs leads to membrane depolarization and Ca^{2+} influx [245], which initiate a range of different cell-signalling pathways [246, 247].

In the midbrain, the nAChR are involved in the dopaminergic system via dopaminergic and GABAergic neurons. Activation of the heteromeric nAChR by nicotine in doses obtained by cigarette smoking leads to excitation of dopaminergic neuron and release of dopamine, while the GABAergic neurons synapse into and inhibit the dopaminergic neurons [237, 238]. The balance between the excitation and inhibition of the different neurons is therefore important [248]. The heteromeric nAChRs are readily desensitised upon exposure to nicotine. However, activation of homomeric (α7) receptors enhances excretion of glutamate, which in turn activates the dopaminergic neurons stimulating dopamine release [238].

Increased attention has been turned towards the nAChRs found in lung tissue. Activation of the receptors by nicotine or nicotine metabolites such as NNK has been found to affect signalling pathways of importance in cancer, such as inhibition of apoptosis and cell proliferation. [249]. An example of this is nicotine and NNK induced phosphorylation and thus

activation of Akt by PI3K which promotes cell proliferation or cell survival by inhibition of apoptosis via the NF-κB protein complex [172, 250]. Nicotine has been shown to increase Akt phosphorylation at doses readily achievable in smokers [172]. In addition, nAChRs operate in an indirect manner by altering the synthesis and release of neurotransmitters that in turn regulate the synthesis of growth factors and angiogenic factors [251]. In paper III we discuss the role of nAChR in ND and lung cancer in relation to the SNP investigated.

2 AIMS OF THE STUDY

The overall aims of this work were to:

1. Identify genetic variants that influence the risk of lung cancer using the GWAS approach
2. Investigate result from the initial study further and extend the outcome phenotypes in a large homogeneous population to clarify uncertainties regarding direct vs. indirect effect of lung cancer.
3. Use genome-wide SNPs data generated in the GWAS to uncover population structures, within Nord-Trøndelag and Tromsø, which might result in bias in GWASs.

3 DATA SOURCES AND METHODS

"Biobanks are the research gold of Norway"[5]

Biobanks are highly valuable research sources. In a world where ever large populations are needed to achieve research goals, they are of immense importance. In Norway, several large population studies have been conducted in various regions of the country. A major advantage in Norway is the opportunity to link data to well established health registries through the Norwegian personal identification number. This, together with the willingness to participate, makes Norwegian biobanks an invaluable asset.

3.1 The Nord-Trøndelag Health Study

The Nord-Trøndelag Health Study (HUNT) is a comprehensive multipurpose population based study having collected data of the adult population aged 20 years or more in three surveys, HUNT1 (1984-86), HUNT 2 (1995-97) and HUNT 3 (2006-08). The collection of data and biological material has been described in detail [252] . In short, the studies comprise data from questionnaires, interviews and clinical examination. All participants in HUNT 2 (about 65,000) and HUNT 3 (about 50,000) provided blood samples. DNA has been made available from most participants in HUNT 2 and is stored in the HUNT biobank. Approximately 36,000 participants participated both in the HUNT 2 and HUNT 3 studies [252, 253].

DNA samples isolated from blood samples collected during HUNT 2 were used in the work presented in this thesis (paper I, II, III and IV). Lung cancer cases were identified by linking the HUNT database to the Cancer Registry of Norway via the unique Norwegian personal identification number. Only individuals who developed lung cancer after participation in the HUNT 2 study and who were diagnosed with lung cancer as the primary tumour were included in the analysis. At the time of case selection, cancer diagnoses from the Cancer Registry of Norway was available up to and including January 1st 2004 for paper I and II, and was extended to 31st of December 2009 for paper III.

[5] A view expressed by Kristian Hveem in "Vil hente fram forskningsgullet", by Elin Fugelsnes, Forskning.no, 10.02.2009. Article in Norwegian: *"Biobankene er Norges forskningsgull"* and translated by Maiken Elvestad Gabrielsen

Parallel to the main HUNT studies, a few select phenotypes have been investigated in more detail, among these lung function and bone mineral density. The Lung Study in HUNT invited a random collection/assortment of participants in HUNT 2 (5%, n = 2791) and HUNT 3 (10%, n=5068). In addition, participants in the two main studies reporting having had asthma, COPD or asthma-related symptoms were also invited, totalling 8,150 from HUNT 2 and 7,391 from HUNT 3. All participants were subjected to lung function measurements (spirometry), measurement of bone mineral density, and went through an interview [253, 254]. For paper III, we utilised data from the additional lung study conducted in association with HUNT 3. Spirometry data from the medical examination identified individuals with impaired lung function.

The population in Nord-Trøndelag is relatively homogeneous with less than 3% non-Caucasians and a net annual out migration of only 0.5 % around the time of HUNT 2.

3.2 The Tromsø Study

The Tromsø study is a large comprehensive multipurpose population based study conducted in Tromsø Municipality. Since 1974 data has been collected for various age groups in six surveys, Tromsø 1, ages 20-49 (1974), Tromsø 2, ages 25-54 (1979-80), Tromsø 3, ages 12-67 (1986-87), Tromsø 4, ages 25-97 (1994-95), Tromsø 5 ages 30-89 (2001-02) and Tromsø 6, ages 30-87 (2007-08). The first study was conducted by the University of Tromsø and the last five studies in cooperation with the National Health Screening Service. The study comprises questionnaire data, data from a clinical examination, and biological samples from some of the studies (http://uit.no/ansatte/organisasjon/artikkel?p_menu=42515&p_lang=2&p_document_id=70715&p_dimension_id=88111, accessed 30.10.2012). Tromsø 4 is the largest and most comprehensive of the six surveys comprising 27,158 participants and is described in detail in Jacobsen and Eggen [255]. DNA has been made available from all participants in Tromsø 4 and is stored in the HUNT biobank.

DNA samples from Tromsø 4 were used in papers I, II and IV. Lung cancer cases were identified by linking the Tromsø database to the Cancer Registry of Norway via the Norwegian personal identification number. Only individuals who developed lung cancer after participation in the Tromsø 4 study (1994) and who were diagnosed with lung cancer as the primary tumour were

included in the analysis. At the time of case selection, cancer diagnoses from the Cancer Registry of Norway were available up to and including January 1st 2004.

3.3 Cancer Registry of Norway

The Cancer registry of Norway was established in 1951 and is one of the oldest cancer registries in the world. Every identified cancer case in Norway since January first 1952 has been registered here (http://kreftregisteret.no/en/General/About-the-Cancer-Registry/About-the-organization/History/ accessed 03.10.12). The Registry has three main objectives: data collection and registration of cancer incidences, research and thirdly, information to the general public regarding cancer. All cancer cases are registered with the Norwegian 11-digit personal identification number and the registration is obligatory. Data in the Cancer Registry of Norway is based on morphological diagnosis from all pathology departments in Norway and a written report from the clinical departments [256]. Lung cancer diagnosis from the Cancer Registry of Norway was used to identify cases for paper I-III. Figure 11 shows a schematic representation of data collected by the registry.

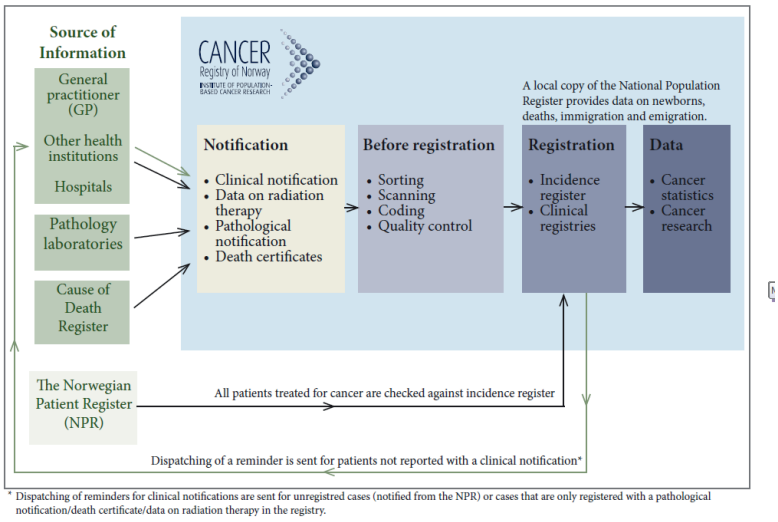


Figure 11. Sources of information and processing of data by the Cancer registry of Norway.[148]. Reproduced with permission from the corresponding author.

3.4 Genome-Wide SNP Arrays

The HumanHap300 and HumanHap370cnv duo and quad bead chips were used for the genome-wide SNP analyses in Paper I and Paper II.

Both the HumanHap 300 and the HumanHap370cnv duo and quad bead chips use the Illumina Infinium II assay [257, 258]. This assay is based on the bead-array technology where silica beads with specific oligonucleotide probes attached to them, are spread across micro-wells on a glass slide. The assay uses whole genome amplification followed by fragmentation and single base enzymatic extension with labelled nucleotides followed by fluorescent staining and imaging (Figure 12).

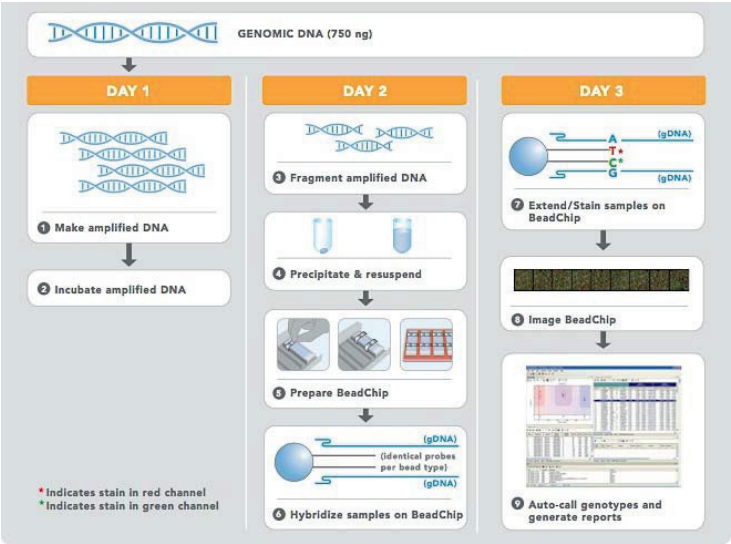


Figure 12. Illumina® Infinium™ II workflow. The Infinium II is a three day protocol. 750 ng of genomic DNA is amplified by whole genome amplification before it is enzymatically fragmented. The fragmented DNA is hybridised to the bead chip followed by a single base extension and staining. The bead chip is scanned using an Illumina BeadArray Reader and the results prepared in the corresponding Illumina software. (Illumina® SNP genotyping, Infinium™ Assay Workflow).

The HumanHap 300 and HumanHap 370CVN chips contain approximately 317,139 and more than 350,000 SNPs respectively. In addition the HumanHap370CNV contains more than 17,000 CNVs. Both are based on data from phase 1 of the HapMap project [29] and utilise tagSNPs to ensure high genomic coverage (approximately 80% of the genetic variation in a population of European ancestry)(Figure 13)[259]. The mean minor allele frequency (MAF) for the European (CEU) population is 0.26 for the HumanHap300 (Illumina SNP genotyping data sheet).

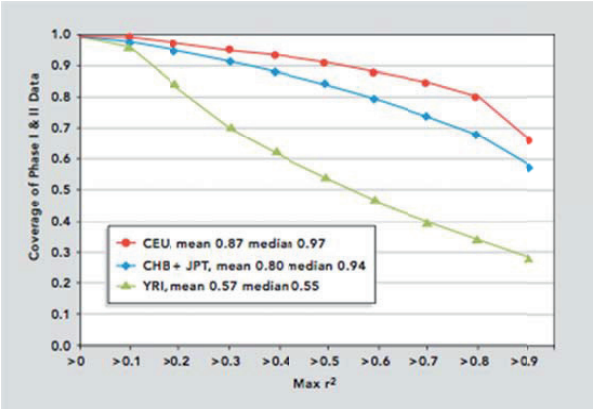


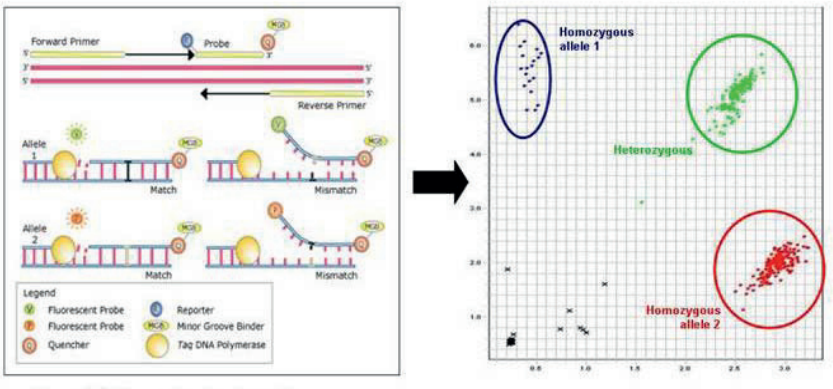
Figure 13. Genomic coverage of the HumanHap 300 bead chip for the three different populations, Europen (CEU), Asian (CHB+JPT) and Yoruban (YRI) (Illumina SNP Genotyping, Sentrix HumanHap300 Genotyoing BeadChip).

3.5 Taq-Man Assays

TaqMan assay is a single SNP genotyping assay using real-time PCR. It relies on the 5´–3´ exonuclease activity of Taq DNA polymerase [260]. In fact, because of this, it is named after the well-known videogame, PacMan (The Real-Time TaqMan PCR and Applications in Veterinary Medicine - From PacMan to TaqMan - a computer game revisited, accessed 30.10.2012). The assay consists of a forward and reverse PCR primer and two allele-specific TaqMan-probes. The TaqMan-probes are labelled with different fluorophores at the 5' end and a quencher molecule at the 3' end. A perfect match results in the Taq polymerase removing the allele specific TaqMan-probe and this releases the fluorophore resulting in a fluorescent signal

(Figure 14)[261]. The mismatch probe will not hybridize and remain intact with the quencher inhibiting the fluorescent signal. Relative strength of the fluorescent signal determines whether the sample is homozygous or heterozygous. The genotypes are determined by plotting the normalized fluorescence intensities [261].

A major advantage of the TaqMan assay is the ease at which it can be implemented. Using 384-well PCR plates, a large number of samples can be run with minor efforts and the amount of DNA needed is low (10ng/sample). However, a disadvantage may be that the multiplexing is low (up to seven SNPs in one reaction) and the proximity of the SNPs in question limited [262].



A B

Figure 14. A) Shows the different steps involved in the TaqMan® genotyping assay for both a match and a mis-match. The primers and fluorescently labelled probe binds to their recognition sequence. In the case of a perfect match the Taq DNA polymerase replicates the strand removing the probe with its exonuclease activity freeing the fluorescent probe. B) Shows the genotype calling and inspection of a cluster, the blue cluster are homozygous for allele 1, the green cluster are heterozygous, and the red cluster are homozygous for allele 2. Black samples are excluded due to failed genotyping or low quality making it impossible to call a genotype. Source: http://medicine.tcd.ie/neuropsychiatric-genetics/functional-genetics-genomics/genotyping.php, accessed: 02.10.2012

In the work presented in this thesis, TaqMan assays were used to genotype three SNPs at chromosome 15, rs16969968, rs8034191 and rs1051730, in paper III. All TaqMan genotyping was performed by qualified personnel at the HUNT biobank (Levanger, Norway).

3.6 Statistical Analysis

Work involving large data sets with hundreds of thousands of SNPs and thousands of individuals require efficient statistical programming. A detailed description of such is well beyond the scope of this thesis. A more detailed discussion on some of the methodological considerations can be found in Chapter 5.

3.6.1 Association Analysis

Association analyses can be carried out in a number ways depending on the number of variables and individuals under investigation. The analysis methods employed are described in the respective papers (papers I-III). In this chapter the statistics involved in Paper I-IV is summarised. Logistic and linear regression models can be employed to examine the association between SNPs and a given dichotomous phenotype or quantitative trait respectively. Logistic and linear regression models offer flexibility in the addition of covariates and are commercially available. The choice of statistical package will depend on the number of variables (single SNP or genome-wide) under investigation. For large genome-wide analyses PLINK [110] offers a flexible and quick analysis. The program is specifically tailored for large data sets and was the program applied for the association analysis in paper I and II. For analysing a smaller number of variants (paper III), the statistical page PWAS-statistics 18 (also known as SPSS) was used. Regression analysis can be performed for a large number of individuals, but is limited in the number of SNPs per analysis. On the other hand, it offers a wide range of analyses tools such as Cox-regression for survival, applied in paper III. A Cox-regression takes into account the time before an event happens and a hazard ration (HR) is calculated.

Paper IV involves analysis of a whole-genome data set associating variance in SNP distribution with geographical locations in Nord-Trøndelag (HUNT cohort) and Tromsø (Tromsø study).

Multiple dimensional scaling (MDS) analysis (principally the same as a PCA) was used to uncover genetic population structures. By transformation of the data a set of principal components are given, depicting the total amount of variation in the data set, with the first component showing the largest possible variance. This can be plotted in scatterplots to visualise the variation seen within the data set [108]. Several packages for analysing population structures exist such as EIGENSTRAT, STRUCTURE and PLINK [110, 263, 264]. Analyses in paper IV were done using PLINK [110].

4 METHODOLOGICAL CONSIDERATIONS

A lot has been said and done and elaborately discussed in regards to GWASs, its pitfalls, failures and triumphs. Covering them all is outside the scope of this thesis. I have therefore chosen a few key points for the following section.

4.1 Study Design.

"*Only when you know the question will you know what the answer means.*[6]*"*

That is true of most things in life. One has to know exactly what one is asking for the answer to make sense. In the above-sited "The Hitchhikers Guide to the Galaxy" they promptly reply: *"So give us the ultimate question then."* In research, there is no "oracle" computer to answer that. We ourselves must do the meticulous work of phrasing the study question. Study design is one of the single most important factors limiting the scope of the questions we can ask [265]. There are several issues concerning study design, 1) the phenotype in question, 2) the characteristics of the study group and number of samples available, 3) the number of loci to genotype and 4) the analytical methods for the association between genotype and phenotype.

Most GWASs and their replications have been conducted using the population based case control design (Reviewed in [87] (for an overview of GWASs conducted see http://www.genome.gov/gwastudies/ accessed 30.10.2012). The case control design is a retrospective study design, which allows for the comparison of allele frequencies between a group of diseased individuals and a group of healthy controls (reviewed in [76]. Family based studies constitute an alternative to population based studies. A major advantage of such a design is the ability to detect rare variants [137], as a family design will enrich for possible rare alleles and allows for detection of co-inheritance with disease in families [266]. This thesis is concerned with population based studies and family studies will not be discussed further in this section.

[6] From "*The Hitchhikers´Guide to the Galaxy"*

4.1.1 Phenotype

Correct definition of the phenotype is of considerable importance. GWASs generally deal with complex diseases dependent not only on one gene but a large number of genes and genetic variants. These are diseases that also show a large variation in phenotype (reviewed in [32]). In a utopian world all cases selected for a study would be as homogeneous as possible, both in regards to disease aetiology and population background. However, in the real world, one does often not have that luxury. Lung cancer is a complex and heterogeneous disease dependent on a large number of genes, environmental- and lifestyle factors (reviewed in [140]). There are also considerable differences in the molecular aetiology of the different histological subtypes and between smokers and non-smokers (reviewed in [153, 267]). However, in order to achieve a larger number of cases many studies analyse these under one large "umbrella"-diagnosis. Though stratifying according to histology would often mean a reduction in power, a few studies have identified histology specific effects [187, 188, 268, 269]. Whilst larger cohorts may have increased power to detect weak genetic associations, the heterogeneity of the phenotype can reduce the power.

Poorly validated diagnosis or misclassification can substantially reduce the power to detect an association between a trait and marker locus [270, 271]. Lung cancer diagnosis in papers I and II was obtained from the Cancer Registry of Norway. As shown in chapter 3.3 this diagnosis is based on information from hospital or general practitioner´s records, pathology reports, the cause of death registry and the Norwegian Patient Register (NPR). Together, this ensures accurate data to researchers. However, in our study another issue could lead to loss of power. Lung cancer generally has a late onset in life. In our study cases and controls were matched according to age. Data from the cancer registry dates back a few years, which makes it possible for controls to develop lung cancer in the meantime or later in life. This is almost certain to be the case in paper III where the entire cohort is genotyped. In studies looking for relatively minute differences in allele frequencies between cases and controls, having individuals who develop the disease later in life in your control group will lead to loss of power.

In paper III we have also included COPD and smoking habits as phenotypes. These are traits with considerable heterogeneity. COPD is a continuous phenotype with varying degrees of severity. An aspect to consider in regards to COPD is the diagnostic criterion being set by pre- or post-bronchodilator spirometry. In paper III we have employed the older definition

according to pre-bronchodilator spirometry FEV1/FVC< 0.7 as this was the available data. The now more commonly used measure is post-bronchodilator which gives about 30% lower prevalence [197]. To reduce the chance of over-diagnosing individuals in our study we included as cases only individuals with moderate to severe COPD adding the criteria FEV1% predicted <80%. Another commonly required criterion for COPD is smoking. In paper III we have chosen not to include smoking as a diagnostic criterion, and therefore refer to the phenotype as loss of lung function equivalent to that of COPD. This could lead to a few cases possibly suffering from asthma instead of COPD, which would reduce the power. However, not using smoking as a selection criterion allows us to analyse for potential differences between smokers and non-smokers in a statistical model.

With regards to the smoking phenotype, a multitude of variation exists and a clear definition of the phenotype might be difficult. A commonly used phenotype is ND, however this is also a phenotype with multiple features [272-275]. Several different scales or indexes have been developed to asses ND such as the Fagerström Test for Nicotine Dependence (FTND) [276] and Wisconsin Inventory of Smoking Dependence Motives (WISDM-68) [273]. We do not hold any information on ND and have analysed smoking habits in the form of cigarettes per day (CPD), number of years smoked and pack-years in our study. It is important to stress that data in paper III is based on self-reported smoking habits. It has been shown that individuals often underreport the true tobacco consumption in large population studies (compared to numbers based on sales) [277] which in turn will reduce the accuracy of the phenotype under investigation leading again to a loss of power.

4.1.2 Study Group and Sample Size

Another part of the study design is considering which samples to collect and how many. Chapter 1.4.1 briefly discusses the impact of population structures on GWASs and the fact that regional differences in allele frequencies can lead to bias. It is therefore desirable to use a genetic homogenous population. The HUNT population studied in this thesis is considered well suited for genetic studies. However, results from paper IV suggest that even within a relatively homogenous population, differences do exist and care should be taken when selecting cases and controls (paper IV). Data from paper IV and others (unpublished) regarding hidden family relations also suggest that family relations should be taken into account when selecting

samples for population based studies from the HUNT cohort. Papers I and II involve large international studies where samples have been collected from different areas. To avoid bias, country of origin was added as a variable in the regression model. In addition to this, a PCA was run to investigate the population structures prior to analysis. Individuals with more than 30% Asian or African ancestry were excluded from the analysis.

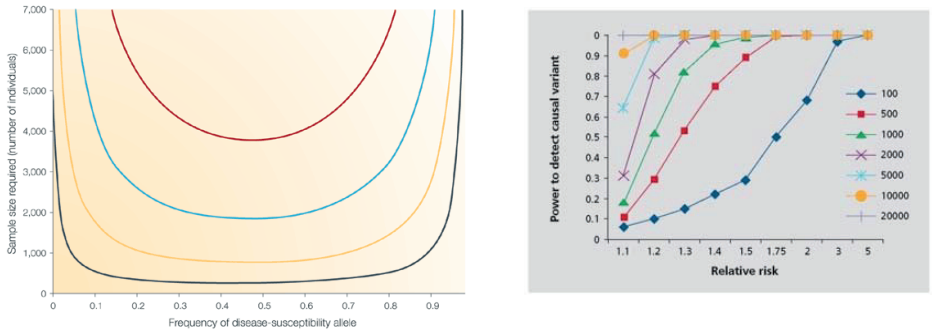
Selection of samples for a study can be either random or targeted. A random sampling would be well suited to investigate underlying genetic variation such as population structures. However, random sampling would come to short when selecting cases and controls for a defined phenotype. Cases for papers I-III were selected based on diagnosis from the Cancer registry of Norway and controls were matched on age and sex (with the exception of paper III where the entire population was genotyped). In paper IV we utilised samples genotyped in paper I and II. This could lead to ascertainment bias if the geographical distribution of cases and controls is not random. Plotting the results of a MDS analysis according to case control status did not show any geographical clustering of either cases or controls.

Sample size is a major concern in GWASs [278, 279]. An insufficient number of samples would lead to loss of power and inability to detect true association. The number of cases and controls needed is dependent on the allele frequency of the disease marker, the risk conveyed by the marker (the allelic OR) and the significance level [279]. Figure 15 (panel a) illustrates these relationships. It is also important to consider the overall disease risk in the population and mode of inheritance (dominant, recessive or additive). However, the latter is generally not known, especially when conducting a GWAS, but can be evaluated during the statistical analysis [280].

4.1.3 Power and Multiple Testing

In GWASs it can be hard to distinguish the true positive signals from the cacophony of all the false positive signals expected when such a large number of SNPs are investigated simultaneously. As a rule of thumbs in statistics, ten individuals are required for every variable tested in the model to achieve statistical significance [281, 282]. In large GWASs where more than 300,000 variables (SNPs) are tested pr. individual it is clear that this requirement is never going to be met (the number of variables tested grossly outnumbers the number of

individuals) and leads to reduced power to detect true associations (statistical issues reviewed in[283]. Power is a complex equation dependent on the set of SNPs, the effect size of the variant and number of samples. It can only be properly addressed through simulation of assumed scenarios [279]. Evaluating the power of a study should be a primary concern in the study design whether conducting a GWAS or a single SNP analysis [279]. Figure 15 (panel b) shows the relationship between the power to detect an association, the effect size and the number of individuals required in GWASs.



a b

Figure 15. a) Shows the relationship between the allele frequency of the disease associated allele and the number of individuals needed (80% power and p-value cut-off = 10^{-6}) to detect an effect size (OR) of 2 (black), 1.5 (yellow), 1.3 (blue) and 1.2 (red). Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics [283], ©(2005). b) Shows the power to detect a causal variant given an effect size (Relative Risk) and a certain number of individuals, assuming dominant model, minor allele frequency=0.2, frequency of disease is 1% and equal numbers of cases and controls [88].

Multiple testing.

GWASs are said to be hypothesis generating, this is true in the sense that one does not have a hypothesis regarding a predetermined genetic locus. In theory (see argument in section 4.3) GWAS analyses are hypothesis testing. The hypothesis tested is that whether differences in allele frequencies exist between a group of cases and controls. A typical null-hypothesis would thus be “allele frequencies are not different between cases and controls” and the alternative

hypothesis would be that there are. The significance level determines the level of type 1 errors, that is, rejecting the null-hypothesis when it is in fact true. This is also referred to as a false positive result. The opposite scenario, type 2 errors, is keeping the null-hypothesis when it is false. This is referred to as the sensitivity of the test. Testing such a large number of variables as in a GWAS, largely increases the chance of type I errors. In GWASs, several methods exist to account for multiple testing, among them Bonferroni corrections [137]. This simple, yet efficient method requires a p-value of 0.05 divided by the number of tests performed to claim statistical significance [137]. This means that if one is using 500,000 SNPs, the significance threshold is a nominal p-value of $1x10^{-7}$. Many consider this to be a too stringent method for GWASs [284]. This is because LD exists between markers on the genotyping array, meaning the markers are not completely independent of each other, hence reducing the number of tests conducted. A consensus has been made for GWASs by the Welcome Trust Case Control Consortium (WTCCC) where a p-value of less than $5x10^{-7}$ is considered statistically significant [285]. For papers I, II and IV the p-value cut-off ($<5x10^{-7}$) suggested by the WTCCC was used.

Permutations are another way of controlling for multiple testing. In its simplest form one can describe permutations as swapping labels for the cases and controls. The null-hypothesis assumes no differences between the group of cases and group of controls. The permutation procedure swaps the status randomly and repeats the statistical test on the permuted data. This is repeated a specified number of times, generally several tens of thousands. The p-value generated by permutation procedures represents “experiment-wide” significance and is generated by the distribution of the best p-value expected in the entire experiment under the null-hypothesis. To show this with an example, if the nominal p-value is 0.001 and in a permutation procedure using 1000 permutations, a p-value of 0.001 is observed 60 times, then the corrected p-value for the entire experiment is 0.06. Permutations are considered to be robust methods for correcting for multiple testing resulting in a low level of type 1 errors while not reducing the power to detect associations. Also the procedure does not require any prior knowledge of the distribution of the variables and traits under investigation and it is available through several genetic software packages such as PLINK [110]. Permutations were applied in paper IV when examining the differences in IBS between different geographical areas. A disadvantage with permutation procedures is the fact that they are computationally very

intensive. The PLINK software manual gives a brief description of different basic permutation procedures [280].

4.1.4 Genotyping and Errors

Practice makes perfect. However, errors do occur. Random genotyping errors lead to a loss of power to detect true associations [286, 287]. The causes for genotyping errors are numerous; poor or low quality DNA, mutations in the sequences involved in the marker detection, poor probe sequence and human errors are some of them (reviewed in [288]. To ensure high quality and accuracy of the genotyping it is recommended to run duplicates [289]. Depending on the genotyping technology used, this can be done by including a set of positive controls on each plate or re-genotyping a set number of samples.

The calling of the genotypes (reviewed in [137]) could also lead to errors. To avoid these errors several quality control (QC) steps, mainly concerned with removing poor quality SNPs, have been made both in the genome-wide genotyping and the single SNP TaqMan genotyping. Illumina´s Genome Studio allows for manual inspection of genotyping clusters (Figure 12) after automated genotype calling, reducing the calling errors. However, inspecting >300,000 clusters is an impossible task, nor necessary, and Illumina provides a set of check-points to help minimize errors [290]. A contributing factor to genotype calling error or low genotyping rate could be the cluster file used to call the genotypes. A cluster file is supplied by Illumina and defines the location and size of a given genotype cluster. Ideally a new cluster file should be made for the population under study to achieve as high as possible genotype calling. A poor concordance between the samples genotyped and the cluster file used to call the genotypes could lead to a lower genotyping success rate. To further enhance the quality of the genotyping data a series of QC steps are performed in PLINK (papers I, II and IV). This includes excluding individuals with call rate lower than 95%, and SNPs with a genotyping rate less than 95%. Monomorphic SNPs and SNPs with a MAF < 1% are also excluded in addition to SNPs with deviation from Hardy-Weinberg equilibrium (HWE). Regardless of QC steps taken prior to analysis, it is always important to inspect the quality of the clusters for SNPs showing an association with the specific trait investigated.

In paper IV we utilised whole-genome SNP data available from papers I and II. These samples were genotyped at three different locations, SNP & SEQ Technology Platform at Uppsala University, Centre National de Génotypage (CNG) Paris and the Genomic Core Facility at NTNU. Differences in the calling of the genotypes between the different sites could lead to systematic bias [291, 292]. To investigate potential systematic bias, results from the MDS analysis were plotted for all four components and the samples labelled according genotyping location. No clustering according to genotype location was observed.

4.2 Effect Size

An important lesson learnt from GWASs is that most common variants found associated with diseases or traits seem to have a low effect size (OR =< 1.5) [71], which is lower than was initially expected. The belief, based on the CDCV hypothesis was that common genetic variants would be held accountable for a majority of the genetic risk factors for human diseases (reviewed in [83]). In retrospect, it is evident that this is not the complete picture (reviewed in [91], different scenarios for rare and common variants are reviewed in [72]). The first successful GWAS published on AMD is an exception to the rule and has an OR between 2.4 and 7.4[66, 293, 294]. Figure 16 shows the relationship between effect size and allele frequency. A possible scenario is that rare alleles with larger effect size are of more importance in determining the individual susceptibility to disease [72, 74, 81, 295]. Rare variants, CNV or other structural variants could be in low LD with the common allele and being responsible for the signal in GWASs [296]. However thorough re-sequencing and further investigation into the genetic basis of common diseases will be needed in order to unravel this.

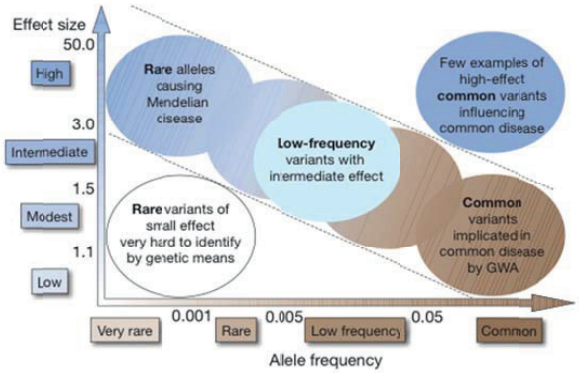


Figure 16. The relation between effect size and frequencies of alleles and the feasibility of identifying genetic variants. Reprinted by permission from Macmillan Publishers Ltd: Nature [74], ©(2009).

4.3 Replication

Independent replication of biological findings is considered an absolute necessity in GWASs and the gold standard for identifying true association signals [297]. The consensus in the field is that the results from a GWAS should not be trusted on their own, and needs additional evidence from independent replication studies that are statistically significant. This is because many do not consider a GWAS a hypothesis test in its own right [297]. A set of guidelines has been made concerning the reporting of genetic associations [292, 297]. David Altshuler and Mark Day [298] list replication as one of the key factors for a successful GWAS.

SNPs differ in allele frequencies between different populations. An associated variant might therefore not be replicable in a second population, and the guidelines outlined by Chanock *et al.* [297] state that failure to replicate in a different population than the initial study should not be regarded as an invalid original finding. Risk variants can be population specific and in this case replication in different populations would fail [299-301]. This can be seen in regards to lung cancer. Four different loci, 3q28, 2129, 13q12.12 and 22.12.2 have been found associated with lung cancer in Asian populations [302-304]. In a recent meta-GWAS neither showed evidence of association in the European population [189]. Similarly, other lung cancer loci showing association in Europeans are not associated with lung cancer risk in Asians [189]. Despite this, it is important to replicate new associations to verify the findings. Follow-up of a

larger number of SNPs might be a way of finding variants not showing genome-wide significance in the initial study [305]. For example, in a large prostate cancer study the four most significant SNPs in the follow-up study was not ranked amongst the top 1,000 SNPs in the initial study [306]. However, at the end of the day fine mapping and mechanistic studies are going to be needed to truly understand the association with the disease or trait in question.

5 MAIN FINDINGS

Paper I

A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25.

A large international GWAS identified a region on chromosome 15q25 in association with the risk of lung cancer. Two SNPs were associated with the risk of lung cancer at the genome-wide level, rs1051730 and rs8034191 ($P = 5x10^{-9}$ and $P = 9x10^{-10}$ respectively) in 1,989 cases and 2,625 controls. These SNPs are located within a region on chromosome 15 containing the nAChR subunits *CHRNA5/A3/B4*. The OR for carrying one copy of the risk allele (C) of rs8034191 was 1.27 (95% CI: 1.11-1.44) and for two copies 1.80 (95% CI: 1.49-2.18). All analyses were adjusted for age, sex and country. In addition, corrections were done for smoking habits, however, this did not change the results. An increase in lung cancer risk was found both in current and former smokers and also in reported never-smokers. The chromosomal region of association was investigated for potential functional variants and a non-synonymous variant (rs16969968) was identified to be in strong LD with rs1051730 and rs8034191. To investigate the specificity of the findings, rs8043191 and rs16969968 were tested for association with cancers of the head and neck. No association was found implying that the association was specific for lung cancer. The findings were replicated in five separate lung cancer studies with an additional 2,513 cases and 4,752 controls.

Paper II

Lung cancer susceptibility locus at 5p15.33.

As an extension of paper I, an additional 1,292 lung cancer cases and 1,561 controls were genotyped to further increase the power to detect association with lung cancer. Genome-wide association analysis adjusted for age, sex and country, was conducted on 3,251 lung cancer cases and 4,159 controls. Eight SNPs exceeded genome-wide significance ($p<5x10^{-7}$) of which seven were located at the 15q25 locus identified in paper I. The most significant SNP was rs1051730 ($p=1x10^{-15}$). A new locus, represented by rs402710 (C allele), on chromosome 5p15.33 also showed genome-wide significance ($p= 2x10^{-7}$). This locus contains two genes,

TERT and *CLPTM1L*. Three additional SNPs, two of which were in strong LD with rs402710 showed marginal significance ($p=5x10^{-6}$). To investigate the association further, rs402710 and rs2736100 ($r^2=0.026$) were genotyped by TaqMan assay in an additional 2,899 lung cancer cases and 5,573 controls. Both SNPs replicated in the independent samples ($p=7x10^5$ for rs402710 and p=0.0016 for rs2736100). The risk allele for rs402710 was the more common C allele with overall OR = 1.4 for homozygous carriers. The risk allele for rs2736100 was the minor G allele with an overall OR= 1.29 for homozygous carriers. The association was found in never, former and current smokers (p= 0.01, p= 0.0007 and p= 0.0001 respectively). Adjustment for smoking exposure did not change the results and no association with smoking intensity was found.

Paper III

Association between 15q25 gene variants, nicotine related habits, lung cancer and COPD in the HUNT study, Norway

Three SNPs (rs16969968, rs1051730 and rs8034191) were genotyped in a large homogenous population cohort, the HUNT-cohort (*n*= 56,307), in an effort to investigate association between the *CHRNA5/A3B4* gene-region, smoking habits and the use of snus, lung cancer and loss of lung function equivalent to that of moderate to severe COPD in more detail. Due to high correlation between the SNPs, only one, rs16969968 was chosen for analysis. A novel association was found between rs16969968 and the use of snus and previously observed associations with lung cancer, COPD and smoking quantity were replicated. No association with lung cancer was found in never-smokers. The novel association with snus showed an increase in snus consumption of approximately 0.51 boxes per month. However, the most interesting association in regards to nicotine addiction was the association between the risk allele (A) of rs16969968 and the motivation behind starting to use snus. It was found that carriers of the risk allele were more likely (OR = 1.17, 95% CI: 1.06-1.29, P= 0.001 per allele) to have started using snus as a means to quit and/or reduce cigarette smoking. This strengthens the possible role for this particular SNP in nicotine addiction.

Paper IV

The genetic structures of stable populations – the HUNT and Tromsø cohorts in Norway.

Genome-wide SNP data from paper I and II was reused to uncover genetic population structures in two large population based health studies, the HUNT cohort in Nord-Trøndelag (n= 884) and the Tromsø cohort in Tromsø (n= 514). Using MDS, population structures were seen both between the two cohorts, and within the HUNT cohort. Even though the differences observed were small, a distinct east-west gradient and north-south gradient could be seen for the samples from the HUNT cohort. Further analyses for the investigation into genetic population structures revealed al larger degree of genetic variation within the Tromsø cohort and subtle differences within the HUNT cohort. To uncover the potential effect on GWASs, a genomic inflation factor, λ, was calculated in simulation experiments assigning different cases control status according to geography. In the worst-case scenario, all cases being from the Tromsø study and all controls from the HUNT study, the genomic inflation factor was above 2. This shows the potential bias from population structures in GWASs if all cases and all controls are selected from different regions, even within one country.

6 DISCUSSION

"The challenges facing the researchers today are at least as daunting as those my colleagues and I faced a decade ago" [7]

Craig Venter 2010

The work presented in this thesis dates back to the beginning of GWASs and was in the planning phase in 2005/2006. Starting up in 2006, this project was planned as a local GWAS using pooled DNA from ~300 lung cancer cases, an equivalent number of controls and the Affymetrix chip available at the time. A follow-up was planned using the TaqMan assay to genotype and investigate further importing findings and also an investigation of SNPs in DNA repair genes. In retrospect one might consider the plans regarding the GWAS a somewhat naïve plan, but with high genotyping costs, limited number of cases and little experience in large scale genomic analyses, this was never the less the starting point. In GWASs very few roads lead to Rome, however, entering into a large international study does in some respect do just that.

6.1 GWASs; What Have We Learnt?

The criticism of GWASs is often directed towards the value that they have produced, especially with regards to the amount of money spent. Three main points are generally criticised; 1) the fact that the detected variants are merely markers of the disease risk and not the true causal variants, 2) GWASs explain a very limited amount of the heritability of the disease or trait and 3) so far they have proved to have limited public health impact [305].

To address the first criticism, the SNP´s identified are merely markers for disease and many of these are outside coding regions or regions with a known biological function. This could seem disappointing, as identifying causal variants that alter a phenotype is the main objective. However, this is based on current knowledge, and large projects such as the ENCODE project might change this in the future. We must also remember that it is really not that long ago that the term "junk-DNA" was frequently heard. However, with basis in the CDCV hypothesis and the elaborate LD structures described by the HapMap project, it was believed that tag-SNPs

[7] Craig Venter 2010, "Multiple personal genomes await", Nature; 464: 676-677

would tag common causal variants in high LD with genotyped SNPs [57]. Many researchers have been opposed to the CDCV hypothesis arguing that common diseases are caused by a large number of rare variants, the CDRV hypothesis [68, 265, 307-309]. In favour of this alternative view, sequencing of susceptibility loci have given little in return as to common causal variants, shedding doubt on the very foundation of GWASs, the CDCV hypothesis [310, 311].

"Ten years and billions of taxpayer dollars later, our once "extreme" position has replaced the mainstream opinion of a decade ago"[8]

McClelland and King argue in a leading edge essay [22] that alleles of significant effect must survive evolutionary forces to persist as polymorphisms in a population and therefore most common variants (though with a number of known exceptions) will be neutral in order not to be diminished [22]. Dickson *et al.* [296] proposed that association signals were due to synthetic associations. A synthetic association is the association to a trait by a common variant due to stochastic association between rare causal variants in the region and the genotyped common allele [296]. Greg Gibson [72] has reviewed the role of rare and common variants and recapitulated them in twenty arguments listing synthetic association as one of the arguments in favour of rare variants. He concludes however that empirically, there is ample support for both classes (rare and common) of effects [72], stating, *"The true debate over the source of genetic variation for disease is not one of "is it caused by rare or common variants?" or even "how much does each class contribute?" but rather "how do they work together?"*

The second disappointment in GWAS, also placing the CDCV hypothesis in a poor light was the realisation that the common variants found, no matter how profoundly replicated, could only explain a small degree of the heritability of the trait. The overall lesson from GWASs has been that common genetic variants explain only 5% of the phenotypic variation [61]. This is disappointingly low considering the expectations. This has led to renewed interest in family-based approaches and linkage studies [266]. Ott *et al.* [266] review the potential of combining family and genome-wide strategies. Family studies have a greater power to detect rare

[8] Terwilliger, J. D. and H. H. Goring (2009). "Update to Terwilliger and Goring's "Gene mapping in the 20th and 21st centuries" (2000): gene mapping when rare variants are common and common variants are rare." Hum Biol 81(5-6): 729-733.

variants than population-based studies given an equivalent sample size [312] because predisposing rare variants will be present at higher frequency in affected relatives [74].

However, it is not all bad news for the common variants. In a study by Yang *et al.* [313] the joint estimate of a large number of common SNPs (294,831) were shown to explain a large proportion of the heredity of height (~45%, using 3,925 unrelated individuals). They replicated their findings in a larger cohort using 14,347 unrelated individuals and 565,040 autosomal SNPs [314]. Here they also estimate heritability for other quantitative traits such as BMI, finding that 17% of autosomal variants can explain the heritability of BMI [314]. They argue that the missing heritability seen in GWASs is due to each variant exerting a small effect rendering it undetectable (in the form of genome-wide significance levels), and incomplete LD between the causal variants and the genotyped variants [313]. The methods used by Yang *et al.* [313] differ from the traditional analysis methods because they do not look for a single association locus but evaluate the contribution of all variants together.

The last point in the chain of criticism is the lack of clinical relevance of the SNPs found to be associated with disease. Despite having identified 1617 published genome-wide associations (GWAs) with a p-value < $5x10^{-8}$ for 249 different traits (as of Sept. 2011) [315] little has been translated into direct clinical relevance for public health. Low heritability and small effect size often seen in GWASs, means assessing common variants for common traits will have little predictive value [22]. An example is a 12 year follow-up of more than 19,000 women for cardiovascular disease. [316]. A risk profile based on known risk factor SNPs for cardiovascular disease had no prediction value in the cohort [316].

"Is the translation of DNA research into medical practice taking longer than expected?"[9]

Despite the limited effect size and predictive value, a myriad of companies are offering personal genome testing over the Internet (two of the largest being 23andMe and deCodeMe; https://www.23andme.com/, http://www.decodeme.com/). These are large array based tests of common SNPs and most of these are based on the result from large GWASs with risk variants with moderate effect size. It is thus argued that they are of little value to an individual's personal health [317-319]. Research based on individuals conducting such a personalised genome tests has identified variants which enables individuals to smell metabolites of

[9] Eliot Marshall, Science 2011; 311: 526-529

asparagus in urine [320], also of little public health impact. However a beneficial and commonly used genetic test (also included by 23andMe) is variation in the *CYP2C19* and *VKORC1* genes. Genetic variants in these genes have been found to be predictive for the response to warfarine, a widely used drug in the prevention of thrombosis and thromboembolism [321].

Green *et al.* [20] describe the transition "from base-pairs to bedside", integrating genetic information into clinical practice, as five domains; understanding the structure of the genome, understanding the biology of the genome, understanding the biology of disease, advancing the science of medicine and finally, improving the effectiveness of healthcare [20]. The timeline suggested by Green *et al*. [20] for these domains stretch well beyond 2020, illustrating the timely task of improving public health and healthcare based on basic scientific research. This is in agreement with the editor of *"Genetics in medicine"* who has said he believe the genomic revolution is going to take decades [322].

6.2 Discussion of Papers

6.2.1 Lung Cancer, COPD and Smoking - papers I-III

Lung cancer and COPD are complex diseases dependent on many genes and environmental factors. Though cigarette smoking is the number one risk factor, there is increasing evidence for a role of inherited genetic factors [140]. Since the initial GWASs on lung cancer [184-186] identified a susceptibility locus at chromosome 15q25, additional susceptibility loci have been identified at chromosome 5p15 [187, 190, 323], 6p21 [190, 323], 22q12 [324, 325], 15q15.2 [326-328], in addition to three loci identified in Asian populations; 13q12.12, 22q12.2 [302] and 3q28 [303].

Susceptibility locus 15q25 - paper I & III

Papers I, II and III included in this thesis are among a number of papers confirming an association between the *CHRNA5/A3/B4* gene cluster on chromosome 15q25, with smoking habits/nicotine addiction (paper III), lung cancer (papers I, II and III) and COPD (paper III). Several SNPs have been reported to be associated with the various phenotypic outcomes

mentioned above, the most studied being rs16969968, rs1051730 and rs8043191. These three SNPs are found in *CHRNA5, CHRNA3* and *AGPHD1,* respectively, and are in high LD with each other (correlation coefficient 0.95-0.99 in the HUNT cohort). From the very beginning, the publication of the first three GWASs [184-186], there has been disagreement on whether the variants identified convey a direct effect on the risk of lung cancer or if they have an indirect effect through an increased risk of ND. Thorgeirsson *et al*. [185] argued already in their original GWAS (2008) and later in a commentary [329] that the association was related to ND. In the 2010 commentary [329] they argued that based on the Doll and Peto equation [330] the low effect size conveyed by the variants could be accounted for by a prolonged duration of smoking. In paper I and III the OR/HR for a homozygous carrier of the risk allele (A) was 1.77 and 2.08 respectively. For the individual this would mean a 77-100% increased risk of lung cancer. Considering the low risk of lung cancer, 3 - 4.5% before the age of 75 in the total population (based on number from the Cancer Registry of Norway http://kreftregisteret.no/en/General/Fakta-om-kreft-test/Lungekreft/ accessed 20.10.2012) even doubling the risk due to genetic factors would be easily outweighed by the increased risk associated with smoking (male current smokers ~15% cumulative risk of death of lung cancer before age 75 compared to 0.2% for never smokers) [145]. However, the genetic contribution is not insignificant and has been estimated to account for 14% (attributable risk) of lung cancer cases (paper I). In paper III we conclude that, at least in regard to rs16969968, the association could be explained through increased smoking, supporting the Stefanssons and Thorgeirssons argument. nAChRs are well known to be involved in ND (reviewed in [331, 332]) and the *CHRNA5/A3/B4* gene cluster was identified in association with ND in both GWAS and candidate gene studies before GWASs found it associated with lung cancer [333, 334]. In paper III we also find an association with the quantity of snus and importantly the motivation for starting to use snus being related to smoking reduction or cessation. These findings strengthen the possible role of rs16969968 in ND.

When discussing the issues regarding the direct or indirect effect of the rs16969968 it is important to keep in mind that the effect sizes of most common SNPs are low and the individual contribution of a single SNP to the trait is in most cases rather small. Neither can one be sure that the associated SNP is the causal SNP. It might be that rs16969968 exerts its effect on the quantity of cigarettes smoked or ND, though other variants might influence the direct effect on lung cancer. In recent years the role of nAChR in carcinogenesis has been

extensively reviewed [251, 335-339]. nAChR have been found to be overexpressed in SCLC [340] and research has also uncovered genetic variants associated with increased expression of *CHRNA5/A3/B4* [341, 342].

In paper III we also find an association between the rs16969968 variant and the risk of COPD. COPD is known to increase the risk of lung cancer and it has been speculated whether COPD has in fact been a confounder in previous studies [343]. Another possibility is the association between rs16969968 and ND and CPD. Cigarette smoking is a common denominator between COPD and lung cancer and an increase in the number of cigarettes smoked or the number of years smoked would increase the risk of both diseases.

In our targeted analysis of the 15q25 region (paper III) the HR and OR is somewhat higher than published elsewhere (HR = 1.45, OR = 1.36 for lung cancer and COPD respectively). A possible explanation for this, not discussed in the paper might be a possible confounder by occupation. Agriculture is the largest profession in Nord-Trøndelag County, and every fourth man-labour year is connected to the farming industry (http://www.bondelaget.no/getfile.php/Bilder%20fylker/Nord%20-%20Tr%C3%B8ndelag/Dokumenter/090310-Brosjyre-Viktigste-N%C3%A6ring-LAVoppl%C3%B8selig.pdf accessed 30.10.2012). It is well acknowledged that farmers have an increased risk of COPD due to exposure in their working environment such as organic dusts (grain, straw, hay), fertilizers and silage [344-346]. It is possible in theory, that an overrepresentation of farmers in our study population, particularly in the case group could introduce a bias and potentially lead to an inflated OR. In our study population (*n*=56,000) > 10,000 individuals report farming as their occupation or the occupation of their spouse. Further research is needed to determine whether this introduces a confounding factor in our data.

Susceptibility locus 5p15 – paper II

The increase in sample size from paper I allowed for the detection of a second locus associated with lung cancer in paper II. As opposed to the 15q25 locus, the 5p15 locus showed association in both smokers and non-smokers indicating a direct effect on lung cancer [187]. The locus has been identified in several other studies [188, 190, 268, 269, 347]. Of the two known genes, *TERT* and *CLPTM1L,* found within the locus, *TERT* is a plausible functional candidate based on

its involvement in lung cancer as mentioned in chapter 1.5.1. However, there is still no evidence for the variants found being causal [168]. It has also been speculated whether the association is independent of *TERT* biology [348]. Zienolddiny *et al.* [348] found that variants in the *TERT-CLPTM1L* locus were associated with higher DNA adduct formation in the lung. The locus has been found associated with a number of different cancer types in a study conducted by deCode genetics [347] thus indicating an impact on cancer aetiology in general.

A recent meta-GWAS gathered data from 16 GWASs totalling 14,900 lung cancer cases and 19,485 controls of European descent confirmed the association with 15q25, 5p15 and 6p21 [189]. The increased power gained from this study enabled the detection of a novel risk locus and the demonstration of histological specificities for 5p15, 6p21 and 12q13 [189]. The 5p15 locus has previously been associated with increased risk of adenocarcinoma [188, 268]. The recent meta-GWAS confirmed this finding and also showed a stronger association in never smokers than ever smokers [189].

It is clear that further research is needed to determine the causal variants at both the 15q25 and 5p15 loci. Moving into an era of whole-genome sequencing will potentially uncover rare variants or family specific variants exerting a larger effect on the disease risk. A growing number of studies have used new sequencing technologies to identify mutations and variants in tumour tissue [175, 177, 349], and a cancer genome for SCLC has been created [177]. Though these studies identify mutations in lung cancer tumours and are not primarily concerned with the risk of developing lung cancer, the increase in knowledge gained through such whole-genome sequencing with respect to the molecular aetiology of the disease is important. Together with GWASs and whole genome sequencing this might allow for a more precise predication of individual risk in the future [19, 350].

6.2.2 Population Structures- paper IV

There has been a tremendous increase in the research on genetic population structures as a result of large scale genome-wide data sets becoming available. Paper IV is a pilot study with regards to population structures in Norway. To our knowledge no research into genetic population structures in Norway has previously been conducted. Norwegian samples have

been part of studies investigating population structures within Europe [95, 96]; however research concerning differences within Norway is lacking. In this pilot study we used the samples available to us through papers I and II. The paper is in some regards a "proof of principle" showing that population structures do exist and that further research is needed to uncover the extent of it.

A very interesting observation besides the population structures seen in paper IV is the extent of hidden and background family relations seen. A large degree (> 80%) is related to at least one other individual to a degree of 4^{th} to 5^{th} degree relative (cousins two to three times removed) (data not shown). Initial analyses show elaborate family structures within the HUNT cohort (Holmen, Kongsgård, Gabrielsen; unpublished). 17,000 families are found with the largest one containing more than 30,000 individuals. The HUNT cohort is considered well suited for genetic studies because of its homogeneity. However, large families increase the risk of selecting close relatives when independent samples for case control studies are what are sought-after. It will be of great importance for large scale genetic studies to have a complete picture of family structures within the HUNT cohort as it represents both a strength and challenge. A challenge, if it becomes custom to exclude individuals down to somewhere between 5 and 10 generations of relatives [313](Håvard Kongsgård personal communication) and a strength as it can allow for detection of larger families which can be combined with traditional GWAS methods.

The findings in paper IV also create new and challenging ethical problems. Especially for the pilot study (paper IV) presented in this thesis, the number of individuals from some geographic areas is rather low and could therefor contribute to inflate some of the values found (especially for ROH and inbreeding factor) based on the fact that by chance, or the degree of hidden relatedness, the individuals from the region are distantly related. When discussing parameters such as ROH and inbreeding factor it is of outmost importance not to create misunderstandings in the general public. Degrees of inbreeding and consanguinity could easily contribute to stigmatisation of a region. It is especially challenging when some of the geographic areas under investigation is scarcely populated. I do not believe however that this should be used as a caution to investigate the HUNT cohort in more detail. On the contrary, I believe including a larger number of individuals will give a more balanced view and diminish the concerns as to burdening regions with unwanted stigma.

Our research in paper IV shows that careful selection of cases and controls is of outmost importance. A study will never be better than its design and avoiding close relatives and an unbalanced geographical distribution when selecting individuals for a study is important. However relying on statistical data is not always accurate enough to avoid selecting close relatives for a study as data from Statistics Norway (ssb.no) diminishes rapidly for individuals born before 1960 (ssb.no). A thorough investigation into the full extent of the population structures and hidden relatedness in the HUNT cohort will be valuable asset for future genetic studies using biological material and data from the cohort.

7 CONCLUDING REMARKS AND FUTURE PERSPECTIVES

"Identification of all DNA variants in the human genome should make it ultimately possible to link all genetic phenotypes to their genic basis. With this understanding it will be possible to diagnose effectively diseases and disease risk, to develop and selectively apply therapeutics to relieve disease symptoms, to treat disease progression and ultimately to prevent disease onset"[10]

I guess we are not quite there yet; though as they say: nothing ventured, nothing gained, and looking back, the expectations were high. Approximately $250 million has been spent on GWASs in the past 5 years [67]. Many will possibly argue that the value for money has been poor and that what we have achieved is merely disproof of the CDCV hypothesis [22, 91]. However, more than 2000 new disease associated loci have been identified, which is substantially more than the candidate gene and linkage studies managed in the time before the GWAS-era. Considering the short time frame, 5-7 years, this should be considered an important achievement [67]. Our research has contributed to the list of these loci through papers I, II and III. Results from the GWASs have contributed to the focus of nAChRs and their role in lung carcinogenesis. Further research into the role of these receptors in lung cancer aetiology could add to a more complete picture of the disease. The other loci identified in association with the risk of lung cancer highlights the diverse nature of cancer risk. Our last paper (paper IV) shed light on the rather unexplored area of genetic population structures in Norway.

The exploration of human genetics and diseases or traits does not end with GWAS. Rather they can be considered important stepping-stones on the road to increased knowledge. The ever reduced prices of sequencing and the 1000 genomes project [351] will allow for large studies to investigate the role of rare variants. What will be important to unravel is the relationship between rare and common variants and how they affect the risk of common diseases and traits. Frazer said in 2009 in regards to common variants that: *It will probably be shown that they* (common variants, editorial note) *do not account for familial concentration of phenotypic traits but rather that they modify the penetrance of causal rare variants with large effect size"*. Today, we cannot answer the question on whether the associations found in GWAS are a result

[10] Schafer, A. J. and J. R. Hawkins (1998). "DNA variation and the future of human genetics." Nat Biotechnol 16(1): 33-39.

of rare variants with larger effect sizes in low LD with the common variant genotyped or more common variants with smaller effect size in high LD with the genotyped marker [67]. However no matter how many GWAS one performs, no matter how statistically significant the results are, I have to agree with a quote in Ledford 2010: *"It´s going to take good old-fashioned biology to really determine what these mutations are doing"* [352]. Though the quote refers to mutations uncovered in the cancer genome project, ultimately the principle will apply to genetic variants as well.

Integrative approaches using several high throughput methods uniting findings not just from genomics but also from epigenomics, transcriptomics and proteomics will be important to obtain a more complete picture [353, 354]. The ENCODE project recently published a landmark article [355]. The ENCODE project was designed to pick up where the HGP left of and aims to characterise all the functional elements of our DNA. This is an enormous task which some say is somewhat of a never-ending story [356]. Together with whole genome sequencing projects, this will bring us closer to the understanding of the human genome. *"First they sequenced it. Now they've surveyed its hinterlands. But no-one knows how much more information the human genome holds, or when to stop looking for it."* [356].

The work involved in this thesis started at a time when hopes were high for GWASs. I will not say that GWASs have failed miserably as considerable scientific knowledge has been gained [67]. In that regard lung cancer and nicotine addiction have been two of the more successful diseases and traits studied by GWASs. However, judging the success or failure of GWASs is more of a personal opinion and somewhat dependent on your initial expectations to them.

In this ever more complex world, no matter what direction we move in, it is unequivocally true that:

"The more we know, the more we realize there is to know"[11]

[11] Jenifer Doudna, biochemist at Uni. California, Berkley, quoted in "Life is Complicated" 2010 by Erika Check Hayden, Nature; 464: 664-667

8 REFERENCES

1. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.
2. *World Health Organization. WHO Report on the Global Tobacco Epidemic, 2008, The MPOWER package*, 2008: Geneva. p. 329.
3. Mendel, J.G., *Versuche über Pflanzenhybriden.* Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, 1866. **Abhandlungen, 3–47**.
4. Watson, J.D. and F.H. Crick, *The structure of DNA.* Cold Spring Harb Symp Quant Biol, 1953. **18**: p. 123-31.
5. Watson, J.D. and F.H.C. Crick, *Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid.* Nature, 1953. **171**(4356): p. 737-738.
6. Franklin, R.E. and R.G. Gosling, *Molecular configuration in sodium thymonucleate.* Nature, 1953. **171**(4356): p. 740-1.
7. Watson, J.D. and F.H. Crick, *Genetical implications of the structure of deoxyribonucleic acid.* Nature, 1953. **171**(4361): p. 964-7.
8. Dahm, R., *Discovering DNA: Friedrich Miescher and the early years of nucleic acid research.* Human Genetics, 2008. **122**(6): p. 565-581.
9. Miescher, F., *Über die chemische Zusammensetzung der Eiterzellen.* Medicinische-chemische Untersuchungen 1871. **4**: p. 441-460.
10. Avery, O.T., C.M. MacLeod, and M. McCarty, *STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES.* The Journal of Experimental Medicine, 1944. **79**(2): p. 137-158.
11. Hershey, A.D. and M. Chase, *INDEPENDENT FUNCTIONS OF VIRAL PROTEIN AND NUCLEIC ACID IN GROWTH OF BACTERIOPHAGE.* The Journal of General Physiology, 1952. **36**(1): p. 39-56.
12. Khorana, H.G., *Synthetic nucleic acids and the genetic code.* JAMA, 1968. **206**(9): p. 1978-82.
13. Friedberg, E.C., *DNA damage and repair.* Nature, 2003. **421**(6921): p. 436-40.
14. Crick, F., *The double helix: a personal view.* Nature, 1974. **248**(5451): p. 766-9.
15. Internation Human Genome Sequencing Consortium, *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
16. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
17. Internation Human Genome Sequencing Consortium, *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-945.
18. *The human genome at ten.* Nature, 2010. **464**(7289): p. 649-650.
19. Lander, E.S., *Initial impact of the sequencing of the human genome.* Nature, 2011. **470**(7333): p. 187-97.
20. Green, E.D. and M.S. Guyer, *Charting a course for genomic medicine from base pairs to bedside.* Nature, 2011. **470**(7333): p. 204-13.
21. Venter, J.C., *Multiple personal genomes await.* Nature, 2010. **464**(7289): p. 676-677.
22. McClellan, J. and M.C. King, *Genetic heterogeneity in human disease.* Cell, 2010. **141**(2): p. 210-7.
23. Cavalli-Sforza, L.L. and M.W. Feldman, *The application of molecular genetic approaches to the study of human evolution.* Nat Genet, 2003. **33 Suppl**: p. 266-75.
24. Cavalli-Sforza, L.L., P. Menozzi, and P. A., *History and Geography of Human Genes*. 1994, Princeton, NJ: Princeton University Press.
25. Xue, Y., et al., *Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree.* Curr Biol, 2009. **19**(17): p. 1453-7.
26. Nachman, M.W. and S.L. Crowell, *Estimate of the mutation rate per nucleotide in humans.* Genetics, 2000. **156**(1): p. 297-304.
27. Krokan, H.E., B. Kavli, and G. Slupphaug, *Novel aspects of macromolecular repair and relationship to human disease.* J Mol Med, 2004. **82**(5): p. 280-97.
28. Kimura, M., *Evolutionary rate at the molecular level.* Nature, 1968. **217**(5129): p. 624-6.
29. *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.

30. Hirszfeld, L. and H. Hirszfeld, *Essau dapplication des methods au probleme des races.* Anthropologie, 1919. **29**: p. 505-537.
31. Eichler, E.E., et al., *Completing the map of human genetic variation.* Nature, 2007. **447**(7141): p. 161-5.
32. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.
33. Kidd, J.M., et al., *Mapping and sequencing of structural variation from eight human genomes.* Nature, 2008. **453**(7191): p. 56-64.
34. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-54.
35. Henrichsen, C.N., E. Chaignat, and A. Reymond, *Copy number variants, diseases and gene expression.* Hum Mol Genet, 2009. **18**(R1): p. R1-8.
36. Shlien, A. and D. Malkin, *Copy number variations and cancer susceptibility.* Curr Opin Oncol, 2010. **22**(1): p. 55-63.
37. *Rare chromosomal deletions and duplications increase risk of schizophrenia.* Nature, 2008. **455**(7210): p. 237-41.
38. Stefansson, H., et al., *Large recurrent microdeletions associated with schizophrenia.* Nature, 2008. **455**(7210): p. 232-6.
39. Marshall, C.R., et al., *Structural variation of chromosomes in autism spectrum disorder.* Am J Hum Genet, 2008. **82**(2): p. 477-88.
40. Sebat, J., et al., *Strong association of de novo copy number mutations with autism.* Science, 2007. **316**(5823): p. 445-9.
41. McCarroll, S.A., et al., *Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease.* Nat Genet, 2008. **40**(9): p. 1107-12.
42. Walsh, T., et al., *Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia.* Science, 2008. **320**(5875): p. 539-43.
43. Levy, S., et al., *The diploid genome sequence of an individual human.* PLoS Biol, 2007. **5**(10): p. e254.
44. Yamaguchi-Kabata, Y., et al., *Distribution and effects of nonsense polymorphisms in human genes.* PLoS ONE, 2008. **3**(10): p. e3393.
45. Sachidanandam, R., et al., *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.* Nature, 2001. **409**(6822): p. 928-33.
46. Lewontin, R.C. and K.-i. Kojima, *The Evolutionary Dynamics of Complex Polymorphisms.* Evolution, 1960. **14**(4): p. 458-472.
47. Gabriel, S.B., et al., *The structure of haplotype blocks in the human genome.* Science, 2002. **296**(5576): p. 2225-9.
48. Abecasis, G.R., et al., *Extent and distribution of linkage disequilibrium in three genomic regions.* Am J Hum Genet, 2001. **68**(1): p. 191-197.
49. Daly, M.J., et al., *High-resolution haplotype structure in the human genome.* Nat Genet, 2001. **29**(2): p. 229-32.
50. Dawson, E., et al., *A first-generation linkage disequilibrium map of human chromosome 22.* Nature, 2002. **418**(6897): p. 544-8.
51. Patil, N., et al., *Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21.* Science, 2001. **294**(5547): p. 1719-23.
52. Phillips, M.S., et al., *Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots.* Nat Genet, 2003. **33**(3): p. 382-7.
53. Reich, D.E., et al., *Linkage disequilibrium in the human genome.* Nature, 2001. **411**(6834): p. 199-204.
54. Taillon-Miller, P., et al., *Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28.* Nat Genet, 2000. **25**(3): p. 324-8.
55. McVean, G., C.C. Spencer, and R. Chaix, *Perspectives on human genetic variation from the HapMap Project.* PLoS Genet, 2005. **1**(4): p. e54.
56. *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
57. Hirschhorn, J.N. and M.J. Daly, *Genome-wide association studies for common diseases and complex traits.* Nat Rev Genet, 2005. **6**(2): p. 95-108.

58. Lander, E.S. and N.J. Schork, *Genetic dissection of complex traits.* Science, 1994. **265**(5181): p. 2037-48.
59. Chial, H., *Mendelian genetics: Patterns of inheritance and single-gene disorders.* Nature Education, 2008. **1**(1).
60. Cooper, D.N. and J. Schmidtke, *Molecular genetic approaches to the analysis and diagnosis of human inherited disease: an overview.* Ann Med, 1992. **24**(1): p. 29-42.
61. Antonarakis, S.E., et al., *Mendelian disorders and multifactorial traits: the big divide or one for all?* Nat Rev Genet, 2010. **11**(5): p. 380-4.
62. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.
63. Collins, F.S., M.S. Guyer, and A. Charkravarti, *Variations on a theme: cataloging human DNA sequence variation.* Science, 1997. **278**(5343): p. 1580-1.
64. Lander, E.S., *The new genomics: global views of biology.* Science, 1996. **274**(5287): p. 536-9.
65. Wang, D.G., et al., *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.* Science, 1998. **280**(5366): p. 1077-82.
66. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.
67. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.
68. Weiss, K.M. and J.D. Terwilliger, *How many diseases does it take to map a gene with SNPs?* Nat Genet, 2000. **26**(2): p. 151-7.
69. Chakravarti, A., *Population genetics--making sense out of sequence.* Nat Genet, 1999. **21**(1 Suppl): p. 56-60.
70. Reich, D.E. and E.S. Lander, *On the allelic spectrum of human disease.* Trends Genet, 2001. **17**(9): p. 502-10.
71. Bodmer, W. and C. Bonilla, *Common and rare variants in multifactorial susceptibility to common diseases.* Nat Genet, 2008. **40**(6): p. 695-701.
72. Gibson, G., *Rare and common variants: twenty arguments.* Nat Rev Genet, 2012. **13**(2): p. 135-145.
73. Maher, B., *Personal genomes: The case of the missing heritability.* Nature, 2008. **456**(7218): p. 18-21.
74. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
75. Zeggini, E., et al., *Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.* Nat Genet, 2008. **40**(5): p. 638-45.
76. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. **9**(5): p. 356-69.
77. Fearnhead, N.S., B. Winney, and W.F. Bodmer, *Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model.* Cell Cycle, 2005. **4**(4): p. 521-5.
78. Cohen, J.C., et al., *Multiple rare alleles contribute to low plasma levels of HDL cholesterol.* Science, 2004. **305**(5685): p. 869-72.
79. Romeo, S., et al., *Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL.* Nat Genet, 2007. **39**(4): p. 513-6.
80. Franke, A., et al., *Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.* Nat Genet, 2010. **42**(12): p. 1118-25.
81. Schork, N.J., et al., *Common vs. rare allele hypotheses for complex diseases.* Curr Opin Genet Dev, 2009. **19**(3): p. 212-9.
82. Baker, M., *Genomics: The search for association.* Nature, 2010. **467**(7319): p. 1135-8.
83. Altshuler, D., M.J. Daly, and E.S. Lander, *Genetic mapping in human disease.* Science, 2008. **322**(5903): p. 881-8.
84. Cambien, F., *Heritability, weak effects, and rare variants in genomewide association studies.* Clin Chem, 2011. **57**(9): p. 1263-6.
85. Kitsios, G.D. and E. Zintzaras, *Genomic convergence of genome-wide investigations for complex traits.* Ann Hum Genet, 2009. **73**(Pt 5): p. 514-9.
86. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease.* N Engl J Med, 2010. **363**(2): p. 166-76.

87. Marian, A.J., *Molecular genetic studies of complex phenotypes.* Transl Res, 2012. **159**(2): p. 64-79.
88. Need, A.C. and D.B. Goldstein, *Whole genome association studies in complex diseases: where do we stand?* Dialogues Clin Neurosci, 2010. **12**(1): p. 37-46.
89. Seng, K.C. and C.K. Seng, *The success of the genome-wide association approach: a brief story of a long struggle.* Eur J Hum Genet, 2008. **16**(5): p. 554-564.
90. Stranger, B.E., E.A. Stahl, and T. Raj, *Progress and promise of genome-wide association studies for human complex trait genetics.* Genetics, 2011. **187**(2): p. 367-83.
91. Terwilliger, J.D. and H.H. Goring, *Update to Terwilliger and Goring's "Gene mapping in the 20th and 21st centuries" (2000): gene mapping when rare variants are common and common variants are rare.* Hum Biol, 2009. **81**(5-6): p. 729-33.
92. Cargill, M., et al., *Characterization of single-nucleotide polymorphisms in coding regions of human genes.* Nat Genet, 1999. **22**(3): p. 231-8.
93. Halushka, M.K., et al., *Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis.* Nat Genet, 1999. **22**(3): p. 239-47.
94. Zietkiewicz, E., et al., *Nuclear DNA diversity in worldwide distributed human populations.* Gene, 1997. **205**(1-2): p. 161-71.
95. Heath, S.C., et al., *Investigation of the fine structure of European populations with applications to disease association studies.* Eur J Hum Genet, 2008. **16**(12): p. 1413-29.
96. Lao, O., et al., *Correlation between genetic and geographic structure in Europe.* Curr Biol, 2008. **18**(16): p. 1241-8.
97. Nelis, M., et al., *Genetic structure of Europeans: a view from the North-East.* PLoS ONE, 2009. **4**(5): p. e5472.
98. Novembre, J., et al., *Genes mirror geography within Europe.* Nature, 2008. **456**(7218): p. 98-101.
99. O'Dushlaine, C.T., et al., *Population structure and genome-wide patterns of variation in Ireland and Britain.* Eur J Hum Genet, 2010. **18**(11): p. 1248-54.
100. Paschou, P., et al., *Tracing sub-structure in the European American population with PCA-informative markers.* PLoS Genet, 2008. **4**(7): p. e1000114.
101. Salmela, E., et al., *Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe.* PLoS ONE, 2008. **3**(10): p. e3519.
102. Tian, C., et al., *Analysis and application of European genetic substructure using 300 K SNP information.* PLoS Genet, 2008. **4**(1): p. e4.
103. Helgason, A., et al., *An Icelandic example of the impact of population structure on association studies.* Nat Genet, 2005. **37**(1): p. 90-5.
104. Huyghe, J.R., et al., *A genome-wide analysis of population structure in the Finnish Saami with implications for genetic association studies.* Eur J Hum Genet, 2011. **19**(3): p. 347-52.
105. Leu, M., et al., *NordicDB: a Nordic pool and portal for genome-wide control data.* Eur J Hum Genet, 2010. **18**(12): p. 1322-6.
106. Price, A.L., et al., *The Impact of Divergence Time on the Nature of Population Structure: An Example from Iceland.* PLoS Genet, 2009. **5**(6): p. e1000505.
107. Salmela, E., et al., *Swedish population substructure revealed by genome-wide single nucleotide polymorphism data.* PLoS ONE, 2011. **6**(2): p. e16747.
108. Patterson, N., A.L. Price, and D. Reich, *Population structure and eigenanalysis.* PLoS Genet, 2006. **2**(12): p. e190.
109. Reich, D., A.L. Price, and N. Patterson, *Principal component analysis of genetic data.* Nat Genet, 2008. **40**(5): p. 491-2.
110. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
111. McQuillan, R., et al., *Runs of homozygosity in European populations.* Am J Hum Genet, 2008. **83**(3): p. 359-72.
112. Nothnagel, M., et al., *Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans.* Hum Mol Genet, 2010. **19**(15): p. 2927-35.
113. Broman, K.W. and J.L. Weber, *Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain.* Am J Hum Genet, 1999. **65**(6): p. 1493-500.

114. Auton, A., et al., *Global distribution of genomic diversity underscores rich complex history of continental human populations.* Genome Res, 2009. **19**(5): p. 795-803.
115. Gibson, J., N.E. Morton, and A. Collins, *Extended tracts of homozygosity in outbred human populations.* Hum Mol Genet, 2006. **15**(5): p. 789-95.
116. Li, L.H., et al., *Long contiguous stretches of homozygosity in the human genome.* Hum Mutat, 2006. **27**(11): p. 1115-21.
117. Hildebrandt, F., et al., *A systematic approach to mapping recessive disease genes in individuals from outbred populations.* PLoS Genet, 2009. **5**(1): p. e1000353.
118. Lander, E.S. and D. Botstein, *Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children.* Science, 1987. **236**(4808): p. 1567-70.
119. Miano, M.G., et al., *Pitfalls in homozygosity mapping.* Am J Hum Genet, 2000. **67**(5): p. 1348-51.
120. Seelow, D., et al., *HomozygosityMapper--an interactive approach to homozygosity mapping.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W593-9.
121. Wang, S., et al., *Genome-wide autozygosity mapping in human populations.* Genet Epidemiol, 2009. **33**(2): p. 172-80.
122. Woods, C.G., et al., *Quantification of homozygosity in consanguineous individuals with autosomal recessive disease.* Am J Hum Genet, 2006. **78**(5): p. 889-96.
123. Curtis, D., A.E. Vine, and J. Knight, *Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations.* Ann Hum Genet, 2008. **72**(Pt 2): p. 261-78.
124. Kirin, M., et al., *Genomic runs of homozygosity record population history and consanguinity.* PLoS ONE, 2010. **5**(11): p. e13996.
125. Hinds, D.A., et al., *Whole-genome patterns of common DNA variation in three human populations.* Science, 2005. **307**(5712): p. 1072-9.
126. Service, S., et al., *Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies.* Nat Genet, 2006. **38**(5): p. 556-60.
127. Campbell, C.D., et al., *Demonstrating stratification in a European American population.* Nat Genet, 2005. **37**(8): p. 868-72.
128. Cardon, L.R. and L.J. Palmer, *Population stratification and spurious allelic association.* Lancet, 2003. **361**(9357): p. 598-604.
129. Freedman, M.L., et al., *Assessing the impact of population stratification on genetic association studies.* Nat Genet, 2004. **36**(4): p. 388-93.
130. Marchini, J., et al., *The effects of human population structure on large genetic association studies.* Nat Genet, 2004. **36**(5): p. 512-7.
131. Pritchard, J.K. and N.A. Rosenberg, *Use of unlinked genetic markers to detect population stratification in association studies.* Am J Hum Genet, 1999. **65**(1): p. 220-8.
132. Reich, D.E. and D.B. Goldstein, *Detecting association in a case-control study while correcting for population stratification.* Genet Epidemiol, 2001. **20**(1): p. 4-16.
133. Wacholder, S., N. Rothman, and N. Caporaso, *Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias.* J Natl Cancer Inst, 2000. **92**(14): p. 1151-8.
134. Yamaguchi-Kabata, Y., et al., *Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies.* Am J Hum Genet, 2008. **83**(4): p. 445-56.
135. Stevens, E.L., et al., *Inference of relationships in population data using identity-by-descent and identity-by-state.* PLoS Genet, 2011. **7**(9): p. e1002287.
136. Voight, B.F. and J.K. Pritchard, *Confounding from cryptic relatedness in case-control association studies.* PLoS Genet, 2005. **1**(3): p. e32.
137. Teo, Y.Y., *Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure.* Curr Opin Lipidol, 2008. **19**(2): p. 133-43.
138. Yu, K., et al., *Population substructure and control selection in genome-wide association studies.* PLoS ONE, 2008. **3**(7): p. e2551.

139. *IARC monographs on the evaluation of carcinogenic risk to humans: Tobacco smoke and invountary smoking.* International Agency for Research on Cancer, Lynon France, 2004. **vol 83**.
140. Brennan, P., P. Hainaut, and P. Boffetta, *Genetics of lung-cancer susceptibility.* Lancet Oncol, 2011. **12**(4): p. 399-408.
141. Curado MP, E.B., Shin HR, Storm H, Ferlay J, Heanue M, Boyle P, *Cancer incidences in five continents, vol IX.* IARC Scientific publications, no 160. Lyon: International Agency for Research on Cancer, 2007.
142. Ferlay J, S.H., Bray F, Forman D, Mathers C, Parkin DM, *GLOBOCAN 2008: cancer incidence and mortality worldwide.* IARC Cancer Base, no 10. Lyon: International Agency for Reseach on Cancer 2010.
143. Crispo, A., et al., *The cumulative risk of lung cancer among current, ex- and never-smokers in European men.* Br J Cancer, 2004. **91**(7): p. 1280-6.
144. Peto, R., et al., *Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies.* BMJ, 2000. **321**(7257): p. 323-9.
145. Brennan, P., et al., *High cumulative risk of lung cancer death among smokers and nonsmokers in Central and Eastern Europe.* Am J Epidemiol, 2006. **164**(12): p. 1233-41.
146. DeMarini, D.M., et al., *Lung tumor KRAS and TP53 mutations in nonsmokers reflect exposure to PAH-rich coal combustion emissions.* Cancer Res, 2001. **61**(18): p. 6679-81.
147. Parkin, M., et al., *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart.*, in *Chapter 1, Tumours of the Lung*, B.E. Travis W.D., Muller-Hermelink H.K., Harris C.C. (Eds.), Editor 2004, IARC Press: Lyon. p. 12-15.
148. *Cancer in Norway 2009. Special issue: Cancer screening in Norway*, T. Haldorsen, Editor 2011, Cancer Registry of Norway: Oslo.
149. Begum, S., *Molecular changes in smoking-related lung cancer.* Expert Rev Mol Diagn, 2012. **12**(1): p. 93-106.
150. Mao, L., *Molecular abnormalities in lung carcinogenesis and their potential clinical implications.* Lung Cancer, 2001. **34 Suppl 2**: p. S27-34.
151. Minna, J.D., J.A. Roth, and A.F. Gazdar, *Focus on lung cancer.* Cancer Cell, 2002. **1**(1): p. 49-52.
152. Yokota, J. and T. Kohno, *Molecular footprints of human lung cancer progression.* Cancer Sci, 2004. **95**(3): p. 197-204.
153. Gazdar, A., et al., *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart.*, in *Chapter 1, Tumours of the Lung*, B.E. Travis W.D., Muller-Hermelink H.K., Harris C.C. (Eds.), Editor 2004, IARC Press: Lyon. p. 21-23.
154. Hussain, S.P., L.J. Hofseth, and C.C. Harris, *Tumor suppressor genes: at the crossroads of molecular carcinogenesis, molecular epidemiology and human risk assessment.* Lung Cancer, 2001. **34 Suppl 2**: p. S7-15.
155. Takahashi, T., et al., *p53: a frequent target for genetic abnormalities in lung cancer.* Science, 1989. **246**(4929): p. 491-4.
156. Hainaut, P. and G.P. Pfeifer, *Patterns of p53 G-->T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke.* Carcinogenesis, 2001. **22**(3): p. 367-74.
157. Le Calvez, F., et al., *TP53 and KRAS mutation load and types in lung cancers in relation to tobacco smoke: distinct patterns in never, former, and current smokers.* Cancer Res, 2005. **65**(12): p. 5076-83.
158. Pfeifer, G.P., et al., *Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers.* Oncogene, 2002. **21**(48): p. 7435-51.
159. Pfeifer, G.P. and P. Hainaut, *On the origin of G --> T transversions in lung cancer.* Mutat Res, 2003. **526**(1-2): p. 39-43.
160. Fukuyama, Y., et al., *K-ras and p53 mutations are an independent unfavourable prognostic indicator in patients with non-small-cell lung cancer.* Br J Cancer, 1997. **75**(8): p. 1125-30.
161. Westra, W.H., *Early glandular neoplasia of the lung.* Respir Res, 2000. **1**(3): p. 163-9.

162. Tam, I.Y., et al., *Distinct epidermal growth factor receptor and KRAS mutation patterns in non-small cell lung cancer patients with different tobacco exposure and clinicopathologic features.* Clin Cancer Res, 2006. **12**(5): p. 1647-53.
163. Zochbauer-Muller, S., et al., *Aberrant promoter methylation of multiple genes in non-small cell lung cancers.* Cancer Res, 2001. **61**(1): p. 249-55.
164. Belinsky, S.A., et al., *Aberrant methylation of p16(INK4a) is an early event in lung cancer and a potential biomarker for early diagnosis.* Proc Natl Acad Sci U S A, 1998. **95**(20): p. 11891-6.
165. Gazzeri, S., et al., *Mechanisms of p16INK4A inactivation in non small-cell lung cancers.* Oncogene, 1998. **16**(4): p. 497-504.
166. Merlo, A., et al., *5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers.* Nat Med, 1995. **1**(7): p. 686-92.
167. Inoue, H., et al., *Sequence of the FRA3B common fragile region: implications for the mechanism of FHIT deletion.* Proc Natl Acad Sci U S A, 1997. **94**(26): p. 14584-9.
168. Baird, D.M., *Variation at the TERT locus and predisposition for cancer.* Expert Rev Mol Med, 2010. **12**: p. e16.
169. Mocellin, S., et al., *Telomerase reverse transcriptase locus polymorphisms and cancer risk: a field synopsis and meta-analysis.* J Natl Cancer Inst, 2012. **104**(11): p. 840-54.
170. Weir, B.A., et al., *Characterizing the cancer genome in lung adenocarcinoma.* Nature, 2007. **450**(7171): p. 893-8.
171. Kang, J.U., et al., *Gain at chromosomal region 5p15.33, containing TERT, is the most frequent genetic event in early stages of non-small cell lung cancer.* Cancer Genet Cytogenet, 2008. **182**(1): p. 1-11.
172. West, K.A., et al., *Rapid Akt activation by nicotine and a tobacco carcinogen modulates the phenotype of normal human airway epithelial cells.* J Clin Invest, 2003. **111**(1): p. 81-90.
173. Govindan, R., et al., *Genomic landscape of non-small cell lung cancer in smokers and never-smokers.* Cell, 2012. **150**(6): p. 1121-34.
174. Kalari, K.R., et al., *Deep Sequence Analysis of Non-Small Cell Lung Cancer: Integrated Analysis of Gene Expression, Alternative Splicing, and Single Nucleotide Variations in Lung Adenocarcinomas with and without Oncogenic KRAS Mutations.* Front Oncol, 2012. **2**: p. 12.
175. Lee, W., et al., *The mutation spectrum revealed by paired genome sequences from a lung cancer patient.* Nature, 2010. **465**(7297): p. 473-7.
176. Liu, J., et al., *Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events.* Genome Res, 2012.
177. Pleasance, E.D., et al., *A small-cell lung cancer genome with complex signatures of tobacco exposure.* Nature, 2010. **463**(7278): p. 184-90.
178. Seo, J.S., et al., *The transcriptional landscape and mutational profile of lung adenocarcinoma.* Genome Res, 2012.
179. Jonsson, S., et al., *Familial risk of lung carcinoma in the Icelandic population.* Jama, 2004. **292**(24): p. 2977-83.
180. Li, X. and K. Hemminki, *Familial and second lung cancers: a nation-wide epidemiologic study from Sweden.* Lung Cancer, 2003. **39**(3): p. 255-63.
181. Amos, C.I., W. Xu, and M.R. Spitz, *Is there a genetic basis for lung cancer susceptibility?* Recent Results Cancer Res, 1999. **151**: p. 3-12.
182. Sellers, T.A., et al., *Evidence for mendelian inheritance in the pathogenesis of lung cancer.* J Natl Cancer Inst, 1990. **82**(15): p. 1272-9.
183. Bartsch, H., et al., *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart.*, in *Chapter 1, Tumours of the Lung*, B.E. Travis W.D., Muller-Hermelink H.K., Harris C.C. (Eds.), Editor 2004, IARC Press: Lyon. p. 24-25.
184. Amos, C.I., et al., *Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1.* Nat Genet, 2008. **40**(5): p. 616-22.
185. Thorgeirsson, T.E., et al., *A variant associated with nicotine dependence, lung cancer and peripheral arterial disease.* Nature, 2008. **452**(7187): p. 638-42.
186. Hung, R.J., et al., *A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25.* Nature, 2008. **452**(7187): p. 633-7.

187. McKay, J.D., et al., *Lung cancer susceptibility locus at 5p15.33.* Nat Genet, 2008. **40**(12): p. 1404-6.
188. Landi, M.T., et al., *A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma.* Am J Hum Genet, 2009. **85**(5): p. 679-91.
189. Timofeeva, M.N., et al., *Influence of Common Genetic Variation on Lung Cancer Risk: Meta-Analysis of 14,900 Cases and 29,485 Controls.* Hum Mol Genet, 2012.
190. Wang, Y., et al., *Common 5p15.33 and 6p21.33 variants influence lung cancer risk.* Nat Genet, 2008. **40**(12): p. 1407-9.
191. Global Initiativ for Chronic Obstructuve Lung Disease, *Global strategy for the diagnosis, management, and prevention of COPD: updated 2010*, 2010.
192. Lopez, A.D. and C.D. Mathers, *Measuring the global burden of disease and epidemiological transitions: 2002-2030.* Ann Trop Med Parasitol, 2006. **100**(5-6): p. 481-99.
193. Jemal, A., et al., *Trends in the leading causes of death in the United States, 1970-2002.* Jama, 2005. **294**(10): p. 1255-9.
194. Eisner, M.D., et al., *An official American Thoracic Society public policy statement: Novel risk factors and the global burden of chronic obstructive pulmonary disease.* Am J Respir Crit Care Med, 2010. **182**(5): p. 693-718.
195. Mannino, D.M. and A.S. Buist, *Global burden of COPD: risk factors, prevalence, and future trends.* Lancet, 2007. **370**(9589): p. 765-73.
196. Gooptu, B., U.I. Ekeowa, and D.A. Lomas, *Mechanisms of emphysema in alpha1-antitrypsin deficiency: molecular and cellular insights.* Eur Respir J, 2009. **34**(2): p. 475-88.
197. Johannessen, A., et al., *Incidence of GOLD-defined chronic obstructive pulmonary disease in a general adult population.* Int J Tuberc Lung Dis, 2005. **9**(8): p. 926-32.
198. Barnes, P.J., S.D. Shapiro, and R.A. Pauwels, *Chronic obstructive pulmonary disease: molecular and cellular mechanisms.* Eur Respir J, 2003. **22**(4): p. 672-88.
199. Pauwels, R.A., et al., *Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary.* Am J Respir Crit Care Med, 2001. **163**(5): p. 1256-76.
200. Opitz, B., et al., *Innate immune recognition in infectious and noninfectious diseases of the lung.* Am J Respir Crit Care Med, 2010. **181**(12): p. 1294-309.
201. Hogg, J.C., et al., *What drives the peripheral lung-remodeling process in chronic obstructive pulmonary disease?* Proc Am Thorac Soc, 2009. **6**(8): p. 668-72.
202. Brusselle, G.G., G.F. Joos, and K.R. Bracke, *New insights into the immunology of chronic obstructive pulmonary disease.* Lancet, 2011. **378**(9795): p. 1015-26.
203. Cosio, M.G., M. Saetta, and A. Agusti, *Immunologic aspects of chronic obstructive pulmonary disease.* N Engl J Med, 2009. **360**(23): p. 2445-54.
204. Yao, H. and I. Rahman, *Current concepts on the role of inflammation in COPD and lung cancer.* Curr Opin Pharmacol, 2009. **9**(4): p. 375-83.
205. Laurell, C. and S. Eriksson, *The electrophoretic pattern alpha I - globulin pattern of serum in alpha I - antitrypsin deficiency.* Scandinavian Journal of Clinical & Laboratory Investigation, 1963. **15**: p. 132-140.
206. Decramer, M., W. Janssens, and M. Miravitlles, *Chronic obstructive pulmonary disease.* Lancet, 2012. **379**(9823): p. 1341-51.
207. Baugh, R.J. and J. Travis, *Human leukocyte granule elastase: rapid isolation and characterization.* Biochemistry, 1976. **15**(4): p. 836-41.
208. Bosse, Y., *Genetics of chronic obstructive pulmonary disease: a succinct review, future avenues and prospective clinical applications.* Pharmacogenomics, 2009. **10**(4): p. 655-67.
209. Kalsheker, N. and S. Chappell, *The new genetics and chronic obstructive pulmonary disease.* COPD, 2008. **5**(4): p. 257-64.
210. Molfino, N.A., *Current thinking on genetics of chronic obstructive pulmonary disease.* Curr Opin Pulm Med, 2007. **13**(2): p. 107-13.
211. Nakamura, H., *Genetics of COPD.* Allergol Int, 2011. **60**(3): p. 253-8.
212. Sandford, A.J., L. Joos, and P.D. Pare, *Genetic risk factors for chronic obstructive pulmonary disease.* Curr Opin Pulm Med, 2002. **8**(2): p. 87-94.

213. Seifart, C. and A. Plagens, *Genetics of chronic obstructive pulmonary disease.* Int J Chron Obstruct Pulmon Dis, 2007. **2**(4): p. 541-50.
214. Silverman, E.K., A. Spira, and P.D. Pare, *Genetics and genomics of chronic obstructive pulmonary disease.* Proc Am Thorac Soc, 2009. **6**(6): p. 539-42.
215. Hancock, D.B., et al., *Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function.* Nat Genet, 2010. **42**(1): p. 45-52.
216. Pillai, S.G., et al., *A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci.* PLoS Genet, 2009. **5**(3): p. e1000421.
217. Saccone, N.L., et al., *Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD.* PLoS Genet, 2010. **6**(8).
218. Wilk, J.B., et al., *A genome-wide association study of pulmonary function measures in the Framingham Heart Study.* PLoS Genet, 2009. **5**(3): p. e1000429.
219. Zhang, J., et al., *Nicotinic acetylcholine receptor variants associated with susceptibility to chronic obstructive pulmonary disease: a meta-analysis.* Respir Res, 2011. **12**: p. 158.
220. Changeux, J.P., *Nicotinic receptors and nicotine addiction.* C R Biol, 2009. **332**(5): p. 421-5.
221. Benowitz, N.L., *Nicotine and coronary heart disease.* Trends Cardiovasc Med, 1991. **1**(8): p. 315-21.
222. Hecht, S.S., *Tobacco smoke carcinogens and lung cancer.* J Natl Cancer Inst, 1999. **91**(14): p. 1194-210.
223. Hecht, S.S., *Progress and challenges in selected areas of tobacco carcinogenesis.* Chem Res Toxicol, 2008. **21**(1): p. 160-71.
224. Hoffmann, D., I. Hoffmann, and K. El-Bayoumy, *The less harmful cigarette: a controversial issue. a tribute to Ernst L. Wynder.* Chem Res Toxicol, 2001. **14**(7): p. 767-90.
225. Hecht, S.S., *DNA adduct formation from tobacco-specific N-nitrosamines.* Mutat Res, 1999. **424**(1-2): p. 127-42.
226. Schuller, H.M., *Nitrosamines as nicotinic receptor ligands.* Life Sci, 2007. **80**(24-25): p. 2274-80.
227. Preussmann, R. and B.W. Stewart, *N-Nitroso carcinogens*. Chemical Carcinogens, ACS Monograph 182, ed. C.E. Searle. Vol. Vol 2. 1984, Washington, DC: American Chemical Society.
228. Hecht, S.S., *Biochemistry, biology, and carcinogenicity of tobacco-specific N-nitrosamines.* Chem Res Toxicol, 1998. **11**(6): p. 559-603.
229. Hecht, S.S., *Recent studies on mechanisms of bioactivation and detoxification of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK), a tobacco-specific lung carcinogen.* Crit Rev Toxicol, 1996. **26**(2): p. 163-81.
230. Hoffmann, D., A. Rivenson, and S.S. Hecht, *The biological significance of tobacco-specific N-nitrosamines: smoking and adenocarcinoma of the lung.* Crit Rev Toxicol, 1996. **26**(2): p. 199-211.
231. Wogan, G.N., et al., *Environmental and chemical carcinogenesis.* Semin Cancer Biol, 2004. **14**(6): p. 473-86.
232. Foulds, J., et al., *Effect of smokeless tobacco (snus) on smoking and public health in Sweden.* Tob Control, 2003. **12**(4): p. 349-59.
233. Daniel Roth, H., A.B. Roth, and X. Liu, *Health risks of smoking compared to Swedish snus.* Inhal Toxicol, 2005. **17**(13): p. 741-8.
234. Holm, H., et al., *Nicotine intake and dependence in Swedish snuff takers.* Psychopharmacology, 1992. **108**(4): p. 507-11.
235. Lee, P.N., *Summary of the epidemiological evidence relating snus to health.* Regul Toxicol Pharmacol, 2011. **59**(2): p. 197-214.
236. Jones, S., S. Sudweeks, and J.L. Yakel, *Nicotinic receptors in the brain: correlating physiology with function.* Trends Neurosci, 1999. **22**(12): p. 555-61.
237. Laviolette, S.R. and D. van der Kooy, *The neurobiology of nicotine addiction: bridging the gap from molecules to behaviour.* Nat Rev Neurosci, 2004. **5**(1): p. 55-65.
238. Mansvelder, H.D. and D.S. McGehee, *Cellular and synaptic mechanisms of nicotine addiction.* J Neurobiol, 2002. **53**(4): p. 606-17.

239. Lindstrom, J., et al., *Structure and function of neuronal nicotinic acetylcholine receptors.* Prog Brain Res, 1996. **109**: p. 125-37.
240. Schuller, H.M. and M. Orloff, *Tobacco-specific carcinogenic nitrosamines. Ligands for nicotinic acetylcholine receptors in human lung cancer cells.* Biochem Pharmacol, 1998. **55**(9): p. 1377-84.
241. Portugal, G.S. and T.J. Gould, *Genetic variability in nicotinic acetylcholine receptors and nicotine addiction: converging evidence from human and animal research.* Behav Brain Res, 2008. **193**(1): p. 1-16.
242. O'Brien, R.D., M.E. Eldefrawi, and A.T. Eldefrawi, *Isolation of acetylcholine receptors.* Annu Rev Pharmacol, 1972. **12**: p. 19-34.
243. Sobel, A., M. Weber, and J.P. Changeux, *Large-scale purification of the acetylcholine-receptor protein in its membrane-bound and detergent-extracted forms from Torpedo marmorata electric organ.* Eur J Biochem, 1977. **80**(1): p. 215-24.
244. Wessler, I. and C.J. Kirkpatrick, *Acetylcholine beyond neurons: the non-neuronal cholinergic system in humans.* Br J Pharmacol, 2008. **154**(8): p. 1558-71.
245. Le Novere, N. and J.P. Changeux, *Molecular evolution of the nicotinic acetylcholine receptor: an example of multigene family in excitable cells.* J Mol Evol, 1995. **40**(2): p. 155-72.
246. Kunzelmann, K., *Ion channels and cancer.* J Membr Biol, 2005. **205**(3): p. 159-73.
247. Roderick, H.L. and S.J. Cook, *Ca2+ signalling checkpoints in cancer: remodelling Ca2+ for cancer cell proliferation and survival.* Nat Rev Cancer, 2008. **8**(5): p. 361-75.
248. Giniatullin, R., A. Nistri, and J.L. Yakel, *Desensitization of nicotinic ACh receptors: shaping cholinergic signaling.* Trends Neurosci, 2005. **28**(7): p. 371-8.
249. Schuller, H.M., *Cell type specific, receptor-mediated modulation of growth kinetics in human lung cancer cell lines by nicotine and tobacco-related nitrosamines.* Biochem Pharmacol, 1989. **38**(20): p. 3439-42.
250. Tsurutani, J., et al., *Tobacco components stimulate Akt-dependent proliferation and NFkappaB-dependent survival in lung cancer cells.* Carcinogenesis, 2005. **26**(7): p. 1182-95.
251. Schuller, H.M., *Is cancer triggered by altered signalling of nicotinic acetylcholine receptors?* Nat Rev Cancer, 2009. **9**(3): p. 195-205.
252. Holmen, J., K. Midthjell, Ø. Krüger, A. Langhammer,T. Lingaas Holmen, G. H. Bratberg, L. Vatten and P. G. Lund-Larsen, *The Nord-Trøndelag Health Study 1995-97 (HUNT 2): Objectives, contents, methods and participation.* Norsk Epidemiol, 2003. **13**(1): p. 19-22.
253. Krokstad S, L.A., Hveem K, Holmen TL, Midthjell K, Stene TR, Bratberg G, Heggland J, Holmen J, *Cohort Profile: The HUNT Study, Norway.* Int J Epidemiol, 2012.
254. Langhammer, A., et al., *Sex differences in lung vulnerability to tobacco smoking.* Eur Respir J, 2003. **21**(6): p. 1017-23.
255. Jacobsen, B.K., et al., *Cohort profile: The Tromso Study.* Int J Epidemiol, 2011.
256. Rostad, H., et al., *[Is the treatment of lung cancer in Norway adequate?].* Tidsskr Nor Laegeforen, 2002. **122**(23): p. 2258-62.
257. Gunderson, K.L., et al., *A genome-wide scalable SNP genotyping assay using microarray technology.* Nat Genet, 2005. **37**(5): p. 549-54.
258. Steemers, F.J., et al., *Whole-genome genotyping with the single-base extension assay.* Nat Methods, 2006. **3**(1): p. 31-3.
259. Barrett, J.C. and L.R. Cardon, *Evaluating coverage of genome-wide association studies.* Nat Genet, 2006. **38**(6): p. 659-62.
260. Livak, K.J., *SNP genotyping by the 5'-nuclease reaction.* Methods Mol Biol, 2003. **212**: p. 129-47.
261. Ragoussis, J., *Genotyping technologies for all.* Drug Discovery Today: Technologies, 2006. **3**(2): p. 115-122.
262. Syvanen, A.C., *Accessing genetic variation: genotyping single nucleotide polymorphisms.* Nat Rev Genet, 2001. **2**(12): p. 930-42.
263. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
264. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.

265. Terwilliger, J.D. and H.H. Goring, *Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design.* Hum Biol, 2000. **72**(1): p. 63-132.
266. Ott, J., Y. Kamatani, and M. Lathrop, *Family-based designs for genome-wide association studies.* Nat Rev Genet, 2011. **12**(7): p. 465-74.
267. Sun, S., J.H. Schiller, and A.F. Gazdar, *Lung cancer in never smokers--a different disease.* Nat Rev Cancer, 2007. **7**(10): p. 778-90.
268. Hsiung, C.A., et al., *The 5p15.33 locus is associated with risk of lung adenocarcinoma in never-smoking females in Asia.* PLoS Genet, 2010. **6**(8).
269. Jin, G., et al., *Common genetic variants on 5p15.33 contribute to risk of lung adenocarcinoma in a Chinese population.* Carcinogenesis, 2009. **30**(6): p. 987-90.
270. Edwards, B.J., et al., *Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies.* BMC Genet, 2005. **6**: p. 18.
271. Zheng, G. and X. Tian, *The impact of diagnostic error on testing genetic association in case-control studies.* Stat Med, 2005. **24**(6): p. 869-82.
272. Hudmon, K.S., et al., *A multidimensional model for characterizing tobacco dependence.* Nicotine Tob Res, 2003. **5**(5): p. 655-64.
273. Piper, M.E., *A multiple motives approach to tobacco dependence: the Wisconsin Inventory of Smoking Dependence Motives (WISDM-68).* J. Consult. Clin. Psychol., 2004. **72**: p. 139-154.
274. Piper, M.E., D.E. McCarthy, and T.B. Baker, *Assessing Tobacco Dependence: A Guide to Measure Evaluation and Selection.* Nicotine & Tobacco Research, 2006. **8**(3): p. 339-351.
275. Shiffman, S., A. Waters, and M. Hickcox, *The nicotine dependence syndrome scale: a multidimensional measure of nicotine dependence.* Nicotine Tob. Res., 2004. **6**: p. 327-348.
276. Heatherton, T.F., et al., *The Fagerstrom test for nicotine dependence: a revision of the Fagerstrom tolerance questionnaire.* Br. J. Addict., 1991. **86**: p. 1119-1127.
277. Slama, K., *Active Smoking.* European Respiratory Monograph, 2000. **5**: p. 305-321.
278. Ioannidis, J.P., et al., *Genetic associations in large versus small studies: an empirical assessment.* Lancet, 2003. **361**(9357): p. 567-71.
279. Spencer, C.C., et al., *Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip.* PLoS Genet, 2009. **5**(5): p. e1000477.
280. Purcell, S., *PLINK (1.07) Docuemntation*, 2010.
281. Concato, J., et al., *Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy.* J Clin Epidemiol, 1995. **48**(12): p. 1495-501.
282. Peduzzi, P., et al., *Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates.* J Clin Epidemiol, 1995. **48**(12): p. 1503-10.
283. Wang, W.Y., et al., *Genome-wide association studies: theoretical and practical concerns.* Nat Rev Genet, 2005. **6**(2): p. 109-18.
284. Nyholt, D.R., *A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other.* Am J Hum Genet, 2004. **74**(4): p. 765-9.
285. *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
286. Gordon, D., et al., *Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms.* Hum Hered, 2002. **54**(1): p. 22-33.
287. Kang, S.J., D. Gordon, and S.J. Finch, *What SNP genotyping errors are most costly for genetic association studies?* Genet Epidemiol, 2004. **26**(2): p. 132-41.
288. Pompanon, F., et al., *Genotyping errors: causes, consequences and solutions.* Nat Rev Genet, 2005. **6**(11): p. 847-59.
289. Laurie, C.C., et al., *Quality control and quality assurance in genotypic data for genome-wide association studies.* Genet Epidemiol, 2010. **34**(6): p. 591-602.
290. Illumina *Infinium genotyping Data Analysis.* Technical Note: DNA analysis.
291. Clayton, D.G., et al., *Population structure, differential bias and genomic control in a large-scale, case-control association study.* Nat Genet, 2005. **37**(11): p. 1243-6.
292. Little, J., et al., *STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement.* PLoS Med, 2009. **6**(2): p. e22.

293. Edwards, A.O., et al., *Complement factor H polymorphism and age-related macular degeneration.* Science, 2005. **308**(5720): p. 421-4.
294. Zareparsi, S., et al., *Strong association of the Y402H variant in complement factor H at 1q32 with susceptibility to age-related macular degeneration.* Am J Hum Genet, 2005. **77**(1): p. 149-53.
295. Cirulli, E.T. and D.B. Goldstein, *Uncovering the roles of rare variants in common disease through whole-genome sequencing.* Nat Rev Genet, 2010. **11**(6): p. 415-25.
296. Dickson, S.P., et al., *Rare variants create synthetic genome-wide associations.* PLoS Biol, 2010. **8**(1): p. e1000294.
297. Chanock, S.J., et al., *Replicating genotype-phenotype associations.* Nature, 2007. **447**(7145): p. 655-60.
298. Altshuler, D. and M. Daly, *Guilt beyond a reasonable doubt.* Nat Genet, 2007. **39**(7): p. 813-5.
299. Liu, Y.J., et al., *Is replication the gold standard for validating genome-wide association findings?* PLoS ONE, 2008. **3**(12): p. e4037.
300. Myles, S., et al., *Worldwide population differentiation at disease-associated SNPs.* BMC Med Genomics, 2008. **1**: p. 22.
301. Shriner, D., et al., *Problems with genome-wide association studies.* Science, 2007. **316**(5833): p. 1840-2.
302. Hu, Z., et al., *A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese.* Nat Genet, 2011. **43**(8): p. 792-6.
303. Miki, D., et al., *Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations.* Nat Genet, 2010. **42**(10): p. 893-6.
304. Yoon, K.A., et al., *A genome-wide association study reveals susceptibility variants for non-small cell lung cancer in the Korean population.* Hum Mol Genet, 2010. **19**(24): p. 4948-54.
305. Witte, J.S., *Genome-wide association studies and beyond.* Annu Rev Public Health, 2010. **31**: p. 9-20 4 p following 20.
306. Thomas, G., et al., *Multiple loci identified in a genome-wide association study of prostate cancer.* Nat Genet, 2008. **40**(3): p. 310-5.
307. Goldstein, D.B., *Common genetic variation and human traits.* N Engl J Med, 2009. **360**(17): p. 1696-8.
308. Pritchard, J.K., *Are rare variants responsible for susceptibility to complex diseases?* Am J Hum Genet, 2001. **69**(1): p. 124-37.
309. Terwilliger, J.D. and T. Hiekkalinna, *An utter refutation of the "fundamental theorem of the HapMap".* Eur J Hum Genet, 2006. **14**(4): p. 426-37.
310. Hafler, J.P., et al., *CD226 Gly307Ser association with multiple autoimmune diseases.* Genes Immun, 2009. **10**(1): p. 5-10.
311. McCarthy, M.I. and J.N. Hirschhorn, *Genome-wide association studies: potential next steps on a genetic journey.* Hum Mol Genet, 2008. **17**(R2): p. R156-65.
312. Li, M., M. Boehnke, and G.R. Abecasis, *Efficient study designs for test of genetic association using sibship data and unrelated cases and controls.* Am J Hum Genet, 2006. **78**(5): p. 778-92.
313. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nat Genet, 2010. **42**(7): p. 565-9.
314. Yang, J., et al., *Genome partitioning of genetic variation for complex traits using common SNPs.* Nat Genet, 2011. **43**(6): p. 519-25.
315. Hindorff, L.A., et al. *A Catalog of Published Genome-Wide Association Studies*. 2011.
316. Paynter, N.P., et al., *Association between a literature-based genetic risk score and cardiovascular events in women.* Jama, 2010. **303**(7): p. 631-7.
317. Bellcross, C.A., P.Z. Page, and D. Meaney-Delman, *Direct-to-Consumer Personal Genome Testing and Cancer Risk Prediction.* Cancer J, 2012. **18**(4): p. 293-302.
318. Hamburg, M.A. and F.S. Collins, *The path to personalized medicine.* N Engl J Med, 2010. **363**(4): p. 301-4.
319. Janssens, A.C. and C.M. van Duijn, *An epidemiological perspective on the future of direct-to-consumer personal genome testing.* Investig Genet, 2010. **1**(1): p. 10.
320. Eriksson, N., et al., *Web-based, participant-driven studies yield novel genetic associations for common traits.* PLoS Genet, 2010. **6**(6): p. e1000993.

321. Wadelius, M. and M. Pirmohamed, *Pharmacogenetics of warfarin: current status and future challenges.* Pharmacogenomics J, 2007. **7**(2): p. 99-111.
322. Marshall, E., *Waiting for the Revolution.* Science, 2011. **331**(6017): p. 526-529.
323. Broderick, P., et al., *Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study.* Cancer Res, 2009. **69**(16): p. 6633-41.
324. Brennan, P., et al., *Uncommon CHEK2 mis-sense variant and reduced risk of tobacco-related cancers: case control study.* Hum Mol Genet, 2007. **16**(15): p. 1794-801.
325. Cybulski, C., et al., *Constitutional CHEK2 mutations are associated with a decreased risk of lung and laryngeal cancers.* Carcinogenesis, 2008. **29**(4): p. 762-5.
326. Rafnar, T., et al., *Genome-wide significant association between a sequence variant at 15q15.2 and lung cancer risk.* Cancer Res, 2011. **71**(4): p. 1356-61.
327. Rudd, M.F., et al., *Variants in the GH-IGF axis confer susceptibility to lung cancer.* Genome Res, 2006. **16**(6): p. 693-701.
328. Truong, T., et al., *Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium.* J Natl Cancer Inst, 2010. **102**(13): p. 959-71.
329. Thorgeirsson, T.E. and K. Stefansson, *Commentary: gene-environment interactions and smoking-related cancers.* Int J Epidemiol, 2010. **39**(2): p. 577-9.
330. Doll, R. and R. Peto, *Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers.* J Epidemiol Community Health, 1978. **32**(4): p. 303-13.
331. Dani, J.A., D. Ji, and F.M. Zhou, *Synaptic plasticity and nicotine addiction.* Neuron, 2001. **31**(3): p. 349-52.
332. Hogg, R.C., M. Raggenbass, and D. Bertrand, *Nicotinic acetylcholine receptors: from structure to brain function.* Rev Physiol Biochem Pharmacol, 2003. **147**: p. 1-46.
333. Bierut, L.J., *Novel genes identified in a high-density genome wide association study for nicotine dependence.* Hum. Mol. Genet., 2007. **16**: p. 24-35.
334. Saccone, S.F., *Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3,713 SNPs.* Hum. Mol. Genet., 2007. **16**: p. 36-49.
335. Catassi, A., et al., *Multiple roles of nicotine on cell proliferation and inhibition of apoptosis: implications on lung carcinogenesis.* Mutat Res, 2008. **659**(3): p. 221-31.
336. Egleton, R.D., K.C. Brown, and P. Dasgupta, *Nicotinic acetylcholine receptors in cancer: multiple roles in proliferation and inhibition of apoptosis.* Trends Pharmacol Sci, 2008. **29**(3): p. 151-8.
337. Improgo, M.R., et al., *From smoking to lung cancer: the CHRNA5/A3/B4 connection.* Oncogene, 2010. **29**(35): p. 4874-84.
338. Improgo, M.R., et al., *The nicotinic acetylcholine receptor CHRNA5/A3/B4 gene cluster: dual role in nicotine addiction and lung cancer.* Prog Neurobiol, 2010. **92**(2): p. 212-26.
339. Improgo, M.R., A.R. Tapper, and P.D. Gardner, *Nicotinic acetylcholine receptor-mediated mechanisms in lung cancer.* Biochem Pharmacol, 2011. **82**(8): p. 1015-21.
340. Improgo, M.R., et al., *ASCL1 regulates the expression of the CHRNA5/A3/B4 lung cancer susceptibility locus.* Mol Cancer Res, 2010. **8**(2): p. 194-203.
341. Smith, R.M., et al., *Nicotinic alpha5 receptor subunit mRNA expression is associated with distant 5' upstream polymorphisms.* Eur J Hum Genet, 2011. **19**(1): p. 76-83.
342. Wang, J.C., et al., *Risk for nicotine dependence and lung cancer is conferred by mRNA expression levels and amino acid change in CHRNA5.* Hum Mol Genet, 2009. **18**(16): p. 3125-35.
343. Young, R.P., et al., *Lung cancer gene associated with COPD: triple whammy or possible confounding effect?* Eur Respir J, 2008. **32**(5): p. 1158-64.
344. *Respiratory health hazards in agriculture.* Am J Respir Crit Care Med, 1998. **158**(5 Pt 2): p. S1-S76.
345. Linaker, C. and J. Smedley, *Respiratory illness in agricultural workers.* Occup Med (Lond), 2002. **52**(8): p. 451-9.
346. Omland, O., *Exposure and respiratory health in farming in temperate zones--a review of the literature.* Ann Agric Environ Med, 2002. **9**(2): p. 119-36.

347. Rafnar, T., et al., *Sequence variants at the TERT-CLPTM1L locus associate with many cancer types.* Nat Genet, 2009. **41**(2): p. 221-7.
348. Zienolddiny, S., et al., *The TERT-CLPTM1L lung cancer susceptibility variant associates with higher DNA adduct formation in the lung.* Carcinogenesis, 2009. **30**(8): p. 1368-71.
349. Ju, Y.S., et al., *A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing.* Genome Res, 2012. **22**(3): p. 436-45.
350. Daniels, M., et al., *Whole genome sequencing for lung cancer.* J Thorac Dis, 2012. **4**(2): p. 155-63.
351. Consortium., G.P., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-1073.
352. Ledford, H., *Big science: The cancer genome challenge.* Nature, 2010. **464**(7291): p. 972-4.
353. Hawkins, R.D., G.C. Hon, and B. Ren, *Next-generation genomics: an integrative approach.* Nat Rev Genet, 2010. **11**(7): p. 476-86.
354. Rakyan, V.K., et al., *Epigenome-wide association studies for common human diseases.* Nat Rev Genet, 2011. **12**(8): p. 529-41.
355. Bernstein, B.E., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
356. Maher, B., *ENCODE: The human encyclopaedia.* Nature, 2012. **489**(7414): p. 46-8.

Yes, yes, I thought it over quite thoroughly,

it is, it´s 42.

PAPERS I-IV

I

Is not included due to copyright

II

Is not included due to copyright

III

Association between a 15q25 gene variant, nicotine related habits, lung cancer and COPD among 56 307 individuals from the HUNT study in Norway

Maiken E. Gabrielsen[1], Pål Romundstad[2], Arnulf Langhammer[2], Hans E. Krokan[1], Frank Skorpen[3]

1. Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

2. Department of Public Health and General Practice, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

3. Department of Laboratory Medicine, Children's and Women's Health, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

Corresponding author: Frank Skorpen, Department of Laboratory Medicine, Children's and Women's Health, Faculty of Medicine, NTNU, PO-Box 8905, N-7491 Trondheim, Norway

Fax: +47 72 57 64 00; e-mail: frank.skorpen@ntnu.no

RUNNING TITLE: CHRNA5 gene polymorphism and nicotine dependence

Keywords: genetic association, lung cancer, nicotine addiction, COPD, snus

Abstract.

Genetic studies have shown an association between single nucleotide polymorphisms on chromosome 15q25 and smoking-related phenotypes such as quantity of smoking, lung cancer and chronic obstructive pulmonary disease. A discussion has centered on the variants and their effects being directly disease related or indirect via nicotine addiction. To address these discrepancies, we genotyped three single nucleotide polymorphisms (rs16969968, rs8034191 and rs1051730) in the *CHRNA5/A3/B4* gene cluster at chromosome 15q25, in 56,307 individuals from a large homogenous population based cohort, The North Trøndelag health study (HUNT) in Norway. Because of high linkage disequilibrium between markers (r^2>0.95), only one marker, rs16969968, was examined further in relation to four different phenotypes: lung cancer, loss of lung function equivalent to that of chronic obstructive pulmonary disease, smoking behaviour, and the use of smokeless tobacco (snus). Novel associations were found between rs16969968 and the motivational factor for starting to use snus, and the quantity of snus used. Our results also confirm and extend previous findings for associations between rs16969968 and lung cancer, loss of lung function equivalent to that of chronic obstructive pulmonary disease, and smoking quantity. Our data suggest a role for rs16969968 primarily in nicotine addiction and the novel association with snus strengthens this conclusion.

Introduction

Tobacco related deaths reached 100 million individuals during the 20th century. It is estimated to reach 1 billion deaths during the 21st century and each year 5.4 million deaths world-wide can be attributed to cigarette smoking [1]. In Norway, a steady decline in daily smoking has been observed since the mid-1990s and to day 17% of the adult population smoke on a daily or occasional basis (http://www.ssb.no/royk/).

Lung cancer and chronic obstructive pulmonary disease (COPD) are both strongly associated with tobacco smoking [2]. Lung cancer is the leading cause of cancer death worldwide with approximately 1.1 million deaths per year[1] , while COPD is the 4th leading cause of death[3], killing 2.75 million people world-wide in 2002[4]. A gene region on chromosome 15q25, containing the nicotine-acetylcholine receptor (nAChRs) subunits *CHRNA5/A3/B4*, has been found to be associated with lung cancer in several genome-wide association studies (GWAS) [5-9] and replication studies [10-12]. GWAS have also shown association with COPD at the same loci[13]. A number of large studies also report an association of this region with smoking related traits and nicotine addiction[14-24]. Associations with several SNPs and distinct loci within the *CHRNA5/A3/B4* region have been reported in these studies[15,16,25].

The *CHRNA5/A3/B4* genes encode subunits of nAChRs. These are ligand gated ion channels classified into two main categories, neuronal and muscular. They are activated both by the endogenous neurotransmitter acetylcholine and chemicals such as nicotine and its metabolites including nicotine specific nitrosamines. The receptors, believed to play a role in nicotine dependence, lead to nicotine-mediated increase of dopamine in the nucleus accumbens (reviewed in [26]). nAChRs have also been found to be expressed in lung tissue where subsequent activation may promote cell proliferation and inhibition of apoptosis [27-29]. Current evidence points to plausible biological associations of nAChR with both nicotine dependence and lung cancer.

In addition to cigarettes, a different nicotine-containing tobacco product, snus (often referred to as Swedish snus), is available in Norway. This is a moist smokeless tobacco product typically placed under the upper lip and kept there (without chewing) [30]. The use of snus in Norway has steadily increased, especially amongst the younger population. Data from Statistics Norway (http://www.ssb.no/royk/) show that around 8% of the adult population use snus on a daily or occasional basis. Among the youngest age group (16-24 years), as many as 25% of males use snus daily. Although snus contains many of the same harmful substances as cigarettes, it is considered less harmful as it does not affect the lungs as cigarettes do, but is believed to be similar in producing nicotine dependence [30].

In this study we report a novel association between the rs16969968 polymorphism in *CHRNA5* and the use of snus. We detect two distinct associations related to the use of snus; one with the quantity of snus used per month and a second to whether the reason for starting to use snus was related to an effort to reduce or quit smoking. We also replicate previously reported associations with the *CHRNA5/A3/B4* gene cluster by examining the rs16969968 polymorphism in relation to lung cancer risk, smoking quantity and loss of lung function equivalent to that of COPD in a large homogenous population cohort (the HUNT cohort of the North Trøndelag County, Norway).

Materials and Methods

Populations studied

The Nord-Trøndelag Health Study (HUNT) is a comprehensive population based study having collected data of the entire adult population aged 20 years or above in three consecutive surveys, HUNT 1 (1984-86), HUNT 2 (1995-97) [31] and HUNT 3 (2006-08) [32]. The studies comprise data from questionnaires, interviews and clinical examinations. All participants in HUNT 2 (n = 65 237) and HUNT 3 (n = 50 807) provided blood samples.

DNA has been prepared from peripheral blood leucocytes from all participants in HUNT 2 and is stored in the HUNT biobank. Approximately 36 000 individuals participated both in the HUNT 2 and HUNT 3 studies (Figure 1) [31,32].

The Lung Study in HUNT invited random samples of participants in HUNT 2 (5%, *n* = 2 791) and HUNT 3 (10%, *n* = 5 068). In addition, participants in the two studies reporting having had asthma, COPD or asthma-related symptoms were invited, totalling 8 150 from HUNT 2 and 7 391 from HUNT 3. All participants were subjected to lung function measurements (spirometry), measurement of bone mineral density, and went through an interview[33].

Phenotype characteristics

Lung cancer phenotype

Lung cancer diagnosis was available from the Cancer Registry of Norway. Data in the Cancer Registry of Norway is based on morphological diagnosis from all pathology departments in Norway and a written report from the clinical departments [34]. Cases were identified by linking the HUNT data-base to the Cancer Registry of Norway via the unique national personal identity number. Only individuals who developed lung cancer after participation in the HUNT 2 study (1995) (Figure 1) and who were diagnosed with lung cancer as the primary tumour were included in the analysis. Only de-identified data were available for researchers.

Loss of lung function phenotype

The loss of lung function phenotype was based on spirometric data from the HUNT 3 lung study (Figure 1). Individuals with loss of lung function equivalent to moderate or severe COPD were identified based on the following standard criteria: prebronchodilator FEV_1/FVC <0.7 and FEV1 % predicted< 80 and/or having received the diagnosis COPD from their medical doctor. Controls were individuals with lung function FEV_1/FVC >0.7 and FEV1 %

predicted> 80. In the present study reference equations developed from the same region was used[35].

Smoking phenotype

Smoking status was categorised into never, former and current smoker based on answers to the HUNT 2 main questionnaire. Never-smokers reported "I have never smoked daily" and had not reported any other smoking related information. Former smokers reported having previously smoked and/or years since smoking cessation, whereas current smokers reported smoking daily and/or reported a number of cigarettes smoked daily. The variable ever-smoker was computed combining current and former smokers. Individuals were also asked to report the number of cigarettes smoked per day or used to smoke per day if quitted smoking. Smoking burden in pack-years was calculated by smoking duration multiplied with daily number of cigarettes divided by twenty.

Snus phenotype

Snus phenotype was categorised into never, former and current users according to answers to the HUNT 3 main questionnaire. Questions on this subject were not included in HUNT 2. Never snus users reported "No, I have never used snus". Former snus users reported having previously used snus, while current snus users reported using snus on a daily or occasional basis. Individuals reporting their age when starting to use snus, snus consumption per month or a motivational factor for starting to use snus were also classified as current snus user. Individuals were also asked to report the number of boxes of snus consumed per month and this variable was used in the snus consumption analysis.

Genotyping

Three SNPs, rs16969968, rs8034191 and rs1051730, from the *CHRNA5/A3/B4* gene cluster on 15q25 were genotyped. All SNPs were genotyped at the HUNT biobank (Levanger,

Norway) using TaqMan genotyping assays (Applied Biosystems, Foster City, CA, USA) and performed on an Applied Biosystems 7900HT Fast real-Time PCR System using 10 ng of genomic DNA. Each 384-well plate contained four negative and four positive controls. Four samples were used as quality controls for genotype consistency and were included on every plate (384-wells) genotyped. The call rate cut-off was set to 90%, and the genotype frequencies were in agreement with HapMap data. Genotyping was performed for all individuals with available DNA (56 307) and laboratory personnel were blinded to any phenotypic status.

Statistical analysis

All analyses were performed in PAWS Statistics 18. Binary outcomes were analysed using logistic regression, continuous outcomes using linear regression and Cox regression was used to estimate hazard ratios (HR) for lung cancer.

Both a genotype specific and per allele model was calculated and adjusted for age and sex, and in an additional model also for cigarettes per day (CPD). In additional analyses we stratified on smoking and sex, and a p-value for trend was calculated for the per allele model. For the Cox-regression analysis the end of follow-up (EOF) date was the 31st of December.2009. "Person-time" was calculated by subtracting the date of participation in HUNT 2 or date of diagnosis for controls and cases respectively from the end of follow-up (EOF) date and dividing the number of days by 365.25. Only ever smokers and snus users were included in the analysis of the smoking and snus phenotype. Heterogeneity between groups was tested by adding an interaction term to a separate regression analysis. A two-sided p-value < 0.05 was considered statistically significant.

Statistical power analysis

A priori power calculations ad modum Lalouel and Rhorwasser [36] for the genotyped SNPs demonstrated > 80% power to detect an effect size (OR) difference of 1.3 for all lung cancer

cases (n=484). A relevant range of minor allele frequencies (38–43%) [National Centre for Biotechnology Information (NCBI) SNP database] was used.

Ethics

This study has been approved by the Regional Committees for Medical and Health Research Ethics (REC). A written consent was signed by all participants in the HUNT study.

Results

There was a very strong correlation (correlation coefficient 0.95-0.99) between rs16969968, rs8034191 and rs1051730 genotypes in the HUNT population. This strongly indicates that these SNPs belong to the same haplotype block. Further analyses were therefore performed on rs16969968 only. Table 1 gives an overall overview of the HUNT 2 and HUNT 3 cohorts included in the present study. Numbers of individuals are given according to genotype, smoking and snus status, lung cancer and loss of lung function.

Lung cancer and loss of lung function

A statistical significant association was found between rs16969968 and the risk of lung cancer in the Cox regression with a hazard ratio (HR) of 1.45 (95% CI: 1.25-1.67, P= 4.60E-07) per allele (A) adjusted for age, sex and CPD (Table 2). Sex was not a significant variable in the regression analysis (Table 2) and when analyses were also run stratified by sex (adjusted for age and CPD), no statistical significant heterogeneity was observed between sexes (P=0.096; data not shown). When stratified by smoking status statistical significant association with lung cancer was seen only in current smokers (HR= 1.51, 95% CI: 1.29-1.77, P= 2.95E-7) while a minor non-significant effect was observed in former smokers (HR=1.25, 95%CI: 0.97-1.62, P= 0.088) and no association was observed in never-smokers. Heterogeneity was observed between the groups (P-het 0.036) (Supplementary table 1).

A statistical significant association was found between the variant allele (A) and the loss of lung function equivalent to COPD (OR= 1.36, 95% CI: 1.19-1.55, P=4.25E-6) (Table 2). Sex was a significant variable in the regression analysis (Table 2). However, when analyses were also run stratified by sex the interaction term was not significant (P= 0.137) (data not shown). A significant association was found in current and former smokers (OR=1.48, 95% CI: 1.25-1.76, P= 4.84E-6 and OR=1.25, 95% CI: 1.06-1.48, P= 0.007, respectively) whereas no association was seen in never-smokers (Supplementary table 2). Heterogeneity was observed between the smoking status groups (P-het. 0.011).

Smoking and snus phenotype

Individuals homozygous for the variant A allele, when compared to non-carriers, smoked on average 1.11 cigarettes more per day (P-trend=3.15E-25) (Table 3), had smoked on average 0.83 years longer (P-trend= 1.11E-6) (Supplementary table 3) and had smoked on average 1.81 pack-years more (P-trend=3.01E-23) when adjusted for age and sex (Supplementary table 4).

A significant association was found between the variant A allele and monthly snus consumption. Individuals homozygous for the A allele used on average 0.51 boxes of snus more per month compared to individuals not carrying the A allele (P-trend= 4.29E-3) (Table 3). Carriers of the A allele were also more likely to have started to use snus in order to reduce or quit smoking (P = 0.001) (Table 4).

Discussion

In this study we demonstrate novel associations between rs16969968 and the reason for starting to use snus being related to smoking reduction or cessation, and to the quantity of snus used. Our results also confirm and extend previous findings of association between the rs16969968 A-allele with increased risk of lung cancer, loss of lung function equivalent to

COPD, and with increased tobacco consumption [10,12,13]. Previous studies have speculated whether the lung cancer association is confounded by COPD. [37]. Due to a limited number of lung cancer cases participating in the HUNT Lung Study, we had insufficient power to detect potential confounding by COPD.

Smoking is the major contributor to the risk of lung cancer. Lips *et al*. [10] argue that the 1.2 CPD increase found for homozygous carriers of the A allele cannot account for the increased risk of lung cancer conveyed by rs16969968. However, an increase in the number of years smoked or pack-years could increase the lung cancer risk substantially more [38]. In the present study, individuals homozygous for the A allele had an average of 1.8 pack-years more and had smoked 0.83 years longer than non-carriers which may contribute substantially to the lung cancer risk.

Previous research has shown that the consumption of snus is associated with an increased probability of being a former smoker [39]. The novel association between the A allele of rs 16969968 and the motivation for starting to use snus being related to smoking reduction or cessation can be seen as a proxy for nicotine dependence as it is likely that individuals with a stronger nicotine dependence substitute cigarettes with smokeless tobacco in order to reduce or quit smoking. A Swedish study from 2003 showed that 30% of former smokers in Sweden used snus while quitting smoking [30]. This fits well with one of the important hallmarks of nicotine addiction, namely the tendency to relapse to tobacco use [40]. The findings in this study strengthen the evidence for an association between rs16969968 and nicotine dependence.

rs16969968 is a non-synonymous SNP, introducing a substitution of aspartic acid (D) with asparagine (N) at amino acid position 398 (D398N) of the CHRNA5 protein. It is a likely candidate to mediate a functional effect, although other SNP variants and haplotypes [20,41-43] in the 15q25 region might modulate the effect on lung cancer and nicotine dependence[15,16,25].

Research by Bierut *et al*. shows that the variant A allele of rs16969968 leads to reduced receptor activity and that individuals carrying the A allele may require larger amounts of nicotine to achieve the same level of dopamine release [15]. This is in concordance with our finding that individuals carrying the A allele tend to smoke more (1.1 CPD more for AA homozygotes) and also continue to smoke for longer period of time (0.83 years for AA homozygotes). This increase in smoking load is likely to greatly increase the risk of lung cancer. Thorgeirsson and Stefansson[38] argue that, based on the Doll–Peto equation, a 5% increase in smoking duration (e.g. from 20 to 21 years) would bring about an ~30% increase in lung cancer risk, strengthening the possible role of the polymorphism in nicotine addiction and smoking behaviour but does not exclude an independent risk on lung cancer in never-smokers.

Based on the findings in this and related studies together with the knowledge of the function of nAChRs it is reasonable to conclude that the SNP rs16969968 has an effect on smoking behaviour linked to nicotine dependence. The increased risk of the A-allele with lung cancer seems in our study to be restricted to current and perhaps former smokers. Even though this is a large population based study, the number of lung cancer patients, especially among never-smokers, is low and gives limited power to detect association in never-smokers as they constitute a minority of the lung cancer patients. However, several larger studies fail to detect an association in never-smokers [12 44 11] and collectively one could possibly argue that the variant allele mediates its effect on lung cancer risk by increasing the tendency to smoke more. During recent years several researchers have investigated or reviewed the role of nAChRs in lung cancer[26,45-49]. Nicotine-derived nitrosamines are capable of activating nAChRs [50] promoting cell proliferation and apoptotic inhibition [51] and both nicotine and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanon (NNK) may stimulate Akt-dependent

proliferation and NFkB-dependent reduction in apoptosis[52,53], providing plausible mechanisms for nicotine and its metabolites to promote disease development

In conclusion, there is convincing evidence that the *CHRNA5/A3/B4* gene cluster plays an important role in both nicotine dependence, lung cancer and the loss of lung function. Our data suggest a role of rs16969968 in nicotine dependence rather than a direct effect on lung cancer risk and loss of lung function. However, as lung cancer is rare in never smokers, this hypothesis is difficult to test and a comprehensive meta-analysis will be required to obtain a sufficient sample size. To uncover the role of the *CHRNA5/A3/B4* gene cluster the genetic variation in this cluster needs to be investigated in more detail possibly by sequencing in order to identify novel variants. To elucidate the role in lung carcinogenesis more functional studies of variant receptors need to be conducted.

Funding

This work was supported by The Norwegian Cancer Society, The Cancer Fund at St. Olavs Hospital and Svanhild and Arne Must Fund for Medical Research.

Acknowledgments:

"The study has used data from the Cancer Registry of Norway. The interpretation and reporting of these data are the sole responsibility of the authors, and no endorsement by the Cancer Registry of Norway is intended nor should be inferred."

The Nord-Trøndelag Health Study (The HUNT Study) is a collaboration between HUNT Research Centre (Faculty of Medicine, Norwegian University of Science and Technology NTNU), Nord-Trøndelag County Council, Central Norway Health Authority, and the Norwegian Institute of Public Health. The lung study in HUNT 2 and 3 received non-demanding funding from AstraZeneca, Norway.

Conflict of interest statement

The authors declare no conflict of interest.

References:

1. WHO Report on the Global Tobacco Epidemic, 2008. The MPOWER package, Geneva, World Health Organization 2008.

2. Wasswa-Kintu S, Gan WQ, Man SFP, Pare PD, Sin DD: Relationship between reduced forced expiratory volume in one second and the risk of lung cancer: a systematic review and meta-analysis. *Thorax* 2005; **60:** 570-575.

3. Global Initiative for Chronic Obstructive Lung Disease : Global strategy for the diagnosis, management, and prevention of COPD: updated 2010, http://www.goldcopd.org/uploads/users/files/GOLD_Pocket_2010Mar31.pdf.

4. Lopez AD, Mathers CD: Measuring the global burden of disease and epidemiological transitions: 2002-2030. *Ann Trop Med Parasitol* 2006; **100:** 481-499.

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

6. Hung RJ: A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008; **452:** 633-637.

7. Thorgeirsson TE, Geller F, Sulem P *et al*: A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008; **452:** 638-642.

8. Wang Y, Broderick P, Webb E *et al*: Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* 2008; **40:** 1407-1409.

9. Broderick P, Wang Y, Vijayakrishnan J *et al*: Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res* 2009; **69:** 6633-6641.

10. Lips EH, Gaborieau V, McKay JD *et al*: Association between a 15q25 gene variant, smoking quantity and tobacco-related cancers among 17 000 individuals. *Int J Epidemiol* 2010; **39:** 563-577.

11. Spitz MR, Amos CI, Dong Q, Lin J, Wu X: The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. *J Natl Cancer Inst* 2008; **100:** 1552-1556.

12. Truong T, Hung RJ, Amos CI *et al*: Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. *J Natl Cancer Inst* 2010; **102:** 959-971.

13. Pillai SG, Ge D, Zhu G *et al*: A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 2009; **5:** e1000421.

14. Bierut LJ, Madden PA, Breslau N *et al*: Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet* 2007; **16:** 24-35.

15. Bierut LJ, Stitzel JA, Wang JC *et al*: Variants in nicotinic receptors and risk for nicotine dependence. *Am J Psychiatry* 2008; **165:** 1163-1171.

16. Saccone NL, Saccone SF, Hinrichs AL *et al*: 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* 2009; **150B:** 453-466.

17. Saccone NL, Schwantes-An TH, Wang JC *et al*: Multiple cholinergic nicotinic receptor genes affect nicotine dependence risk in African and European Americans. *Genes Brain Behav* 2010; **9:** 741-750.

18. Saccone NL, Wang JC, Breslau N *et al*: The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans. *Cancer Res* 2009; **69:** 6848-6856.

19. Saccone SF, Hinrichs AL, Saccone NL *et al*: Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet* 2007; **16:** 36-49.

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

21. Ware JJ, van den Bree MB, Munafo MR: Association of the CHRNA5-A3-B4 gene cluster with heaviness of smoking: a meta-analysis. *Nicotine Tob Res* 2011; **13:** 1167-1175.

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

23. Liu JZ, Tozzi F, Waterworth DM *et al*: Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010; **42:** 436-440.

24. Thorgeirsson TE, Gudbjartsson DF, Surakka I *et al*: Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* 2010; **42:** 448-453.

25. Saccone NL, Culverhouse RC, Schwantes-An TH *et al*: Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet* 2010; **6**.

26. Improgo MR, Scofield MD, Tapper AR, Gardner PD: From smoking to lung cancer: the CHRNA5/A3/B4 connection. *Oncogene* 2010; **29:** 4874-4884.

27. Dasgupta P, Chellappan SP: Nicotine-mediated cell proliferation and angiogenesis: new twists to an old story. *Cell Cycle* 2006; **5:** 2324-2328.

28. Maneckjee R, Minna JD: Opioids induce while nicotine suppresses apoptosis in human lung cancer cells. *Cell growth & differentiation : the molecular biology journal of the American Association for Cancer Research* 1994; **5:** 1033-1040.

29. Wright SC, Zhong J, Zheng H, Larrick JW: Nicotine inhibition of apoptosis suggests a role in tumor promotion. *FASEB J* 1993; **7:** 1045-1051.

30. Foulds J, Ramstrom L, Burke M, Fagerstrom K: Effect of smokeless tobacco (snus) on smoking and public health in Sweden. *Tob Control* 2003; **12:** 349-359.

31. Holmen J, K. Midthjell, Ø. Krüger, A. Langhammer,T. Lingaas Holmen, G. H. Bratberg, L. Vatten and P. G. Lund-Larsen: The Nord-Trøndelag Health Study 1995-97 (HUNT 2): Objectives, contents, methods and participation. *Norsk Epidemiol* 2003; **13:** 19-22.

32. Krokstad S LA, Hveem K, Holmen TL, Midthjell K, Stene TR, Bratberg G, Heggland J, Holmen J: Cohort Profile: The HUNT Study, Norway. *Int J Epidemiol* 2012.

33. Langhammer A, Johnsen R, Gulsvik A, Holmen TL, Bjermer L: Sex differences in lung vulnerability to tobacco smoking. *Eur Respir J* 2003; **21:** 1017-1023.

34. Rostad H, Naalsund A, Norstein J, Jacobsen R, Aalokken TM: [Is the treatment of lung cancer in Norway adequate?]. *Tidsskr Nor Laegeforen* 2002; **122:** 2258-2262.

35. Langhammer A, Johnsen R, Gulsvik A, Holmen TL, Bjermer L: Forced spirometry reference values for Norwegian adults: the Bronchial Obstruction in Nord-Trondelag Study. *Eur Respir J* 2001; **18:** 770-779.

36. Lalouel JM, Rohrwasser A: Power and replication in case-control studies. *Am J Hypertens* 2002; **15:** 201-205.

37. Young RP, Hopkins RJ, Hay BA, Epton MJ, Black PN, Gamble GD: Lung cancer gene associated with COPD: triple whammy or possible confounding effect? *Eur Respir J* 2008; **32:** 1158-1164.

38. Thorgeirsson TE, Stefansson K: Commentary: gene-environment interactions and smoking-related cancers. *Int J Epidemiol* 2010; **39:** 577-579.

39. Lund KE, Scheffels J, McNeill A: The association between use of snus and quit rates for smoking: results from seven Norwegian cross-sectional studies. *Addiction* 2011; **106:** 162-167.

40. Piper ME, McCarthy DE, Baker TB: Assessing Tobacco Dependence: A Guide to Measure Evaluation and Selection. *Nicotine & Tobacco Research* 2006; **8:** 339-351.

41. Baker TB, Weiss RB, Bolt D *et al*: Human neuronal acetylcholine receptor A5-A3-B4 haplotypes are associated with multiple nicotine dependence phenotypes. *Nicotine Tob Res* 2009; **11:** 785-796.

42. Berrettini W, Yuan X, Tozzi F *et al*: Alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol Psychiatry* 2008; **13:** 368-373.

43. Hansen HM, Xiao Y, Rice T *et al*: Fine mapping of chromosome 15q25.1 lung cancer susceptibility in African-Americans. *Hum Mol Genet* 2010; **19:** 3652-3661.

44. Wang Y, Broderick P, Matakidou A, Eisen T, Houlston RS: Chromosome 15q25 (CHRNA3-CHRNA5) variation impacts indirectly on lung cancer risk. *PLoS ONE* 2011; **6:** e19085.

45. Catassi A, Servent D, Paleari L, Cesario A, Russo P: Multiple roles of nicotine on cell proliferation and inhibition of apoptosis: implications on lung carcinogenesis. *Mutat Res* 2008; **659:** 221-231.

46. Egleton RD, Brown KC, Dasgupta P: Nicotinic acetylcholine receptors in cancer: multiple roles in proliferation and inhibition of apoptosis. *Trends Pharmacol Sci* 2008; **29:** 151-158.

47. Improgo MR, Scofield MD, Tapper AR, Gardner PD: The nicotinic acetylcholine receptor CHRNA5/A3/B4 gene cluster: dual role in nicotine addiction and lung cancer. *Prog Neurobiol* 2010; **92:** 212-226.

48. Improgo MR, Schlichting NA, Cortes RY, Zhao-Shea R, Tapper AR, Gardner PD: ASCL1 regulates the expression of the CHRNA5/A3/B4 lung cancer susceptibility locus. *Molecular cancer research : MCR* 2010; **8:** 194-203.

49. Improgo MR, Tapper AR, Gardner PD: Nicotinic acetylcholine receptor-mediated mechanisms in lung cancer. *Biochem Pharmacol* 2011; **82:** 1015-1021.

50. Schuller HM, Orloff M: Tobacco-specific carcinogenic nitrosamines. Ligands for nicotinic acetylcholine receptors in human lung cancer cells. *Biochem Pharmacol* 1998; **55:** 1377-1384.

51. Schuller HM: Is cancer triggered by altered signalling of nicotinic acetylcholine receptors? *Nat Rev Cancer* 2009; **9:** 195-205.

52. Tsurutani J, Castillo SS, Brognard J *et al*: Tobacco components stimulate Akt-dependent proliferation and NFkappaB-dependent survival in lung cancer cells. *Carcinogenesis* 2005; **26:** 1182-1195.

53. West KA, Brognard J, Clark AS *et al*: Rapid Akt activation by nicotine and a tobacco carcinogen modulates the phenotype of normal human airway epithelial cells. *J Clin Invest* 2003; **111:** 81-90.

Titles and legends to figures

Figure 1

Flow-chart visualising the number of individuals for the different phenotypes selected from the HUNT 2 and HUNT 3 study.

Figure 1

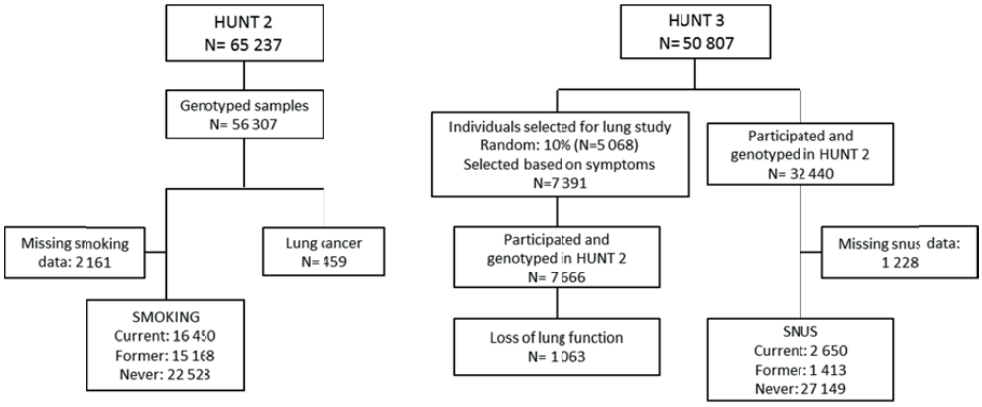

HUNT 2
N= 65 237
Genotyped samples
N= 56 307
Missing smoking data: 2 161
Lung cancer N= 459
SMOKING
Current: 16 450
Former: 15 168
Never: 22 523
HUNT 3
N= 50 807
Individuals selected for lung study
Random: 10% (N=5 068)
Selected based on symptoms
N=7 391
Participated and genotyped in HUNT 2
N= 32 440
Participated and genotyped in HUNT 2
N= 7 566
Missing snus data: 1 228
Loss of lung function
N= 1 063
SNUS
Current: 2 650
Former: 1 413
Never: 27 149

Table 1. Characteristics of the HUNT 2 and HUNT 3 population, overall and per genotype. Questions on the use of snus and spirometry data was available from the HUNT 3 study only. The % of smoking and snus status is calculated from their total population (*n*) respectively. Missing data for smoking and use of snus was 3.8%.

				Genotype	
		Overall	**GG**	**GA**	**AA**
HUNT 2	*n*	56 307	24 800	25 035	6 472
	Males (%)	26 839 (47.7)	11 868 (44.2)	11 914 (44.4)	3 057 (11.4)
	Female (%)	29 468 (52.3)	12 932 (43.9)	13 121 (44.5)	3 405 (11.6)
	Mean age	50.0	50.0	49.9	49.6
Smokers	Never (%)	22 528 (40.0)	10 032 (44.5)	9 933 (44.1)	2 563 (11.4)
	Former (%)	15 168 (26.9)	6 848 (45.2)	6 665 (43.9)	1 655 (10.9)
	Current (%)	16 450 (29.2)	6 919 (42.1)	7 491 (45.5)	2 040 (12.4)
Lung cancer	Cases (%)	459	155 (33.8)	227 (49.4)	77 (16.8)
	Controls (%)	55 823	24 634 (44.1)	24 798 (44.4)	6 391 (11.5)
HUNT 3	*n*	32 440	14 384	14 372	3 684
	Males (%)	14 775 (45.5)	6 548 (44.3)	6 551 (44.3)	1 676 (11.3)
	Females (%)	17 665 (54.5)	7 836 (44.4)	7 821 (44.3)	2008 (11.4)
	Mean age	58.2	58.4	58.2	57.7
Snus users	Never (%)	27 149 (83.7)	12 024 (44.3)	12 045 (44.4)	3 080 (11.3)
	Former (%)	1 413 (4.4)	623 (44.1)	634 (44.9)	156 (11.0)
	Current (%)	2 650 (8.2)	1 195 (45.1)	1 134 (42.8)	321 (12.1)
Loss of lung function	Cases (%)	1 063	412 (38.8)	499 (46.9)	152 (14.3)
	Controls (%)	5 301	2 420 (45.6)	2 289 (43.2)	592 (11.2)

Table 2. Hazard ratio (HR) for lung cancer and Odds ratio (OR) for of loss of lung function equivalent to COPD according to rs16969968 allele distribution.

Lung cancer	Case subjects	Control subjects	HR unadj[a]	HR[b]	95.0% CI for HR		P-value
					Lower	Upper	
rs16969968 per allele	383[d]	28 369[d]	1.48	1.45	1.25	1.67	4.60E-07
rs16969968 GG	125[d]	12 386[d]	Ref	Ref	-	-	3.30E-06
rs16969968 GA	189[d]	12 685[d]	1.51	1.47	1.17	1.84	8.85E-04
rs16969968 AA	69[d]	3 298[d]	2.16	2.08	1.55	2.79	1.07E-06
Sex[c]				0.94	0.760	1.169	0.592
Age				1.06	1.049	1.064	1.67E-51
CPD				1.03	1.022	1.046	5.87E-09
Loss of lung function	**Case subjects**	**Control subjects**	**OR unadj[a]**	**OR[b]**	**95% C.I. for OR**		**P-value**
					Lower	**Upper**	
rs16969968 per allele	715[e]	2 253[e]	1.38	1.36	1.19	1.55	4.25E-06
rs16969968 GG	264[e]	1 018[e]	Ref	Ref	-	-	2.47E-05
rs16969968 GA	335[e]	986[e]	1.37	1.35	1.11	1.64	0.003
rs16969968 AA	116[e]	249[e]	1.90	1.86	1.41	2.46	1.09E-05
Sex[c]				1.41	1.17	1.70	3.00E-04
Age				1.07	1.06	1.08	3.94E-58
CPD				1.03	1.02	1.05	1.27E-07

Complete Cox regression analysis model for lung cancer and logistic regression analysis for loss of lung function.

[a] Adjusted for sex and age only.

[b] HR and OR are adjusted for age, sex and CPD.

[c] Reference sex is female.

[d] Only individuals with valid data for smoking quantity (CPD) were included in the analysis.

[e] Only individuals with valid person-time were included in the analysis.

Sex, age and CPD are covariates in the regression analysis. The p-value shows whether these variables are significant in the model and the HR and OR, respectively, show their contribution to disease risk.

Table 3. Association of rs16969968 with smoking quantity in CPD and snus quantity in boxes per month (BPM)

SMOKING QUANTITY			95% CI mean CPD		
Genotype	***n*[a]**	**Mean CPD**	**Lower**	**Upper**	**P-trend**
GG	12 520	11.02	10.91	11.13	
GA	12 882	11.66	11.55	11.77	
AA	3 370	12.13	11.92	12.35	
Abs diff between homozygous		1.11			3.15E-25
By sex					
Men					
GG	6 468	12.60	12.42	12.78	
GA	6 504	13.29	13.11	13.47	
AA	1 698	13.78	13.43	14.14	
Abs diff between homozygous		1.18			8.11E-12
Women					
GG	6 052	9.38	9.25	9.51	
GA	6 378	9.98	9.85	10.10	
AA	1 672	10.42	10.18	10.67	
Abs diff between homozygous		1.04			2.79E-17
SNUS QUANTITY			**95% CI mean BPM**		
Genotype	***n*[b]**	**Mean BPM**	**Lower**	**Upper**	**P-trend**
GG	1 662	5.34	5.12	5.58	
GA	1 606	5.87	5.60	6.13	
AA	438	5.85	5.37	6.30	
Abs diff between homozygous		0.51			4.29E-03
By sex					
Men	3 341	5.82	5.65	5.99	5.91E-03
Women	365	3.91	3.44	4.39	0.462

Multiple linear regression model for CPD and snus used in boxes per month, means are adjusted for age and sex.

[a] Only individuals with valid data for smoking quantity (CPD) were included.

[b] Only individuals with valid data for snus quantity used per month were included.

Table 4. Motivation for start using snus related to smoking, yes/no

Genotype	*n*No (%)	*n*Yes (%)	OR	95% CI for OR		P-value
				Lower	Upper	
Per allele (A)	2 383	1 560	1.17	1.06	1.29	0.001
GG	1 083 (45.4)	641 (41.1)	Ref	-		-
GA	1 050 (44.1)	702 (45.0)	1.09	0.95	1.26	0.218
AA	250 (10.5)	217 (13.9)	1.46	1.18	1.81	4.55E-04

Logistic regression analysis for the association between rs16969968 and the motivation behind starting to use snus was performed adjusted for age and sex. Only individuals reporting a motivational factor for starting to use snus were included in the analysis.

Supplementary tables

S1. Relative risk for lung cancer (HR) according to the A allele of rs16969968 stratified by smoking status

	Case subjects	Control subjects	HR	95% CI for HR		P-value
				Lower	Upper	
Never (per allele)	22	22 504	0.74	0.38	1.41	0.382
Former (per allele)	122	15 027	1.25	0.97	1.62	0.088
Current (per allele)	313	16 134	1.51	1.29	1.77	2.95E-07

Cox regression stratified by smoking status (P-het.= 0.036). Only individuals with valid smoking status and person-time were included in the analysis.

S2 Logistic regression for the association between the A allele of rs16969968 and loss of lung function equivalent to COPD, stratified by smoking status

	Case subjects	Control subjects	OR	95% CI for OR		P-value
				Lower	Upper	
Never (per allele)	148	2 189	0.98	0.76	1.27	0.88
Former (per allele)	443	1 661	1.25	1.06	1.48	0.007
Current (per allele)	451	1 367	1.48	1.25	1.76	4.84E-06

Stratified by smoking status (P-het = 0.018). Only individuals with valid smoking status were included in the analysis.

S3. Number of years smoked

Genotype	*n*	Mean no. of years smoked	95% CI		P-trend
			Lower	Upper	
GG	13 407	22.67	22.49	22.84	
GA	13 795	23.12	22.94	23.29	
AA	3 601	23.50	23.16	23.84	
Abs diff between homozygous		0.83			1.11E-06

Multiple linear regression model for number of years smoked, means were adjusted for age and sex. Only individuals with valid data for the number of years smoked were included in the analysis

S4. Association of rs16969968 with pack-years

Genotype	n	Mean	95% CI		P-trend
			Lower	Upper	
GG	12 404	12.69	12.51	12.88	
GA	12 760	13.72	13.53	13.9	
AA	3 339	14.50	14.14	14.86	
Abs diff between homozygous		1.81			3.01E-23
By sex					
Men					
GG	6 438	15.33	15.03	15.84	
GA	6 454	16.53	16.22	16.84	
AA	1 690	17.48	16.88	18.08	
Abs diff between homozygous		2.15			5.95E-13
Women					
GG	5 966	9.91	9.70	10.12	
GA	6 306	10.78	10.58	10.98	
AA	1 649	11.42	11.02	11.81	
Abs diff between homozygous		1.51			2.80E-14

Generalised linear model for smoking quantity in pack-years, means are adjusted for age and sex. Pack-years were calculated as follow: (CPD × number of years smoked)/20. Only individuals with valid data on number of years smoked and CPD were included in the analysis.

IV

Is not included due to copyright