From
DEPARTMENT OF CLINICAL SCIENCE,
INTERVENTION AND TECHNOLOGY
DIVISION OF EAR, NOSE AND THROAT-DISEASES
Karolinska Institutet, Stockholm, Sweden

GENETIC ASSOCIATION STUDIES IN ALLERGIC RHINITIS

Daniel Carlberg

Karolinska Institutet

Stockholm 2014

Published by Karolinska Institutet.
Printed by Media-Tryck

ISBN 978-91-7549-756-3

Genetic association studies in allergic rhinitis

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i Petrénsalen, Nobels väg 12B, KI Solna

Torsdagen den 11 december 2014, kl. 13.15

av

Daniel Carlberg

Huvudhandledare:
Professor Lars-Olaf Cardell
Karolinska Institutet
Department of Clinical Science, Intervention and Technology
Division of ENT Diseases

Bihandledare:
Professor Christer Halldén
Kristianstad University
Biomedicine
Division of Education and Environment

Fakultetsopponent:
Docent Pär-Ola Bendahl
Lund University
Department of Clinical Sciences
Division of Oncology

Betygsnämnd:
Docent Karin Broberg Palmgren
Karolinska Institutet
Institute of Environmental Medicine

Professor Arne Egesten
Skåne University Hospital in Lund
Department of Clinical Sciences
Division of Respiratory Medicine and Allergology

Professor Åke Davidsson
Örebro University
School of Medicine

ABSTRACT

Allergic rhinitis (AR) is a global health problem that causes major disability worldwide. Nasal obstruction, secretion and itching are characterizing features of the disease. The development and severity of AR are determined by a complex interaction between environmental and genetic factors and the heritability for AR has been estimated to be high. Genetic association studies are commonly used to investigate complex disease. Most association studies have focused on common variation in the form of single nucleotide polymorphisms (SNPs), where allele and genotype frequencies are compared between cases and controls. More than 100 SNPs have previously been associated with AR. In the first study of this thesis, the general reproducibility of previous AR associations were evaluated. The overall result showed that very few of the investigated SNPs were associated with AR in the study populations. The study also indicated that odds ratios were inflated in most of the original studies, and concordance of risk alleles between the studies was low. Since the genetic background of asthma has been far more investigated compared to the genetics of AR and since asthma is closely related to AR, the second study investigated genetic variation in 21 highly replicated asthma-associated genes for associations also in AR. Only two genes were identified as potential links between AR and asthma, indicating the different genetic make-up of these diseases. Three meta studies of genome-wide association studies (mGWAS) recently identified a total of 47 index SNPs associated with different AR phenotypes. In the third study, these SNPs were investigated in a replication study using the same SNPs and the same phenotype definitions as in the mGWAS. Two out of four loci (*TLR1-TLR6* and *HLA-DQA1-HLA-DQB1*) identified by all three original studies were also detected in the replication study. This suggests a central role of these loci in the epidemiology of allergic disease. In addition, associations between genetic variation in the *SSTR1-MIPOL1* and *TSLP-SLC25A46* loci and age at which the allergic symptoms started was also identified. This was the first report of age at onset effects in AR. The Toll-like receptors (TLRs) have earlier been investigated for their involvement in different allergic diseases. In the fourth study, common genetic variation in the TLR genes was investigated for

association with AR in two ethnically different populations. The *TLR7-TLR8* locus was identified as associated with AR in both populations in a sex-specific manner. In addition, weak associations were also observed for *TLR1* and *TLR6*. In the fifth study of this thesis, rare variation in both the coding sequences and the putative promoter regions of the TLR genes were investigated using sequence data from 288 AR patients and a European subset (EUR) of the 1000 Genomes project. The promotor region of *TLR10* showed a significant accumulation of SNPs with minor allele frequency ≤ 1% in the AR population compared to the EUR population. Another potential accumulation was a nonsense mutation, S324* in *TLR1*, estimated to 5 copies in the AR population but none in the EUR populations. This indicates that both common and rare SNPs in the TLR genes contribute to AR. In summary, this thesis demonstrates that both rare and common variation are associated with AR and highlights the importance of the *TLR10-TLR1-TLR6* locus in the development of the disease. It also emphazises the need for large and well-characterized populations in association and replication studies.

LIST OF SCIENTIFIC PAPERS

I. Nilsson D, Andiappan AK, Halldén C, Tim CF, Säll T, Wang De Y, Cardell LO.
Poor reproducibility of allergic rhinitis SNP associations.
PLoS ONE 2013; 8:e53975.

II. Andiappan AK, Nilsson D, Halldén C, Yun WD, Säll T, Cardell LO, Tim CF.
Investigating highly replicated asthma genes as candidate genes for allergic rhinitis.
BMC Med Genet 2013; 14:51.

III. Nilsson D, Henmyr V, Halldén C, Säll T, Kull I, Wickman M, Melén E, Cardell LO.
Replication of genomewide associations with allergic sensitization and allergic rhinitis.
Allergy 2014; 69:1506-14.

IV. Nilsson D, Andiappan AK, Halldén C, Yun WD, Säll T, Tim CF, Cardell LO.
Toll-like receptor gene polymorphisms are associated with allergic rhinitis: a case control study.
BMC Med Genet 2012; 13:66.

V. Carlberg D*, Henmyr V, Manderstedt E, Lind-Halldén C, Säll T, Cardell LO, Halldén C.
Discovery of rare and common variants in the Toll-like receptor genes (*TLR1-TLR10*) and their association with allergic rhinitis.
Manuscript.

*Change of family name from Nilsson to Carlberg (2014)

CONTENTS

LIST OF ABBREVIATIONS

AFR	African
ASN	Asian
AMR	South American
AR	Allergic rhinitis
bp	Base pair
CDCV	Common disease, common variant
CDRV	Common disease, rare variant
CEU	Utah Residents with Northern and Western European Ancestry
CNV	Copy number variant
EUR	European
FDR	False-discovery rate
GWAS	Genome-wide association study
HGP	The Human Genome Project
HWE	Hardy-Weinberg equilibrium
LD	Linkage disequilibrium
MAF	Minor allele frequency
mGWAS	Meta genome-wide association studies
OR	Odds ratios
SNP	Single nucleotide polymorphism
SPT	Skin prick test
TLR	Toll-like receptor
tSNP	Tag singel nucleotide polymorphism

1 INTRODUCTION

1.1 THE HUMAN GENOME

The human genome is composed of 22 pairs of autosomal chromosomes and two sex chromosomes (X and Y chromosomes). Somatic cells are diploid, meaning that they contain two copies of each chromosome, one copy inherited from the father and one from the mother resulting in a total of 46 chromosomes in each cell. The Human Genome Project (HGP), initiated in 1990, aimed to produce a draft sequence of the human genome. This was by far the largest genome to be sequenced at that time. In 2001 the first draft sequence covering approximately 94% of the human genome was released [1]. The same year Celera genomics, which was a private initiative led by J. Craig Venter, released a 2.98 billion base-pair assembly of the human genome. Celera used a different strategy than HGP and also used the information released by HGP to assemble their sequence data into an almost complete genome sequence [2]. The latest assembly of the human genome consists of 3.4 billion base pairs (bp) and encodes approximately 20 500 genes according to the Genome Reference Consortium Human Build 38 (GRCh38) Assembly.

1.2 GENETIC VARIATION

Most sequence variants of an organism are neutral or nearly neutral. A mutation can modify a gene or its expression or even delete an entire gene. Here the term mutation is used when sequence variants are harmful for the organism, otherwise the term variant or polymorphism is more appropriate to use. Such detrimental mutations are normally rare in the population and often affects the function or expression level of the protein. There are a number of different types of genetic variants like single nucleotide polymorphisms (SNPs), insertions or deletions of single nucleotides or nucleotide-stretches of various length, length variations of tandem repeated sequences (micro and mini satellites) and copy number variants (CNVs). Mutation of DNA is a random event and different types of mutations have different mutation rates, i.e. the frequencies are type-dependent. SNPs are the most abundant form of genetic variation and are also the most common form

of variation analyzed in genetic association studies. A SNP is a variant position in the DNA sequence. It typically has two alleles, with the least abundant allele present at a frequency ≥ 1% in the population. It has been estimated that there is on average one SNP every 300 bp in the human genome when the total human population is investigated. If instead a single individual is compared with the reference sequence, a total of ca. $3x10^6$ SNPs are present equivalent to one SNP every 1000 bp in the genome. With the exception of sex chromosomes in males, humans have two copies of each chromosome and both alleles can therefore be present in the same individual when looking at a specific SNP. An individual has one allele on one chromosome and one allele on the other, which is referred to as a genotype. Consequently, there are three different states for a genotype: homozygous for the common allele (e.g. AA), heterozygote (AG) and homozygous for the rare allele (GG).

1.2.1 Haplotypes

The set of alleles residing on a specific chromosome or a specific part of a chromosome is called a haplotype. For a sub-region of a chromosome with SNP alleles at a number of different loci, the alleles define the possible haplotypes for that region. For example, for two SNPs with the alleles A/G and C/T there are four different possible haplotypes for that region: A-C, A-T, G-C and G-T. The original haplotype is changed first at one of the positions and then at the other position resulting in three different haplotypes. The fourth haplotype is in most cases created through recombination between the two positions. The frequencies of the haplotypes depend on the frequencies of each allele in the population. New haplotypes for a chromosome region arise by the accumulation of additional mutations and recombination events.

1.2.2 Linkage disequilibrium

Alleles located close to each other are more likely to be co-inherited than alleles that are located farther apart. Such non-random association of alleles is known as linkage disequilibrium (LD). Each time recombination occurs between two loci, LD is also weakened between them. LD is calculated as the difference between

the frequency of haplotypes carrying the alleles *A* and *B* at two different loci and the product of the frequencies of those alleles, i.e. $D_{AB} = p_{AB} - p_A p_B$. However, *D* is often not the most convenient statistic to use for describing LD since the range of values for *D* depends on the allele frequencies and is therefore not optimal when comparing LD at different pairs of loci. A number of additional statistics to describe LD has been proposed, among them *D´* and r^2 which are both commonly used. *D´* is the ratio of *D* to the largest possible value of *D* given the allele frequencies (the smaller of $p_A(1-p_B)$ and $p_B(1-p_A)$). The other commonly used statistic, r^2, represents the correlation coefficient between two loci and is calculated as: $r^2 = D^2 / p_A(1-p_A)p_B(1-p_B)$. *D´* is a convenient measure since *D´*=1 indicates that no historical recombination has occurred between the loci in the ancestors of the sample, which is often referred to as perfect LD. However, *D´* is defined in such a way that *D´* = 1 if two or three of the possible haplotypes are observed and it is < 1 if all four haplotypes are present. The r^2 statistic on the other hand takes the value of 1 only when two of the possible haplotypes are present [3,4].

1.2.3 The HapMap project

In 2005 the International HapMap Consortium released the first haplotype map of the human genome. Phase I of HapMap describes the common patterns of genetic variation and LD patterns in 269 individuals from four different populations: 90 individuals from Utah with Northern and Western European ancestry, 90 Yoruban individuals from Ibadan, 44 Japanese individuals from Tokyo and 45 Han Chinese individuals from Beijing. The project genotyped more than 1 million SNPs with minor allele frequencies (MAFs) ≥ 5%. The release of HapMap data was a great resource for designing genotyping efforts in genetic association studies. Since different SNPs are associated through LD, tag-SNPs (tSNPs) can be used to capture a large fraction of the variation in any given region. Thus, tSNPs can serve as proxies for SNPs not selected for genotyping. There are a number of different algorithms for the selection of tSNPs. Pairwise-tagging is a commonly used method, where tSNPs are selected until all common SNPs are highly correlated (ex. $r^2 > 0.8$) with the selected

tSNPs. For example, to capture all common SNPs in the Chinese and Japanese samples of the HapMap Phase I data would require 260 000 tSNPs by using a r^2-value of 0.8. By using a lower r^2-value of 0.5 the number of tSNPs decreases to 159 000 for the same populations [5]. In Phase II of HapMap an additional 2.1 million SNPs were genotyped in the same populations. The resulting SNP density of Phase II was approximately one SNP per kilobase-pair and provides a better representation of rare variants compared to Phase I [6]. The knowledge of human genetic variation was still limited with respect to variant type, frequency and population diversity. Even though rare variants are considered as major contributors for common diseases, focus has been on common variation. In Phase III of the HapMap project, 1.6 million SNPs were genotyped in 1184 individuals from 11 different populations and also ten 100 kbp regions were sequenced in 692 individuals in order to create a high-resolution map of genetic variation including rare variants [7].

1.2.4 Sequencing projects

Ten years after the first draft human genome sequence was published, the 1000 Genomes Project Consortium published the results from the pilot phase of an extensive sequencing project [8]. Following the introduction of next-generation sequencing technology, the cost of sequencing decreased dramatically. The 1000 Genomes Project was the first project to provide a comprehensive resource on human genetic variation determined by whole-genome sequencing in a large number of individuals. The aim of the project was to characterize over 95% of sequence variation with an allele frequency of ≥1% in five different populations: European, East Asian, South Asian, West Africans and Americans. The pilot phase of the project was divided in three different projects. In the first project low-coverage whole-genome sequencing (2-4X) was performed in 179 individuals. The second and third project included deep sequencing of six individuals in two families and exon sequencing of 8140 exons in 697 individuals [8]. In 2012, the 1000 Genomes Project Consortium released the complete genomes of 1092 individuals from 14 different populations capturing up to 98% of accessible SNPs at a frequency of ≥1%. The data was generated

using a combination of low-coverage (2-6X) whole-genome sequence data and high coverage (50-100X) exome sequence data and SNP genotyping for validation. The sequencing detected 38 million SNPs, 1.4 million short insertions and deletions and more than 14 000 large deletions [9]. In 2014 the full project was released containing sequence data from 2535 individuals from 26 different populations (http://www.1000genomes.org). UK10K is another large sequencing project that was initiated in 2010. The primary goal of the project was to sequence the whole genome of 4000 samples from the TwinsUK and ALSPAC cohorts to 6X sequencing depth and 6000 exomes of extreme phenotypes for a number of specific conditions (http://www.uk10k.org/). At the American Society of Human Genetics conference in October 2014, the Haplotype Reference Consortium presented a database containing about 50 million genetic variants collected from whole-genome sequence data generated by 23 research collaborations. At the same time, the Exome Aggregation Consortium released a public database containing the exome sequences of approximately 63 000 samples [10].

1.3 DISEASE INHERITANCE

Genetic diseases can be inherited in a number of different ways. When a disease is due to a single mutation in a gene, the disease is referred to as monogenic or Mendelian. The inheritance pattern of monogenic diseases follows Mendel's laws and can be traced in familial pedigrees. Polygenic diseases have a complicated pattern of inheritance since disease is due to mutations in multiple genes, and in complex diseases different environmental factors also acts on disease status.

1.3.1 Monogenic diseases

Monogenic diseases are classified according to their inheritance patterns. In dominant inherited diseases, one copy of the defective gene is enough to cause disease. Even if an individual has one healthy copy of the gene and one defective copy, the individual will suffer from disease. A disease inherited in an autosomal recessive manner requires that both copies of the gene are defective. Autosomal

recessive diseases are generally rare in the population and the defective genes can hide in carriers harboring only one copy of the defective gene. In recessive X-linked diseases males are more frequently affected than females, since males only have one copy of the X-chromosome. Dominant X-linked diseases requires only one defective copy just as in the case of autosomal dominant diseases, but the inheritance pattern differ since males only have one copy of the X-chromosome whereas females have two.

1.3.2 Complex diseases

In complex disease like diabetes, asthma and allergic rhinitis, the inheritance pattern is not as straight forward as in monogenic disease even though the disease tends to accumulate in families. Contrary to in the case of monogenic disease where the effect of one mutation is enough to cause the disease, the combined effect of mutations in multiple genes in combination with different environmental factors causes disease in the case of complex disease. This implies that the effect of a single mutation is small and therefore difficult to detect in genetic studies. The primary focus for the investigation of complex disease has for long been on investigating common variation according to the Common Disease, Common Variant (CDCV) hypothesis, which states that common complex diseases are due to common variants each with small to modest effects. The information gathered within the HapMap project in concert with technical advances made it possible to conduct genome-wide association studies (GWAS), which investigates common variation [11]. According to the Catalog of Published Genome-Wide Association Studies (http://www.genome.gov/gwastudies/) more than 2000 GWAS studies have been reported in the literature. Even though GWAS have been a successful strategy in identifying associations with various diseases, the collective associations only explain a minor fraction of the phenotypic variation in the population [11]. This lack of explanation for a large proportion of the phenotypic variation in complex diseases is referred to as the missing heritability and has been widely debated over the years. Epigenetic effects, rare variations and also entirely unforeseen sources have been suggested to account for this missing heritability [12]. In

contrast to the CDCV hypothesis, the Common Disease, Rare Variant (CDRV) hypothesis argues that multiple rare variants with relatively strong individual effects are the major contributors to common disease. Rare genetic variants in the form of CNVs have been associated with a number of diseases, especially neuropsychiatric diseases [11]. In the case of autism spectrum disorders a large number of rare CNVs have been identified. In contrast to SNPs identified through GWAS, the effect sizes have been reported to be 3 times higher for CNVs [13]. Multiple rare SNPs have also been linked to a number of different diseases including type 1 diabetes, colorectal cancer and tuberculosis [14-16].

1.4 MOLECULAR GENETICS

The polymerase chain reaction (PCR) invented by Kary Mullis in the mid 80's is the basis for many of the techniques used in molecular biology today. PCR uses a heat-stable polymerase to synthesize copies of the target sequence in a cyclic three-step process involving a denaturation step where all DNA molecules are made single-stranded, an annealing step where the primers that specifies the target sequence is binding to template DNA and finally a synthesis step where a copy is made of each template strand. Since copies are made of both strands simultaneously the target sequence is amplified exponentially. The specificity of this process is determined by the single stranded primers that is designed to bind to a specific complementary sequence in the template DNA. The end result after 25-35 cycles of the three-step process is an amplification of a specific DNA-region to many millions of copies. This amplification is a prerequisite for many of the analysis techniques used within molecular genetics today [17].

Genotyping can be made using a great variety of different techniques. When a few SNPs are analyzed in many individuals a commonly used system is the TaqMan system. This system relies on two dual-labelled probes, one for each allele of a specific SNP. During PCR amplification of a DNA sequence that contains the SNP in question the two probes can bind to their respective complementary sequences and upon binding the probes are digested by the polymerase releasing a fluorophore. This fluorophore is quenched as long as it is bound to the probe, but after release and exposure to light of a suitable

wavelength it fluoresces. The fluorescence of the allele-specific fluorophores is registered and the genotypes of the samples determined. This system is highly precise and is often used to confirm genotype data generated by other techniques [17].

Another common genotyping system is the Sequenom multiplex primer-extension method. Also this system relies on PCR amplification of target sequences containing the interrogated SNPs. After purification of the primary PCR products thereby eliminating the primers and dNTPs of the first reaction, an extension primer is elongated into the polymorphic site. This extension is made in the presence of a mixture of nucleotides where some of the nucleotides are elongating and some are terminating the growing chain of nucleotides. This creates allele-specific extensions of the primer and these differently sized molecules can then be detected by MALDI-TOF mass spectrometry. This whole process can be made in a multiplex fashion simultaneously recording the genotypes of up to 36 different SNPs. This genotyping system is much used in candidate gene studies and in replication studies where a couple of hundred SNPs are analyzed [18,19].

Microarray hybridization is used to interrogate many SNPs in parallel in a given sample. This is the technique typically chosen for GWAS since it can reach the very high numbers of SNPs (0.5-5 million) that are necessary to cover the human genome at a SNP density that is sufficient for this type of analysis. A microarray consists of millions of different unlabelled oligonucleotide probe populations that have been fixed to a surface within a high density grid format. Each square in this grid is covered with millions of identical copies of just one probe sequence. A test sample is genotyped on this array by first labelling the DNA fragments of the sample and then hybridizing denatured fragments onto the array. After a washing step that decreases non-specific binding of the labelled fragments, the remaining sample molecules are detected using a laser scanner. The emitted signals are then analyzed using digital image software to distinguish the binding of DNA fragments that differs in only one position, i.e. the different alleles of the interrogated SNP. There are a number of strategies to achieve the

high sequence specificity that is necessary to discriminate DNA sequences that are almost identical [20].

The ultimate test for polymorphisms is DNA sequencing [17]. For many years different variants of the dideoxy DNA sequencing strategy originally described by Fredric Sanger has been the primary sequencing technique. Also this technique relies on the initial PCR amplification of the studied DNA sequence. After purification of the PCR products with regard to primers and deoxy-nucleotides, the actual sequencing reaction takes place. This reaction contains one sequencing primer that defines where DNA sequencing will take place and a mixture of deoxy- and dideoxy-nucleotides (elongators and terminators) that will be incorporated into the growing chains. All of the four bases are present both as elongators and terminators with the elongators present at a 100-fold surplus compared to the terminators. The four different terminator-bases are also labelled with four different base-specific fluorophores. Since termination is a random process, a set of DNA fragments will be created that are terminated at each position of the sequence under analysis. All of the fragment copies of a given length will be terminated in the same base and carry the same fluorophore. The fragment population will then be separated according to size using capillary electrophoresis and the fluorescent signals of the fluorophores detected as they pass in front of a detector after being separated on the capillary. After base-calling the resulting DNA sequence is aligned to a reference sequence and interpreted for differences compared to the reference sequence. The analysis of both strands of a given DNA molecule results in high quality of the resulting sequence. This type of DNA sequencing, often called Sanger sequencing, is now mostly used to analyze the promoter and the exons of one or a few genes since the method is time-consuming and costly [17] .

In the years 2005-2008 a number of new sequencing methods were developed. They are termed next-generation sequencing methods to acknowledge the fact that they are all massively-parallel and quite different from the Sanger method. The Ion Torrent sequencing technique, for example, is based on a complex mixture of DNA fragments where the individual fragments are amplified by

emulsion PCR producing millions of copies of individual DNA sequences each bound to single spheres. The DNA sequences of the spheres are then sequenced in parallel using flows of each base over vast collections of spheres. Since a hydrogen ion is released for each base that is incorporated into a growing chain the incorporation of bases can be determined by measuring the pH of the solution surrounding the sphere. This is made by a semiconductor pH sensor device located in close proximity to each sphere. This method allow very large numbers of sequences to be determined in parallel and lowers the costs for sequencing substantially [21,22].

1.5 GENETIC ANALYSIS

Different types of study design and statistical methods have been developed and adopted for the identification of disease-causing genes or chromosomal regions harboring genetic variation influencing disease.

1.5.1 Linkage studies

Linkage studies uses related individuals to map genetic loci that predispose to disease. Two loci are linked if they are co-transmitted from parent to offspring more often than expected under independent inheritance. Another common way of expressing this is that, for two linked loci the recombination frequency between them is less than 0.5. In brief, by studying the segregation of genotyped markers through pedigrees and calculating recombination frequencies the relative positions of those markers can be inferred. Using this strategy, disease loci can be mapped relative to the genotyped markers. Linkage analysis is useful for the identification of wide regions harboring disease-causing genes that can be further investigated using other methods [23].

1.5.2 Population association studies

Population association studies use, in contrast to linkage studies, unrelated individuals for the identification of genetic variation that vary systematically between diseased individuals and control individuals. An observed association is either direct, i.e. the SNP itself influences the trait in question, or indirect, i.e. the

SNP is in LD with a genetic variant that influences the trait. Association studies are a major tool in identifying risk loci in complex diseases. Most association studies have focused on common variation in the form of SNPs, where allele and genotype frequencies are compared between cases and controls. A significantly higher frequency of one or the other allele in cases relative to in controls indicates an increased risk for disease in the presence of the specific allele [24]. Since association studies compare allele and genotype frequencies, it is important to ensure a common genetic background for cases and controls so that any observed differences are due to presence or absence of disease and not to population stratification. In the selection of cases a strict phenotypic definition is important to ensure a homogeneous population. To improve statistical power the selection of cases can be restricted to include those who are likely to have a high genetic load, e.g. extreme phenotypes or early onset of disease. Controls can be randomly selected from the general population with unknown disease status or selected on the criteria to be healthy from the disease under study. The use of healthy controls in association studies of common diseases is preferable due to the loss of power using randomly selected individuals from the general population [24,25]. The completion of the International HapMap project encouraged the development of SNP-based GWAS. GWAS interrogates the whole genome at once using tSNPs and selected non-synonymous SNPs and have become a popular method for the investigation of common variants. Illumina (http://www.illumina.com/) provides genotyping arrays containing up to 5 million genetic markers that can be simultaneously analyzed. Since all of these genetic markers are all investigated for associations using different statistical methods, a large number of associations are expected purely by chance using the conventional threshold for significance at $P < 0.05$. Thus, multiple testing is one of the major issues with GWAS and stringent levels for significance have therefore been adopted [26]. In contrast to GWAS, candidate gene studies are hypothesis driven and the selection of genes for study can be made using several different strategies, e.g. genes selected from relevant biochemical pathways or genes implicated in previous studies. Even though GWAS have become very popular since their introduction, candidate gene studies are frequently reported in the literature. The strategy to investigate a

small number of candidate genes has some benefits compared to GWAS that might explain the frequent use of this approach. For example, the issue with multiple testing is less pronounced since the numbers of SNPs are much smaller compared to in GWAS. The SNPs analyzed in candidate gene studies are also to a larger extent more dependent through LD compared to in GWAS, and the correction for multiple tests can thus be less strict. It is also possible to choose a more complete set of functionally relevant SNPs. This makes the association results easier to interpret compared with SNPs located in intergenic regions far from known genes, which is often the case in GWAS. Lastly, the lower cost for genotyping a much smaller total number of SNPs is of course also a major factor speaking in favor of candidate gene studies.

1.5.3 Hardy-Weinberg equilibrium

The rediscovery of Mendel's laws in 1900 raised the question of their relevance for the genetics of populations. This lead to the development of modern population genetics. One of the first important results within this discipline appeared in 1908, when the English mathematician G. H. Hardy and the German physician W. Weinberg concluded that the expected genotypes of a random mating population can be calculated using allele frequencies of that specific population [27]. The allele frequencies of alleles *A* and *a* is p and $1-p = q$, respectively. Thus, $p+q = 1$ and the frequencies of the genotypes *AA*, *Aa* and *aa* under Hardy-Weinberg equilibrium (HWE) are expected to be p^2, $2pq$ and q^2, respectively. A test for HWE using a chi-square test or Fishers exact test is primarily made as a quality control of the genotype data in association studies. Departures from HWE can be due to genotyping problems where genotypes are being miss-interpreted. Also the existence of sub-population structures or actual association signals can lead to departures from HWE, even though the test have limited power to detect these circumstances [24,28].

1.5.4 Association analysis

The allele and genotype counts of both cases and controls can be tabulated in contingency tables and standard statistical tests can be used to test the null

hypothesis of no association. The chi-squared 2-degree of freedom test is often used when investigating the distribution of genotypes among cases and controls. The test compares the observed distribution of genotypes among cases and controls with the expected distribution. The test has a reasonably high power to detect associations regardless of the underlying risks. If the underlying genotype risks are additive, a chi-squared 1-degree of freedom test to consider the distribution of alleles among cases and controls is more powerful. However, using this approach requires an assumption of HWE in the combined samples of cases and controls. An alternative test to the chi-squared 1-degree of freedom test that do not rely on the assumptions of HWE is the Cochran-Armitage test [28].

1.5.5 Multiple testing

In association studies that involve a large number of SNPs, false positive results are expected. Given the conventional significance threshold of 0.05, one test out of 20 is expected to be significantly associated by chance only, i.e. false positive associations are a common phenomenon when many tests are performed. Therefore the significance level of the individual tests, which is the probability for making a type-I error, has to be adjusted. In GWAS where the number of tests can be in the order of millions, a threshold for significance has been widely adopted at $5x10^{-8}$ to reduce the number of false positive associations. One way to lower the significant threshold is by applying a Bonferroni correction, in which the conventional significance level of 0.05 is divided by the number of tests. Since SNPs can be correlated through LD, the corresponding tests for associations will not be independent. Consequently, the Bonferroni correction is very conservative and the significance threshold will be overcorrected [26]. Instead of lowering the significance threshold, an alternative approach is to estimate the false-discovery rate (FDR), which is the proportion of false positive results among all tests identified as significant. The *q*-value is an extension of the FDR and is similar to the *P*-value. Each test is associated with a *q*-value that measure the proportion of false positives when considering all *P*-values equal to or more extreme than the associated *P*-value for the test under consideration [29].

1.5.6 Odds ratio

The *P*-value of an association test between marker and phenotype gives information of the statistical evidence of the association. This is only one aspect of the data. An equally important aspect is the risk for disease given the effect of specific SNP alleles or genotypes, if a SNP is considered to be associated with the disease. This can be calculated through odds ratios (ORs). ORs are calculated as the odds for cases having a specific allele over the odds for controls having the same allele. It is common practice to calculate accompanying 95% confidence intervals for ORs that describes the probability of the estimates [25].

1.6 ALLERGIC AIRWAY INFLAMMATION

1.6.1 Allergic rhinitis

Allergic rhinitis (AR) is a global health problem that causes major disability worldwide. It is due to an allergen induced IgE-mediated inflammation of the membranes lining the upper airways. Nasal obstruction, secretion and itching are characterizing features of the disease, which is often associated with eye symptoms, fatigue and asthma. The impact of AR on health-related quality of life, work and school performance are well recognized [30]. The widely debated hygiene hypothesis as origin for the development of allergy is based on the observation that early childhood exposure to parasitic infections and other microorganisms has been reduced due to better standard of living. Such infections encourage a normal development of the immune responses. As the exposure to such infections has been reduced in the modern society, there is a tendency for certain individuals to develop immune responses against harmless environmental allergens [31]. The development and severity of AR are determined by a complex interaction between both environmental and genetic factors and are not fully understood.

1.6.2 Asthma

Asthma is a common chronic inflammatory disease of the airways characterized by variable and recurring symptoms, reversible airflow obstruction and bronchospasm. Symptoms include wheezing, coughing, chest tightness, and

shortness of breath. It includes hyper reactivity to a variety of stimuli and remodeling of the airways [32,33]. The development of asthma is due to a complex interaction between both genetic and environmental factors and the heritability of asthma has been estimated to 0.5-0.6 in twin and family studies [34].

1.6.3 Genetics of allergic rhinitis and asthma

Although AR does not demonstrate Mendelian inheritance, the heritability for AR is relatively high and has been estimated to 0.66-0.78 [35,36]. Although more than 100 SNPs have been associated with the disease, the genetic basis for AR is poorly understood. Genetic findings in AR has primarily been identified through candidate gene studies, but the associations have been difficult to replicate in later studies [37,38]. To date, four GWAS have been published investigating AR and allergic sensitization, i.e. the presence of allergen-specific IgE. In a GWAS of a Singapore Chinese population, analysis of 456 AR cases and 486 controls identified no genome-wide significant SNPs for AR. However, after replication of 77 SNPs in a replication cohort of 676 AR cases and 511 controls suggestive associations were identified for two SNPs in the *BCAP* and *MRPL4* genes [39]. A meta-study of GWAS (mGWAS) investigating self-reported AR and grass sensitization in 6248 cases and 18 997 controls identified three loci (*HLA-DRB4*, *C11ORF39* and *TMEM232*) at a genome-wide significance level ($P < 5x10^{-8}$) and an additional 12 loci with suggestive associations ($5x10^{-8} < P < 5x10^{-6}$). Using the data set in a candidate gene approach, identified an additional three genes as associated with the investigate phenotypes (*TSLP*, *TLR6* and *NOD1*) [40]. A second mGWAS investigated allergic sensitization in 11 903 cases and 19 976 controls of European ancestry. The study identified a total of 10 loci at a genome-wide significance level ($P < 5x10^{-8}$), including *TLR6*, *C11orf30*, *STAT6*, *SLC25A46*, *HLA-DQB1*, *IL1RL1*, *LPP*, *MYC*, *IL2* and *HLA-B* [41]. A third mGWAS investigated self-reported cat, dust-mite and pollen allergies in 27 551 cases and 26 311 controls. A total of 16 genome-wide significant loci ($P < 5x10^{-8}$) and an additional 6 loci with suggestive evidence of association ($5x10^{-8} < P < 1x10^{-6}$) were identified in the

study. The detected associations were located in the regions of *TLR1-TLR6*, *WDR36-CAMK4*, *C11orf30-LRRC32*, *IL1RL2-IL1RL1*, *HLA-DQA1-HLA-DQB1*, *HLA-C-MICA*, *PTGER4*, *PLCL1*, *LPP*, *RANBP6-IL33*, *NFATC2*, *GSDMB*, *SMAD3*, *GATA3*, *ADAD1*, *FOXA1-TTC6*, *TPD52-ZBTB10*, *ID2*, *CLEC16A*, *IL4R-IL21R*, *PEX14* and *ETS1*. Several of these loci have previously been associated with other immunity-related phenotypes including asthma [42]. Allergic disease like AR and atopic asthma have an overlapping etiology in for example increased levels of IgE which may indicate a common genetic background (at least in part) for these diseases. In addition, patients with AR have a 5-6 fold increased risk of developing asthma [43-45]. In asthma, the genetic background has been more extensively studied compared to the genetics of AR. GWAS investigating asthma phenotypes have identified more than 18 susceptibility loci for the disease and the *ORMDL3*/*GSDML*, *HLA-DR*/*DQ*, *IL33* and *IL18R1*/*IL1RL1* loci have been consistently identified in multiple asthma GWAS [34]. A large number of asthma-associated genes from different types of studies have been positively associated multiple times. The *ARDB2* gene have for example been identified more than 40 times, and *IL13*, *HLA-DRB1* and *IL4R* have been associated with asthma phenotypes more than 25 times [46].

2 AIMS

The overall aim of this thesis has been to identify genetic variants associated with allergic rhinitis (AR).

When I started my PhD studies in 2009, genetic association studies in AR were mostly carried out as candidate gene studies based on relatively small sample sizes and few of the identified associations had been replicated. Accordingly, **paper I** of this thesis aimed to investigate the reproducibility of such previously reported candidate gene associations in AR.

The genetic background for asthma has been far more investigated compared to the genetics of AR and many asthma-associated genes have been successfully replicated. It is also well known that asthma is closely related to AR. Thus, **paper II** aimed to examine genetic variation in highly replicated asthma-associated genes for associations also in AR.

Three mGWAS investigating AR were published in 2011-2014. These studies made important contributions to the field of AR genetics since all used large numbers of cases and controls and identified highly significant loci that were also detected in several of the studies. Therefore, **paper III** aimed to investigate the reproducibility of these associations. In addition, the study also investigated these associations for age at onset effects.

The Toll-like receptors (TLRs) have been extensively studied in various allergic diseases including AR and a number of studies have identified associations with variants in the TLR genes. **Paper IV** aimed to search for common genetic variants associated with AR in the TLR genes.

Common variants are not able to explain the phenotypic variability of complex disease more than to a minor degree. This missing heritability is often argued to be due to the presence of detrimental rare variants present in addition to the common factors. **Paper V** aimed to characterize rare genetic variants in the TLR genes and to investigate the possibility for an accumulation of rare and detrimental variants in AR patients.

3 MATERIALS AND METHODS

3.1 SUBJECTS

Three different study populations have been investigated in the present work, two Swedish populations and one Singapore Chinese population.

3.1.1 Swedish Malmö population

The first population is a Swedish study population from southern Sweden. It was recruited at Malmö University hospital in 2003-2009 and consists of unrelated subjects from the general population. Both patients and controls were of Caucasian origin, with both parents born in Sweden. The diagnosis of birch and/or grass pollen induced AR was based on a positive history of intermittent AR for at least 2 years and a positive skin prick test (SPT; mixture of 11 common airborne allergens, ALK-Abelló, Hørsholm, Denmark SPT) or Phadiatop test (mixture with 8 different inhalant allergens, Pharmacia Upjohn, Uppsala, Sweden) to birch and/or grass. SPT were performed on the volar side of the forearm with saline buffer as negative and histamine chloride (10 mg/ml) as positive controls. A wheal reaction diameter of ≥3 mm was considered a positive SPT response. SPT was only performed if the AR cases had not taken any anti-allergic drugs for at least 3 days prior to the test. All patients were classified as having severe symptoms (itchy nose and eyes, sneezing, nasal secretion and nasal blockage) during pollen season and they had all been treated with antihistamines and nasal steroids during pollen seasons previous years. Controls had no history of AR or any other atopic disease and had a negative SPT or Phadiatop test. Data for a total of 1061 individuals were included in the analysis. The study was approved by the Ethics Committee of the Medical Faculty, Lund University, and written informed consent was obtained from all subjects.

3.1.2 Swedish BAMSE population

The second Swedish population was derived from the Stockholm area as part of the BAMSE study [47,48]. The BAMSE study is an unselected population-based birth cohort and consists of 4089 children recruited between February 1994 and

November 1996. All participants were living in the northern and central parts of Stockholm, Sweden, and were of mixed social economic status and from urban, suburban and inner city areas. Each participant and parents and/or legal guardians were given questionnaires concerning allergies, social economic status, housing and birth place of the parents. Follow-up studies were made at 1, 2, 4, 8, 12 and 16 years of age. Allergy data were obtained at 4, 8 and 16 years of age. Sera were screened with Phadiatop and fx5 (mixture with 5 different food allergens, Pharmacia Diagnostics AB) at each evaluation. Sera with immunoglobulin E-value for Phadiatop and/or fx5 $\geq$ 0.35 kU_A/l were then further analyzed for reactivity to the single allergens. Data from the 8 and 16 year evaluations for a total of 2153 children were included in the analysis. The BAMSE study was approved by the Ethics Committee at Karolinska Institutet, Stockholm, Sweden, and written informed consent was obtained from parents and/or legal guardians.

3.1.3 Singapore Chinese population

The third population was collected in Singapore using multiple recruitment-drives and consists of unrelated Chinese individuals. All AR cases exhibited symptomatic house dust mite induced AR whereas the participating controls had no atopy and allergic symptomology. Diagnostic procedure included personal interview of medical history using a standardized questionnaire and SPT performed using standard panels of common allergens such as *Dermatophagoides pteronyssinus* and *Blomia tropicalis*. Data for a total of 2221 individuals were included in the analysis. This study was approved by the Institutional Review Board (IRB, Reference - NUS07-023) of National University of Singapore and is also in compliance with the Helsinki declaration. All DNA samples were collected following standard protocols of informed consent.

3.2 DNA EXTRACTION

DNA was extracted from blood or buccal cells. DNA concentrations were determined by fluorometry using PicoGreen (Molecular Probes, Eugene, OR, USA) or by Nanodrop (Thermo Fisher Scientific Inc, Wilmington, DE, USA).

3.3 GENOTYPING

Determination of small and medium numbers of SNP genotypes were either performed using a TaqMan-based strategy or a Sequenom MassARRAY MALDI-TOF strategy. Dual-labelled TaqMan-assays were carried out at Kristianstad University according to the manufacturer's protocol using either ABI PRISM 7900HT (Applied Biosystems, Foster City, CA, USA) or CFX384 (Bio-Rad, Hercules, CA, USA) genotyping systems. The Sequenom MassARRAY MALDI-TOF system (Sequenom Inc, San Diego, CA, USA) analyzed allele-specific primer extension products using mass spectrometry. Assay design was made using the MassARRAY Assay Design ver. 2.0 software and primers were obtained from Metabion GmbH (Martinsried, Germany). Sequenom genotyping was made at the Mutation Analysis Facility at Karolinska Institutet. Determination of large numbers of genotypes, i.e. whole genome genotyping was performed using the Illumina HumanHap 550 k BeadChip version 3 (Illumina, San Diego, CA, USA) at the Genome Institute of Singapore.

3.4 SANGER SEQUENCING

Primers were designed using Primer-BLAST (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) and purchased from DNA Technology A/S (Risskov, Denmark). Big Dye Terminator Sanger sequencing was performed in both directions using a 3130XL Genetic Analyzer (Applied Biosystems). Primary PCR was performed using KAPA Taq Extra HS PCR Kit (KAPA Biosystems, Cape Town, South Africa) and primary PCR products were treated with ExoSAP-IT® (Applied Biosystems) according to the manufacturer's instructions. DNA sequencing was subsequently performed in a total volume of 5 µl containing 0.5X Big Dye sequencing ready reaction premix (Big Dye Terminator v 2.0, Applied Biosystems), 0.5X Big Dye Sequencing buffer and

3.2 pmol of the sequencing primer. The sequencing reactions were purified using Xterminator (Applied Biosystems) according to the manufacturer's instructions. Sequences were interpreted and all polymorphisms were identified using SeqScape ver. 2.5 and confirmed by manual inspection.

3.5 ION TORRENT SEQUENCING

The primers were designed using Ion AmpliSeq Designer, pipeline version 2.0.3 (http://www.ampliseq.com). Template DNA was pooled such that each pool contained equimolar amounts of DNA from 12 individuals. Initial amplification of the targeted regions was performed using the Ion AmpliSeq™ Library Kit 2.0 (Life Technologies, Carlsbad, CA, USA) in 10 µl PCR reactions using 20 ng of template DNA. Adapters were subsequently ligated to each pool of amplicons and clean-up and size selection was performed using Agencourt Ampure XP beads (Beckman Coulter, Indianapolis, IN, USA). DNA concentration and fragment size distribution of each library were determined by capillary electrophoresis on a Fragment Analyzer (Advanced Analytical Technologies Inc, Ames, IA, USA). The libraries were then diluted and emulsion PCR performed using a OneTouch 2 machine (Life Technologies) with the Ion PGM Template OT2 200 kit (Life Technologies). Templated spheres were recovered using Ion PGM Enrichment Beads (Life Technologies). The samples were centrifuged and the resuspended pellets were mixed with sequencing primer from Ion PGM Sequencing 200 kit v2 (Life Technologies). The sequencing primer was annealed by incubation followed by the addition of sequencing polymerase. The samples were loaded onto an Ion PGM 314 chip v2 (Life Technologies) and sequencing performed on an Ion Torrent PGM (Life Technologies) using the default flow order. The sequences were aligned against the human reference sequence (build GRCh37) using Torrent Suite 3.6 (Life Technologies) and primer sequences were trimmed off. Variant calling was performed using parameters tuned for high sensitivity. Annotation of the detected SNPs was made using SeattleSeq Annotation 137 (http://snp.gs.washington.edu/SeattleSeqAnnotation137/).

3.6 STATISTICAL ANALYSIS

Statistical analyses were made using R statistical software [49,50] and PLINK v1.07 [51]. Genotype frequencies were calculated and tested for Hardy-Weinberg equilibrium in both cases and controls. Allele and genotype frequencies were then investigated for association with AR using χ^2-homogeneity tests. Odds ratios (ORs) and 95% confidence intervals were estimated by using the most common allele as the referent and are reported for each minor allele. Associations between SPT response and genotype were analyzed using Kruskal-Wallis rank sum test. False discovery rate was quantified using the *q*-value introduced by Storey [52] and calculated using the R package qvalue v.1.32 [53]. Power calculations were made using simulations for the association tests of allele effects. For each SNP a data set was simulated with the actual numbers of cases and controls, using the published ORs and the allele frequencies observed in our populations. A total of 10 000 runs were made in each case. For these, a full data set was simulated and the χ^2-value was calculated. The number of times the test quantity exceeded the critical value (0.05 and 0.001) was subsequently scored. Age at onset effects were investigated using a likelihood ratio test investigating the heterogeneity of age classes when AR cases were stratified into those with early onset (≤ 8 years of age) and those with late onset (> 8 years of age). The potential impact of asthma was evaluated by partitioning the difference in allele frequencies observed between cases and controls into three different components. The first component (C1) being the difference in risk allele frequency between cases and controls in the absence of asthma, i.e. the difference that we aim to investigate. The second component (C2) represents the direct confounding due to the combined effect of an excess of asthmatics among cases and the difference in allele frequencies between cases with or without asthma. The third component (C3) is due to the interaction between allergy and asthma on their effect on the risk allele frequency.

3.7 BIOINFORMATICS

tSNPs were selected using HapMap (http://www.HapMap.org) CEU data and the tagger function implemented in Haploview v.4.2 [54] and missense and

nonsense variants were extracted from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). Polymorphisms were extracted from dbSNP and 1000Genomes (http://www.1000genomes.org/) databases. The variants were obtained from the Integrated Variant Set of the 1000Genomes Project (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/) release April 2012. Variants were extracted for 1092 individuals using tabix [55]. Using VCFtools [56] this data set was then subdivided into four separate populations; individuals of European (EUR; 379 individuals), African (AFR; 246), Asian (ASN; 286) and South American origin (AMR; 181). Missense mutations identified in the study population and in the 1000Genomes population were investigated using SIFT [57] and PolyPhen-2 [58].

4 RESULTS

4.1 PAPER I

More than 100 SNPs have been associated with AR, predominantly through candidate gene studies. However, few of these associations have been replicated. Replication of previous associations is considered to be an important step in validating associations and an association is not to be considered reliable until it has been properly replicated.

Paper I presents an investigation of the general reproducibility of associations between AR and previously identified candidate SNPs. A total of 49 SNP markers were genotyped in one Swedish and one Chinese population. The Swedish study population consisted of 352 AR patients and 709 healthy control individuals, and the Chinese population of 948 AR patients and 580 control individuals.

According to the published ORs most of the previously identified candidate SNPs showed high power to detect significant AR associations in both populations of the present study. However, the overall result of the association study in **paper I** indicated that very few of the investigated SNPs were associated with AR. In the Swedish population, only 2 SNPs showed *P*-values below 0.05 and in the Chinese population there were 4. A FDR test was used to investigate the validity of the predictions. All of the indicated *P*-values gave rise to high false-discovery rates except 1 SNP in the Swedish population. Two other approaches were used in an attempt to identify additional potential associations. Correlation coefficients of ORs between the present and the original studies were calculated and the number of concordant and discordant SNPs in the direction of the ORs (>1 or <1) was also calculated. These approaches failed to detect convincing association signals for the investigated SNPs.

Paper I summarizes previously reported AR associations and also emphasizes the difficulties in replicating associations and the need for large and well-

characterized case and control materials to be able to successfully identify and replicate candidate SNPs in AR.

4.2 PAPER II

It is well known that asthma is closely related to AR. Compared to the genetics of AR the genetic background for asthma has been thoroughly investigated, and many asthma-associated genes have been successfully replicated. Thus, using genes that have been replicated multiple times in asthma as candidate genes for AR could be a successful approach to detect AR-associated genes given the co-occurrence and partly shared etiology of the two diseases.

Paper II uses the data presented for asthma in a study by Ober and Hoffjan [59] to search for AR associations. They reported 25 genes positively associated with asthma/atopy in six or more independent studies. Twenty-one of these genes were analyzed for association with AR in **paper II** using a total of 192 tSNPs covering these genes. Sequenom genotyping was made in a Swedish population consisting of 246 AR patients and 431 controls. Additionally, genotypes of 429 tSNPs from the same genes were extracted from a Singapore Chinese GWAS cohort consisting of 456 AR cases and 486 controls.

The overall association results corresponded well to the expectations in the absence of an effect for most markers. However, in the Chinese population the number of significant *P*-values exceeded the expectations. The strongest association signals were found for SNPs in *CTLA4* and *NPSR1*. In each of these genes, more than one SNP showed *P*-values <0.05 with corresponding false-discovery rates <0.05. In the *NPSR1* gene some *P*-values were lower than the Bonferroni correction level. This result indicates that *NPSR1* could be a genetic link between AR and asthma. In the Swedish population, weaker indications were found for *IL13* and *GSTP1* with respect to sensitization to birch pollen.

The use of asthma genes as candidate genes for AR associations was an alternative way to identify genes involved in the development of AR. The fact that a few potential associations were found indicates that this was a successful

strategy. Yet, the number of detected associations is low and the difficulties of finding associations largely remain.

4.3 PAPER III

Three large mGWAS have been reported in different European and North American study populations. Together these studies defined a set of 47 SNPs associated with allergic sensitization and self-reported AR.

In **paper III**, the well-characterized BAMSE cohort was used to investigate the reproducibility of the SNP associations detected in the three mGWAS. The replication was made using the same phenotype definitions as in the original studies. In addition, a more strict AR definition was used.

An absolute majority of the observed risk alleles and the allele frequency differences between cases and controls were concordant with the results of the original studies and thus supports the previously reported associations. Two out of the four loci that were identified in all three original studies were also identified in the present study. The *TLR1-TLR6* locus was identified by 3 index-SNPs and the *HLA-DQA1-HLA-DQB1* locus by 2 SNPs. In addition, the *TSLP-SLC25A46* locus that was identified in two of the original studies was identified also in the present study by 2 SNPs. These more frequently identified loci are strong candidates for harboring genetic variation truly associated with allergic disease. The tests for AR association using a more strict AR definition clearly identified the *TLR1-TLR6* locus. Thus, this more clinically relevant definition did also detect this highly replicated locus. Two additional loci, *SSTR1-MIPOL1* and *TSLP-SLC25A46*, were identified as being involved in the early onset of AR.

4.4 PAPER IV

The TLRs are important in activating immune responses upon recognition of different pathogens. Since they are central in regulating the immune response, the TLRs have been investigated for their involvement in different allergic diseases.

In **paper IV** a total of 73 SNPs covering 9 of the 10 TLR genes were genotyped in 182 AR patients and 378 control individuals in a first screen for association with AR. The association analysis revealed 1 significantly associated SNP in each of the *TLR1*, *TLR6* and *TLR7* genes, and *TLR8* showed 3 SNPs with *P*-values below 0.05. Based on the results from the association analysis, an additional 24 SNPs in the *TLR7-TLR8* gene region were analyzed in one Swedish population with 352 patients and 709 controls and one Singapore Chinese population with 948 cases and 580 controls.

Subsequent analysis of the 24 SNPs from the *TLR7-TLR8* gene region identified 7 and 5 significant SNPs in the Swedish and Chinese populations, respectively. The corresponding risk-associated haplotypes were significant after Bonferroni correction and were the most common haplotypes in both populations. The associations were primarily detected in females in the Swedish population, whereas it was seen in males in the Chinese population. Further independent support for the involvement of this region in AR was obtained from quantitative analysis of SPT data generated in both populations.

4.5 PAPER V

In **paper V** the putative promoter regions and coding sequences of the TLR genes (*TLR1-TLR10*) were searched for rare and common polymorphisms. The primary aims of the study were to search for an accumulation of rare variants in patients relative to controls and to identify AR associated-candidate mutations.

Sequencing of 288 AR patients detected a total of 156 SNPs of which 37 where located in the promoter regions. In the coding sequence, 69 were missense mutations of which 30% were classified as damaging using both SIFT and Polyphen-2 predictions, 3 were nonsense mutations and 47 were synonymous polymorphisms. The distribution of variants among the TLR genes was highly uneven. *TLR10*, *TLR1* and *TLR6* that reside in the same chromosomal locus showed a high level of variation, whereas *TLR7* and *TLR8* which are located in a narrow region on the X-chromosome showed a low level of variation.

The number of rare variants (MAF ≤1%), AR-specific variants and damaging nonsense and missense variants in the investigated 288 AR patients were compared to the numbers obtained for 379 EUR individuals of the 1000Genomes project (EUR). The overall results indicated no accumulation of rare variants in the AR population. Neither was there any overrepresentation of mutations classified as damaging by SIFT or PolyPhen-2 in the coding sequences. The only exception was the promoter region of *TLR10* where a group of 6 variants were unique to the AR population. This result was supported using a simulation test (P=0.00009). Another potential exception was a nonsense mutation, S324* in *TLR1*, estimated to 5 copies in the AR population but none in the EUR population, which is a clear overrepresentation. These results in combination with the results of **paper IV** clearly indicate the major importance of the *TLR10-TLR1-TLR6* locus for the development of AR.

5 DISCUSSION

A number of different strategies have been used in this thesis to identify genetic variation associated with AR. The first study examined the reproducibility of previously reported associations identified in small candidate gene studies. There were two main reasons for choosing this strategy. The first reason was to characterize the populations with respect to HWE and their ability to detect previous associations given different effect sizes. Another important reason for performing this study was the lack of published general replication attempts in AR. The strategy itself was of limited success since the reproducibility turned out to be very low. There are a number of reasons for this, but the major factors are probably that many of the early genetic studies in AR had small population sizes with heterogeneous phenotype definitions, resulting in false positive associations. There may of course be a number of true associations among the investigated SNPs, but their effect sizes are in that case very small resulting in a lack of statistical power for the association analysis in paper I. This illustrates the phenomenon that many early investigations in an area often are of limited value when it comes to the actual results, they have served merely as a test of methodology and concept.

The second study investigated tSNPs from highly replicated candidate genes of a related disease. The genetic background in asthma is more thoroughly investigated than in AR and since asthma and AR have in part a shared etiology, the strategy to investigate these genes also for AR is not controversial. However, the genetic backgrounds for asthma and AR do not appear to overlap to any considerable degree, since only a few of the investigated genes were identified as candidate genes also for AR. Other studies have shown that some genes and loci are associated with both diseases, e.g. *IL33* and *HLA-DQ* [38,42]. This indicates that these two diseases are related and share a common but limited genetic background.

The third strategy was analogous to the first strategy, but instead of investigating risk SNPs identified in small candidate gene studies, this study investigated the reproducibility of risk SNPs from well-powered GWAS. The replication attempt

was made in a study population with limited statistical power, but with an extensive phenotypic characterization. This was advantageous since this made it possible to investigate these associations using exactly the same phenotype definitions that were used in the original studies. This approach was clearly the most successful so far, since the majority of the observed risk alleles and the allele frequency differences between cases and controls were concordant with the results of the original studies and 2 out of 4 loci that were identified in all three original studies were also identified in the replication study. This highlights the robustness of associations detected in well-powered studies. Since the two loci (*HLA-DQA1-HLA-DQB1* and *TLR1-TLR6*) detected in all three original studies were detected also in the replication study this indicates the importance and central role of these loci in the development of AR.

Another strategy used in this thesis was to investigate genes likely to be involved in the development of AR on the basis of their known biochemical and physiological functions. The fourth study investigated common variants in the TLR genes and resulted in a positive association of the *TLR7-TLR8* locus in two different populations. In addition, weaker associations for the *TLR1-TLR6* locus were also observed in the Swedish Malmö population. Since these two loci have been indicated in a number of previous studies in AR [60,61,40-42], these loci are highly likely to be involved in the development of AR. Given this result the *TLR1-TLR10* genes were further investigated for rare variants. The fifth study identified an accumulation of rare variants in AR patients for the previously indicated locus *TLR10-TLR1-TLR6* despite the fact that a limited number of AR patients were investigated. An obvious extension to this study would be to re-sequence the entire *TLR10-TLR1-TLR6* locus in additional AR patients and controls from the same population to search for additional accumulation of rare variants in AR patients. In light of the missing heritability encountered in many complex diseases the strategy to investigate patients for rare variants has become a popular avenue following the introduction of next-generation sequencing [10].

An interesting aspect of association studies is the almost philosophical question of when an association is to be considered a true association. One can of course

not be certain until repeated functional biochemical studies of a specific variant has been thoroughly examined, the actual mechanism has been explained and the effect measured. But when is it appropriate to take this step forward? Is it when the *P*-value of an association reaches a specific threshold or the association have been replicated in additional populations a specified number of times? In my opinion, the likelihood for an association to be true increases for every time an association has been replicated showing the same risk allele. If both common and rare variants of both SNPs and CNVs can be shown to be disease-associated, the likelihood increases even more. Based on the data analyzed in this thesis, the highly indicated locus *TLR10-TLR1-TLR6* qualifies for this criterion, since it has been implicated in three large mGWAS [40-41] and in three different studies of this thesis investigating both common and rare variation.

There is a general problem with the interpretation of replications in light of publication bias. We can tabulate the number of times a given association has been reported as successfully replicated in the literature. We do not know, however, how many times this particular replication has been attempted, but been unsuccessful and therefore in many cases unreported. A replication is still a replication, but since the underlying statistics of exactly how many attempts that have been made is unknown to us it is difficult to judge the true value of this type of results. The value of a replication attempt is also quite different depending on how it is made. If a single association is being replicated this says something about that particular association, but not so much more. However, if all claims for associations to a specific disease are investigated in parallel this will generate other types of information in addition to the level of replication of the individual associations. Obviously, it will report also the negative results where claimed associations will not replicate, but in addition to this it can be evaluated if certain associations are replicated more often than others. Such highly replicated associations may indicate the existence of pathways or genes that are more often involved in the development of the disease, in contrast to those that are less often replicated that can be contributing to the disease in a small number of cases only. In my opinion this latter type of replication studies is therefore much more valuable than the former.

Also the quality of the phenotyping and the size of the study material are very important factors. The classification of the material into cases and controls can be made in a number of ways depending on the exact definitions of cases and controls. For example, cases with severe disease can be used together with so called “super-controls” that are classified as very healthy according to some logic. This may be a more efficient way to identify genes that are associated with the disease in question. The exact clinical definitions often also vary, at least slightly, between different studies and this may of course affect the results. The size of the study population is directly related to the power of the study so a larger size is of course better. But also in this case it is important to control a number of other factors. An eventual subpopulation structure may cloud the picture and it is important to control for it or even select a homogenous study material. Cases collected as patients coming to one or several clinics are a convenient source of cases, but raises questions of how to obtain well-matched controls and may lack precision in the phenotyping of the patients. Cohort studies of newborns like in the BAMSE study have a lot of different advantages as for example; definition of study material before appearance of disease, very good phenotyping with follow-up over time and possibility to control for many socio-economic and environmental factors through the evaluation of questionnaires.

Assuming now that the *TLR10-TLR1-TLR6* locus is indeed involved in the development of AR, how can we proceed? A logical first step would be to characterize the genetic variation of this complete genomic region in a large enough number of patients. Since both common variants showing frequency differences between cases and controls and rare variants showing an accumulation in cases relative to controls are identified, it seems that an unbiased investigation for all types of common and rare variants in the whole locus is the way to go forward. Thus, covering the locus with overlapping long range PCR amplicons would allow a direct screen for deletions after analysis using agarose gel electrophoresis and in combination with a scaffold of TaqMan-based copy-number assays and relative quantitative PCR analysis this strategy would detect a majority of large deletions and duplications of this region. Also smaller

deletions would be detected using this scheme. The long-range PCR products could also serve as substrate in the re-sequencing of the complete region using Ion Torrent sequencing. If this re-sequencing uses pooling of individuals this would allow screening for polymorphisms in a large number of individuals. It is difficult to know exactly how many individuals that it is necessary to screen to obtain a large enough number of candidate risk variants, but screening 1000 patients is clearly possible and would make a first evaluation possible.

Assuming that this locus indeed is harboring risk-associated variants at some frequency, this gene may serve as a beach-head directing further studies to genes and proteins directly interacting with the genes and gene products of this locus. Such a strategy would allow analysis for epistatic interactions that are possibly generated between these genes and gene products. There is a possibility that such interactions give rise to strong effects that will explain considerably more of the variation of the trait than what the single factors do. Another possible approach for further studies would be to re-sequence the exomes of as many patients as possible. This is a more unbiased approach since it do not rely on any selection of genes for study. But is also a more demanding approach in terms of the number of individuals that must be analyzed to obtain a large enough data set to allow identification of the disease-associated genes. However, re-sequencing of exomes, much like GWAS, is suitable for meta-analysis and one obvious way forward is that researchers in the field join forces, each re-sequencing as many exomes that there funding will allow and then this data could be merged to one large data set with sufficient power to identify the elusive disease-associated genes that we are so eager to identify.

6 CONCLUSIONS

Since the development of AR depends on both environmental and genetic factors, the individual effect of a single polymorphism is expected to be small. Large and well-characterized populations are therefore crucial for the identification of robust associations. Conversely, smaller populations with less stringent phenotyping are less likely to identify true associations. This is illustrated in **paper I** where 49 previously AR-associated SNPs were tested in a replication study. The original studies were all based on relatively small populations and the definition of AR also differed between the studies. The overall result showed a low replication rate and that very few of the investigated SNPs were associated with AR in the two populations used in the replication attempt. The study also indicated that the odds-ratios were inflated in most of the original studies, and a low concordance of risk alleles between the studies was observed.

Replication studies are very important for validating reported associations. The genetic background of asthma has been extensively studied and replicating previous associations have been far more successful compared to AR. AR and asthma are considered to be closely related and genes that have been replicated multiple times in asthma might also turn out to be candidate genes for AR. In **paper II** this strategy was evaluated by investigating 21 genes highly replicated in asthma. The *NPSR1* and *CTLA4* genes were identified as potential links between AR and asthma, with *NPSR1* as the strongest candidate. However, the majority of the investigated genes showed no or a modest level of association. Since the strategy to use asthma genes as candidate genes for AR only had limited success, it is tempting to suggest that genetics in AR and asthma might not share a common etiology to the extent that was previously argued for.

GWAS have been a successful approach to identify potential candidate genes in complex diseases. Even though the identified variants only explain a small portion of the disease, they provide leads to relevant genes and biochemical pathways involved in the disease mechanisms. Three meta-GWAS investigating AR phenotypes successfully identified a total of 47 index SNPs at a genome-

wide or suggestive association level. In **paper III**, these SNPs were investigated in a replication study using the same SNPs and the same phenotype definitions as in the mGWAS. The results showed a much higher replication rate than what was observed in **paper I**, indicating more robust associations. Two out of four loci (*TLR6-TLR1* and *HLA-DQA1-HLA-DQB1*) identified by all three original studies were also detected in the replication study using the same phenotype definitions and a more strictly defined AR phenotype. These two loci are likely to have a central role in the epidemiology of allergic disease. In addition, associations between genetic variation in the *SSTR1-MIPOL1* and *TSLP-SLC25A46* loci and age at which the allergic symptoms started was also identified. Notably, this was the first report of age at onset effects in allergic rhinitis.

The Toll-like receptors induce a prompt immune response upon pathogen recognition. Since they are central in regulating the immune response, the TLRs have been investigated for their involvement in different allergic diseases. In **paper IV**, common genetic variation in the TLR-genes was investigated for association with AR in two ethnically different populations. The *TLR7-TLR8* locus was identified as associated with AR in both populations in a sex-specific manner. However, different sexes were associated in the respective populations probably reflecting the loss of power when investigating X chromosomal SNPs in men. In addition, weak associations were also observed for *TLR1* and *TLR6*.

Since the TLR genes has been found to harbor genetic variants associated with AR in a number of different studies it is highly likely that these genes contribute to the development of AR. In the previous studies, focus has primarily been on common variation. In **paper V** the focus is shifted towards rare variation in both the coding sequence and the putative promotor regions. Comparisons of TLR sequence data from 288 AR patients with a European subset of the 1000Genomes project showed limited signs of accumulation of rare variants in the AR population. Also, no signs of any excess of damaging missense mutations were observed. The promotor region of *TLR10* was an exception from this general trend, where a total of six SNPs with MAF < 1% were detected in

the AR population. These SNPs were not present in the European subset or any of the other populations of the 1000Genomes project. Another potential exception was a nonsense mutation, S324* in *TLR1*, estimated to 5 copies in the AR population but none in the EUR populations, which is a clear overrepresentation. This indicates that both common and rare SNPs in the TLR genes contribute to AR.

7 ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to everybody that have contributed to this thesis. In particular, I would like to thank:

Lars Olaf Cardell, my main supervisor, for giving me the opportunity to do my PhD in your group and for your invaluable input on allergic diseases.

Christer Halldén, my co-supervisor, for believing in me and taking me under your wings, for your guidance and support, and for all the valuable input on my papers. You have an ability to always get the logic right and are a never ending source of inspiration.

Torbjörn Säll, my co-author, for your extremely valuable input on population genetics and statistics, for always helping and supporting me, and for all the amusing stories you have told me.

Annika Lidén, for valuable help in the lab, for organizing samples and making sense of disorganized databases, and for your perfectionism in reviewing manuscripts and this thesis, you always know when a comma is missing!

Viktor Henmyr, my co-author and colleague, for sharing all this time behind our computers analyzing and exploring the data in paper IV and V, for the competitive atmosphere during the Bioinformatics algorithms course, and for great conversations during coffee breaks.

Eric Manderstedt, my co-author and colleague, for all the help with NGS and input in paper V, for your controversial ideas and amusing discussions.

Christina Lind-Halldén, my co-author and colleague, for supporting me, for always having a joyful attitude, and for producing such nice sequencing data. The variant calling process is less painful with such high-quality data!

Terese Hylander, *Camilla Rydberg Millrud* and *Susanna Kumlien Georén*, my supportive and helpful colleagues, for all your help guiding me in the special situation it is to be a PhD-student.

Rolf Uddman, for taking the time to proofread this thesis

Agneta Wittlock, for all your help with administrative questions and concerns.

Anand Kumar Andiappan, *Chew Fook Tim* and *Wang De Yun*, my co-authors and collaborators, for your contributions on paper I, II and III.

My former co-workers at Department of Clinical Chemistry in Malmö, for all help and guidance in the lab.

My colleagues at Kristianstad University, for nice conversations during coffee breaks.

Barbro and *Pia*, my former teachers at Forum Ystad, for encouraging me to further studies

Mom, *dad*, *Sven*, *Gertie*, *Christian*, *Cecilia*, *Sara* and *Lina*, for always being there for me and supporting me.

Elsa and *Oscar*, my children, for being the most wonderful children one could ever whish, I love you very much.

Sofia, my wife, for your support and patience, I know it has been a tough time lately with two small children about 6 months old. You have done an amazing job in taking care of them, the house, the horses, the chickens, the cats and me while I have been busy working on this thesis. I love you!

8 REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al.; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001; 409:860-921. Erratum in: Nature 2001; 412:565. Nature 2001; 411:720.

2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001; 291:1304-51. Erratum in: Science 2001; 292:1838.

3. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 2003; 4:587-97. Review.

4. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet 2008; 9:477-85. Review.

5. International HapMap Consortium. A haplotype map of the human genome. Nature 2005; 437:1299-320.

6. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007; 449:851-61.

7. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature 2010; 467:52-8.

8. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature 2010; 467:1061-73. Erratum in: Nature. 2011; 473:544.

9. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491:56-65.

10. Check Hayden E. Giant gene banks take on disease. Nature 2014; 514:282.

11. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 2009; 19:212-9. Review.

12. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 2010; 11:446-50.

13. Riggs ER, Ledbetter DH, Martin CL. Genomic Variation: Lessons learned from whole-genome CNV analysis. Curr Genet Med Rep 2014; 2:146-150. Review.

14. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 2009; 324:387-9.

15. Azzopardi D, Dallosso AR, Eliason K, Hendrickson BC, Jones N, Rawstorne E, et al. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. Cancer Res 2008; 68:358-63.

16. Ma X, Liu Y, Gowen BB, Graviss EA, Clark AG, Musser JM. Full-exon resequencing reveals Toll-like receptor variants contribute to human susceptibility to tuberculosis disease. PLoS ONE 2007; 2:e1318.

17. Van Pelt-Verkuil E, van Belkum A, Hays JP. Principles and technical aspects of PCR amplification. Springer Science + Business Media B.V. 2008. ISBN 978-1-4020-6240-7.

18. Tost J, Gut IG. Genotyping single nucleotide polymorphisms by MALDI mass spectrometry in clinical applications. Clin Biochem 2005; 38:335-50. Review.

19. Meyer K, Ueland PM. Use of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for multiplex genotyping. Adv Clin Chem 2011; 53:1-29. Review.

20. Brown S. Microarray. In: Concise Encyclopaedia of bioinformatics and computational biology (eds JM Hancock and MJ Zvelebil) Wiley Blackwell 2014; pp 420-423. ISBN 978-0-4709-7871-9.

21. Rothberg JM1, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature 2011; 475:348-52.

22. Merriman B, Ion Torrent R&D Team, Rothberg JM. Progress in ion torrent semiconductor chip based sequencing. Electrophoresis 2012;33:3397-417.

23. Dawn Teare M, Barrett JH. Genetic linkage studies. Lancet 2005; 366:1036-44. Review.

24. Lewis CM. Genetic association studies: design, analysis and interpretation. Brief Bioinform 2002; 3:146-53. Review.

25. Lewis CM, Knight J. Introduction to genetic association studies. Cold Spring Harb Protoc 2012; 2012:297-306. Review.

26. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet 2014; 15:335-46. Review.

27. Edwards AW. G. H. Hardy (1908) and Hardy-Weinberg equilibrium. Genetics 2008; 179:1143-50.

28. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet 2006; 7:781-91. Review.

29. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003; 100:9440-5.

30. Brozek JL, Bousquet J, Baena-Cagnani CE, Bonini S, Canonica GW, Casale TB, et al.; Global Allergy and Asthma European Network; Grading of Recommendations Assessment, Development and Evaluation Working Group. Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines: 2010 revision. J Allergy Clin Immunol 2010; 126:466-76.

31. Galli SJ, Tsai M, Piliponsky AM. The development of allergic inflammation. Nature 2008; 454:445-54. Review.

32. Asher I, Pearce N. Global burden of asthma among children. Int J Tuberc Lung Dis 2014; 18:1269-78.

33. Maddox L, Schwartz DA. The pathophysiology of asthma. Annu Rev Med 2002; 53:477-98. Review.

34. Koppelman GH, Hall IP. Asthma genetics 2014: reaching for high-hanging fruit. Clin Exp Allergy 2014; 44:1296-8.

35. Fagnani C, Annesi-Maesano I, Brescianini S, D'Ippolito C, Medda E, Nisticò L, et al. Heritability and shared genetic effects of asthma and hay fever: An Italian study of young twins. Twin Res Hum Genet 2008; 11:121-31.

36. Willemsen G, van Beijsterveldt TC, van Baal CG, Postma D, Boomsma DI. Heritability of self reported asthma and allergy: A study in adult Dutch twins, siblings and parents. Twin Res Hum Genet 2008; 11:132-142.

37. Nilsson D, Andiappan AK, Halldén C, Tim CF, Säll T, Wang DY, et al. Poor reproducibility of allergic rhinitis SNP associations. PLoS ONE 2013;8:e53975.

38. Portelli MA, Hodge E, Sayers I. Genetic risk factors for the development of allergic disease identified by genome wide association. Clin Exp Allergy 2014. doi: 10.1111/cea.12327. [Epub ahead of print]

39. Andiappan AK, Wang de Y, Anantharaman R, Parate PN, Suri BK, Low HQ, et al. Genome-wide association study for atopy and allergic rhinitis in a Singapore Chinese population. PLoS ONE 2011; 6:e19719.

40. Ramasamy A, Curjuric I, Coin LJ, Kumar A, McArdle WL, Imboden M, et al. A genome-wide meta-analysis of genetic variants associated with allergic rhinitis and grass sensitization and their interaction with birth order. J Allergy Clin Immunol 2011; 128:996-1005.

41. Bønnelykke K, Matheson MC, Pers TH, Granell R, Strachan DP, Alves AC, et al. Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. Nat Genet 2013; 45:902-906.

42. Hinds DA, McMahon G, Kiefer AK, Do CB, Eriksson N, Evans DM, et al. A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. Nat Genet 2013; 45:907-911.

43. Sichletidis L, Markou S, Daskalopoulou E, Constantinidis T, Tsiotsios J, Pechlivanidis T. The prevalence of asthma and allergic rhinitis among children in Greece. Am J Respir Crit Care Med 1999; 159:A143-A143.

44. Linneberg A, Nielsen NH, Frolund L, Madsen F, Dirksen A, Jorgensen T. The link between allergic rhinitis and allergic asthma: a prospective population-based study. The Copenhagen Allergy Study. Allergy 2002; 57:1048-1052.

45. Leynaert B, Neukirch C, Kony S, Guenegou A, Bousquet J, Aubier M, et al. Association between asthma and rhinitis according to atopic sensitization in a population-based study. J Allergy Clin Immunol 2004; 113:86-93.

46. Vercelli D. Discovering susceptibility genes for asthma and allergy. Nat Rev Immunol 2008; 8:169-82. Review.

47. Wickman M, Kull I, Pershagen G, Nordvall SL. The BAMSE project: presentation of a prospective longitudinal birth cohort study. Pediatr Allergy Immunol 2002; 13:11-13.

48. Ballardini N, Kull I, Lind T, Hallner E, Almqvist C, Östblom E, et al. Development and comorbidity of eczema, asthma and rhinitis to age 12 – data from the BAMSE birth cohort. Allergy 2012; 67:537-544.

49. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2009 ISBN 3-900051-07-0, URL http://www.R-project.org.

50. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012 ISBN: 3-900051-07-0, URL http://www.R-project.org/

51. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81:559-75.

52. Storey, JD. A direct approach to false discovery rates. J R Stat Soc Series B Stat Methodol 2002; 64:479-98.

53. Dabney A, Storey JD, Warnes GR. qvalue: Q-value estimation for false discovery rate control. R package version 1.32. 2014. http://www.bioconductor.org/packages/release/bioc/html/qvalue.html. Accessed February 12, 2014.

54. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005; 21:263-5.

55. Li H. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 2011; 27:718-9.

56. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011; 27:2156-8.

57. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 2012; doi: 10.1093/nar/gks539.

58. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010; 7:248-9.

59. Ober C, Hoffjan S. Asthma genetics 2006: the long and winding road to gene discovery. Genes Immun 2006; 7:95-100.

60. Haagerup A, Borglum AD, Binderup HG, Kruse TA. Fine-scale mapping of type I allergy candidate loci suggests central susceptibility genes on chromosomes 3q, 4q and Xp. Allergy 2004; 59:88-94.

61. Møller-Larsen S, Nyegaard M, Haagerup A, Vestbo J, Kruse TM, Borglum AD. Association analysis identifies TLR7 and TLR8 as novel risk genes in asthma and related disorders. Thorax 2008; 63:1064-1069.