LUND UNIVERSITY

Genetics of complex disease

Henmyr, Viktor

2017

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Henmyr, V. (2017). *Genetics of complex disease*. [Doctoral Thesis (compilation), Department of Biology]. Lund University, Faculty of Science, Department of Biology.

Total number of authors:
1

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Genetics of complex disease

VIKTOR HENMYR
FACULTY OF SCIENCE | DEPARTMENT OF BIOLOGY | LUND UNIVERSITY

Genetics of complex disease

Genetics of complex disease

Viktor Henmyr

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty of Science, Lund University, Sweden.
To be defended at Biologihuset A, Hörsalen.

Friday, September 29, 2017, 9:00.

Faculty opponent
Assoc. Prof. Staffan Nilsson, Chalmers

Organization LUND UNIVERSITY	Document name DOCTORAL DISSERTATION
	Date of issue
Author(s) Viktor Henmyr	Sponsoring organization: Sven och Lily Lawskis fond för naturvetenskaplig forskning, personal PhD-stipend

Genetics of complex disease

To find true genetic associations in complex diseases, such as allergic rhinitis and chronic rhinosinusitis, is a difficult task. Both are polygenetic diseases where it is believed that genetic variation in several genes make up their phenotypes. In addition, the environment of the affected individuals also play an important role, further obscuring eventual associations. Today most genetic association studies of polygenetic disease perform tests on the whole-genome level, but this has not always been the case. By 2012 a total of 56 genetic association studies had been published in allergic rhinitis and by 2013 a total of 27 studies in chronic rhinosinusitis. In but a few exceptions, all studies of the two diseases targeted candidate genes in small case and control groups. Replication attempts of significant associations from previous studies has shown some promise by pointing out more likely candidates. However, these associations do not explain the heritability of the studied diseases to any great extent. To find this missing heritability researchers have to look beyond common variants and instead focus on other potential targets. Re-sequencing studies of candidate genes could find rare variants with much higher effects sizes and studies of RNA and proteins should also prove to be helpful. New technologies and advances in molecular biology are looking promising but it is still hard to interpret all the new data. Polygenetic disorders are hard to study, low effect sizes makes novel candidate genes hard to discover. However, it is a numbers game, as more and more studies are made, more information will be available and eventually the genetic component, however big or small it is, of complex diseases will be solved.

Key words complex disease, allergic rhinitis, chronic rhinosinusitis, genetics, associations, GWAS, re-sequencing

Classification system and/or index terms (if any)

Supplementary bibliographical information		Language English
ISSN and key title		ISBN 978-91-7753-387-0; 978-91-7753-388-7
Recipient's notes	Number of pages	Price
	Security classification	

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature

Date 2017-08-24

Genetics of complex disease

Viktor Henmyr

LUND
UNIVERSITY

Cover photo by Viktor Henmyr

Copyright Viktor Henmyr

Faculty of Science
Department of Biology

ISBN 978-91-7753-387-0, 978-91-7753-388-7

Printed in Sweden by Media-Tryck, Lund University
Lund 2017

“The laws of genetics apply even if you refuse to learn them”

- *Allison Plowden*

Content

List of scientific papers 10
Contributions 11
Populärvetenskaplig sammanfattning 12
List of abbreviations 14

1. Aims of this thesis 15

2. Introduction 17
2.1 The human genome 17
2.2 Genetic association studies 18
2.2.1 Study design and strategy 19
2.2.2 Haplotypes and linkage disequilibrium 19
2.2.3 Case control studies 20
2.3 Statistical analysis 20
2.4 Replication 22
2.5 Interpretational difficulties 22
2.6 Complex diseases and missing heritability 23
2.7 Allergic disease 25
2.7.1 Allergic rhinitis 26
2.7.2 Genetic association studies in AR 26
2.7.3 Chronic rhinosinusitis 28
2.7.4 Genetic association studies in CRS 29

3. Materials and methods 31
3.1 Subjects 31
3.1.1 Belgian CRS population 31
3.1.2 Malmö AR population 31
3.1.3 BAMSE AR population 32
3.2 Databases 32
3.2.1 The HapMap project 33
3.2.2 The 1000 genomes project 33
3.2.3 Exome Aggregation Consortium 34
3.3 Simulations and permutations 34
3.4 Bioinformatics 35

4. Paper summaries....37
4.1 Paper I....37
4.1.1 Introduction....37
4.1.2 Results....37
4.2 Paper II....38
4.2.1 Introduction....38
4.2.3 Results....38
4.3 Paper III....38
4.3.1 Introduction....38
4.3.2 Results....39
4.4 Paper IV....39
4.4.1 Introduction....39
4.4.2 Results....40
4.5 Paper V....40
4.5.1 Introduction....40
4.5.2 Results....41
5. Discussion....43
6. Conclusions....49
7. Acknowledgements....51
8. References....53

List of scientific papers

I. Henmyr V, Vandeplas G, Halldén C, Säll T, Olze H, Bachert C, Cardell LO.
Replication study of genetic variants associated with chronic rhinosinusitis and nasal polyposis.
J Allergy Clin Immunol. 2014;133(1):273-5.

II. 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(11):1506-14.

III. Henmyr V, Lind-Halldén C, Carlberg D, Halldén C, Melén E, Wickman M, Bergström A, Säll T, Cardell LO.
Characterization of genetic variation in TLR8 in relation to allergic rhinitis.
Allergy. 2016;71(3):333-41.

IV. Henmyr V, Lind-Halldén C, Halldén C, Säll T, Carlberg D, Bachert C, Cardell LO.
Chronic Rhinosinusitis Patients Show Accumulation of Genetic Variants in PARS2.
PLoS One. 2016;11(6):e0158202.

V. Henmyr V, Carlberg D, Manderstedt E, Lind-Halldén C, Säll T, Cardell LO, Halldén C.
Genetic variation of the Toll-like receptors in a Swedish allergic rhinitis case population.
BMC Med Genet. 2017;18(1):18.

Contributions

I. Conceived and designed the experiments: **VH** CH CB LOB. Performed the experiments **VH**, HO. Analyzed the data: **VH** HO TS. Statistical analysis: **VH** TS. Wrote the paper: **VH** CH TS CB LOB.

II. **All** authors contributed to the design of the experiments. DN, **VH**, CH, and TS performed statistical analysis. IK, MW, EM, and LOC acquired phenotypic data, and **all** authors contributed to the writing of the manuscript and approved the final version of the manuscript.

III. **All** authors contributed to the design of the experiments. DN, **VH**, CH, and TS performed statistical analysis. CLH performed Sanger sequencing. MW, EM, and LOC acquired phenotypic data and **all** authors contributed to the writing of the manuscript and approved the final version of the manuscript.

IV. Conceived and designed the experiments: **VH** CLH CH TS DC CB LOC. Performed the experiments: CLH. Analyzed the data: **VH** TS DC. Contributed reagents/materials/analysis tools: CB LOC CH. Wrote the paper: **VH** CLH CH TS DC CB LOC.

V. **All** authors contributed to the design of the experiments. DC, **VH** and TS performed statistical analysis. EM produced Ion Torrent sequencing data and CLH produced Sanger sequencing data. **VH** and DC performed bioinformatics analysis. LOC acquired phenotypic data and **all** authors contributed to the writing of the manuscript and approved the final version of the manuscript.

Populärvetenskaplig sammanfattning

Många sjukdomar som inte beror på virus eller bakterier kopplas ofta till genetik. Vår förståelse kring hur gener påverkar vår hälsa har ökat dramatiskt med molekylärbiologins framfart. Sedan 1930-talet har det utvecklats nya tekniker för att tyda gener och dess DNA. Länge var gener ett ganska abstrakt koncept, man visste att det fanns något som fördes vidare genom generationerna men inte exakt vad. Det gick att observera hur vissa anlag fördes vidare, medan andra försvann eller hoppade över en generation. I det mesta var det enkel genetik; en gen, en egenskap. En egenskap kan syfta på både yttre utseende som kroppslängd eller ögonfärg, eller en sjukdom så som allergier. På 1970-talet kom de första metoderna för att sekvensera DNA, alltså att få information om exakt i vilken följd de fyra olika DNA-baserna, A, C, G och T satt i. Med den vetskapen gick det att beskriva vilka mutationer som orsakade vissa egenskaper. Detta var fortfarande på nivån, en egenskap, en gen.

I början av 2000-talet presenterades hela människans genom. I ett gigantiskt projekt hade varenda en av människans tre miljarder baspar kartlagts. Detta var början på något riktigt stort inom genetik, men det var bara en människas genom, och det hade tagit flera år att sammanställa. Betydligt snabbare metoder men med mindre information började komma i samma veva. I människans tre miljarder stora genom finns det små skillnader mellan individer och även mellan kromosomparen i samma individ. Dessa kallas för varianter. En metod som blev mycket populär var ett litet chip som kunde diagnostisera omkring miljoner av dessa varianter. Metoden kom att kallas genome-wide association studies, genomsträckande associationsstudier. En associations-studie letar efter skillnader mellan friska och sjuka grupper av individer, ofta i form av att jämföra frekvensen av vissa varianter. Varianterna som valdes ut i dessa studier agerade som temperaturmätare, där varianten satt fanns också en gen, och beroende på vilken DNA-bas som satt där, kunde olika egenskaper förväntas. Med dessa associationsstudier förväntades att alla genetiska sjukdomar skulle lösas. Men det var inte så lätt, istället insåg forskare att en egenskap kan bero på flera gener. Trots denna genomsträckande metod kan vi än idag inte enbart med hjälp av dessa varianter förklara varför vissa är längre än andra, och varför vissa får allergier och inte andra.

Istället har utvecklingen gått vidare, och gett forskare ännu fler verktyg i jakten på förklaringsmodeller. Idag går det att sekvensera ett helt människogenom på några timmar och till ett relativt lågt pris, jämfört med det första människogenomet som tog flera år och som dessutom kostade åtskilliga miljoner. Med informationen kring hur en frisk människas DNA ser ut och med hur en sjuks ser ut, borde vi inte veta exakt var som orsakar sjukdomen? Nej, det verkar inte vara så enkelt. Först och

främst är det svårt att avgöra vilka skillnader som är ”normal” variation och vilka som faktiskt orsakar specifika egenskaper. En annan faktor är studiedesignen. Det krävs ofta stora grupper av både friska och sjuka då det ger starkare bevisningskraft. Dessa grupper bör också vara väldefinierade, till exempel att alla har exakt samma typ av allergi, har genomgått samma typer av undersökningar och framförallt har samma etniska ursprung. Det sistnämnda är viktigt då även fast skillnader mellan folkgrupper är små på genomisk skala, så kan små skillnader i ofarliga egenskaper ge utslag som falska positiva, alltså ett resultat som inte alls är kopplad till sjukdomen. De flesta studier inom allergier har och fortsätter fokusera på små grupper av friska och sjuka, och med begränsade metoder med enstaka varianter. Dessutom görs för få försök att återupptäcka samma resultat som tidigare forskare funnit. Detta är ett viktigt koncept inom all forskning, kan inte resultatet upprepas kanske det inte heller var så viktigt.

Genetiska studier i allergier hade tjänat på samarbete, dels genom att då kunna studera större grupper av individer, satsa på mer genom-sträckande studier och då framförallt på helgenomsekvensering. Även fast associationsstudier har varit lärorika tillför de inte lika mycket information som sekvenseringsstudier kan. Genom att följa upp tidigare, validerade, resultat från associationsstudier finns det all möjlighet att sekvenseringsstudier kommer hitta viktiga mutationer som kan behandlas och på så sätt dämpa eller till och med bota allergier. All information som erhålls från sekvenseringsstudier adderar till en fortsatt växande databas av information, och därför bör fältet röra sig åt detta håll.

I min forskning har jag funnit att validering av gamla resultat är oerhört viktigt. Ett ensamt resultat som pekar ut en gen säger ganska lite. Om däremot flera resultat konsekvent pekar ut en och samma gen, oberoende av varandra, kan vi lita mer på resultatet. Vidare, associationer på vanliga varianter alltså de så kallade temperaturmätarna ger ganska lite information om vad som egentligen orsakar sjukdomen. För att hitta förklaringar krävs fler sekvenseringsstudier, precis som de vi utfört. I fallet allergisk rinit där vi följt upp den bästa, mest validerade signalen i *TLR10-TLR1-TLR6* locus har vi också hittat viktiga ovanliga varianter som kan ha en inverkan på vilka som blir sjuka i hösnuva och vilka som inte blir. Däremot krävs det betydligt fler och större studier innan vi vet exakt vad som orsakar sjukdom.

List of abbreviations

AR	Allergic rhinitis
bp	Base pairs
CD-CV	Common disease, common variant
CD-RV	Common disease, rare variant
CI	Confidence interval
CNV	Copy number variation
CRS	Chronic rhinosinusitis
ExAC	Exome Aggregation Consortium
FDR	False discovery rate
GWAS	Genome-wide association studies
IgE	Immunoglobulin E
LD	Linkage disequilibrium
MAF	Minor allele frequency
OR	Odds ratio
PARS2	Prolyl-tRNA synthetase 2
SNP	Single-nucleotide polymorphism
TLR	Toll-like receptor

1. Aims of this thesis

Genetic association studies in complex diseases have evolved as new technology has been introduced, going from few single-nucleotide polymorphisms (SNP) to millions and all the way to full-scale genomic sequencing. Finding associated genes was thought to be easy with the introduction of genome-wide association studies. Still there is an apparent lack of explanation of cause in our current understanding of the genome and common diseases such as allergy.

Even though most allergic diseases show strong inheritance patterns few genes have been directly linked. There are two aims with this thesis, 1) finding likely (gene) candidates through replicating previous findings and 2) further investigating these candidates through re-sequencing, which has never been done in either disease before, to hopefully find causative variants. To our disposal we had two populations suffering from allergic rhinitis and one population suffering from chronic rhinosinusitis. Earlier studies of these phenotypes have been limited and few replication attempts have been done. By going from replication to targeted re-sequencing our hope is to shed some new light to the genetic causes of these diseases.

2. Introduction

2.1 The human genome

DNA, short for deoxyribonucleic acid, is a molecule built of four different bases: adenine (A), cytosine (C), guanine (G) and thymine (T). Together the four bases are organized as strands of bases which pairs with a complementary strand of bases, where A pairs with T and C pairs with G. This is called double-stranded DNA and each pair of bases along the double-strand are called base pairs. The human genome is built of around 3 billion of these base pairs (bp). Different combinations of these base pairs, varying from a few hundred bp to 2.4 million bp, create genes. As of the latest genome assembly GRCh38.p10 there are 20 310 genes that code for proteins. The genes are arranged on larger segments called chromosomes of which humans have 2n = 46. 2n is an indication that humans are diploid and each offspring receives 23 (n) different chromosomes from each parent. Each chromosome is of different size, chromosome 1 being the largest with 2044 genes and chromosome Y being the smallest with 63 genes. Chromosomes 1-22 are called the autosomes which every normal human carry two copies of. The sex chromosomes X and Y are carried in different amounts depending of biological sex. Females carry two X chromosomes whereas males carry one X and one Y chromosome. Loss or gain of whole or parts of chromosomes is usually detrimental for the offspring and most fetuses with abnormalities are miscarried. However there are cases such as Down's syndrome (three chromosome 21) and Klinefelter Syndrome (XXY males) that give rise to live offspring that can live a relatively normal life.

The DNA of genes are not by themselves involved in processes such as metabolism but rather it is the proteins of which the genes code for. This does however mean that a change in DNA can alter the protein, and this can in turn alter how the body functions and even alter how we appear, this is frequently termed as our phenotype. Overall most humans are nearly identical in the DNA of their genes. Small changes, polymorphisms, in the DNA have, and still do happen which has led to the phenotypic variability seen in humans. The genetic composition of an individual is often termed the genotype. When studying the genotype, changes are usually in form of single nucleotide polymorphisms (SNP), i.e. a change in a single base pair. When referring to the frequency of a polymorphism it is common to denote the minor-allele frequency (MAF). There are in most cases only two alleles possible at one

location in the genome, and the MAF is the less common of the two, and by knowing the MAF you also know the frequency of the other allele (1-MAF). Changes such as curly or flat hair, brown or white skin or the color of our irises are mostly benign and do not impact the overall survival rate of the individual. However, some changes can have adverse effects such as the singe mutation to one base pair in the cystic fibrosis transmembrane regulator gene, which causes the life-threatening disorder cystic fibrosis. Thanks to genetic association studies we now know exactly where this mutation is located and can screen for it[1].

2.2 Genetic association studies

Genetic studies of disease are an ever-changing field of science. As technical advances are made within molecular biology and computing, changes are made in how genetic association studies are performed. There are two major blocks of genetic studies, one looking at monogenetic diseases and one looking at polygenetic diseases. The monogenetic diseases, one gene, one disease, were the first to be studied. The first studies used small repetitive DNA sequences called micro-satellites as genetic markers and observed their co-segregation with disease in family materials. This strategy could pinpoint genomic regions and eventually also specific genes. With technical advances came the possibility to re-sequence large pieces of the genome and the strategy now focused on finding specific deleterious mutations in specific genes, also in family materials[2, 3]. The research of polygenetic diseases, one disease, many genes, started off later, as more and more monogenetic diseases were being resolved. At first, the strategy consisted of looking at candidate genes in small populations of cases and controls. These found many associations but also many false positives creating a lot of noise. The introduction of chip technology containing millions of genetic markers in combination with ever increasing population sizes significantly improved the signal to noise ratios. However, it has proven difficult to determine most causes of disease through these methods. The latest era of studies of polygenetic diseases has become an endless search for the missing pieces of the puzzle through re-sequencing of cases and controls, similar to the monogenetic diseases, but on the genomic level[4].

Genetic association studies try to correlate genetic variation in genomic regions or specific genes to a specific disease or phenotype. This is usually achieved through direct comparison of allele and/or genotype frequencies of common SNPs between groups of healthy and affected individuals. Additionally, genetic variation such as copy number variants (CNV) can also be analyzed. Genetic association studies are especially powerful in diseases with a complex nature, discussed later in this thesis.

In such diseases both genetic and environmental components play important roles[4].

2.2.1 Study design and strategy

There are two strategies that are utilized, candidate genes studies and genome-wide association studies (GWAS). Candidate gene studies use *a priori* knowledge to hypothesize what genes might be involved in the disease under study. There are a few different means of selecting candidate genes: 1) looking at the function of the gene, for example a gene regulating cell division could be important for cancer research, 2) looking at a selection of genes within a biological pathway, 3) looking at genes previously associated to similar phenotypes and 4) replication of previous findings. GWAS, on the other hand considers the complete genome and are thereby hypothesis free. By utilizing large study populations and SNP chips that contain > 500 000 SNPs, the small effects (odds ratios of 1.2-5.0, see section 2.3) often observed in polygenetic diseases are detectable. This strategy results in a scenario of extreme multiple testing, i.e. in a chip of 500 000 SNPs you would by chance expect around 25 000 significant SNPs at the nominal 5% level. Therefore a more stringent level of significance, $P < 1.0\times10^{-8}$ or lower, is applied in most studies. In addition, there are other significance correction methods relevant not only to GWAS but also to candidate gene studies. Two of these are; 1) Bonferroni correction, which is calculated by dividing the *P*-value by the total number of tests performed, and 2) false positive estimates such as *Q*-values which take the *P*-value distribution of all tests into account[5, 6].

2.2.2 Haplotypes and linkage disequilibrium

The associated alleles are often not the cause of the disease itself. However, the causal genetic variant is often located close to the associated variant. Detecting the causal variant through nearby located SNPs is achieved by making use of a genetic property of alleles, namely their linkage disequilibrium (LD). Two loci are said to be in LD when one locus with a specific allele can predict the allele at another locus, i.e. the alleles of the two loci are not randomly associated. This is mostly due to recombination as there is a small chance for a recombination event to occur in between two closely located loci. LD can also be disrupted by mutations and genetic drift and the LD structure is often different across world populations. A haplotype is defined by the alleles along one chromosome and loci with common alleles and the LD structure they make up define the frequency and extent of the haplotypes. Instead of trying to look right at the deleterious variant, researchers investigate common variants in the genome that are in LD with a causative variant. This will

narrow down the search window to specific haplotypes of specific genes where deleterious mutations might be present[5, 7]. However, in rare variation studies (re-sequencing) it is the actual associated variant which also is the causing variant, more on this later in the thesis.

2.2.3 Case control studies

Whether a candidate gene or GWAS strategy is applied, the most straight-forward study design is the case-control study which consists of two sets of individuals, the cases and the controls. The two groups should be as well matched as possible both ethnically and with respect to sex and age. The cases are a group of individuals selected for a specific trait, often a common disease. The controls can be either recruited specifically as disease-negative, i.e. phenotypically not carrying the disease or they can be recruited as background population with the corresponding expected population frequency of the disease. The major difference between those alternatives comes down to power, as using disease-negative controls increases the chance of detecting actual differences in allele frequencies between cases and controls. However, the increase in power often comes with an increase in cost as disease-negative controls need to be clinically phenotyped. Regardless of the recruitment strategy of the control population, the controls should be ethnically matched to the cases to avoid false positive signals due to population differences. The better the match of cases to controls, also called mapping 1:1, the less heterogeneity will be created. Heterogeneity amongst cases and controls can cause false positives, where random benign genetic variation is being falsely associated with the disease[4].

2.3 Statistical analysis

Testing in case-control studies is usually done on the level of the SNP. A single SNP with two alleles can generate three different genotypes. From this information two different contingency tables can be created: the 2 x 3 table of the three genotypes over the two groups of individuals and the 2 x 2 table for the two alleles over the two groups of individuals. The test for association is carried out by a Chi-Square test for each SNP separately. Table 1 shows the typical layout of a contingency table for a case-control study. The expected count for X11, E(X11), would be X1.*X.1/2n and similarly X2.*X.2/2n for E(X22). The general formula to calculate the association would then be; $\chi 2 = (X11-E(X11))^2/E(X11)+\ldots+(X22-E(X22))^2/E(X22)$, where the test value is two-tailed with one degree of freedom.

This same principle applies to genotype testing as well but results in a two-tailed test value with two degrees of freedom.

Table 1: Contingency table for case-control studies at the allele level.

Group/Allele	a	A	Total
Cases	X11	X12	X1.
Controls	X21	X22	X2.
Total	X.1	X.2	2n

It is also common to quantify the effect of the associated allele. This is done through calculating the odds and is usually referred to as an odds ratio (OR). This can also be calculated using the contingency table. The equation is X12*X21/X11*X22. The ratio acquired through this method is the increased risk, meaning that an OR value of 2.0 doubles the risk of having the disease given the presence of the risk allele[4, 8]. ORs are usually accompanied with the 95% confidence interval (CI) and will be done so throughout this thesis.

Another often-used statistical method is logistic regression. This is a regression model with a binary outcome, 0 or 1, sick or healthy, where the dependent variable is categorical and the independent variable is explanatory. In genetic association studies this can be used in multivariate analysis of several genetic markers (SNPs) to evaluate their impact on the disease, for example to define a haplotype, a set of specific alleles, that is likely to be causing disease. It can also be used to predict disease based upon immunoglobulin E (IgE) levels. This is often used when dividing cases from controls in cohort studies, at a certain level of IgE in the blood it is very likely that you are suffering from an allergic disease[3].

A typical study could be performed as follows: Cases are recruited at a hospital, for instance patients remitted to a specific clinic with a similar phenotype. Blood samples are taken and different phenotypic tests are performed. Cases are often also requested to fill out questionnaires with specific questions regarding factors that are considered important for the evaluation of the results. Blood donors are often used as a control group and are considered the equivalent of a normal background population. The control group could be issued a questionnaire as well, but is often not tested for disease-specific phenotypes. Typically, the ratio between cases and controls varies from 1:1 to 1:3 or 1:4. The more individuals present in each group the better, as higher number of individuals decreases the risk of false positives due to low power. DNA is extracted from each participant and then genotyped. If little is known about the genetics of the selected disease, a GWAS approach is chosen. Quality control of the genotype data is performed where SNPs and individuals with high frequencies of missing data are removed. Statistical testing of alleles and genotypes for association as well as testing for deviations from Hardy-Weinberg proportions, which are the expected proportions of genotypes given no selection, is

performed using tools such as PLINK[9]. The associations are then scrutinized for false positives, corrected for multiple testing and an OR is calculated for each risk allele. If the disease can be quantified, i.e. through varying levels of a biomarker, other tests such as Kruskal-Wallis can be performed. This will then associate genetic variants with the varying level of biomarkers or the severity of the disease.

2.4 Replication

When SNP associations are first detected, the SNP is usually one out of many SNPs tested. Performing a lot of tests increases the chance of obtaining false positives. In addition, the same SNP might not be associated to the same phenotype in other tests. By consistently showing an association of a specific SNP to a specific phenotype, the probability of the association is increased. In other words, it is a numbers game. Genetic association studies rely on replication to confirm or reject previous findings. Replication is strictly defined as finding the same result, i.e. same risk allele, in a new study looking at an independent sample from the same population[8]. There is a growing concern that positive results, defined as studies that show a significant association to disease, are favored for publication. In other words, there is a lack of published negative results. Furthermore, when associations are first reported in the literature they tend to show stronger associations and odds ratios than in later replication attempts[10]. This phenomenon is usually referred to as "The winner's curse" and can for example be due to the fact that markers detected by GWAS need to go through very stringent significance thresholds. This also makes it difficult to replicate these associations, at least when the threshold is set at the nominal $P < 0.05$ significance level. In one in third instances, two studies at 80% power could have conflicting results, where one reports a variant as significant and the other does not[11]. In a study of replication rates, it was shown that most GWAS fail to replicate across independent cohorts. This is in part due to the winner's curse but also due to differences in ancestry which will have varying LD structures[12].

2.5 Interpretational difficulties

There are a couple of limitations in the field of genetic association studies. As implied in the section above, a larger number of positive results rather than negative results are being published. This is mostly due to publication biases, as scientific journals favor publications that show strong effects and low *P*-values. Furthermore, there are no good options for researchers to publish their negative results, so instead these results never reach the scientific community. This can create a problem as one

published positive result could potentially actually have several never seen negative results, diminishing its impact. The early genetic studies of complex disease in candidate genes used hypothesis driven strategies and very small population sizes. The trend now is to include more information, going not only from single markers to many markers in the targeted genomic regions, but also to greatly increased population sizes. These types of studies find, not only in cases but also controls, a lot of rare missense variants, synonymous variants and intergenic variants. So, when has enough ground been covered? And when is the population size big enough? When can we really trust the associations found?

There are tools to answer these questions. One is replication; as genetic association studies are a numbers game, the more studies pointing towards the same explanation, the more likely this explanation becomes. The use of public databases can give clues to what is normal variation and what is deleterious. The Exome Aggregation Consortium[13] and the 1000Genomes project[14] are excellent databases for "normal" variation (more on these two in section 3.2), and databases like Factor VIII are great for predicting harmful mutations in specific diseases. Furthermore, there are freely available softwares that can, for example, predict the eventual harmful effects of missense mutations, such as PolyPhen2[15] and SIFT[16]. In addition, making use of a combination of both quantitative studies such as case-controls studies, and qualitative studies with variable degrees of disease, it is possible to determine what causes the disease and what is affecting the severity of the disease. Additionally, although beyond the scope of genetic association studies, testing the associations in physiological studies and determining the actual function of the associated genetic markers and not only predicting their outcome is a very important tool to understand the underlying mechanisms.

2.6 Complex diseases and missing heritability

Genetic diseases may arise through single mutational events and follow what is called Mendelian inheritance patterns. This type of monogenetic disease is typically very rare and can be classified as either recessive or dominant in nature. Although by itself, each monogenetic disease is rare, many different monogenetic diseases exist, which together makes this group of diseases relatively common. One of the more well-known is Hemophilia A which was explained genetically by maker-based analysis of families segregating for the disease[17]. Other diseases show a more complex form of inheritance because they are either oligo- (few genes), or polygenetic (many genes) in nature. These are often called complex diseases and they are often more common in the population. Typical examples of complex diseases are type 2 diabetes, stroke, Alzheimer's disease and allergic diseases. The

general model is, unlike in monogenetic diseases, that the combined effect of several mutations in several or many genes in combination with one or several environmental factors cause the phenotypic change. There are two leading hypotheses for what is causing disease. The common disease common variant (CD-CV) hypothesis, proposes that for a common disease, the genetic variants causing the phenotypic change are also common. Common variants are defined as variants having at least a 1% allele frequency in the population. The rare variants with population frequencies well below 1% are present in the other end of the allele frequency spectrum. These variants are the focus in the competing hypothesis, namely the common disease rare variant (CD-RV) hypothesis. During the population expansion in recent centuries, the human lineage has acquired several mildly deleterious mutations. A supporter of the CD-RV hypothesis, Pritchard, argued that it is more likely that such rare variants explain complex diseases compared to common variants that have been subjected to potential selective forces for a very long time[18, 19].

Two strategies were mentioned in the beginning of this thesis, candidate gene studies and GWAS. These are typical examples of methods performed assuming that the CD-CV hypothesis is correct. However, even after several hundreds of published GWAS, there is an apparent lack of explanatory power of their findings. As an example, in 2009 a total of 40 different loci had been associated to human height but the combined effect of these genetic factors could only explain 5% of the heritability. Similarly, a total of 18 loci had been associated with risk of type 2 diabetes that in total explained 6% of the heritability. Although in some cases, such as age-related macular degeneration, the combined effect of 5 loci could explain up to 50% of the heritability. One problem with CD-CV is that if each single variant only has a small effect on the phenotype, they will be hard to discover, even in a study with large population sizes. Still, the GWAS and in essence the CD-CV hypothesis fail to explain the majority of the heritability of complex disease. This is often referred to as the missing heritability[20, 21].

The missing heritability can be explained by the existence of rare variants. Rare variation cannot be effectively targeted by GWAS analyzing common variants, and as such this type of variation is missed. It is possible that the combined effects of many semi-deleterious rare variants could explain the phenotypic change. To search for this variation, targeted re-sequencing of specific genes or genomic regions can be made. There are also more comprehensive methods like whole-exome sequencing, targeting the protein coding regions, and whole-genome sequencing analyzing the complete genome. Many different projects around the globe try to find this rare variation using the aforementioned methods. One of these is the 1000Genomes project, which has shown what can be expected in a healthy background population. They found an average of 2500 non-synonymous variants, 20-40 damaging variants in conserved sites and 150 variants causing loss of function

in each individual[14, 22]. Indeed, a majority of these variants would never be detected using a GWAS approach and it is apparent that the rare variants found by whole-genome sequencing may play a crucial role in finding the missing heritability. However, GWAS studies could be used to point towards regions that could possibly harbor genetic variants and as such the CD-CV hypothesis and CD-RV could work in unison. As mentioned, the CD-CV hypothesis proposes that individuals with a common disease share common variants, and these variants in turn make up the haplotypes. The major haplotype, i.e. the most common haplotype, has an increased risk of carrying rare deleterious mutations by chance alone. Thus, by identifying probable loci, using candidate gene studies or preferably GWAS one can more effectively search for possible rare damaging variants and the missing heritability[20, 21].

2.7 Allergic disease

Allergic diseases are typical examples of complex diseases. Not only are they not explained solely through single mutational events, the cause of the disease is also believed to be heavily influenced by the environment. The term allergic disease encompasses several distinct phenotypes, such as different food allergies, atopic dermatitis and airway diseases like asthma, allergic rhinitis and chronic rhinosinusitis. The latter two will be given extra attention throughout the rest of this thesis. Allergic disease can generally be summarized as an imbalance in T-helper cell response to a common harmless substance. There are two types of T-helper cells, TH1 and TH2, and it is the latter that starts the inflammatory process leading to the allergic disease. TH2 cells aid in activating other immune cells that in turn produce IgE antibodies. This response is measurable and can be used to clinically determine allergic disease[23, 24]. In general individuals suffering from allergic disease tend to suffer first from eczema in young ages, then asthma and allergic rhinitis as they get older. This phenomenon is called the atopic march and it is highly discussed whether or not eczema causes the latter two conditions, or if later changes in the developing immune system affects the IgE-response[25]. This thesis mainly discusses genetics, but there are a few other risk factors proposed to cause allergic disease. Not only do we inherit our DNA from our parents, we for the most part also share their environment. One topic heavily influenced by the environment is the hygiene hypothesis. The hypothesis proposes that as we as a species get less exposed to microorganisms and viruses and as a results our under-stimulated immune system turns on itself[26]. In contrast, another potential risk factor proposed for allergic disease is the constitution of our microbiome, which are the bacteria in and on our bodies, where different microbiomes would confer different risk[27]. Another factor

influenced by the environment is the birth-order of siblings. Studies have shown that the higher your birth-order, the less likely you are to get allergies[28].

2.7.1 Allergic rhinitis

Allergic rhinitis (AR) is an often life-long condition and common disease affecting 10% to 20% of the world population. It is an example of an extremely common complex disease. It is characterized by runny nose and eyes after allergen exposure and is often associated with an IgE-mediated inflammatory response. Typical allergens are grass pollen, birch pollen, dust-mites and different furs. Estimates of economic burden, due to patients not given proper care, ranges between 55 and 151 billion euros each year in the European Union. Studies show that, given proper care, these costs could be reduced by 95%. The costs mainly stem from lowered productivity and days of sick leave[25, 29]. As AR is believed to be a complex disease both the environment and the genetics play an important role. Studies in twins show estimates of heritability of AR in and around 0.6-0.8[30, 31]. Furthermore, other allergic disorders such as asthma, atopic dermatitis and chronic rhinosinusitis (CRS) are often diagnosed in patients with AR. Although no causal relationship has been shown between CRS and AR, the comorbidity of asthma and AR is very high, with more than 80% of asthmatics also having AR and up to 40% of AR patients also having asthma[25, 29]. Comparing this to the expected comorbidity by chance of asthma and AR of between 0.07% and 0.14%, where asthma affects around 7%[32] of the westernized world, it is apparent that asthma and AR have some kind a causal relationship.

2.7.2 Genetic association studies in AR

By 2012 around 52 studies together had reported 116 SNP associations to AR, of which the majority were candidate gene studies and a handful family based studies. These studies were in most cases characterized by small study populations with a low power to detect associations, a high risk of false positives and they often reported high ORs. Reproducibility of these findings has been analyzed in a study using both a Swedish and a Chinese population. In this study the authors used a subset of 49 randomly selected SNPs from the previously reported 116. The authors concluded that, given the power for detecting associations calculated from previously reported ORs, coupled with the low number of significant replicated SNPs, many of the previous associations are likely to be either false positives or have smaller effect sizes than reported[33]. The same group of authors later looked at previously reported asthma genes and used them as candidate genes for AR as the two diseases have high comorbidity. 192 SNPs previously reported as asthma

associated were genotyped in a Swedish population. The authors concluded that genes associated to asthma to a large degree cannot explain the genetics of AR as very few associations were found[34].

In recent years four large meta-GWAS have been reported. In 2011 Ramasamy et al published their work using four different European study cohorts. Their AR definitions were based on questionnaire and clinical data on grass sensitization. In total 3933 AR cases and 8965 AR controls and 2315 sensitized to grass and 10 032 not sensitized grass were genotyped. Three loci reached the nominal significance level for GWAS, the rs7775228 SNP in the HLA region, the rs2155219 SNP close to *C11orf30* and the rs17513503 SNP close to *TMEM232*. Other SNPs of note were for example rs1898671 in *TSLP* and rs3860069 close to *TLR6*[35].

In 2013 Bønnelykke et al published their meta-analysis of allergic sensitization. The study, which was a two staged meta-GWAS of allergic sensitization, included data from 16 different study populations. Positive skin prick tests (SPT) or elevated allergen-specific IgE levels in blood was used to define sensitization status. This included both common inhalant and food allergens. The first stage of the study, including 5798 sensitized cases and 10 056 controls that were not sensitized, found 26 loci of either genome-wide significance (5 loci) or suggestive evidence (21 loci). The second stage replication, using 6114 independent cases and 9920 independent controls replicated 10 loci, all reaching genome-wide significance. The SNP reaching the strongest association in the replication was the rs17616434 located in the *TLR10-TLR1-TLR6* locus. Other notable loci were two different HLA loci and the *C11orf30* locus[36].

Hinds et al performed an independent companion study to that of Bønnelykke et al in 2013. It is to this date the largest GWAS published analyzing allergic disease. This study looked at 53 862 individuals (27 551 cases, 26 311 controls) with self-reported allergies from two different study populations (23andMe and ALSPAC). There were three different phenotypes: cat allergy, pollen allergy and dust-mite allergy. Using generalized estimating equations, a subset of 3725 markers were found at a nominal evidence of association for at least one allergen. In the meta-analysis of shared effects in the different allergies 16 genome-wide significant SNPs were found. The SNP with the lowest *P*-value was the rs2101521 located in the *TLR10-TLR1-TLR6* locus. Other notable findings include two HLA loci and the *C11orf30* locus among various other effector and receptor proteins[37].

The fourth was a specialized study by Ferreira et al, looking at the combination of asthma and AR phenotypes in a meta-analysis. The study included 4 study populations totaling 6685 cases with physicians diagnosis of asthma and AR and 14 219 controls with no diagnosis of AR nor asthma. The authors hypothesized that by including both phenotypes, the associations would pinpoint to genes involved in a broader phenotype of allergic disease. The study performed as a two-staged meta-

analysis, used the four different populations in a discovery phase. In the second stage another part of the 23andMe population was used in a replication study of the variants discovered from stage 1. Eleven SNPs were found to be significant on the genome-wide level. Among them were the *HLA* and the *TLR10-TLR1-TLR6* locus[38].

Following the first three large meta-GWAS a replication study of the three was performed. The study showed high replication rate with regard to risk alleles and consistently replicated the *TLR10-TLR1-TLR6* locus, more details of this study in section 4.2[39].

To date there are only two re-sequencing studies published on AR. The first was based on previous candidate gene studies that focused on Toll-like receptor (TLR) genes in general[40] and the two X-chromosomal genes *TLR7* and *TLR8* in particular[41]. The study of Nilsson et al found *TLR8* in particular to be associated with AR for both a Chinese and a Swedish population. However, the associations differed between the two populations as different haplotype and sex effects were found. The authors concluded that the answer could lie within rare variation on the different major haplotypes[40]. The re-sequencing study focused on *TLR8*, more details of this study can be found in section 4.3. The second re-sequencing study focused more broadly on the TLRs, as the evidence from the meta-GWAS all pointed out the *TLR10-TLR1-TLR6* locus. More on this study in section 4.5.

2.7.3 Chronic rhinosinusitis

Chronic rhinosinusitis (CRS) is an allergic disease characterized by chronic inflammation of the mucosal tissue in the sinuses. It is diagnosed as having symptoms such as rhinorrhea, sinus pressure, nasal congestion and loss of smell lasting longer than 12 weeks. The prevalence of CRS varies across world populations. In North America estimates range around 13% whereas data suggests lower prevalence in both Europe and Asia with ranges around 10% and 7%, respectively. CRS is generally sub-divided into two different phenotypes, CRS with nasal polyps (CRSwNP) and CRS without nasal polyps (CRSsNP). CRSwNP is diagnosed through nasal endoscopy and patients of this phenotype are more prone to symptoms such as loss of smell and nasal obstruction[42, 43]. Few studies have looked at the heritability of CRS. However, in one study looking at 1638 CRSwNP and 24 200 CRSsNP the authors demonstrate that first degree relatives carry a 4.1-fold increased risk of also developing the CRS phenotype, compared to controls[44]. Patients with CRS in general, and CRSwNP in particular, carry an increased risk of asthma by up to 2.8x as well as an increased risk of AR by up to 2.6x. Other risk factors for CRS includes smoking, other inflammation diseases and the general air-quality in patients environment[45].

2.7.4 Genetic association studies in CRS

Genetic studies in CRS have almost exclusively been candidate gene studies. Before 2013, a total of 27 studies had been published, one of these was a pooled GWAS and only a handful were replication attempts. Most studies used population sizes in the size ranges 35-200 cases, with only a few studies using > 600 cases and controls. The previous studies presents 24 loci significantly associated to the CRSsNP phenotype and 11 loci significantly associated to the CRSwNP phenotype, in total 53 different SNPs. The candidate gene studies focused on interleukins and their receptors and other immune system-associated genes such as *IRAK-4*. The strongest associations before any replication were made, were the *IL22RA1* for the CRSsNP phenotype and *IL33* for the CRSwNP phenotype[46].

The pooled GWAS, conducted in a Canadian population of 173 cases and 130 controls, found no SNP on the genome-wide significance level. However, considering the limited population size and the fact that the samples were pooled, the power to find anything below such a threshold is low. The two most significant findings were the *LAMA2*, a gene responsible for organizing cells into tissues during embryonic development and *PARS2* which is a gene responsible for charging proline to tRNA[47].

Replication of previous signals in CRS has been very limited. By 2013 two SNPs had been replicated for the CRSwNP phenotype, rs17561 in *IL1A* and rs1800629 in *TNFA*, the latter replicated twice, although at much reduced significance. To date, two replication studies looking at most previous reported associations in the literature have been published, one Chinese study and one Swedish study looking at Belgian cases. The Chinese study, consisting of 638 cases and 315 controls, looked at 41 previously reported SNPs plus adding SNPs located in the *IRAK-4* gene. They replicated four loci, among them *AOAH*, previously identified by the pooled GWAS, as the strongest association for both of the CRS phenotypes. Interestingly, although not reported in the literature before, rs4532099 in *RYBP* was found to be significantly associated to the CRSsNP phenotype[46, 48]. The Swedish study, consisting of 613 Belgian CRS cases and 1588 background population controls, looked at 53 previously reported SNPs. More on this study in section 4.1[46].

To date, only one targeted re-sequencing study has been performed in the field of CRS genetics. This study focused on the best signal, *PARS2*, from previous association studies, more on this study in section 4.4[49].

3. Materials and methods

3.1 Subjects

Three primary study populations have been utilized throughout this work, one Belgian CRS case-only population and two Swedish AR case and control populations.

3.1.1 Belgian CRS population

This population which was cases-only, consisted of consecutive patients at the Ear, Nose, and Throat Department, University Hospital, Ghent, Belgium. A total of 613 patients were included and of these 365 were male and 248 were female. All patients included were of Caucasian origin. Two distinct sub-phenotypes existed, with 275 suffering from nasal polyps and 338 without any polyps (CRSwNP and CRSsNP respectively). Patients were diagnosed using historical data, nasal endoscopy, clinical examination and scans of sinuses using computed tomographics. The subdivision of phenotypes into CRSwNP and CRSsNP were done according to the criteria of the EPOS Guidelines 2007. As CRS is an allergic disease which is common amongst other patients with different allergic disease, tests for atopic status and asthma occurrence were performed . The atopic status of the patients was determined using skin prick tests for the most common inhalant allergens. Occurrence of asthma was confirmed according to the Global Initiative for Asthma 2006 guidelines by a trained chest physician based on tests of pulmonary function and symptoms of asthmatic disease. Ethical approval was attained from the Ethics Committee of Ghent University Hospital, Belgium, and written informed consent was obtained from each patient before inclusion in the study.

3.1.2 Malmö AR population

One of the two Swedish AR populations, with unrelated patients from the general population of southern Sweden. Patients were recruited at Malmö University hospital in 2003-2009 and were all of Caucasian origin with both parents born in Sweden. It consisted of 360 AR cases, 191 males and 169 females, and 720 controls,

426 males and 294 females. A smaller selection of 288 randomly picked patients were used for re-sequencing studies. Criteria for positive birch- or grass-induced AR diagnosis was based on having intermittent AR for at least 2 years combined with a positive skin prick test (ALK-Abelló, Hørsholm, Denmark) or Phadiatop test (Pharmacia Upjohn, Uppsala, Sweden) to birch or grass allergens. A wheal reaction of ≥ 3mm in patients that had not used anti-allergic drugs 3 days prior to the test, was considered as a positive response. Patients included in the study were all considered, during pollen season, to suffer from severe symptoms such as itchy nose, runny eyes, nasal secretions, sneezing and nasal blockage. All patients had previously been treated with nasal steroids and antihistamines. Contrarily, controls showed no reactions to skin prick tests or Phadiatop tests. The study was approved by the Ethics Committee of the Medical Faculty, Lund University. Written informed consent was obtained from all subjects.

3.1.3 BAMSE AR population

The BAMSE population is an unselected population based birth cohort that consists of 4089 children which was recruited between 1994 and 1996 in the Stockholm area in Sweden. Participants were of different social economic status and were living in both urban and inner city areas. A questionnaire was given to each participant and their parents regarding allergies, households, birth location of parents and their social economic status. Follow-up studies were made at the ages 1, 2, 4, 8, 12 and 16. At each evaluation of allergy status, at ages 4, 8 and 16, sera were collected and screened with Phadiatop and fx5 (Phadia AB, Uppsala, Sweden). Sera containing ≥ 0.35 kUA/l immunoglobulin E were further analyzed for reactivity to specific allergens. In this work data from 8 and 16 year evaluations were analyzed, which included 2153 children, counting both cases and controls. The BAMSE study was approved by the Ethics Committee at Karolinska Institutet, Stockholm, Sweden. Written informed consent was obtained from parents.

3.2 Databases

Throughout my thesis public databases have provided me with background population data as well as important population genetics data. As technologies advanced going from singe polymorphisms to millions and all the way up to whole genomes, so did the public databases. An important aspect of genetic research is knowledge, what can we expect to find? In what frequencies? Do the frequencies differ in different populations? Are the different variants connected through linkage disequilibrium? Are the variants damaging? Is the variation located in a regulatory

element, coding sequence or something else? These types of questions have gotten much easier to answer thanks to the use of public databases. To a great degree, almost all genetic variation found is being reported into databases. As these databases grow, so does our knowledge of the very complex nature of the human genome. Below follows a couple of databases that have been invaluable during my research.

3.2.1 The HapMap project

The now retired HapMap project, short for haplotype map, project was developed to create a map of human haplotypes. The database has been fundamental in discovery of diseases and their treatment. The project included 270 samples from four different subpopulations, people of Western European ancestry living in Utah, USA, Han Chinese living in Beijing, China, Japanese living in Tokyo, Japan and Yoruba people living in Ibadan, Nigeria. SNP genotyping of all samples was made sure to cover SNPs with < 5% MAF and every other 5000 base pairs, with some regions having denser genotyping. The data covered 99% of all haplotypes above 5% in frequency and thus accounted for almost all common variation. The database provided information of LD, haplotype structures and allele frequencies for all different populations. The power of knowing haplotype blocks allows for higher precision for researchers in selecting relevant genomic regions for further investigation[6].

3.2.2 The 1000 genomes project

Similar to the HapMap project, the 1000 genomes project aims to provide a genetic resource to aid in disease discovery. The big difference however, is that the 1000 genomes project uses newer technology which captures most genetic information by whole-genome sequencing. This has been done in different phases, where in phase 1 1092 individuals were included, originating from 14 different populations across the globe. In this phase, intergenic parts of the genome were sequenced at low coverage (~2-6x) whereas parts of the genome containing genes were sequenced at a much higher coverage (~50x). Overall they detected 38 million SNPs, 1.4 million short insertion or deletions and > 14 000 large deletions. They estimated to cover 98% of all variation ≥ 1% in the investigated populations[22]. While phase 2 did not expand further on populations, phase 3 did adding up to a total of 2504 genomes now spread over 26 different populations[14]. The data gathered from 1000 genomes has gained novel insights as to what one can expect to find in one human genome. Although all participants were considered healthy a remarkable amount of loss of function variants could be found, for example several

individuals carry a homozygous nonsense mutation that completely truncates TLR5 rendering it useless. Furthermore, all data is on the individual level, and is phased, making it extraordinarily suitable for genetic studies. A drawback of the dataset is the use of Illumina technology only, which uses short reads and thus will miss out on a lot of structural variation.

3.2.3 Exome Aggregation Consortium

The exome aggregation consortium (ExAC) is, as the name implies, a database of whole-exome data, meaning it covers only genes. While it is smaller in scope to the 1000 genomes project when it comes to coverage, it boasts a much larger population size. It consists of > 60 000 individuals from different population genetic studies. Access of data is more restricted, and can only be analyzed population-based and not in individual level. However, it is an excellent tool for comparison of damaging rare variation and allele-frequency data[13].

3.3 Simulations and permutations

Throughout my work I have utilized simulations and permutations. In all re-sequencing projects, papers III-V, only cases were included. As background populations, the 1000 genomes project and ExAC were used. In the permutation tests we used known chromosome counts of known population sizes of our cases and the 1000 genomes project individuals. We pooled all alleles of selected frequency, for example MAF < 1% if assessing rare variation, of both populations. The variants were then randomly assigned to either of the populations. The chance of getting a variant was in direct correlation to population size, and thus the larger population would by chance receive more variants. We then tallied each time our case population received more variants than the observed number for the case population, this was calculated 100 000 times. The resulting *P*-value of this one-sided test would then be number of times the case population received a greater number of alleles than the actual, observed population, divided by number of iterations (100 000).

Different types of simulations have been utilized. The most prevalent uses the non-Finish population of ExAC. In these simulations we only assess the coding variants of our case populations. For each gene simulated, a list of variants with their respective MAF was collected. From this list an identical number of chromosomes to the case population at hand were randomly created. This was done by using the MAF of each variant as the chance of a chromosome harboring that specific variant. The total number variants were tallied, and calculated 100 000 times. This was then

compared to the observed counts of the case population at hand. Each time the simulated dataset from ExAC had an excess of variants compared to the case population a score was tallied. This score was then divided by number of iterations (100 000) to give the test quantity (*P*-value).

3.4 Bioinformatics

Haplotype information was gathered from HapMap (retired) and 1000 genomes project (https://www.internationalgenome.org). Allele information was gathered from dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/) and 1000 genomes. All information from the 1000 genomes project was gathered from the Integrated Variant Set (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/) released April 2012. Extraction of data from the 1000 genomes was done by tabix[50] and VCFtools[51]. Effect of missense variation was evaluated using SIFT[16] and PolyPhen2[15]. All genetic coordinates presented uses the GRCh37 assembly.

4. Paper summaries

4.1 Paper I

4.1.1 Introduction

The first paper assessed the, then current, state of genetics in CRS. A total of 53 SNPs associated to CRS phenotypes were found in the literature. The aim of the study was investigate the reproducibility of these SNPs. Our colleagues in Belgium had a patient material of around 700 individuals suffering from CRS with or without nasal polyps. The material did not have matching controls and also included patients of non-European heritage. After genotyping, 613 CRS patients remained, of these 275 had CRSsNP and 338 had CRSwNP. As controls we used a background population from Illumina. Filtering for European heritage and genotyping panels that included as many SNP markers as possible a set of 1588 controls were attained. Association tests were performed with all patients vs all controls as a test for a general CRS phneotype and with CRSsNP vs CRSwNP as a test for a specific nasal polyp association.

4.1.2 Results

Following primer design and genotyping using a Sequenom MassARRAY MALDI-TOF platform 43 SNPs produced high quality genotyping data. In the test for the general CRS phenotype seven SNPs reached a significance of $P < 0.05$. The SNP in *PARS2* had the lowest *P*-value of 0.00022 and also the highest OR of 1.29 (1.19-1.42, 95% CI). With the exception to *PARS2* and Discoidin (*DCBLD2*) all other associated SNPs were located in genes involved in inflammatory responses. PARS2 is responsible for charging proline to tRNA. In the test for nasal polyp specific genes, the acyloxacyl hydrolase gene (*AOAH*) had an uncorrected *P*-value of 0.022 and an OR of 1.32(1.14-1.56, 95% CI). Although significant and a likely candidate due to previous replication by a Chinese study and links to other allergic diseases, it had a high false discovery rate of $q = 0.68$. No significant confounding was found due to AR and asthma sub phenotypes.

4.2 Paper II

4.2.1 Introduction

Similar to the approach in paper I, this paper aimed to assess the replication rate of previous findings in the literature, this time AR instead of CRS. Moreover, the state of genetics in AR was explored to a much greater degree than CRS, with large ambitious genome-wide association studies already performed. Three meta-GWAS had been published during the years 2012-2013 which included from around 25 000 individuals in the smallest study to around 53 000 individuals in the largest study. These three studies were a mix of some clinical data such as skin prick tests, as well as self-reported allergies. Although more modest in size, the BAMSE cohort, which we used in this replication study had good phenotyping and thus, we were able to exactly match the phenotypes of the three meta-GWAS.

4.2.3 Results

In total 44 one-sided tests were done for the exact matching phenotypes and of these 12 were significant at the $P < 0.05$ level. Furthermore, 36 out of the 44 tests also showed concordant risk alleles to that of the original studies. In the three original studies there were a great degree of overlap of genetic loci. Four loci were significantly associated in all three studies, and out of these, two were replicated in the BAMSE cohort; the *TLR10-TLR1-TLR6* locus and the *HLA-DQA2-HLA-DQA1* locus. In another test, using our own clinically defined AR phenotype, the strongest associations were seen for the SNPs located in the *TLR10-TLR1-TLR6* locus. Finally, since data was available from both 8 and 16 years of age, an age at onset analysis was done. The *SSTR1-MIPOL1* and *TSLP-SCLC25A46* loci showed significant association to early onset AR.

4.3 Paper III

4.3.1 Introduction

This study was motivated by previous candidate gene studies of the TLR genes in relation to AR. Both linkage studies and common SNP association studies had pointed towards a locus on the X-chromosome on which *TLR7* and *TLR8* resides in tandem, with stronger associations towards *TLR8*. In the SNP association studies by

Nilsson *et al* conflicting associations were found for the two populations included in the study[40]. The Swedish Malmö population showed associations for females whereas the Chinese Singaporean population showed associations for males. To understand the associations, we performed targeted re-sequencing of a putative promotor of *TLR8* and its coding sequence in the same Swedish Malmö population in which the SNP associations were found. A replication attempt in another Swedish AR population was also performed, using the BAMSE cohort as was used in Paper II.

4.3.2 Results

Sanger sequencing detected 13 polymorphisms, three in the promotor and 10 in the coding sequence. Four of the 10 coding polymorphisms had MAF < 1% and of these three were novel. Simulations using background population data from the 1000 genomes project and ExAC revealed no evidence of accumulation of rare variants, nor did SIFT or PolyPhen2 reveal any evidence of an excess of damaging variants. In the replication attempt using the BAMSE cohort five SNPs were found at $P < 0.05$ as well as four significant associations towards birch pollen specific IgE. These results were in stark contrast to the discovery in the Malmö population, with opposing risk alleles, different sex effects and different allergens and thus was not truly a replication. The associations of the Malmö, BAMSE and Chinese Singaporean populations were re-evaluated, using simulations of randomly generated populations with similar haplotype structure, sex ratios and recombination frequencies as the original populations. Results of the simulations found that the associations are likely to be due to random associations perhaps due to X-chromosome genetics difficulties.

4.4 Paper IV

4.4.1 Introduction

This study was made to follow up the leads from Paper I in which a replication attempt of previous genetic markers was performed. The SNP rs2873551 showed the strongest association towards CRS in a test of 43 previously associated SNPs. This SNP is located within the haploblock together with *PARS2*, a gene which encodes a protein responsible with charging proline to tRNA. The hypothesis was; the common haplotype of the *PARS2* gene, found to be accumulated in the case population, could potentially carry an excess of rare variants. Sanger sequencing of

310 CRS patients for a putative promotor and coding sequence of *PARS2* was performed, as well as long range PCR over the region to detect copy number variation. This was the very first re-sequencing project done in CRS research. Two overlapping long range-PCR systems were also designed to cover the gene and surrounding areas. Simulations of accumulation of rare variation was done using the two background populations, European ancestry 1000 genomes individuals and ExAC data.

4.4.2 Results

The long range PCR did not discover any insertions or deletion in and around *PARS2* in the 310 cases. Sanger sequencing detected 10 variants in the promotor region of which six had MAF < 0.01 all present on single chromosomes. In the coding region 11 variants were detected, of these 7 were of MAF < 0.01 and present in only one or two chromosomes. Similar patterns were seen in 1000 genomes individuals of European ancestry. Simulations revealed weak associations towards an accumulation of rare variation in CRS patients compared to the background populations. Analysis of inferred CRS haplotypes compared to that of 1000 genomes individuals showed a tendency towards more rare haplotypes in the CRS population.

4.5 Paper V

4.5.1 Introduction

This study was primarily motivated by the results of Paper II. Although *TLR8* had been shown to not harbor an excess of rare variation in Paper III the other nine TLRs had not been investigated for rare variation. The *TLR10-TLR1-TLR6* locus was the strongest association found in Paper II and thus a strong candidate for further studies. Furthermore, considering the TLRs central role in the immune system, acting as a bridge between the innate and adaptive immunity they all are especially interesting candidates. In this study we used an ION Torrent sequencing ampli-seq approach to sequence all coding sequences of all 10 TLRs. All 10 putative promotors of the TLRs were also sequenced using Sanger sequencing. The study population consisted of 288 Malmö AR patients. This was by far the biggest re-sequencing project done in AR research.

4.5.2 Results

TLR8 sequencing using ION Torrent sequencing produced similar results compared to *TLR8* Sanger sequencing data used in paper III. A total of 37 promoter polymorphisms and 119 coding sequence polymorphisms were detected, of these 14 respectively 68 were considered rare polymorphisms. *TLR10-TLR1-TLR6* locus was the most variable and the *TLR7-TLR8* locus was the least variable. The overall variation of all 10 genes coincided with what was reported by the 1000 genomes project. Analysis using SIFT and PolyPhen2 of the coding variants showed no indication of excess damaging variation in AR cases compared to background populations. Loss of functions variants were detected, and one in particular was the S324* nonsense mutation in *TLR1* which was clearly overrepresented in the AR cases with 4 copies in 576 chromosomes in the Malmö population, none in 758 chromosomes in the 1000 genomes population and 1 in > 60,000 chromosomes in the ExAC population. Accumulation of rare variants was assessed using simulations with data from background populations. *TLR10* promotor showed a high level of variation in AR cases ($P < 0.00009$). *TLR1*, *TLR5*, *TLR7*, and *TLR9* coding sequences showed accumulation of rare variation. Overall the strongest indication was towards the *TLR10-TLR1-TLR6* locus.

5. Discussion

In essence, genetic association studies can answer three different types of questions. Firstly, what is the genetic mechanism underlying disease? This is probably the most basic scientific question; understanding which genetic alterations cause the disease, and how these alterations operate in the genome. Secondly, when the mechanisms of the genetic variation in specific diseases are known, can it be used to diagnose the diseases? This creates the possibility to give personalized treatment and medication which results in much better disease control. The third, and ultimate, goal of this type of research is the prevention of the disease. If we know the causes of the disease, we can also screen for the genetic variants at an early age. This gives ample opportunity to prevent diseases at an early stage, greatly reducing or abolishing complications in the patients.

Most of published results regarding complex diseases are still dealing with the first question. There are a lot of weak candidate and GWAS signals and these do not in any significant way explain why people get the disease at the individual level. There are a couple of ways to battle this lack of power to detect genetic effects. First, clinically selected and much larger disease populations that are collected through collaboration between research groups. To recruit *en masse* with a strict phenotype definition from the beginning will result in a much higher chance to detect strong effects. These populations could be part of large cohorts where the data on many phenotypic variables are available. This creates a much better testing scenario where subdivision in to specific sub-phenotypes is possible. Using these big populations could also bring clarity to previous findings through replication. This creates subsets of highly probable loci that can be tested further. Another approach, common in monogenetic disease, is the formation of comprehensive databases. These would include not only all positive findings for one specific disease, but also give an opportunity for researchers to report their negative findings. This will deter other researchers from studying weak or negative findings and instead focus the research on findings of greater relevance. As mentioned previously in the thesis, the field is driven towards whole exome and genome studies as these methods have become possible through technical advances that has significantly lowered the costs. These studies can be performed using the big population sizes mentioned earlier, and skip the various small independent studies and thus diminishing the need for meta-analysis. In the end, to find all the missing heritability of the candidate gene and

GWAS, the focus has to go beyond DNA studies. There are other levels of regulation in the genome. RNA sequencing looks at the mRNA transcribed in cells and can give answers to differing levels of gene expression and splice variants. Regulatory features such as promoter regions and methylation of DNA may contribute. Varying levels of methylation in certain genes could very well cause different levels of gene expression between individuals. Although greatly influenced by their coding DNA, protein-to-protein interaction and varying levels of the protein are possible venues. More complex interactions are possible, both of sense and anti-sense genes and the interactions of gene products to other genes.

I mention at the start of this section that the genetic association findings in the end could be used in the clinic, via diagnosis and prevention of disease, but what is the clinical relevance of these studies? Overall the GWAS signals only present small effect sizes, and as such their relevance to the clinic is reduced. The ones that do show strong effects are however good targets for diagnosis and screening in the general population. There are of course different situations for different diseases. The need to screen for a relatively mild diseases such as allergic rhinitis is not as important as for disease with life threatening outcomes such as diabetes type 2 and stroke. Then again, as the technical advances are driving the field of genetics towards whole-genome sequencing there may be no need to sub-divide on severity of disease. By instead sequencing the full genome of each individual, clinicians can predict disease outcomes and prevent them early on. This does not come without ethical problems however, as this could be sensitive information if it were to come into the wrong hands. Furthermore, does everyone want to know what could possibly kill them? Even if the ethical issues could be solved, one human genome at 50x coverage takes on average up 120 gigabytes of storage space creating an issue of massive storage halls and the data safety of these.

The contribution of my research is mostly that of basic science, trying to answer the question of what are the genetic mechanisms underlying disease. Paper I and II focused on understanding the enemy. I achieved this through replication, taking the best leads in AR with the meta-GWAS and in CRS with various genetic association studies and determining their reproducibility. In both AR and CRS there have been a lot of small studies, with very few trying to replicate previous findings. My studies have tried to find probable genes that could be good candidates for further evaluation. In AR we found the *TLR10-TLR1-TLR6* locus that in turn was detected by all four large meta-GWAS. Additionally, my colleagues have previously found strong evidence pointing towards the *TLR7-TLR8* locus[40]. Not only were these two loci highly replicated at level of common variation, the genes themselves are an important part of the immune system. The TLRs operate at the border of the innate and adaptive immune system. The activation of TLRs by virus or bacteria starts a cascade which will create an immune response, recruiting effector proteins. Thus, any disruption of the TLRs, either increasing or decreasing the expression levels of

the genes could have implications of how strongly the immune system reacts to the environment[52, 53].

We pursued *TLR8* first, as there were some unanswered questions from the previous work, where the results showed conflicting MAF and sex effects between the Chinese and the Swedish Malmö populations. It was proposed that this could be due to geographic reasons and by using another Swedish population in a replication study. In paper III we tried to replicate the results in the BAMSE population, but again there were conflicting results, both different alleles and allergens. This study also did Sanger sequencing of *TLR8*. The running hypothesis was that rare variation could explain the association, since in both the Chinese and Swedish Malmö population it was the major haplotype of respective population which was associated, and by chance they would acquire more rare variation. The Malmö population did not show any signs of accumulating rare variation in *TLR8*. We then asked ourselves, can we really trust any of the previous associations? To test this we simulated similar population sizes with equal sex ratios and haplotype structures. Could we find equally many associations by chance in randomly simulated populations? The answer seems to be, yes. In all three populations, Malmö, BAMSE and the Chinese, there was no robust evidence of strong associations, as we could by chance alone in otherwise healthy populations find association patterns similar to that of the three original study populations. *TLR8* is located on the X-chromosome and doing association studies on this chromosome comes with a few problems. First, in many GWAS this chromosome is completely ignored, as it requires additional computations, due to different Hardy-Weinberg proportions as men only have one X-chromosome. Second, the fact that men have half the (X) chromosomes of females greatly reduces power of a study as men are now only contributing to that of a half test-subject[54].

Although *TLR8* had not shown any signs of accumulating rare variation, the *TLR10-TLR1-TLR6* locus had consistently shown strong associations towards the same major haplotype in both our replication study and in the three meta-GWAS. Thus, we expanded the search for rare variation, not only in this locus, but for all ten TLR genes in paper V. *TLR8* was also included in this paper, as we now used another sequencing method and the previous Sanger data from paper III would serve as an excellent positive control. We targeted all coding sequence of all ten genes, as well as a putative promotor. We defined the promotor as 50 bp into exon 1 and 500 bp upstream for all ten genes. There are more sophisticated ways to define the promotor, especially through Ensembls regulatory tracks, but there are no easy ways to exactly define endpoints in any direction. We argued that, by targeting the base pairs closest to the start of the transcript, we would capture valuable information and not create any bias between the genes.

Comparison of the *TLR8* Sanger data and *TLR8* ION Torrent data revealed no discrepancies in the 288 case samples. We then compared total counts of variants, allele frequencies, impact of missense variation and accumulation of rare variation to European ancestry in 1000 genomes and ExAC. Overall the ten genes showed similar variation patterns, with more variants in *TLR1*, *TLR5*, *TLR6* and *TLR10*, and less in *TLR7* and *TLR8*. However, when looking at the rare variants, above all the promoter of *TLR10* had significantly more variation compared to 1000 genomes. Furthermore, *TLR1*, *TLR5*, *TLR7* and *TLR9* showed increased accumulation of rare variation compared to the ExAC population. In addition, a nonsense mutation in *TLR1* which truncates the protein was found on four chromosomes of 576, compared to one in 60 000 in the ExAC database. Taken together this is notable, especially for the *TLR10-TLR1-TLR6* locus, which keeps popping up. In recent years *TLR10* has gone from being of unknown function (even pseudogene status) to being an important regulator of the TLR2 pathway. It has been proposed that it can suppress the pathway either by competitive binding to TLR2 instead of TLR1 or TLR6, or by activating a different pathway which reduces the inflammatory response[53]. Having altered expression of *TLR10* through variation disrupting regulatory elements in the promotor could influence the inflammatory response. Furthermore, variation in *TLR1* could further shift this balance by having more or less available TLR1 favoring alternative pathways. The results of *TLR5* are harder to interpret, as the gene is riddled with missense and even nonsense variation even in healthy individuals. *TLR7* is somewhat more interesting, taken together with earlier studies that motivated paper III, could it be that the variation responsible for the associations to *TLR8* was in fact more closely related to *TLR7*? It is definitely worth further investigations. TLR9 has similar functions to TLR7 as in intracellular receptor, as well as phylogenetically its most similar protein.

CRS is a disorder often associated with AR but still a distinct phenotype. Genetic studies of CRS are still a fair bit behind those of AR. There is an apparent lack of GWAS performed in CRS, and the only one made so far used a pooling strategy of less than 200 cases and less than 200 controls. Even still, it is the largest study using a hypothesis-free driven approach. It is from this study *PARS2* was found to be associated with the CRS phenotype, which I later replicate in a larger Belgian CRS population. My replication study lacked a matching control population, but we argue that since both the cases and the background population I used were of European ancestry and did not show any sub-population structures, they should therefore not be causing any spurious associations. In fact, since prevalence of CRS could be as high as 10%, the power of being able to detect any signal would be reduced as one would expect similar frequencies in the background population. We decided to follow up on *PARS2*, both since it had at least been replicated once with the strongest association and OR, and also because it is a relatively small gene. However, the function of the gene in relation to the phenotype remains somewhat of a mystery.

Where in the case of AR and the TLRs a link can quite easily be established, *PARS2* is a housekeeping gene that charges proline onto tRNA in which very few cases loss-of-function mutations have been reported. In fact, in the few cases reported it has been coupled with adverse effects on the brain[55]. We still wanted to follow the same strategy as in AR, following up on our best leads in each disease and re-sequence the genes that were associated, and later also replicated.

Using a similar approach as in the studies of the TLRs, we re-sequenced both the coding sequence and a putative promoter, although this time aiming to include most of the regulatory elements presented on Ensembl and again comparing our results to the European ancestry of 1000 genomes and ExAC. The mutation spectrum looked similar to 1000 genomes, with only few rare mutations and the rest were in high frequencies (> 10% MAF) with an apparent lack of variants in 1-5% MAF. Simulations revealed a slight increase of variants in the below 5% MAF category for the CRS cases, but the effect diminished when looking at rare variation (< 1% MAF). The housekeeping role of the gene, the fact that it is under negative selection and combined with the lack of strong evidence to CRS cases accumulating an excess of rare variation, the gene seems less plausible to be a strong contender for causing the CRS phenotype. Overall, the evidence for *PARS2* for CRS was much sparser than TLRs for AR going into re-sequencing. The TLRs were both plausible genes as well as replicated many times in much larger study samples.

Genetic association studies in allergic disease are complex and might be so because genes only play a limited effect. We inherit our genes from our parents, but we also inherit their environment. According to the hygiene hypothesis we come in contact with less pathogens and are thus keeping our immune system less occupied. There is a small chance that all I have achieved through the work of my thesis, is to create a lot of data. Furthermore, even though our understanding of the human genome has greatly increased, especially since the first draft in 2001, we still do not know nearly enough. We still do not know to what extent CNVs affects the genome and how common it is. Sequencing methods today operate on short reads, around 400 bp of length, but to discover truly impactful copy number variation we need to be able to read lengths of 20 000-50 000 bp. There are emergent technologies which can do this, but they are still error prone. Combinations of short reads which have high accuracy with the information of the long reads could be the answer. Even so, there is information that we have, but do not know how to interpret, namely most of the non-coding DNA. What effects do variation have on regulatory elements? Is all non-coding, non-regulatory DNA junk?

6. Conclusions

Finding true associations for complex diseases is, by definition, truly complex. It requires big study samples and genome-wide methods that cover all the variation of the genome. It has been interesting to see the difference in AR and CRS research. In AR we had a lot of information going into re-sequencing studies, where both we and other groups had pointed towards the *TLR10-TLR1-TLR6* locus. In contrast, it might have been premature to do targeted re-sequencing in CRS as only one GWAS had been performed and all studies, even our own, had small population sizes with limited power. The field of CRS genetics would really benefit from collaborations between groups, or perhaps gathering newer and bigger case populations with matched controls and then continue in the fashion of AR research with meta-GWAS, trying to replicate their own results. Alternatively, to go to whole-exome or whole-genome sequencing as the costs of these methods are dropping.

I raised a concern in the discussion: are genetic association studies in AR and CRS just an exercise in data collection? The short answer would probably be no. In AR I strongly believe that the *TLR10-TLR1-TLR6* locus is a key component to understand AR pathogenesis. The answer does not have to be black or white, perhaps altered expressions of the TLRs renders some more prone to develop AR in certain conditions, be it too clean living, few siblings or any other environmental factor. There is definitely an effect of the nonsense mutation in *TLR1* and the accumulation of rare variation in *TLR10* promotor. The associations from the meta-GWAS resides on a haploblock of around 100 000 bp, far greater than the around 10 000 bp we so far have re-sequenced. There is a high probability that more rare variation in terms of SNPs and even CNVs exists in this genomic area. The road to screen for or cure AR and CRS is long, my research has only touched upon understanding the underlying mechanisms of the diseases, and as it is a polygenetic disease either genes in each respective disease can only explain a small part of the mechanisms.

Doing whole-genome sequencing is becoming more and more common, for example the 100 000 British genomes project which has sequenced 100 000 genomes. Projects like this have immense amounts of power to detect underlying genetic mechanisms. In a not too distant future most newborns will have their genomes sequenced. Comparing the amount of information this will give genetic researchers in the future to what we have available to us now is exciting but it could be another situation like when the GWAS was first introduced. We keep getting

better and better tools to unravel the complexity of our genome and the answers always seem to be in reach within the next technological advance but to really understand what we are seeing in our DNA, focus has to be on study design first and foremost. By going forward with sequencing studies, at least any small study can contribute with their findings in an ever-expanding database of genetic variation and as such perhaps the days of common variation association studies are coming to an end.

7. Acknowledgements

I would like to thank everyone who has in any way contributed to the work of my thesis. Above all, I would like to thank:

Sven och Lily Lawskis fond för naturvetenskaplig forskning, for financial support and as such making this thesis possible.

Torbjörn Säll, my supervisor, for all the help with population genetics and statistics. You really got me hooked on genetics from the first lecture back in 2007. Thanks for all the support!

Christer Halldén, my supervisor, a man of many visions, for a never ending supply of ideas, for incredible problem solving and for always seeing opportunities in even the trickiest of situations. Thank you so much for believing in me and for all the hard lessons learnt!

Daniel Carlberg, my co-author and colleague, for the countless of hours writing manuscripts together, for the inspiration and the motivation to become better at programming and bioinformatics, and for keeping the dullest of days entertaining!

Lars-Olaf Cardell, my co-author, for you invaluable input on the world of allergies and for an always cheerful attitude.

Christina Lind-Halldén, my co-author, for all your help during my teaching hours, for your excellent organizational skills, for the impeccable Sanger sequencing data and not least your always positive attitude (and cookies).

Annika Lidén, for invaluable help in the lab, for excellent organizational skills and excellent notes. You never missed a single error, from the biggest of tables to the smallest of commas!

Eric Manderstedt, my co-author and colleague, for your intelligent solutions and input, for the ever inspiring discussions, ranging from single atoms to the whole universe, nothing is too big or small to discuss. You are a never ending source of inspiration and you have changed my views on many topics (that is an absolute truth!).

Guðný Björk Óðinsdóttir, my colleague, for your competitive nature, especially outside of natural sciences, and of course for the coffee break banter!

Markus Brorsson and Pernilla Lindahl, my colleagues, for the completing the infamous puzzle, and for all the fun we had during lunch breaks!

Claus Bachert with colleagues at Ghent Hosptial, for all the help with paper I and IV.

Erik Melén, *Magnus Wickman* and all others at the BAMSE project at Karolinska Institutet, for the help with paper II and III.

My colleagues at Kristianstad University, for all the wonderful inter-scientific discussions during the lunch breaks!

Kati, Håkan, Albert, Ronja, Tove Molly and relatives and friends, for always being there and supporting me throughout life!

Elsa, my partner and best friend, for always supporting me no matter what, for your excellent input both scientific and linguistic, for your incredible patience during both ups and downs. You are truly amazing and I love you with all of my heart!

8. References

1. Griffiths AJF, Wessler SR, Lewontin RC, Carroll SB (2008) INTRODUCTION to GENETIC ANALYSIS, 9th ed. W.H. Freeman and Company
2. Ott J, Wang J, Leal SM (2015) Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet 16:275–284. doi: 10.1038/nrg3908
3. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45–61. doi: 10.109700125817-200203000-00002
4. Lewis CM, Knight J (2012) Introduction to genetic association studies. Cold Spring Harb Protoc 7:297–306. doi: 10.1101/pdb.top068163
5. Hardy J, Singleton A (2009) Genomewide association studies and human disease. N Engl J Med 360:1759–1768. doi: NEJMra0808700 [pii]\r10.1056/NEJMra0808700
6. Frazer KA, Ballinger DG, Cox DR, et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–61. doi: 10.1038/nature06258
7. Dudbridge F (2016) Polygenic Epidemiology. Genet Epidemiol 40:268–272. doi: 10.1002/gepi.21966
8. Clarke GM, Anderson C a, Pettersson FH, et al (2011) Europe PMC Funders Group Basic statistical analysis in genetic case-control studies. Nat Protoc 6:121–133. doi: 10.1038/nprot.2010.182.Basic
9. Purcell S, Neale B, Todd-Brown K, et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575. doi: 10.1086/519795
10. Ioannidis JPA. (2005) Why most published research findings are false. PLoS Med 2:e124. doi: 10.1371/journal.pmed.0020124
11. Amrhein V, Korner-Nievergelt F, Roth T (2017) The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ 5:e3544. doi: 10.7717/peerj.3544
12. Palmer C, Pe'er I (2017) Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. PLOS Genet 13:e1006916.
13. Lek M, Karczewski K, Minikel E, et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv
14. Auton A, Abecasis GR, Altshuler DM, et al (2015) A global reference for human genetic variation. Nature 526:68–74. doi: 10.1038/nature15393

15. Adzhubei IA, Schmidt S, Peshkin L, et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249. doi: 10.1038/nmeth0410-248

16. Sim N-L, Kumar P, Hu J, et al (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40:W452-7. doi: 10.1093/nar/gks539

17. ROBERTSON JH, TRUEMAN RG (1964) COMBINED HEMOPHILIA AND CHRISTMAS DISEASE. Blood 24:281–288.

18. Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19:212–219. doi: 10.1016/j.gde.2009.04.010

19. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137. doi: 10.1086/321272

20. Eichler EE, Flint J, Gibson G, et al (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–50. doi: 10.1038/nrg2809

21. Manolio TA, Collins FS, Cox NJ, et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–53. doi: 10.1038/nature08494

22. McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. doi: 10.1038/nature11632

23. Wawrzyniak P, Akdis CA, Finkelman FD, Rothenberg ME (2016) Advances and highlights in mechanisms of allergic disease in 2015. J Allergy Clin Immunol 1–16. doi: 10.1016/j.jaci.2016.02.010

24. Campbell DE, Boyle RJ, Thornton CA, Prescott SL (2015) Mechanisms of allergic disease - environmental and genetic determinants for the development of allergy. Clin Exp Allergy 45:844–858. doi: 10.1111/cea.12531

25. Ng CL, Wang DY (2015) Latest developments in Allergic Rhinitis in Allergy for Clinicians and Researchers. Allergy 70:1521–1530. doi: 10.1111/all.12782

26. von Mutius E (2007) Allergies, infections and the hygiene hypothesis - The epidemiological evidence. Immunobiology 212:433–439. doi: 10.1016/j.imbio.2007.03.002

27. Ipci K, Altıntoprak N, Muluk NB, et al (2016) The possible mechanisms of the human microbiome in allergic diseases. Eur Arch Oto-Rhino-Laryngology. doi: 10.1007/s00405-016-4058-6

28. Kusunoki T, Mukaida K, Morimoto T, et al (2012) Birth order effect on childhood food allergy. Pediatr Allergy Immunol 23:250–254. doi: 10.1111/j.1399-3038.2011.01246.x

29. Bousquet J, Khaltaev N, Cruz AA, et al (2008) Allergic Rhinitis and its Impact on Asthma (ARIA) 2008 update (in collaboration with the World Health Organization, GA2LEN and AllerGen). Allergy Eur J Allergy Clin Immunol 63:8–160. doi: 10.1111/j.1398-9995.2007.01620.x

30. Fagnani C, Annesi-Maesano I, Brescianini S, et al (2008) Heritability and shared genetic effects of asthma and hay fever: an Italian study of young twins. Twin Res Hum Genet 11:121–131. doi: 10.1375/twin.11.2.121

31. Willemsen G, van Beijsterveldt TCEM, van Baal CGCM, et al (2008) Heritability of self-reported asthma and allergy: a study in adult Dutch twins, siblings and parents. Twin Res Hum Genet 11:132–142. doi: 10.1375/twin.11.2.132

32. Fanta CH (2013) Asthma. N Engl J Med 360:1002–1014. doi: 10.1056/NEJMra0804579

33. Nilsson D, Andiappan AK, Halld??n C, et al (2013) Poor Reproducibility of Allergic Rhinitis SNP Associations. PLoS One. doi: 10.1371/journal.pone.0053975

34. Andiappan AK, Nilsson D, Halldén C, et al (2013) Investigating highly replicated asthma genes as candidate genes for allergic rhinitis. BMC Med Genet 14:51. doi: 10.1186/1471-2350-14-51

35. Ramasamy A, Curjuric I, Coin LJ, et al (2011) 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 128:996–1005. doi: 10.1016/j.jaci.2011.08.030

36. Bønnelykke K, Matheson MC, Pers TH, et al (2013) Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. Nat Genet 45:902–6. doi: 10.1038/ng.2694

37. Hinds DA, McMahon G, Kiefer AK, et al (2013) A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. Nat Genet 45:907–11. doi: 10.1038/ng.2686

38. Ferreira MAR, Matheson MC, Tang CS, et al (2014) Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. J Allergy Clin Immunol 133:1564–71. doi: 10.1016/j.jaci.2013.10.030

39. Nilsson D, Henmyr V, Halld??n C, et al (2014) Replication of genomewide associations with allergic sensitization and allergic rhinitis. Allergy Eur J Allergy Clin Immunol 69:1506–1514. doi: 10.1111/all.12495

40. Nilsson D, Andiappan AK, Halldén C, et al (2012) Toll-like receptor gene polymorphisms are associated with allergic rhinitis: a case control study. BMC Med Genet 13:66. doi: 10.1186/1471-2350-13-66

41. Møller-Larsen S, Nyegaard M, Haagerup a, et al (2008) Association analysis identifies TLR7 and TLR8 as novel risk genes in asthma and related disorders. Thorax 63:1064–1069. doi: 10.1136/thx.2007.094128

42. Beule AG (2015) [Epidemiology of chronic rhinosinusitis, selected risk factors, comorbidities and economic burden]. Laryngorhinootologie 94 Suppl 1:S1–S23. doi: 10.1055/s-0034-1396869

43. Stevens WW, Lee RJ, Schleimer RP, Cohen NA (2015) Chronic rhinosinusitis pathogenesis. J Allergy Clin Immunol 136:1442–1453. doi: 10.1016/j.jaci.2015.10.009

44. Oakley G, Curtin K, Orb Q, et al (2015) Familial risk of chronic rhinosinusitis with and without nasal polyposis: Genetics or environment. Int Forum Allergy Rhinol 5:276–282. doi: 10.1002/alr.21469

45. Min J-Y, K.Tan B (2015) Risk Factors For Chronic Rhinosinusitis. 8:1699–1712. doi: 10.1016/j.rasd.2014.08.015.Social

46. Henmyr V, Vandeplas G, Halldén C, et al (2014) Replication study of genetic variants associated with chronic rhinosinusitis and nasal polyposis. J Allergy Clin Immunol 133:273–275. doi: 10.1016/j.jaci.2013.08.011

47. Bossé Y, Bacot F, Montpetit A, et al (2009) Identification of susceptibility genes for complex diseases using pooling-based genome-wide association scans. Hum Genet 125:305–318. doi: 10.1007/s00439-009-0626-9

48. Zhang Y, Endam LM, Filali-Mouhim A, et al (2012) Polymorphisms in RYBP and AOAH genes are associated with chronic rhinosinusitis in a Chinese population: A replication study. PLoS One 7:6–13. doi: 10.1371/journal.pone.0039247

49. Henmyr V, Lind-Halldén C, Halldén C, et al (2016) Chronic Rhinosinusitis Patients Show Accumulation of Genetic Variants in PARS2. PLoS One 11:e0158202.

50. Li H (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27:718–719.

51. Danecek P, Auton A, Abecasis G, et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158.

52. Tesse R, Pandey RC, Kabesch M (2011) Genetic variations in toll-like receptor pathway genes influence asthma and atopy. Allergy 66:307–316. doi: 10.1111/j.1398-9995.2010.02489.x

53. Oosting M, Cheng S-C, Bolscher JM, et al (2014) Human TLR10 is an anti-inflammatory pattern-recognition receptor. Proc Natl Acad Sci U S A 111:E4478-84. doi: 10.1073/pnas.1410293111

54. Zheng G, Joo J, Zhang C, Geller NL (2007) Testing association for markers on the X chromosome. Genet Epidemiol 31:834–843. doi: 10.1002/gepi.20244

55. Sofou K, Kollberg G, Holmström M, et al (2015) Whole exome sequencing reveals mutations in NARS2 and PARS2, encoding the mitochondrial asparaginyl-tRNA synthetase and prolyl-tRNA synthetase, in patients with Alpers syndrome. Mol Genet Genomic Med 3:59–68. doi: 10.1002/mgg3.115

Printed by Media-Tryck, Lund 2017
NORDIC ECOLABEL, 3041 0903

LUND UNIVERSITY

LUND UNIVERSITY
Faculty of Science
Department of Biology

ISBN 978-91-7753-387-0