Complex disease genetics

Utilising targeted sequencing and homogeneous ancestry

ARGYRI MATHIOUDAKI
Abstract

The complex disease investigations presented in this thesis aimed to provide new information regarding underlying genetics by using targeted sequencing and ethnically homogeneous cohorts. This work moved past current methodologies and addressed data stratification issues, that might have been hindering new findings. The results contribute to a more comprehensive view of the genetics of ankylosing spondylitis (AS) and breast cancer (BC), in Sweden.

Paper-I presents a sex-stratified analysis of a Swedish AS cohort that incorporated both common and rare variants. Single variant and aggregate tests both showed different signals in AS male and female patients, previously masked. Specifically, the RUNX3 locus in males (univariate test: rs7414934, OR=2.58, \( p=1.7 \times 10^{-5} \)) and MICB in females (SKAT: 27 variants, \( p=1.2 \times 10^{-5} \); rs3828903, OR=4.62, \( p=6.2 \times 10^{-13} \)) exceeded discovery thresholds. In the functional follow up of these loci, risk alleles appear to regulate the expression of genes in multiple tissues. Also, the results highlight the importance of disease regulation from different haplotypes and loci breakdown proved that Sweden’s genetic architecture might be critical for AS studies.

Paper-II is a replication study, in our modest-sized Swedish cohort, of AS associations, previously discovered in populations of British origin. Initially, power calculations assessed that the Swedish cohort had the power to replicate only published associated markers with high effect (OR > 7), e.g., HLA-B but the replication analysis revealed three associated loci (OR\(_{\text{range}}\):1.9-2.7). Notably, the multiplicated HLA-B marker (rs4349859) was not in HWE equilibrium. Population structure differences could not explain this replication pattern. However, sequencing resolution revealed fine-scale differences with repositioned association signals in the known loci. Specifically, the identification of two CCHCR1 protective haplotypes (OR: 0.14/0.3) that affect other MHC gene expression through eQTLs, provided the first suggestion of the differential function of known associated loci with cis gene regulation.

Paper-III provides the first fingerprint of the somatic mutation profile of Swedish BC. The significantly mutated genes were PIK3CA (28%), TP53 (21%) and CDH1 (16%) while histone-modifying genes (e.g., KMT2C and ARID1A: together 28%) exhibited an increased somatic mutation prevalence, not observed previously. Additionally, within the patients that did not receive neoadjuvant treatment, there were distinct age groups with different mutational profiles and differential APOBEC signature driving genes.

Taken together, these studies emphasize the contribution to the underlying genetics deriving from smaller ethnic populations, when assessed with a shift in methodology to account for biological bias, like sex and age. The results will hopefully assist and guide other genetic studies of human complex disease.

Keywords: ankylosing spondylitis, breast cancer, targeted sequencing, Sweden, genetics, population stratification.

Argyri Mathioudaki, Department of Medical Biochemistry and Microbiology, Box 582, Uppsala University, SE-75123 Uppsala, Sweden.

© Argyri Mathioudaki 2019

ISSN 1651-6206
ISBN 978-91-513-0710-7
urn:nbn:se:uu:diva-390457 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-390457)
As you set out for Ithaka, hope your road is a long one, full of adventure, full of discovery.

K. Kavafis

The undeciphered Phaistos disc from the Minoan era (ca 1700 B.C)
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Related work from the author not included in the thesis:


Contents

Introduction ...............................................................................................11
From genetics to genomics .................................................................12
Mapping of disease .................................................................................13
Complex diseases: genetic basis and challenges...............................14
Technologies assisting complex disease mapping ..................................15
  Next-generation sequencing .............................................................17
  Association testing in complex disease .............................................20
  Challenges in genetic studies .........................................................21
Autoinflammation: an example of immune-related complex disease ....25
  Immunity and disease .....................................................................25
  Autoinflammatory disease: an overview .........................................25
  Ankylosing spondylitis: the prototype of all SpA .........................26
Cancer: ultimate complex disease .......................................................31
  Breast cancer .................................................................................31

Aim of thesis .............................................................................................34
Present investigations ................................................................................35
  Paper I. The sex-stratified genetic architecture of ankylosing spondylitis ........................................................................35
    Aims and background ....................................................................35
    Methods .......................................................................................36
    Discussion ....................................................................................37
  Paper II. Replication and fine mapping of ankylosing spondylitis replicated loci in the Swedish population reveal different CCHCR1 protective haplotypes. .........................................................38
    Aims and background ....................................................................38
    Methods .......................................................................................38
    Results ..........................................................................................39
    Discussion ....................................................................................39
  Paper III. Targeted sequencing reveals the somatic mutation landscape in a Swedish breast cancer cohort .................................................41
    Aims and background ....................................................................41
    Methods .......................................................................................41
    Discussion ....................................................................................42

Concluding remarks ...................................................................................44
Future perspectives ................................................................. 46
Acknowledgements .............................................................. 49
References ................................................................................. 53
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AID</td>
<td>autoinflammatory disease</td>
</tr>
<tr>
<td>AS</td>
<td>ankylosing spondylitis</td>
</tr>
<tr>
<td>APC</td>
<td>antigen presenting cell</td>
</tr>
<tr>
<td>BC</td>
<td>breast cancer</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>CD</td>
<td>cluster of differentiation</td>
</tr>
<tr>
<td>CDCV</td>
<td>common disease-common variant</td>
</tr>
<tr>
<td>CNA</td>
<td>copy number aberration</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DMARD</td>
<td>disease-modifying anti-rheumatic drug</td>
</tr>
<tr>
<td>FMF</td>
<td>familial Mediterranean fever</td>
</tr>
<tr>
<td>GWAS</td>
<td>genome-wide association studies</td>
</tr>
<tr>
<td>HGP</td>
<td>human genome project</td>
</tr>
<tr>
<td>HLA</td>
<td>human leukocyte antigen</td>
</tr>
<tr>
<td>IBD</td>
<td>inflammatory bowel disease</td>
</tr>
<tr>
<td>IBP</td>
<td>inflammatory back pain</td>
</tr>
<tr>
<td>IGV</td>
<td>integral genome viewer</td>
</tr>
<tr>
<td>IL</td>
<td>interleukin</td>
</tr>
<tr>
<td>indel</td>
<td>insertion or deletion</td>
</tr>
<tr>
<td>Kb</td>
<td>kilobase (10^3 bases)</td>
</tr>
<tr>
<td>LD</td>
<td>linkage disequilibrium</td>
</tr>
<tr>
<td>MAF</td>
<td>minor allele frequency</td>
</tr>
<tr>
<td>Mb</td>
<td>megabase (10^6 bases)</td>
</tr>
<tr>
<td>MHC</td>
<td>major histocompatibility complex</td>
</tr>
<tr>
<td>MRI</td>
<td>magnetic resonance imaging</td>
</tr>
<tr>
<td>NGS</td>
<td>next-generation sequencing</td>
</tr>
<tr>
<td>NF-κβ</td>
<td>nuclear factor, κ-chain enhancer of activated B-cells</td>
</tr>
<tr>
<td>NK cell</td>
<td>natural killer cell</td>
</tr>
<tr>
<td>NSAID</td>
<td>non-steroidal anti-inflammatory drug</td>
</tr>
<tr>
<td>OR</td>
<td>odds ratio</td>
</tr>
<tr>
<td>PsA</td>
<td>psoriatic arthritis</td>
</tr>
<tr>
<td>ReA</td>
<td>reactive arthritis</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>SpA</td>
<td>spondyloarthropathy(ies)</td>
</tr>
<tr>
<td>SLE</td>
<td>systemic lupus erythematosus</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>Th</td>
<td>T-helper cells</td>
</tr>
<tr>
<td>TNF</td>
<td>tumor necrosis factor</td>
</tr>
<tr>
<td>WES</td>
<td>whole exome sequencing</td>
</tr>
<tr>
<td>WGS</td>
<td>whole genome sequencing</td>
</tr>
</tbody>
</table>
Clinical genetics aims to explain how traits and disorders that have medical implications are inherited. The puzzle of some monogenic disorders was solved, but the mapping techniques used were not sufficient to decipher complex disorders, regulated from multiple genetic loci and by specific environmental triggers. New technologies and analysis methods, such as genotyping, next-generation sequencing, and various association models, have assisted the discovery of many “low-hanging fruit” predisposing loci.

However, as these practices matured, new hurdles came to light. The majority of complex diseases are difficult to phenotype, and often one diagnosis consists of several diseases. Global statistics also indicated a distinct disease prevalence across continents, and many independent cohorts did not replicate the same associations. In addition, associations do not reveal the causative mechanism, and the exact contribution of these variants in pathogenesis scenarios is not always clear. Non-coding signals occur at a broad set of genomic locations and have no apparent impact at the protein level. Their contribution to the phenotype is difficult to measure accurately, as their role in pathogenesis is still difficult to comprehend. Additionally, prioritization for functional validation remains challenging, as more and more information becomes available for the three-dimensional structure and plasticity of the genome across tissues. Apart from their location, mutations predisposing for disease do not exhibit a specific allele frequency range, and at a population level, these mutations can be both common and rare. Moreover, genetic variation databases demonstrate the polymorphisms and haplotype structure across different populations, and recent studies have revealed a significant number of private variants within different ancestries and ethnical groups.

This thesis presents an effort to gain insights into addressing some of the challenges that are still relevant to current research. Complex trait genomics is aimed to account for all the sources of data stratification that possibly mask the correct genetic components. This work, summarized in three studies, provides an in-depth exploration of germline and somatic mutations in two complex diseases, ankylosing spondylitis (AS) and breast cancer (BC). Before delving into the methodology and findings, we will commence by laying the basis for this thesis: What constitutes a complex, hereditary disease and which technologies and statistical models are appropriate for genetic association studies and subsequent analyses? Then, we will define the two illnesses separately in terms of their known clinical and genetic features and address the
From genetics to genomics

Genetics has existed since the birth of civilization. People use expressions, such as “it lies in their blood” or “runs within the family,” to convey their understanding of how certain traits and maladies are inherited. Empirically, people have even selected for specific desired characteristics, e.g., the large diversity of dog or chicken breeds, during plant crossing and animal breeding. However, the processes underlying these observations have not been scientifically studied for some time and are still not understood.

An entire century was required, from the time of Mendel’s principles (1) and the first DNA isolation from Miescher, to learn that DNA is the “inheritance molecule,” depict its properties (2) and structure (3) and finally define genes and their regulation (4). Subsequently, a DNA sequencing methodology was developed (5,6), and its applications in research seemed promising and of critical importance. Today, genetics, characterized by rapid advances and numerous findings, has reached far beyond the research community. Many elements of our culture reflect genomics progress, and our vocabulary consists of terms, such as the genome, DNA, and mutations, which mainly refer to human disease.

The birth of this genetic revolution was undoubtedly the announcement and completion of the Human Genome Project (HGP (7–9)), which is still one of the most significant collaborative research efforts to date. The HGP promised and delivered the whole human euchromatic DNA sequence and the first human gene count, along with the physical locations and functional annotations of the genes. This more complete view of the human genome, in terms of structure, organization, and function, was publicly available and created the field of genomics. The HGP led to the rapid development of new technologies, the improvement of tools for the analysis of this type of genetic data, and the implementation of these tools in current research. Soon daughter projects and fields were shedding light on various aspects using the genome as a starting point. Comparative genomics was a natural consequence of the availability of mapped genomes of model organisms that were also sequenced as part of the HGP (e.g., Escherichia coli, Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster, Mus musculus). Evolution and conservation genomics informed the tree of life and showed that almost 5.5% of the genomes are highly conserved, spanning both coding and non-coding elements (10). The vital role of the non-coding human genome, which includes intronic and promoter regions, small and long ncRNAs, and enhancer and insulator regions, was established by the Encyclopedia of DNA Elements Consortium (ENCODE), proving that only a small portion (1.2%) of the human genome
actually encodes mRNA, shifting the interest in the regulatory potential of the rest of the genome and how this regulation varies across different cell types (11). Information about the differential conformation, accessibility, transcription factor affinity, and organization of the genome into topologies that interact during chromosome folding across different tissues has provided an additional three-dimensional layer of genome regulation (12). Additionally, the development of visualizing tools, such as the UCSC (13) and the Ensembl browser (14), has facilitated the incorporation of this plethora of data in research routines.

Based on the HGP, it was clear that genomes within a species are not identical. Genetic changes occur throughout the genome due to mutations and recombination. In humans, genetic variation varies in type and size, and the main categories are as follows: i) single nucleotide polymorphisms (SNP, 1 bp changes), which are the most straightforward and most common mutations; ii) small insertions and deletions (indels, 1-5 bp gained or lost); and iii) larger structural variations, which include copy number variations (gain or loss of sequence copies) and chromosomal rearrangement events (structure change of native chromosomes). Significant projects that catalogued variation were the 1000 genomes project (7) and all phases of the haplotype map project (HapMap (15)). Together, these projects provided the mapping tools necessary for more complex association studies in humans. Most importantly, these projects revealed the common variations (minor allele frequency (MAF) > 0.05) across populations and the tag-SNPs to fingerprint the human genome, based on their linkage disequilibrium (LD) with human haplotype blocks.

Thus, if HGP provided the cartography of the genes, then all the other tools and projects together have provided the coordinates and the compass, while the visualization tools offered the actual human genome map.

Mapping of disease

To map disease, it is necessary to scrutinize the genome map using the tools mentioned above to pinpoint the genetic elements responsible.

Medical or clinical genetics focuses on explaining the genetic basis of human syndromes and phenotypes with medical relevance. The particular focus of genetics is defining the heritability of illnesses, recognizing and listing disease predisposition, clarifying disease mechanisms and assisting diagnosis, prognosis and treatment response prediction. Despite being feasible, genetic disease mapping still faces many obstacles, such as the varied background, the various patterns of inheritance, phenotype penetrance, and expressivity differences.

Accordingly, some rare yet highly penetrant monogenic diseases were the first to be deciphered. As only one genetic locus was underlying the phenotype, the inheritance of these diseases followed Mendel’s laws and showed
specific patterns (recessive, dominant, X-linked). The linkage between these diseases and a specific locus could be drawn by studying extensive pedigrees and the manner in which specific markers segregate with the disease phenotype, while dissecting the segregating locus further could lead to the identification of the causal mutation. This mapping strategy relied on the size of the pedigrees and the fact that alleles located on the same chromosome are linked, so a recombination event was unlikely (16).

In the majority of Mendelian disorders, the mutations are detrimental and were found at low frequencies because purifying selection tends to eliminate these anomalies from the genetic pool of the population. Typical examples of diseases mapped with such methodology are autosomal recessive cystic fibrosis, where linkage studies on extensive pedigrees revealed mutations in the \textit{CFTR} gene (17), and the X-linked recessive disorder haemophilia B (known as “royal disease” (18)). Currently, according to the statistics presented in Online Mendelian Inheritance in Man (OMIM), for almost 6,500 catalogued Mendelian phenotypes, the molecular basis is known, (19).

Additionally, some other chronic, frequent and noninfectious diseases, even though they run more often in families, exhibited a different, more complex inheritance modality and remain a challenge for disease mapping.

Complex diseases: genetic basis and challenges

In reality, the majority of diseases with clinical relevance are, in fact, multifactorial, polygenic, or complex. It is hypothesized that complex diseases are not caused by a single mutation at a single locus. Instead, these diseases are the result of a complex interaction of alterations in multiple genes that work in conjunction with certain environmental stimuli. Examples of such complex diseases are cancer, heart disease, diabetes, psychiatric disorders, and immunological diseases. These diseases have been a high focus of genetic research over the years due to their impact not only on the individual by affecting the quality of life and longevity but also on global health, economy, and society.

In general, mapping complex disease is more effective when there is a significant contributing genetic component. Cases with a young age of onset are characterized by heavy family history and therefore a more significant genetic burden, and their segregation often resembles Mendelian inheritance. A typical example is early-onset Alzheimer’s disease, inherited as an autosomal dominant trait, where mutations with a high effect in three loci have been implicated (\textit{APP}, \textit{PSEN1} and \textit{PSEN2}), (20–22). However, for the vast majority of complex disorders, defining heritability is necessary before attempting to map their genetic basis. Heritability, the part of the disease phenotype defined and controlled by genetics, is defined as the ratio of the genetic component to the total phenotypic spectrum of the disease (Table 1, (23)). Heritability
ranges from 0 to 1, is a population characteristic at a given time and is usually calculated by monozygotic twin studies (24).

Apart from polygenic inheritance, other challenges in mapping complex diseases include the diversity of clinical manifestations and the varying severity spectrum, which makes it challenging to define a descriptive disease phenotype.

Table 1. Heritability estimation for complex disease

<table>
<thead>
<tr>
<th></th>
<th>( h^2 = \frac{\sigma^2 A}{\sigma^2 P} = \frac{\sigma^2 A}{\sigma^2 G + \sigma^2 E} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma^2 A ): variance of observed phenotypes, ( P ): observed phenotypes, ( G ): unobserved genotype, ( E ): unobserved environmental factors</td>
<td></td>
</tr>
</tbody>
</table>

In addition, such illnesses do not have a specific age of onset, and in combination with the incomplete penetrance of low impact predisposing factors, an accurate diagnosis might be evasive for many years. Additionally, a pheno- copy often occurs in complex traits, such as neurological disorders (25) and cancer (26), where the phenotype might be observed in an individual, but there are no inherited genetic mutations (24).

Technologies assisting complex disease mapping

![Figure 1. Type of variants underlying disease targeted by the different mapping methodologies. Adapted from Manolio et al. with permission (27)](image)

The methods designed to map complex diseases have matured and were shaped according to precise needs. The genetics of complex traits should reflect their heterogeneity. The different approaches to discovering genetic
predisposition are roughly divided into linkage and association studies. Historically, linkage studies, which have deciphered a plethora of monogenic disorders, were unsuccessful for complex diseases and were replaced by candidate-gene association studies. However, the a priori hypothesis of these studies was incompatible with the novel findings, thus the genome-wide association studies (GWAS) methodology was proposed to resolve this issue. GWAS dominated until next-generation sequencing (NGS) was affordable enough for more complete association population studies, based on targeted, whole-exome and whole-genome sequencing data (Figure 1).

**Linkage and candidate-gene association studies**

The numerous genetic loci behind complex phenotypes do not follow Mendelian inheritance patterns, explaining why mapping these loci with linkage methodology has not been so entirely successful, with the exception of the large effect NOD2 mutations in Crohn’s disease (28–30).

Linkage studies and their extensive pedigrees have great power to pinpoint rare variants with high penetrance and marked effects, so these methodologies have been rapidly replaced with association studies for complex phenotypes (31). With the latter method, instead of investigating whether the disease follows certain chromosomal regions within a family, allele or genotype frequencies in affected individuals are compared to those in healthy controls (32). Association studies became a reality due to the availability of appropriate mapping tools, including human reference, comprehensive SNP and common haplotype catalogues, analysis models and techniques. Initially, association studies were limited to analysing markers spanning a few selected candidate genes. This a priori hypothesis of the disease was mostly based on previous observations and epidemiological studies.

Candidate-gene studies tend to have high statistical power, but their results were not consistent and reproducible due to gene selection, misclassification of the samples, and recall bias (33). Despite their pitfalls, candidate-gene association studies have contributed to important findings, such as CAPN10 in type 2 diabetes (34) or the marker in the non-coding region of APOC3 for atherosclerosis (35). Candidate-gene association slowly catalysed this drastic design change from family-based to population-based genetic studies.

**Genome-wide association studies**

The idea of searching for associations across the whole genome, the cornerstone of GWAS, subsequently arose. Without any prior hypothesis, novel loci with intermediate effects could show associations and shed light on phenotype genetics. Whole-genome sequences were prohibitively expensive at the time, so genotyping SNP-chips were developed containing established, evenly spaced and polymorphic SNPs, which taking advantage of the LD across the human genome were tagging specific haplotypes (≈ 50 kb (36)). The ability to
appropriately and rapidly fingerprint the genome of thousands of individuals (>100,000 samples in a well-powered GWAS) facilitated the discovery of common alleles associated with many complex diseases at the population level.

Here, the higher the significance (p-value) of the associated marker was, the higher the chance that the allele was associated with the disease. Associated markers point to the locus (or multiple loci) where the cases and controls differ genetically and possibly contribute to disease (37). Following a “common disease-common variant” (CDCV model) hypothesis (38–40), associations were revealed for many complex diseases. Characteristic examples are Crohn’s disease with 71 associated loci (41), type 2 diabetes with mutations in \( UBE2E2 \) and \( C2CD4A-C2CD4B \) (42), and for both prostate and breast cancer, elements of their susceptibility have been partly clarified (43,44). Apart from disease, complex traits, such as the body mass index (BMI), were also investigated with this methodology (45). The summary statistics in the NHGRI-EBI catalogue of published GWAS findings show that, from 2006 until now, almost 5,150 studies have discovered 70,000 associations (46). In total, GWAS has been a valuable tool to uncover inherited predisposing factors.

Next-generation sequencing

Following the success of GWAS, the continuously growing need for larger-scale human genetic studies, the missing heritability problem, where single markers cannot explain much of the heritability, and the fact that a large part of the genome has not been studied have led to the rapid development of new sequencing technologies. Among these techniques, DNA sequencing is at the core of genomics. This method determines the exact order of the nitrogenous bases (A, T, G or C) that constitute the nucleotides in a DNA molecule. As this technology advanced, sequencing genomic loci larger than the bacterial genome or a single human gene became possible. After completing the HGP, surpassing Sanger and 454 sequencing technologies, NGS offered higher sensitivity, lower error rate, and great speed, with thousands of sequencing reactions being performed “in parallel.” By enabling the analysis of complete germline and somatic genomes, exomes and transcriptomes, NGS became a valuable personalized medicine tool (47,48) with seemingly endless uses.

Human genomic DNA is very large and is still hard sequence in one piece, so in practice, NGS libraries are created by ligating sheared genomic DNA fragments with specific adaptors. Subsequently, sequencing reactions are performed in an instrument that determines the exact sequence. The output, as millions of reads, is then computationally aligned to the reference genome. Then, the aligned reads can be used for different analyses, mainly identifying the differences between the sequenced and reference genomes, such as mutations (individual) and polymorphisms (population). With a cost of less than
1,000 USD per genome, the implication of NGS in biomedical research is profound (49).

In contrast to older sequencing methods, NGS data require more storage and computational resources, and its analysis is highly dependent on the specific software used. Raw data generated from different library designs require specific quality control processing before their downstream analysis. Additionally, the investigations of germline and somatic genomes are quite similar, but constitutional and somatic mutations have distinct characteristics. Despite these differences, all sequencing data aiming to uncover the mutations underlying complex traits follow the same necessary steps: i) sequence read mapping and alignment of the sequence reads to the reference genome, performing all stages of quality control (e.g., sequencing errors or artefacts), ii) variant calling depending on the type of genome (germline vs. somatic), iii) quality control on the variant and individual level and iv) variant annotation (Figure 2).

![Figure 2. Schematic representation of pipeline for both germline and somatic genomes](image)

**Platforms and library design**

Many NGS platforms currently exist, with Illumina HiSeq short pair-end sequencing being the most popular and economical for human sequencing studies (47). New long-read technologies, such as single-molecule real-time (SMRT) sequencing, are preferred for sequencing full genomes to perform a de novo assembly or for capturing full-length transcripts or “difficult” regions
(e.g., GC-rich sequences) (50). Certainly, these technologies will contribute more to understanding complex diseases, as the focus of the investigation shifts from SNPs to structural rearrangements across the genome.

Regarding library design, there are three main categories: i) whole-genome sequencing (WGS), ii) targeted sequencing and iii) whole-exome sequencing (WES) (Table 2). WGS provides a sequence for the majority of the elements in the genome on both coding and non-coding regions. Despite being a great resource that facilitates the discovery of novel genomic events, WGS for large-scale studies remains expensive. High depth coverage comes with additional costs, so WGS depth coverage is usually $\approx 30x$, inhibiting the discovery of low frequency events. Additionally, WGS requires additional infrastructure to store and analyse the data and is computationally demanding.

Alternatively, targeted sequencing provides the full sequences for certain genomic loci without necessarily comprising depth. Target selection can be focused around genes or pathways of interest or can even aim to capture longer regions (e.g., chromosomes). Deep targeted sequencing can be performed up to 1000x, especially if it captures only a few genes. This method usually involves a flexible design (41), and the amount of generated data is more manageable. However, deep targeted sequencing relies on an a priori hypothesis for target selection.

One type of targeted sequencing is gene panels, typically utilized in the clinics for screening and WES, which have emerged to be the most popular among all NGS library designs (52). Gene panels are a type of targeted sequencing designed to capture all the exons, resulting in sequencing 1.5% of the human genome. This strategy incorporates the best characteristics from WGS and targeted sequencing and takes advantage of the fact that our understanding of non-coding mutations is still limited. WES has been successful in identifying causal variants for some rare Mendelian disorders, such as mutations in the DHODH locus, which encodes a key enzyme in the pyrimidine de novo biosynthesis pathway, for the Miller syndrome (Postaxial acrofacial dysostosis syndrome) (53) and MLL2 mutations causing the rare autosomal dominant Kabuki syndrome (54).

Table 2. Comparison of GWAS, WGS, WES, and targeted sequencing

<table>
<thead>
<tr>
<th>method</th>
<th>location</th>
<th>effect size</th>
<th>allele frequency</th>
<th>target size</th>
<th>target coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWAS</td>
<td>coding, non-coding</td>
<td>modest</td>
<td>common</td>
<td>SNP-chip</td>
<td>Single SNP</td>
</tr>
<tr>
<td>WGS</td>
<td>all</td>
<td>all</td>
<td>rare, common</td>
<td>$\approx 3Gb$</td>
<td>30x</td>
</tr>
<tr>
<td>WES</td>
<td>only coding</td>
<td>large</td>
<td>rare, common</td>
<td>60Mb</td>
<td>30x</td>
</tr>
<tr>
<td>Targeted seq.</td>
<td>all</td>
<td>modest</td>
<td>rare, common</td>
<td>flexible</td>
<td>$&gt;30x-100x$</td>
</tr>
</tbody>
</table>
Association testing in complex disease

The careful selection of the genetic data (genotypes or sequencing data), with many steps of quality control, is followed by statistical analysis and association estimation.

In complex disease association studies, the underlying pattern of inheritance is not known beforehand. However, in a case-control setting, an additive model is usually adopted, as it carries more power (55), under the assumption that the underlying genetic components are characterized by similar effect size and minor allele frequency (MAF). In this sense, for an individual SNP, the risk of the heterozygote will lie between the risks of the two homozygotes. To estimate the p-values under the $H_0$ null hypothesis for multiple markers, a mixed model of linear regression, that can account for cofounding factors, such as population stratification and cryptic relatedness, is generally adopted (56,57). This model can be described as follows:

$$y = X\beta + u + \epsilon$$

where $y$ is the described phenotype, $X$ is a matrix of the candidate SNP genotypes (including covariates), $\beta$ is the regression coefficient that describes the effect size of the SNP ($X\beta$: fixed effects), and $u$ describes the random effects matrix (genetic relationship matrix) and $\epsilon$ the environmental impact.

Usually, for 2 million genome-wide SNPs, an alpha level of $5 \times 10^{-8}$ is adopted (58). The parallel estimation of p-values for multiple markers increases the chance of incorrectly rejecting the null hypothesis (Type I errors), suggesting that there is increased risk for an association to be discovered by chance. Therefore, the significance threshold needs to be adjusted for the number of hypotheses and corrected for the family-wise error rate (FWER). Usually, a threshold Bonferroni correction for $\alpha=0.05$ can compensate for the FWER (threshold= $\alpha / \text{number of hypotheses}$, e.g., $p = 1 \times 10^{-7}$ for 500,000 SNPs, $p = 5 \times 10^{-8}$ for 1 million SNPs). Larger datasets can counterbalance the Bonferroni-induced decreased statistical power and reduced sensitivity for loci with an intermediate effect to show an association. Alternatively, the false discovery rate (FDR) may be a less conservative approach.

In addition, with NGS data, all types of inherited mutations, regardless of allele frequency, are available and can be assessed for their contribution to the complex phenotype. Rare variant testing, apart from large sample sizes, requires a different set of statistical methods since single variant tests are underpowered (59). Aggregate tests are commonly used for rare variant testing; instead of assessing each variant individually, these tests collapse the effect of the rare variants in a gene or region. Direction of effect can inform the statistical method choice, with burden testing selected in case of variant effect uniformity or variance-component testing, in case of effect variability.
**Associations in cancer?**

NGS technology has been widely applied in cancer and has revolutionized cancer genomics, with many cancer genomes being analysed at a high resolution. Germline mutations in cancer syndromes are studied in the same fashion as any of the other complex disorders mentioned above. However, tumours, characterized by genomic instability and not diploid genomes, cannot be investigated with the same rules and principles.

Overall, true somatic variants are more challenging to call than germline mutations are and require deeper sequencing for accurate analysis. Cancer genome sequencing uses the same technology but involves direct sequencing of the tumour, normal adjacent tissue and often the cells of the normal microenvironment. Since both cancer and normal cells acquire somatic mutations, a comparison with the matched normal tissue sequence is necessary to filter out the normally occurring somatic mutations and distinguish the mutations unique to cancer. This process requires more than 30x (60), but 60x is usually used. The first attempt focused on a set of genes in two types of cancer (61). Different variant callers and filtration steps are employed to remove artefacts from tissue (germline) contamination and distinguish clonal mutations. In somatic data analysis, the goal is to identify recurrently mutated genes, driver mutations that assist rapid proliferation and provide metastatic potential to the tumour cells and other key pathways that assist progression and tissue infiltration. Among these anomalies, mutations in key cancer players, such as tumour suppressor genes and oncogenes, and other somatic mutations that arise at different times compose the landscape of each tumour. Cancer types share key recurrently mutated genes, e.g., NOTCH1 in chronic lymphocytic leukaemia (62), but every tumour is its own entity. The processes behind the various genetic alterations have left their mark on the cancer genome and can be observed as specific mutational signatures (63).

**Challenges in genetic studies**

*The path after association*

The route from discovering an association to fully grasping its role in the underlying biology of a complex phenotype is not straightforward.

The associations derived from the GWAS methodology are rarely the causal mutations. Instead, these studies identify the locus or the haplotype of interest, with putative implication in the disease. Additionally, the vast majority of associations were located in non-coding regions or gene deserts, with no clear functional implication. Presumably, these variants have a substantial regulatory effect, with no clear function inference at the time. As a result, the functional role of only a small number of GWAS associations has been identified.
Currently, the stepwise interpretation of the associated variants in a genetic study remains challenging in the NGS era, despite the different resolution of the genetic data. Once an association is statistically confirmed, especially after all multiple test corrections, it requires careful inspection. First, deconstructing the LD pattern of the locus will reveal whether the association represents an independent signal. Subsequently, careful examination of the haplotype segregation pattern in the locus usually pinpoints a set of variants that can be further functionally inspected. Occasionally, extra fine-mapping of the region might be necessary, where a different, higher resolution data set will reveal the most associated variants and how these variants might act.

In the case of exonic associated variants, many software programs, such as PolyPhen2 (64), SIFT (65), or PROVEAN (66), can provide a report of all the potential effects of the variant on all possible transcripts. Annotation algorithms such as snpEff (67), VEP (68) or Annovar (69) or Oncotator (70) are commonly used depending on the amount of additional annotation required. UniProt (71) is often utilized for a more visual inspection of the domains affected by the variants.

The functional characterization of non-coding associated variants can be partly computational and utilizes additional layers of evidence. The same annotation algorithms can be employed with a wide selection of databases to extract information from. First, chromatin accessibility markers (DNase sensitivity), DNA methylation and histone modification markers provide information about the specific regulatory potential of a locus, whether it is an active promoter, an enhancer or a silencer (72). Additionally, evidence of transcription factor binding site (TFBS) or TFBS regulation is considered, as found in large chromatin immune-precipitation sequencing (Chip-seq) studies, especially in tissues related to the phenotype. Another important feature to be assessed is the conservation score of the region and the individual core markers. Si-Phy LOD (73) or GEPR (74) scores are commonly utilized to distinguish elements under constraint, that is, strong purifying selection. An additional layer of tissue-specific regulation will provide additional information on the locus function and reveal whether variants can affect gene expression through eQTLs (75). Moreover, it is important to recognize whether the associations belong in the same topologically associated domain (TAD) or interact with other distant loci in the nucleus (12).

The abovementioned resources can inform a better selection and prioritization of candidate variants, but their exact functional role can only be determined experimentally. Carefully designed cell-based experiments, such as electrophoretic mobility shift assay (EMSA), supershift or luciferase reporter assays, can shed light on the actual effect of the associated variant, defining the effect of the risk allele. Additionally, it is important to replicate these findings in an independent cohort with resequencing or a different genotyping method.
Complications in genetic studies

Certain issues became clear over the years overflowing with new genetic findings: i) there is a large component of heritability not explained by common variants, ii) there are many reasons causing spurious associations and iii) the vast majority of associations are not consistently replicated across populations.

Missing heritability is a major issue in complex disease genetics (27,76), but so far, many of these techniques have extensively been used for mapping only SNPs. However, genetic variation is not limited to SNPs, and utilizing the remaining the genetic variation features is lacking. Large duplications, inversions and deletions are rare at the individual level but do not vary much across populations (77). Structural variation could contribute substantially in the future, as sequencing has become more efficient, cheaper and many analysis pipelines have been established.

Data stratification and false associations are more often caused by population substructure, suggesting the existence of smaller subgroups within the tested population set. Undetected grouping of the samples is possible due to errors during processing or due to imprecise phenotyping. In the majority of genetic studies, it is assumed that the complex disease phenotype is in fact a single entity—an assumption that is now known to be a simplification. For example, in the case of psychiatric disorders, distinct biomarkers that define the diagnosis do not exist, while in the case of autoimmune diseases, many symptoms overlap. For the latter, it was proposed that the different subgroups based on diagnostic criteria might actually represent different diseases. However, studying these subgroups separately might result in smaller scale studies, which are also prone to spurious results due to the small power to detect genetic factors with intermediate effects.

Apart from disease comorbidities, other factors, such as age of onset or sex, might represent disease bias and lead to data stratification. Especially in cancer, older patients will exhibit a larger amount of somatic mutations, and younger cases will appear with a more severe genetic burden. Age of onset is the primary factor dividing the malignancies into sporadic or hereditary cases that are often studied separately. This effect is supported by the Knudson two hit hypothesis for cancer, where one mutation contributing to malignancy is inherited and the second one is acquired later in life (78). Additionally, some diseases are much more prevalent in one sex than in the other. Typical examples are autoimmune disorders, which are more prevalent in women. For example, systemic lupus erythematosus (SLE) is a complex disease in which 90% of SLE patients are female, with age of onset peaking during their reproductive years (79). These sex differences might be reflected in different underlying genetics.

In large-scale studies, apart from sub-phenotypes and specific selection criteria of the cohorts, the most common reason for population substructure is
neglecting or not accounting for relatedness and ancestry differences. This type of substructure can be translated as a genetic distance and then visually inspected, and the main methods used for this inspection are principal component analysis and multidimensional scaling analysis of SNP genotypes. If present, then both population substructure and relatedness can be incorporated into the association model as covariates.

However, recent global statistics show different disease prevalence across different countries that cannot be explained solely by environmental differences (e.g., multiple sclerosis (MS), whose prevalence differs worldwide and especially within the European continent (80)). The majority of findings from genetic studies that have been performed in populations of European ancestry to evaluate their disease risk are not necessarily transferable to other ancestries (81). In the progression towards personalized medicine and to improve the mapping of complex disease, high-quality population-based control genomes are vital. Based on this idea, whole reference genomes of specific ethnic populations, e.g., the Sardinian (82), the Dutch (83), or the Icelandic genome (84), have been proposed. These references revealed how demographics shaped up the modern populations and regarding disease they showed deleterious mutations established in higher frequencies and at the homozygote state (e.g., Finnish (85)). Recently, the genome of 1000 healthy Swedes was also sequenced and showed a substantial amount of unique variation (86). With all population-specific resources, it is encouraged to place disease genetic predispositions in a wider disease context.

In particular, homogeneous ancestry, age and sex are some of the sources for genetic data substructure that I have particularly taken into consideration when studying the genetics of two complex diseases as described in the following chapters.
Autoinflammation: an example of immune-related complex disease

Immunity and disease

The immune system, a defence mechanism against infection, is multilayered and well-regulated. The immune system can be divided into innate and adaptive immunity, and these two components differ mainly in speed, specificity, and memory. Each immune response, through the careful orchestration of different components, employs different barriers, cell types and secreted proteins. Early stages of infection trigger innate immunity, while the adaptive immune system controls later responses (87). The two immunities work together to guarantee quick and effective protection, utilizing antigen presenting cells (APCs) and natural killer cells (NK) as their cellular bridges.

Despite strict regulation, immune system malfunctions can still occur, resulting in either a suppressed or a hyperactive immune system. Autoimmune diseases are an extensively studied group of illnesses where the immunity mechanisms are on overdrive and tissues that can no longer be recognized as self are targeted and damaged. Another group of immune-related illnesses are autoinflammatory diseases (AID), characterized by an unprovoked inflammation (88), with complete absence of the hallmarks of autoimmunity, circulating high-titre autoantibodies or activated self-reactive T-cells (89).

Autoinflammatory disease: an overview

Clinically, the recently described AID appears with recurring episodes of inflammation that are often accompanied by fever (90). The occurring inflammation is either tissue-specific or of a systemic nature. Depending on the AID type, the episodes usually have a certain duration and frequency, while during the symptom-free periods, C-reaction protein levels (CRP) and erythrocyte sediment rate (ESR) remain elevated within the blood of patients and function as markers for ongoing inflammation. The continuous underlying inflammation can manifest as joint pain or skin rashes but might also lead to more severe complications, such as amyloidosis, an accumulation of misfolded protein complexes within tissues (91). Especially in the kidneys, amyloidosis disrupts the filtering function and ultimately can result in renal failure.

At the microscopic level, auto-inflammation is controlled by the myeloid cells of the innate immune system, but the mechanisms behind this abnormal activation and function can vary and could include: i) inflammasomopathies ii) NF-κB pathway activation iii) protein misfolding iv) complement activation v) continuous cytokine signalling, vi) macrophage activation (88). It is possible that AID is the result of a combination of all the above, with each mechanism contributing up to a certain extent to each different type of AID.
AID include monogenic disorders, such as Familial Mediterranean Fever (FMF), TNF-associated periodic syndromes (TRAPS), and cryoporin-associated periodic syndromes (CAPS), but also polygenic diseases, such as gout, Crohn’s disease and spondyloarthropathies (SpA). A large proportion of AID patients have unknown underlying mutations and do not fall in any of the six previous categories (92). Altogether, it is expected that the underlying genetic background of many AID will be clarified as sequencing technologies are becoming of use in clinical assessment.

SpA is a group of inflammatory diseases of the spine with various phenotypes. Clinically, SpA has a negative serostatus and no evidence of autoantibodies or rheumatoid factors in blood. There are five different SpA subtypes: reactive arthritis (ReA), enteropathic arthritis, psoriatic arthritis (PsA), juvenile SpA, and ankylosing spondylitis (AS). Apart from axial inflammation, symptoms can occur on other articular and non-articular sites, e.g., sacroiliac and peripheral joint involvement, enthesitis, uveitis (93), and due to the wide set of affected organs and tissues, SpA is now considered a systemic disease, with either cooccurring subtypes or gradual disease shifting into another subtype (93).

Ankylosing spondylitis: the prototype of all SpA
Ankylosing spondylitis (AS) is considered to be the prototype SpA disease, with the different subtypes differentiating into AS at some point of the disease. Apart from seronegative and unprovoked inflammation, AS is characterized by irreversible ossification and joint fusion at inflamed sites. AS is often diagnosed at late stages, as the earliest symptom - inflammatory back pain (IBP) is not sufficient for a diagnosis, according to the New York modified criteria (94). It is possible that comorbidities, such as iritis or uveitis, inflammatory bowel disease (IBD), and psoriasis, could be misleading and responsible for many misdiagnoses. The gradually ankylosing spine can only be verified by radio-graphic evidence of sacroiliac erosions, sclerosis, and unclear cortical borders (95), which in practice requires several years until these occurrences become traceable.

Genetics
AS is highly heritable \( h^2=0.9 \) (89), and the risk of developing the disease is 52 times higher for an individual with a diagnosed first degree relative (96). Global prevalence is difficult to calculate due to the differences in diagnostic criteria across countries, but it is estimated to affect 0.5-1.5% of the Caucasian population (93) and predominantly affects men (3:1 ratio) (97). Apart from the very first genetic link, \( HLA-B27 \) on chromosome 6, our current knowledge of AS genetics is based on a GWAS of a population of British individuals (98–101). More than 45 loci have been implicated, but almost 70% of the
heritability remains unexplained. Topographically, the associations can be divided into major histocompatibility complex (MHC) and non-MHC loci. Functionally, these associations can be divided into loci affecting key pathways, such as MHC misfolding and MHC presentation (HLA-B), peptide trimming (ERAP1, ERAP2) and cell activation as a response to endoplasmic reticulum (ER) stress (IL23R, IL12B).

**MHC**

The MHC molecules (also known as Human Leukocyte Antigens (HLA)), are proteins that bind and present antigenic peptides to T-cells. The MHC class I (A, B, C) binds antigens from the cytosol and presents to CD8+ T-cells whereas MHC class II (DR, DP, DQ) recognizes and binds internalized extracellular antigenic peptides and presents to the CD4+ cells (87, 102). However, both MHC class II and I molecules are structurally similar (Figure 3).

**Figure 3. MHC I and MHC II molecules within a leukocyte membrane:** Both molecules have three distinct domains: an extracellular, a trans-membrane and an intracellular part

Despite the new findings, the highest effect is still exhibited by HLA-B27, which was replicated in many studies with multiple markers (87, 101) and consistently associated across different ethnicities (e.g., British origin: rs7743761, p = 5.0 x 10^{-304} (101); Han Chinese origin: rs13202464, p < 5.0 x 10^{-324} (98)). This association is among the strongest signals ever discovered for human disease. The majority (90%) of AS patients test HLA-B27 positive, but only 1-5% of the global HLA-B27 positive individuals will eventually suffer from AS (105). Additionally, epidemiological data showed that AS is more frequent in populations with high HLA-B27 prevalence, such as the Scandinavian
population (106). The observation that HLA-B27 status and disease do not always overlap provided the first insight that AS is polygenic (107). However, apart from HLA-B27, other HLA-B alleles are implicated in the disease. HLA-B60 is also associated with AS in Europeans (108) and was later found in the Han Chinese population (109), confirming its association with the disease in multiple ethnic groups. HLA-B60 acts in an additive way with HLA-B27, and consequently, this allele is 45 times more likely to appear in AS patients than in healthy individuals (110). Other HLA-B alleles that confer risk are HLA-B13:02, -40:01, -0:02, -47:01 and -51:01 (111), with the latter also being the allele mostly associated with Behcet’s syndrome (112). In addition, another MHC I (HLA-A02:01) and MHC II locus (HLA-DPB1, HLA-DRB1) exhibited an independent association signal with AS (111).

The main difference among HLA-B and HLA-B27 alleles is that the latter is misfolded during its synthesis in the ER (113,114) and can form homodimers with no β-microglobulin, part of the MHC1, on the cell surface (115). Regarding misfolding in the ER, when the newly synthesized HLA molecules remain unfolded or uncleaved, they cause ER stress and an unfolded protein response (UPR) (116), which has been implicated in many AID aetiologies (90). Either the accumulation is resolved by dislocation of the degraded protein complexes in the cytosol, or there is activation of adaptation and autophagy signals, oxidative stress, and inflammation (116,117). Additionally, different responses to UPR might cause differential expression of cytokines, such as TNF-α and IFNγ. Regarding the HLA-B27 heavy chain homodimers, unprocessed HLA-B27 can homodimerize and remain on the cell surface because MHC I molecules fold under slower assembly kinetics (118). Nevertheless, these homodimers, known as HC-B27, stimulate NK cells and Th17 cells through an IL-17/IL-23-mediated inflammatory response (119).

MHC I molecules (A, B, C) display antigen peptides to the T-receptors of CD8+ T cells. The antigen peptides are bound to a special structure of the MHC class I molecule, the peptide binding groove. At position 97 in the HLA-B molecule, the amino acid residue that lies exactly in the peptide-binding groove is linked to residue 29 at the C-terminal of the protein. The reference allele has a serine at position 97 (Ser97) and is most commonly found in HLA-B07 and HLA-B08 alleles. Val97 is found in HLA-B57 alleles, which are considered to be protective, since they are mostly found in healthy individuals. On the other hand, Asp97 and Thr97, which are uniquely found in HLA-B27 alleles and HLA-B51 alleles, respectively, are associated with increased AS risk.

**Non-MHC**

Non-MHC-associated markers implicated in AS lie in genes, such as IL23R, ERAP1, KIF21B, and RUNX3, and in the 2p15 and 21q22 intergenic regions. How these genomic regions interact and the mechanisms behind AS are still not clear, but all of these genes contribute to a fuller disease scenario pointing
mechanisms behind AS etiopathology: i) the processes of peptide trimming and antigen peptide presentation, ii) microbes or self-peptides activating the innate system and iii) the activation of the IL-23 pathway.

The associations among four aminopeptidases (ERAP1, ERAP2, LNPEP, NPEPPS) and two ubiquitination enzyme-coding genes (UBE2E3, UBE2L3) suggest that peptide trimming and presentation to MHC class I is crucial. Ubiquitinated proteins are marked so they can be recognized and degraded by the proteasome. After proteasome cleavage, the aminopeptidases cleave more amino acid residues from the N-tail until the peptides reach the optimal length or MHC class I presentation (nine residues). ERAP1 and ERAP2 studies showed that while the ERAP2 association is with HLA-B27-negative patients, ERAP1 can also be found in HLA-B40:01 carriers (111). The theories behind ERAP1 function in AS suggest that it leads to inappropriate MHC class I binding where self-peptide complexes are recognized as foreign, initiating an inflammatory response (arthritogenic theory) (120).

The association of the IL23 pathway with AS was established in many populations, and thus far, markers in six loci of the IL-23/IL-17 pathway have shown a strong association. Activated IL-23R, possibly induced by UPR, drives the differentiation of CD4+ positive Th17 cells that will produce IL-17 and subsequently activate the NF-κβ pathway. With NF-κβ activation and the secretion of pro-IL cytokines, inflammation is promoted and preserved. Another pathway responsible for inflammation, working in combination with the IL23 pathway, is the JAK-STAT pathway, including TYK2, JAK2, and STAT3, which is associated with AS (119). The IL-23 pathway rapidly emerged as an attractive candidate therapeutic target, and antagonist drugs were generated. For some of these new drugs, there is sufficient evidence of efficacy against disease, and some of these drugs are currently under trials (tofacinib; JAK inhibitor, fostamatinib; TYK2 inhibitor).

**Sexual dimorphism**

As mentioned before, men are more affected by AS with an observed 3:1 ratio across many countries (97). The clinical presentation, underlying inflammation and the response to treatment is overall the same between the sexes (121). Usually, men have an earlier disease onset accompanied by more radiographic evidence of inflammation-induced damage. However, male patients self-report lower inflammatory activity and have an overall better outcome with low impact on the quality of life when compared to female patients, who might develop disease later in life but of a more severe nature (122). The hypothesis that oestrogen might delay AS manifestation was shown in a mouse model for arthritis, where the pro-inflammatory cytokine TNFα, along with IFN-γ and IL-17A, reduced activity in oestrogen-treated mice (123). In addition, oestrogen inhibits Th17 cell production (124), which might be the reason behind the increased Th17 in males (125). Until more genetic studies pursue the
clarification of the basis of such observations, the underlying mechanisms will remain hidden.

**Treatment**
Chronic pain affects patient quality of life, and disease progression is accompanied by loss of proper posture, spinal mobility impairment, flexibility loss, osteoporosis (126) as well as high stress levels and chronic depression (127). Treatment focuses mainly on pain relief, monitoring disease progress and halting the underlying, disease-causing inflammation, and the selected regime is tailored according to the patient’s clinical profile (94). The drugs used are non-steroidal anti-inflammatory drugs (NSAID) together with analgesics, disease-modifying antirheumatic drugs (DMARD) and occasionally anti-TNF (anti-tumour necrosing factor) therapy, which is comparatively more expensive. A deeper understanding of the molecular mechanisms underlying AS pathogenesis might reveal new drug targets.

Overall, further understanding of the underlying genetics in AS is required. Guided by genetics, these studies might be able to predict drug response, reveal the mutations that predispose disease, and allow for early screening groups to be established so that treatment is assigned faster.
Cancer: ultimate complex disease

Cancer is a heterogeneous group of numerous disorders, all sharing the characteristic abnormal and uncontrolled cell proliferation, ability to invade the surrounding physiological tissues, evasion of apoptosis and other control mechanisms, such as immune surveillance (128). Organ and tissue of origin are used to distinguish cancers in six main categories: carcinomas (epithelial tissue), sarcomas (mesenchymal tissue), leukaemias (haemopoietic cells), lymphomas and myelomas (immune cells), and central nervous system cancers. Among these categories, carcinomas are the most common in humans (128).

Cancer is a genetic disease, as the functions of genes that regulate normal proliferation or programmed cell death need to be perturbed. Both germline mutations and the step-wise accumulation of somatic mutations are equally important in cancerogenesis.

Breast cancer

Breast cancer (BC) is the most diagnosed malignancy and the first and second cause of death in the developed and underdeveloped world, respectively. It is estimated that 1 in 8 (or 10) women will eventually develop breast cancer in their lifetime. However, during the last 30 years, timely diagnosis and better treatment contributed to reduced lethality (129). The exact aetiology of BC has not been determined, but there are some established risk contributing factors, such as being female, older age, family history, BC in a younger age, early onset menstruation and late onset menopause, belated first full-term pregnancy, hormonal substitution therapy, radiation, increased density (mammogram) and genetic mutation, e.g., BRCA1/2. Nonetheless, apart from sex and age, the mentioned risk factors are correlated with a small percentage of breast cancers.

Pathogenesis

All breast neoplasms will either occur in the epithelium lining the lobules and the lactiferous ducts of the mammary gland, with 85% of carcinomas being ductal and 15% being lobular (130). All breast carcinomas during the in situ phase (pre-invasive carcinoma) are confined within the epithelial compartment. However, when carcinomas become invasive, they breach the basement membrane and the epithelium barrier and infiltrate the surrounding connective tissue. Invasive tumours have metastatic potential, meaning establishing secondary deposits at distant sites through complex biological processes.

At the time of diagnosis, the primary tumour is staged according to size (T), direct invasion to lymph nodes (N) and the existence or absence of metastases (M). Tumours are surgically resected, and sentinel node biopsy for the auxiliary node is performed to more accurately determine the lymph node
status. During recent decades, the categorization of tumours according to molecular traits has become routine in clinics. Oestrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor 2 (HER2) are commonly used to divide the tumours into groups: i) luminal A, which differentiates at a slower rate and has a good prognosis; ii) luminal B, which is slightly more aggressive than luminal A; iii) HER2 enriched, which even though they are poorly differentiated and aggressive, these tumours benefit from targeted therapies; and iv) triple negative, with tumours that lack receptor expression and are truly aggressive (131). According to gene expression and copy number studies, the molecular subcategories might be more common and are shown in Table 3 (132).

Table 3. The most common molecular subtypes of breast cancer

<table>
<thead>
<tr>
<th>Molecular Type</th>
<th>Subtype</th>
<th>Molecular traits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminal</td>
<td>Luminal A</td>
<td>[ER+</td>
</tr>
<tr>
<td></td>
<td>Luminal B</td>
<td>[ER+</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[ER+</td>
</tr>
<tr>
<td>HER2</td>
<td>HER2 overexpression</td>
<td>ER-PR-HER2+</td>
</tr>
<tr>
<td>Triple negative</td>
<td>Basal</td>
<td>ER-PR-HER2+, basal marker+</td>
</tr>
<tr>
<td></td>
<td>Claudin-low</td>
<td>ER-PR-HER2-, EMT+, stem-cell marker+, claudin+</td>
</tr>
<tr>
<td></td>
<td>Metaplastic</td>
<td>ER-PR-HER2-, EMT+, stem-cell marker+</td>
</tr>
<tr>
<td></td>
<td>Interferon-rich</td>
<td>ER-PR-HER2-, interferon-regulated genes +</td>
</tr>
</tbody>
</table>

*ki-67: proliferation marker, EMT: epithelial to mesenchymal transition markers

Apart from the abovementioned criteria, there is the additional distinction of BC in sporadic and familial cases. Familial cancers are defined by certain criteria that include BC diagnosis before the age of 50, triple negative BC before the age of 60, d2 primary tumours, BC in men, ovary cancer, first degree relative with BC (133). Approximately 1/3 of the familial cancers have an identified genetic basis. The majority carry high penetrance mutations in one of the genes: *BRCA1*, *BRCA2*, *PTEN*, *TP53*, *CDH1* and *STK11*. Identified mutations in these genes are accompanied by protocols of sound clinical management. A smaller percentage of BC can be explained by intermediate penetrance, e.g., *CHEK2*, *ATM*, *BRIP1* and *PALB2* (134).

The role of genes in BC was highlighted with the definition and discovery of oncogenes and tumour suppressor genes. Oncogene alterations will cause gain-of-function effects, e.g., *HER2* and *c-MYC*, while tumour suppressor genes will lead to loss-of-function, e.g., *TP53*, both assisting the malignant phenotype. Mutations in these genes could be germline (e.g., Li-Fraumeni syndrome due to *TP53* mutations) but also somatic. Somatic mutations are randomly acquired throughout the lifetime after exposure to various carcinogens or ageing that damage the DNA.
Sequencing analyses in extensive cohorts have portrayed the somatic mutational landscape of tumours and revealed many of the genetic alterations that include copy number aberrations (CNA) and high-frequency substitution and insertion/deletion (indels). Sanger technology showed that the genomic landscape of BC involves the synthesis of a handful of genes mutated at a high frequency and numerous genes that are mutated at a lower frequency. The highly mutated genes are more likely to contain driver mutations, conferring a survival and proliferation advantage to cancer. Driver events are established at a higher frequency because they occur at an early disease stage. Additionally, passenger mutations exhibit a lower frequency (135) WGS analysis revealed a wide spectrum of structural rearrangements in BC along with the discovery of novel in-frame gene fusions \textit{ETV6-ITPR2, NFIA-EHF} and \textit{SLC26A6-PRKAR2A} (136). Recurrent BC mutations in \textit{GATA3, PIK3CA, MAP3K1} and novel mutations in \textit{CBFB} were described with WES. (137,138). In another study, novel recurrent mutations in \textit{AKT2, ARID1B, CASP8, CDKN1B, MAP3K1, MAP3K13, NCOR1, SMARCD1} and TBX3 contributed to a fuller view of the somatic landscape underlying BC phenotypes (139)

Overall, some of the most frequent somatic events include mutations (\textit{TP53, PIK3CA, CDH1, AKT1, GATA3, MAP2K7, MYC}), duplications (\textit{ERBB2}) and deletions (\textit{PTEN} or \textit{MAP2K4}), with mutation rates depending on BC type subtype, study design and cohort characteristics (size, age), (140–143). Pinpointing the exact contribution of these somatic alterations and utilizing these variations for treatment and prognosis remains.

\textit{Treatment}

After surgery, the treatments that can be used are standard endocrine treatment and chemotherapy. Standard endocrine treatment consists of tamoxifen in premenopausal women or an aromatase inhibitor in postmenopausal women. As proliferation increases, the effect of chemotherapy increases. Standard chemotherapy is based on a combination of anthracycline with a taxane. Luminal A is most responsive to endocrine treatment and least responsive to chemotherapy. ER negative subtypes are chemo-sensitive, whereas HER2 positivity indicates a response towards monoclonal antibodies, such as trastuzumab, pertuzumab, and lapatinib. Cancer genetic research has recently focused on the characterization of driver somatic mutations and attempting to link driver mutations to clinical phenotypes. Hopefully, further studies will provide insights into the more effective therapeutic regime selection, prediction of treatment response and possibly new therapeutic targets.
Aim of thesis

The main objective of this thesis was to expand our understanding of the underlying genetics in complex disease. The research presented aimed primarily to move past current methodologies and to study both germline and somatic genetic profiles. Cohorts of Swedish ancestry were enrolled to study Ankylosing spondylitis (AS), that mostly occurs in males and breast cancer (BC), a predominantly female illness.

In the work presented, the following issues were addressed more specifically:

- Characterize AS genetic predisposition in the Swedish population, by utilizing targeted sequencing on a representative cohort, and by analyzing the contributions from both common and rare variants.
- Investigate whether there are differential genetic modifiers for AS between the sexes, in Sweden.
- Characterize and explain the Swedish replication pattern of known AS associations and place Swedish contribution in a wider disease context.
- Provide a fuller somatic profile in Swedish BC patients to identify novel, or confirm known established, mutated genes
- Explore the relationship of the mutational signals within different phenotypic groups.
Present investigations

Paper I. The sex-stratified genetic architecture of ankylosing spondylitis

Aims and background

AS is a heritable, chronic and disabling disease with the hallmarks of unprovoked inflammation and the irreversible ossification of the spine and sacroiliac joints. The phenotypic heterogeneity of AS, with multiple comorbidities on several tissues and a prominent sexual dimorphism (3:1 ratio in males), is mirrored in its genetics, based on findings from recent genetic studies on extensive cohorts. Together, the almost 50 MHC- and non-MHC-associated loci can explain only 30% of the heritability and implicate key immune pathways and mechanisms. The highest risk is conferred by HLA-B variants, especially HLA-B27, with 95% of the AS cases being HLA-B27 positive. From the non-MHC loci, the association of aminopeptidases ERAP1/ERAP2 indicate antigen peptide processing and presentation to antigen presenting cells, while the IL123R association indicates signalling for Th-17 maturation. However, apart from HLA-B27, the remaining associations were not all replicated in other AS populations and in conjunction with additional clinical and diagnostic variabilities, suggesting differential genetics among men and women. The aim of this study was to study AS genetics in Sweden and investigate the role of sex in AS predisposition.
Methods

Individuals of Swedish ancestry were enrolled in the study and analysed with a “beyond the coding regions” targeted sequencing approach as large as a WES capture (1% of the genome). In brief, the tailored 32 Mb array targeted the coding and non-coding regions (3’- and 5’-UTRs, splice sites, promoters) of ~1,900 genes with known involvement in immune pathways. Conserved regions within 100 kb of the genes (SiPhy lod score >7) were also included. After quality control, the cohort included 310 AS cases and 381 healthy controls. Apart from targeted sequencing, both common and rare variants were incorporated in the statistical analyses, with gene- or region-based aggregate tests. A step-wise backwards elimination was employed to pinpoint the variants within the aggregate group that were driving the association signal. Functional validation of selected variants was performed with a luciferase reporter assay and EMSA using AS-relevant cell lines (HaCaT, Jurkat, K-562, SaOS-2). Replication of association was performed for selected variants in an independent cohort from the same Swedish region. Additional databases were utilized to explore the functional and regulatory properties of the variants, e.g., UCSC genome browser, GTEx, ENCODE Roadmap data, and sTRAP for differential TF binding.

- Overall, in a single marker test using all common markers, chromosome 6, containing the HLA-B, exhibited the highest association signal. This MHC signal was consistent when analysing male and female samples separately, while the male subset revealed an association at the RUNX3 locus (rs7414934: ORm (95% CI) = 2.01 (1.52-2.64), pm=1.7x10^-6, pf=0.4; Figure 1a, Table 1, Supplementary Tables S8). The locus also exhibited strong regulatory potential, acting as a distal promoter, residing upstream of a GeneHancer and containing multiple TFBS (Figure 2a).
- Markers were analysed simultaneously in a sex-stratified SKAT analysis, regardless of their MAF. This aggregate test revealed a single significant association in both sexes: male, P3H1 (or LEPRE1; p=1.3x10^-5) and female, MICB (p=1.2x10^-6).
- Haplotype analysis of the 5 associated variants in the RUNX3 locus showed that 33% of male and 20% of female cases carried the risk-associated pair H1.1, and this finding was replicated in an independent Swedish cohort.
- A luciferase reporter assay of the risk and protective haplotypes in the RUNX3 locus was performed in 3 AS-relevant cell lines. In all cell lines, the risk construct (larger fragment) acted as a significantly stronger enhancer (Figure 2b).
- A backwards elimination (SKAT-BE)-assisted method advised the prioritization of variants for subsequent functional inspection from the aggregate-significant loci. Specifically, in MICB, three variants (rs3828903,
rs2534671, and rs2534671) were selected for functional follow up and replication of the association in an independent cohort.

- rs3828903-A showed competitive binding across multiple cell lines, suggesting the variability of genes affected by this variant. Further inspection showed that this gene is part of a GeneHancer element (GH061031493: 31,461,141-31,466,978 bp) and might affect the expression of other genes based on GTEx evidence.

Discussion

Overall, our study proves that there is differential genetic architecture contributing to AS between the sexes, which is hidden when the cohorts are combined.

Autosomal heterogeneity between the sexes is outside the MHC, with the RUNX3 locus being significant in the male set when using common markers.

Further evidence for this sexual dimorphism was provided with the joint analysis of both common and rare markers with aggregate testing, where males showed an association in the P3H1 locus and females showed an association in the MICB locus. This finding suggests that while considering only common markers may increase significance and exhibits less severe correction issues, it may come at the cost of diluted genetic signals and hinder the ability to interpret fully disease predisposition, regarding the effect of all observed variants.

In addition, the observation that the signals in the associated loci were predominantly driven by non-coding variation emphasized the timing and cellular location of gene perturbation as key to understanding disease pathogenesis.

Regarding the insights into AS pathogenesis, our findings suggest mechanisms on the axis of TNF-α expression that are generally found more in male patients, and its production could be triggered by T cells. The availability and differential affinity to NKG2D of the ligands MICA and MICB might regulate T cell activation. The different haplotypes, with both common and rare variants, might influence the affinity, while oestradiol reduces the availability of surface MICB. Additionally, the role of Runx3 in osteogenesis and CD4 and CD8 cell regulation is known from previous studies. Our results suggest that Runx3 overexpression assists CTL differentiation by promoting chromatin accessibility during T cell receptor stimulation.

Overall, dissecting the influence of sex on autosomal variation may offer insight into other heterogeneous complex diseases, especially immune-regulated diseases.
Paper II. Replication and fine mapping of ankylosing spondylitis replicated loci in the Swedish population reveal different \textit{CCHCR1} protective haplotypes.

\textbf{Figure 5. Illustrative summary of Paper II}

\textbf{Aims and background}

Massive GWAS on cohorts of British origin has revealed about the current insights into the genetic background of AS. In these cohorts, the frequency of \textit{HLA-B27} is estimated to be 9.5\%, and the AS prevalence is 0.10\% (144). However, within Europe, the disease prevalence is higher in Sweden 0.18\% (106). Moreover, Sweden underwent unique bottlenecks, shows a longitude-dependent genetic variation (145), and private population variants were recently identified (86). AS prevalence in combination with the special genetic qualities of these individuals makes the Swedish population key for AS studies. The aim of this study was primarily to assess the reproducibility of known AS susceptibility patterns in a modest-sized Swedish cohort and place our findings in a more extensive population setting.

\textbf{Methods}

A previously utilized Swedish cohort was enrolled for this replication study (Paper I), and their \textit{HLA-B27} status was determined as positive or negative with a multiplex PCR (146) method. A literature search for AS genetic predisposition was performed, assessing AS publications up to 2017. The literature set of associations included markers that were: i) part of a large genetic study in cohorts of British or Han Chinese origin (> 2,000 markers), ii) showing association or indication of association with AS. Power calculations were performed for the Swedish cohort to replicate the literature set of selected associations. Replication analysis utilized genotypes from the previously described tailored targeted sequencing analysis of this cohort. Tests of AS
association were performed with Fisher’s exact test for allele count (Bonferroni’s threshold for the replication analysis: p-value < 1.25 x 10^{-3}). The genomic landscape surrounding the replicated variants was assessed, and fine-mapping was performed with a mixed polygenic model to account for genomic inflation (fine-mapping threshold: p-value < 5 x 10^{-8}). A pair-wise Weir and Cockerham’s F_{ST} analysis (147) was performed with data from the Swedish AS controls and individuals from the UK10K Twin registry.

Results

- The Swedish cohort had 80% power to replicate high effect AS associations, e.g., HLA-B variants. Medium effect associations (e.g., IL23R) required a specific allele frequency range, while a low effect (OR ≈1.1, such as ERAP1- rs30187) was not expected to be replicated.
- Thirty-three gene loci were available for replication due to target design and quality control filtering criteria.
- Four markers replicated the association: HLA-B (rs2596501, p-value = 2.8 x 10^{-16}), CCHCR1 (rs1265112, p-value = 3.7 x 10^{-6}), and two markers in IL23R (rs11209026, p-value = 7.15 x 10^{-4} and rs11465804, p-value = 1.17 x 10^{-3}).
- The different replication patterns were not due to subsampling or population differentiation.
- Fine mapping revealed signals repositioned and independent from the previous association signal, with an HLA-B risk haplotype (7 markers).
- In the CCHCR1 locus, two new associated haplotypes regulate the expression of different genes through eQTLs.

Discussion

This replication study aimed to place Swedish genetic predisposition in the context of previously published cohorts of British origin. While the replication pattern was not similar, it was not due to subsampling of the Swedish cohort or genetic substructure between the two European populations. F_{ST} analysis revealed high differentiation at the MHC on chromosome 6, but the differentiated windows did not contain previous AS associations. The fact that the medium and low effect associations could not be replicated in this cohort is in line with what was observed already for replication studies in East Asian or Han Chinese cohorts (e.g., IL23R-rs11209026: not replicated, ERAP1-rs30187 only moderately associated). In addition, RUNX3, despite being associated in UK (148), did not reliably replicate in either other UK cohorts or East Asian populations. A closer examination of the allele frequencies in Sweden, also proved that while a smaller cohort did not have the power to detect published associations, the genotypes determined at sequencing resolution point to mechanisms where their link to disease was not known. Notably, even
though our cohort could not compare in size with the large published cohorts, it represents 3% of the absolute number of AS patients in Sweden (310/11,000) and managed to provide new insights into AS genetics. The protective haplotypes identified in \textit{CCHCR1} were GTEx annotated to upregulate \textit{TCF19}, a known regulator of protein folding and ER transporter mediators (e.g., \textit{BIP}, \textit{p58IPK}, \textit{EDEM1}, and \textit{CALR}) (149). Reasonably, an upregulation of \textit{TCF19} might result in a more effective response to ER stress. The evidence of \textit{POUF51} expression regulation might be linked to gut epithelium comorbidities (150) but warrants further inspection.

Ultimately, the identification of the \textit{cis} regulatory \textit{CCHCR1} protective haplotypes, apart from providing the first suggestion of the differential function of the associated loci, suggests a regulatory role for other associated non-coding variants.
Aims and background
Breast cancer (BC), the most commonly diagnosed malignancy and a leading cause of death in women worldwide, has among the highest reported incidence rates in Northern Europe (151). NGS efforts revealed that somatic events, including CNAs and high-frequency substitutions and insertions/deletions (indels), take place in BC, including mutations in TP53, PIK3CA, CDH1, AKT1, GATA3, MAP2K7, and MYC, ERBB2 duplications and PTEN or MAP2K4 deletions. (140–143). Due to the lack of descriptions and comparisons for a specific European BC population, this study is an effort to examine the somatic landscape of a Scandinavian BC population, utilizing a targeted array designed based on both human and dog genetic studies.

Methods
A custom 20.5 Mb array targeted 765 selected genes and their surrounding regulatory regions based on i) their role in human BC and ii) previous association (analogous loci) in canine mammary tumour (CMT) GWAS (152). Sequencing data were analysed for 61 tumour-normal pairs after ancestry and relatedness analysis. Specific tools were employed for the calling of different types of somatic variants (point mutations, indels and CNAs) and filtered according to best practices. Subsequently, statistical methods were employed for the identification of recurrently somatic events and mutational signatures.
Results

- The cohort was 85% ER+ and showed a significant bimodal distribution of age that divided the cohort into two separate age groups i) under 70 years old (n = 38) and ii) above 70 years old (n = 23).
- The overall mutation rate of 2.7 mutations per Mb and the mean mutation rate of 0.50 mutations per Mb in coding regions were similar to previous studies.
- The significantly mutated genes (SMG, q < 0.01) were $PIK3CA$ (28%), $TP53$ (21%) and $CDH1$ (10%). The genes also appeared to be potential drivers in this cohort.
- Mutations in the second most mutated gene $KMT2C$ (23%) were mutually exclusive with $PIK3CA$ mutations ($p = 1.3 \times 10^{-3}$). Most mutations were predicted to lead to a truncated protein without a PHD domain.
- The significantly mutated genes were $PIK3CA$ (28%), $TP53$ (21%) and $CDH1$ (10%). The genes also appeared to be potential drivers in this cohort.
- Mutations in the second most mutated gene $KMT2C$ (23%) were mutually exclusive with $PIK3CA$ mutations ($p = 1.3 \times 10^{-3}$). Most mutations were predicted to lead to a truncated protein without a PHD domain.
- The histone modifying gene group was the most mutated group in the Swedish BC cohort ($KMT2C$ and $ARID1A$, together 28%)
- The most recurrent deletion (16q24.3, 80%), contained the BC tumour suppressor $CDK10$, and the most prevalent amplification (1q32.1, 80%) encompassed the BC oncogene $MDM4$
- Mutational signatures observed were APOBEC cytidine deaminase signature, signature 5 (substitutions at ApTpN) and defective DNA mismatch repair signature (signature 6)
- $TP53$ (younger group 29% vs older group 9%), $CDH1$ (younger group 5% vs older group 17%), and $CD23$ (younger group 18% vs older group 0%) were differentially mutated in the age groups.
- The APOBEC driving genes are not similar in the two age groups (younger group: $MAP3K1$ and $CDH23$ vs older group $KMT2C$)

Discussion

The somatic BC mutations in Swedish patients have not yet been described, despite the high prevalence of disease, the second highest in Europe (151). This study characterizes the somatic landscape of a mostly ER+ (85%) ser, utilizing a targeted NGS approach and a homogeneous Swedish cohort, by reporting the significantly mutated genes (SMGs), recurrent CNAs, mutational signatures and the age-dependent differences.

Notably, the inclusion of patients who did not receive neoadjuvant chemotherapy may have contributed to the lack of aggressive molecular subtypes, such as HER2-enriched or triple negatives. Despite the small cohort size and the representation of one molecular subtype, our cohort replicates the previously described somatic landscape of breast cancer (140,141,153), as we report $PIK3CA$, $TP53$ and $CDH1$ as SMGs (Fig 1). Interestingly, $CDH1$ was mutated in 10% of the cohort and classified as an SMG, despite being associated mainly with lobular carcinomas (154).
A histone modifying gene less frequently mutated in BC, $KMT2C$, showed the second highest mutation count, which was exclusive to $PIK3CA$ mutations and affected the recruit responsible PHD domain. The results underline the importance of chromatin regulators in the somatic landscape of BC and highlight the necessity for additional functional investigation of the role of $KMT2C$ in ER+ BC. Overall, this somatic fingerprint of the breast cancer landscape in Sweden highlights several potential differences from previous studies and suggests that a larger age-diverse population incorporating more breast cancer molecular subtypes should be studied to elucidate the underlying mechanisms.
Concluding remarks

This thesis presents studies on two complex diseases: ankylosing spondylitis (AS) and breast cancer (BC). Papers I and II focused on AS, a highly heritable chronic inflammatory disease of the axial skeleton and sacroiliac joints with unresolved aetiology and no cure, which is more prevalent in males than in females (3:1 ratio). Paper III studied BC, the most frequently diagnosed cancer among women, with the highest prevalence in Swedish individuals in Northern Europe.

With the main objective to improve methodologies for dissecting these complex diseases, both diseases were studied in Swedish cohorts through a tailored, targeted sequencing approach, and several key biological stratification factors were taken into account when analysing the genetic data.

The significant findings of this work were as follows:

Paper I
- Sex-stratified analysis of a modest AS Swedish cohort revealed associated loci, previously hidden by the standard case-control analysis.
- Incorporation of rare variants and common variants strengthened the differences in the genetic make-up of male and female AS patients.
- Association loci dissection suggested that risk variants might act in concert in haplotypes, e.g., MICB variants together act as enhancers.
- The RUNX3 locus haplotype suggests cis-regulation of other genes in AS.
- Pathogenesis scenarios were suggested along the axis of TNF-α regulation and RUNX3 involvement in T cell activation and recruitment during the inflammatory response.
- First experimental evidence of RUNX3 involvement in the ossification process, a hallmark of AS, through the results in an SaOS cell line.

Paper-II
- The Swedish cohort had the power to replicate only high-effect published associations, e.g., HLA-B.
- The sequencing resolution revealed fine-scale differences with repositioned association signals.
• Two CCHCR1 protective haplotypes may affect the expression of other MHC genes through eQTLs.
• Associated loci highlight different MHC genes through the cis-regulation function of the associated variants

Paper III
• The study represents the first fingerprint of the somatic mutational landscape of Swedish BC studying point mutations, indels, copy number aberrations, and mutational signatures.
• The Swedish cohort exhibits increased somatic mutation prevalence in the histone-modifying genes (e.g., KMT2C and ARID1A) not observed previously.
• Distinct age groups within the patients exhibit different mutational profiles and APOBEC-driving genes.

The studies described in this thesis aimed to provide insights for complex disease while improving current practices. In particular, this work indicates that new genetic information could derive from smaller homogeneous populations when studied at a different resolution (Papers I & II). Moreover, this work revealed that not accounting for data stratification factors when analysing genotypes can dilute the signals and hide valuable information (Papers I & III).
Future perspectives

In line with the initial goals, the work presented in this thesis highlighted the significant shift in the methodology. Utilizing homogeneous ancestry and data from coding and regulatory regions, combining rare and common variants and exploring data stratification within the same phenotype could greatly add to deciphering complex disease.

The sex-driven genetic differences from the presented AS studies provide a valuable perspective not only to AS but also to all immune-related diseases with prominent dimorphism genetics. Collaborative efforts assembling larger cohorts are needed, ensuring the necessary power to reveal associations with a lower effect. Indeed, future WGS analyses will reveal the totality of the differential architecture across the genome, but until these studies are feasible, imputed data could also provide a first broader glimpse. Specifically, for Sweden, a key population for AS studies, collecting a much larger cohort might not be possible, as the already analysed cohort represented 3% of the absolute number of AS patients (310/11,000) with the exclusion of related patients and ensuring homogeneous Swedish ancestry.

More immediate aspects of AS investigations will focus on the further acquisition of data to gain a complete understanding of the pathophysiology of the disease. Investigating the role of the sex chromosomes is still lacking, as the presented work focused on autosomal variation. In addition, it would be necessary to assess the role of MHC under the observed sex bias, as there are indications of differential architecture between men and women. (Paper I). Notably, MHC is a difficult region to analyse due to the high gene density and the observed LD, and its role in AS remains a puzzle. Experimentally, it would be interesting to assess the effect of the discovered \textit{MICB} haplotype \textit{in vitro} in a cell line amenable for inflammation investigation. Another key experiment would be to monitor the level of active disease in female patients who undergo oestrogen replacement therapy and are carriers of the \textit{MICB} haplotype to assess its effect \textit{in vivo}. It would be interesting to monitor cell activation and RUNX3 levels and link these effects with ossification levels in patients.

Similarly, the findings in the \textit{CCHCR1} locus warrant further investigation (Paper-II). Stress in the ER often occurs from the misfolding of HLA-B chains
in HLA-B27+ individuals before antigen presentation; thus, it would be intriguing to investigate this idea further. Specifically, it would be interesting to assess with an experiment whether ER stress and induced inflammation could be counter-balanced by TCF19 upregulation and linking TCF19 levels with the CCHCR1 protective haplotypes.

Genotyping of the missing markers, due to design or filtering criteria (e.g., HLA-B SNP in HWE, markers in intronic regions and gene deserts), will contribute to a fuller replication analysis. In this sense, WGS data would be more informative regarding the rest of the untargeted variation, specifically the structural variation that is not assessed in the current studies. Regarding the global disease setting, it would be intriguing to identify the similarities and differences with other AS-prone populations, e.g., Han Chinese, for which descriptive data were not available at the time of this thesis (p-values, allele frequencies, OR).

Following the BC study presented (Paper-III), there are numerous investigations to pursue. Among these future studies, prioritizing candidate variants for functional validation and assessing non-coding variation will be a priority. Additionally, exploring CNAs with a more reliable method, such as WGS or a specific copy number array, would reveal more information about the structural variation in these samples. Notably, WGS in our cohort would allow for better resolution of all somatic events the investigation of hypermutation (kataegis) points and assessment of all the pathways implicated in BC and cancer, as other studies have previously achieved. Incorporating more tumours in this analysis would improve the statistical power of the existing analysis, and it would also assist in contrasting the evidence of age stratification within the patients with no neo-adjuvant treatment. In the current study, the somatic profile of ER+ tumours was presented, while triple-negative and HER2+ tumours were almost entirely not represented. These more aggressive tumours usually occur in younger women, and their treatment is neoadjuvant therapy. However, incorporating all molecular portraits will result in a more spherical assessment of BC in Sweden. The exciting finding of the highly mutated group of histone-modifying genes (KMT2C, ARID1A) prompt a closer look at the whole set of genes with a histone-modifying function that is currently missing from the tailored, targeted array. Additionally, based on the different methodology presented for germline risk, it would be intriguing to assess this risk in the Swedish population, where patient cohorts from both sporadic and familial cases could be analysed and contrasted. Overall, placing Swedish BC risk in a broader context would be necessary, as Sweden is characterized by high disease prevalence.

A fitting end to this analysis is more general thoughts after studying these two complex diseases. During the last decade, biomedical research has been
sprinting an endless race, with new findings surfacing constantly. Notably, the numerous advances in genomics have led to considerable collaborations to unravel the genetics of complex traits. This spirit of sharing will probably continue to exist as the colossal task remains: identify the functional role of associations and deliver the results to the clinics. The goal of providing new diagnostic criteria, precise methods of early screening, and more effective personalized treatments seems attainable and might have a massive impact on health and socio-economic balance.

Overall, for complex disease, pieces of the puzzle will be gradually obtained by studying different populations and subgroups/phenotypes and assessing in detail the contributions from different layers of the genome. Therefore, this thesis presented an effort towards this direction, as this work attempted to discover novel genetic factors underlying complex diseases in a previously underutilized population while examining all types of variants, with no criteria on genomic topology or MAF.

However, have we come any closer to more personalized medicine? Genetics is gradually more incorporated in the clinics. This technique is used for the diagnosis of Usher’s syndrome, testing for $BRCA1$ and $BRCA2$ mutations, young-onset Alzheimer’s prediction, to analyse cytochrome P450 variants and predict responses to particular foods and medicines. Additionally, understanding genetic disease has led to a decrease in cystic fibrosis and Huntington’s disease rates. However, do all patients have access to this additional layer of information? Next-generation sequencing analyses are not incorporated at the same rate in all countries or at all hospitals within the same country. There are still large gaps between what scientists do in actual clinical practices and what is happening in our society.

Moreover, the majority of the different populations are studied just as subjects. What are the implications of this practice? Commercial genetic testing that promotes knowledge about potential health risk is not currently part of regular healthcare services. Can we handle the information about potential risks to make informed decisions? Additionally, reports of false and severe risks and reports of variants of unspecified significance create distress in potential patients. Is our genome sequence a medical tool that belongs in our medical history? There is still much work to be done, on multiple levels.
Acknowledgements

Oh, hello there! Rumor has it that this is the part that most of you will read, so I want to start by saying that undertaking this Ph.D. has been a truly life-changing experience, and without a doubt, it would not have been possible without the support and guidance that I received from many people.

First and foremost, I would like to thank my main supervisor, Kerstin. Thank you so much for being such a warm and brilliant supervisor! I am grateful for the one to ones in your office discussing this or that. Every time your view of things made me feel so inspired. Your input was valuable, and your support profound in my scientific upbringing. And yes, it was fun!

My deepest gratitude also goes to my associate supervisor, Jennifer. You have been there for me when I needed advice, and you were guiding me at every single turn. Indeed, you have taught me so much! From thinking in flowcharts to avoid rabbit holes, asking the right questions and how never to let the challenges get to me. Bless your cotton socks!

Vikki, wisdom flows freely from you, and I am lucky to have had you as my advisor during this time. Apart from all the science talks, the advice and encouragement you gave me have been transformative! Fără dumneavoastră nu aș fi ajuns aici. Vă mulțumesc din suflet!

Tobias, I was fortunate to have met and work with you during rather challenging times. The way you interact with your group and inspire them to be independent and take the initiative completely changed the way I see science, and for that, I am grateful!

Gerli, you are truly remarkable! Working with you was amazing and genuinely engaging! Thank you for all your advice and teaching/reminding me of Plan B in life. And Plan C. And Plan D.

Sergei, if it weren't for you, I would still despise immunology and cell work. Thank you for all the inspiring discussions and the encouragement you gave me nonstop throughout the years. Go team cell-lab!

To the immuno-group Jessika, Lina, Nina, Matteo, Daniel, Sergei: I am so thankful for all the talks about immuno-databases, aggregate tests, genomic spaces, filtering, and no filtering, experiment setting. It was an excellent time troubleshooting with you!

To the cancer people Viktor, Ginger, Maja, Kate, Chao, and Sharda: so happy that we could talk about all the cancer stuff. I enjoyed our discussions, the book club and honestly, transitioning to somatic analyses was more comfortable knowing that I have you guys to talk to.
Eva and Åsa what would I do without you? You know everything, and you are there to take care of all of us. Thank you for all the help!

Special thanks go out to the SNP&SEQ Sequencing platform for delivering the sequences and UPPMAX crew for all their assistance, the so much needed nobackup space and not shutting our jobs down!

All this work was performed in IMBIM, that has hosted me from my Master’s. A big thank you so much, Alexis, Veronica, Rehné, Malin and to all the administrators for their help and answering all my questions promptly!

Many thanks are owed to all our collaborators for providing with excellent data, sharing great research ideas, and making my work possible.

On a more personal note, I was truly blessed to have the best colleagues and office mates ever.

Doreen, from being my master's degree opponent, you became a dear friend of mine! Thank you for the "Beauty and the Beast" and for being always there. You have always pushed me to follow my dreams no matter what; you are the best! I promise I will always be there to eat the ginger from your sushi.

Jessika, I am incredibly grateful for your rhubarb pie, helping me out with scripts, the awesome panda-time in our office and your friendship. I will always cherish our UK and Paros trips!

Sharda, if there is a person that makes bioinformatics look like a game is you! Thank you so much for your support and teaching me cool tricks! I am grateful that you introduced me to some fantastic audiobooks and of course my new stress-managing habit Etsy.

My dear Fan, our coffee breaks became a habit that I know from now that I will miss deeply. Talking about everything and nothing while staring at the cathedral. Thank you for all the wise advice and your friendship!

Caro Matteo, from the time you cursed your computer I knew! I am delighted to count you among my friends, and thank you for making me laugh even in the most stressful times.

Throughout the application and especially during writing, I have received a great deal of support from the T & T team! Thibaut and Tillman thank you for all the laughter and fun chats! Believe me, they did not go unappreciated.

Erik, Matt, and Andreas thank you for all your advice throughout the years! I was fortunate to have met you, and I appreciate that you have allowed me to blubber about Crete, climbing or how much I want Iron Maiden to sign this thesis…more than once.

Many thanks to all D11:3 people, the old and the new, because you have helped me in countless ways so far. I cherish the brainstorming times as well the fantastic personal interactions with all of you! Thanks for all the fantastic magnets that you have brought for me from all corners of the world!
Outside the lab, I am indebted to all my friends, the family I have chosen for myself here.

Dancing crew, mainly Lennart and Marie for teaching me how to let go of the stress of the day. You taught me to practice, patience, and how to believe in myself! Climbing crew Kevin, Andrea, Philip, Saba thank you for the belays. I have truly put my life in your hands and thank you for always bringing back to earth. To my yoga crew, I will forever cherish the warmth you gave me during the RYT200 training. D&D crew, Axel and Sina thank you for being so much more than my imaginary support team. Thank you for your long-lasting "healing" and "inspire" spells!

Simona, Emanuele, Ben, and Marilina, you mean the world to me. We have done so much together that I can’t even count: games, Greek parties, edges of Norway, forests, canceled concerts, lamb roasts, road trips. And so much more... You have lifted me up and boosted my confidence many times. Thank you for being in my life and being my best buddies!! With you around the floor can never be lava.

Σαμίρα, Ιωσηφίνα, Σταυρούλα, Ελένη…τα λόγια είναι περιττά. Όλες σας, καθεμία με το δίκο της τρόπο, μου δώσατε τόση αγάπη και συμπαράσταση! Από το να πείτε τον κακοφτιαγμένο καφέ μου ή να φάτε το κρακεράκι πίτσα που έφτιαξα, από το να μου στέλνετε μηνύματα στο Ινσταγράμ ή καρτ-ποστάλ, στο να με ποτίζετε επίτηδες για να λυθώ, βραδινά τηλεφωνήματα και να με μαλώνετε για να πάρω αποφάσεις. Σας ευχαριστώ από τα βάθη της ψυχής μου!

Any type of acknowledgment would be genuinely incomplete, without thanking the most significant source of my strength: My family. Μαμά, μπαμπά σας ευχαριστώ για την υποστήριξή και κατανόηση σας αυτά τα χρόνια. Αν και μακριά, ένοιωθα την αγονία και την αγάπη σας. Είμαι ευγνώμον που κάθε φορά που με επισκέπτεστε προσπαθείτε να χωρέσετε το σπίτι σε μια βαλίτσα. Μα σπίτι δεν είναι τα λεμόνια, η φέτα ή το σπιτικό κεικ.. Είστε εσείς.

Μέλια, για μένα έκανες το Λονδίνο/Ουψάλα Πανόρμου/Ευαγγελισμός. Δεν υπάρχει στο κόσμο άλλη αδερφή που να κάνει αυτά που έκανες και κάνεις για μένα. Δεν κατάφερα να κάνω την Σουηδία να ανθίζει τον Νοέμβρη αλλά με σένα κατάφερα αυτό που κρατάς. Σε ευχαριστώ που πάντα πιστεύεσε σε μένα. Είμαι περήφανη που σε βλέπω να γίνεσαι ο άνθρωπος που θες. Κι αν σε αγαπάω...; Όσο όλες τις γάτες του κόσμου μαζί...άπειρα εν ελίγος.

Andris, you have always believed in me! When we hold hands, everything falls in place and I know how to move forward and tackle whatever. Thank you for being my rock through this time.
Άλλης, σε ταλαιπώρησα καιρό. Καταφέραμε πολλά αλλά υπόσχομαι να μην το ξανακάνω. Πάμε._
References


53


109. Wei JCC, Tsai WC, Lin HS, Tsai CY, Chou CT. HLA-B60 and B61 are strongly associated with ankylosing spondylitis in HLA-B27-negative Taiwan Chinese patients. Rheumatology. 2004 Jul 1;43(7):839–42.


Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1588*

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-390457