Genome wide association study in steroid sensitive nephrotic syndrome

Dr Stephanie Dufek-Kamperis

A Thesis Submitted for the Degree of

Doctor of Philosophy

University College London

April 2020

Declaration

I, Stephanie Dufek-Kamperis, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Steroid sensitive nephrotic syndrome (SSNS), the most common form of nephrotic syndrome in childhood is considered a complex disease with the immune system playing a critical role in its development. This is supported by recent molecular findings showing an association with classical human leukocyte antigen; yet, the exact nature of the disease, specifically the genetic architecture outside the HLA region, has not been elucidated.

With this thesis we aimed to explore the genetics of SSNS by performing a genome-wide association study on a cohort of 422 European cases and 5642 ethnically matched controls with more than 5 million high-quality imputed genome-wide markers.

Our results revealed three *loci* achieving genome-wide significance in association with the disease. The strongest association was found within the HLA-DR/DQ region (lead variant rs9273542, $p=1.59\times10^{-43}$, OR=3.39, 95%CI=2.86-4.03) confirming findings of previous GWAS. Moreover, we are the first reporting on two *loci* outside the HLA region on chromosome 6q22.1 and 4q13.3 that are associated with SSNS with genome-wide significance. The region on chromosome 6q contains the gene *CALHM6*, which has been implicated in the regulation of the immune system and is particularly expressed on CD4+ cells and naïve and memory B cells. The identified lead variant (rs2637678, $p=1.27\times10^{-17}$, OR=0.51, 95%CI=0.44-0.60) is a strong expression quantitative trait *locus* (eQTL) for *CALHM6*, with the risk allele predicting lower expression of *CALHM6* on lymphocytes and hence possibly altered immune regulatory responses. The same variant is also an eQTL for the neighbouring gene *DSE,* which codes for an enzyme essential in the dermatan sulfate production. Overexpression of dermatan sulfate has been previously associated with glomerular diseases and could be a potential antigen involved in SSNS.

These findings support the hypothesis that the immune system and its dysregulation play a critical role in the pathogenesis of SSNS.

Impact statement

Despite SSNS being a rare disease, it is the most common form of nephrotic syndrome in childhood and affects approximately 1 in 10,000 children. Children present with heavy proteinuria, oedema and low albumin, and per definition respond to steroid treatment. Until now our understanding of the disease pathomechanisms is mainly built on indirect hints supporting that the immune system plays an important role. By performing the largest genome-wide association study (GWAS) so far reported on SSNS, we shed light on the genetics of this complex disease. This has crucial impacts on unravelling the pathomechanisms and developing the optimal treatment strategies for these patients.

To date, 3 GWAS on SSNS have been published showing an association of the HLA region with the disease, confirming the general involvement of the immune system in the disease development. Arguably, the crucial key for the understanding of the pathogenesis of the disease is the detection of an association outside the HLA region. In this thesis, we were the first to detect such an association outside the HLA region. Our lead variant in the non-HLA region is associated with altered expression of two candidate genes, *CALHM6* and *DSE*.

The first, *CALHM6,* is a gene that has been implicated in immune regulatory processes. It is highly expressed on lymphocytes and regulated by cytokines, mainly IFN-y. Although little is known about the function of the protein encoded by *CALHM6,* this finding strongly supports the concept that immune dysregulation is crucial in the disease pathomechanisms and contributes to our understanding of the aetiology of this rare autoimmune disease.

Moreover, the lead variant identified in this study also alters the expression of the gene *DSE,* which codes for an enzyme involved in dermatan sulfate synthesis. Overexpression of dermatan sulfate has been associated with other glomerulopathies including FSGS. Hence, our study could have possibly contributed to the detection of an antigen involved in the development of SSNS.

Although at the current point of time, these hypotheses need further investigation and the exact pathomechanisms remain to be elucidated, our findings are the first providing more detailed insight into the genetic architecture of this complex disease. By identifying an association outside the HLA region with the disease, we shed light on the possible pathomechanisms and discovered two candidate genes for the disease. Our results are published in a high impact journal and are made available for all researches with interest in this field. The identification of these *loci* open up the field for a whole new line of research into SSNS.

Acknowledgments

With the greatest appreciation, I would like to extend my gratitude to all people who helped me bringing this study into reality, in particular to the following:

To Robert, who has the substance of a genius, for the unique opportunity to perform a PhD in your renowned group and for setting the standards for highly acclaimed research. Without your consistent guidance throughout and your incredible patience in working with me over the past 4 years, this PhD would not have been possible.

To Detlef, who impressively combines clinical excellence and cutting-edge research, for providing indispensable advice and information, for your support on many aspects of my project and for continuously being inspiring in regards to my career and life.

To Horia, the best teacher, your seminars were probably the most fascinating and useful I have ever taken. Thank you for sharing your endless knowledge and for always being supportive and encouraging.

To Prof Magdi Yaqoob and the William Harvey Research Institute, who generously supported this thesis through a William Harvey Paediatric Fellowship.

To the greatest team of colleagues, that I have worked with over the past years, that always had an open ear for my worries and were up for a laughter: Chris, Sanjana, Matt, Monika, Anne, Melanie, Mallory, Vaksha, Mehmet and Anselm. A special thanks to Chris, without your admirable computer skills this thesis would not have been possible.

To my parents, Georg and Elli, there are no words expressing my gratitude for the unconditional love and support you have given me, throughout this thesis, but also throughout my whole life.

Finally, I thank you Kostas, for the encouragement and support you have given me, for patiently dealing with my dissertation frustration and for the uncountable times you read through this thesis. At the same time whilst you were looking after our most precious gift, Paris.

Content

List of figures

List of tables

Abbreviations

1KGO – 1000 Genome Project
3KBC – 1958 British Birth Cohort
A – Adenine
AC – Allele Count
BAT – Basic Allele Test
BCF – Binary Calling Format
BED – Browser Extensible Data
BP – Base Pairs
C – Cytosine
CALHM6 – Calcium Homeostasis Modulator Family Member 6
CR – Call Rate
DBSNP – The Single Nucleotide Polymorphism DataBase
DNA – Deoxyribonucleic Acid
DSE – Dermatan Sulfate Epimerase
FSGS – Focal and Segmental Glomerulosclerosis
G – Guanine
GWAS – Genome Wide Association Study
HAT – Haplotype Association Test
HLA – Human Leukocyte Antigen
HMM – Hidden Markov Model
HWE – Hardy-Weinberg Equilibrium
IBD – Identity By Descent
LD – Linkage Disequilibrium
MAF – Minor Allele Frequency
MCD – Minimal Change Disease
MN – Membranous Nephropathy
NCBI – National Centre for Biotechnology Information
PARM1 – Prostate Androgen-Regulated Mucin-Like Protein 1
PCA – Principal Component Analysis
PLAR2 – Anti-Phospholipase-A2-Receptor

QC – Quality Control

REMEDY – Re-coding For Merging of Genotyping Data

RNA – Ribonucleic Acid

SNP – Single Nucleotide Polymorphism

SSNS – Steroid Sensitive Nephrotic Syndrome

SVS – SNP & Variation Suite

T – Thymine

VCF – Variant Calling Format

WTCCC – Wellcome Trust Case Control Consortium

Part 1: Introduction

Chapter 1. Nephrotic syndrome

Introduction and historical account

Nephrotic syndrome is a renal disease in which increased permeability of the glomerular filtration barrier leads to loss of vast amounts of protein in the urine. It is characterized by a triad of symptoms:

- Proteinuria
- Hypoalbuminemia
- Generalized oedema

The first description of patients with severe proteinuria dates back over 2000 years to Hippocrates, who identified the relationship between "*bubbles settling on the surface of the urine*" and disease of the kidney [1,2]. In 1484 Cornelys Roelans of Belgium described a child with "*whole body swelling*". He further commented on the treatment of this disease by suggesting a mixture of herbs and other remedies. During the 16th to the mid 18th century the concept of "*dropsy*" for generalized oedema was established, but was used as a general term for many types of swellings, with no distinction between potential causes.

The first accurate description of "*dropsy*" related to nephrotic syndrome in children was made by paediatrician Theodore Zwinger of Basel in 1722, who described a range of symptoms including pitting, generalised oedema, thirst, altered bowel motions and affection of the respiratory system [3]. Already at that point, he associated the swelling with renal disease. He noticed that urine output was decreased and related that finding to compression of the tubules in the kidneys.

In the later 18th century Morgagni and his follower William Heberden pursued the concept that the disease might be attributed to a specific organ. The division of dropsy into those dependent on "*morbid viscera*" (liver and heart) versus a "*general, inflammatory*" form was established [1]. During this time, several observers, including

Cotugno, Cruikshank, Wells and Brande noted that urine had a high specific gravity with a high amount of albuminous material [3,4].

In 1827 Richard Bright set a landmark by distinguishing the heat-coagulable albuminous material in the urine as a result of a kidney disease [5]. He further made the association of "*dropsy*" with albuminuria and low albumin in the serum. This was based on findings of his colleague, John Bostock, who used specific gravity to quantify protein in the serum. He observed that the specific gravity of the serum is the lowest, if there is a high amount of albumin in the urine. This was soon after confirmed by Robert Christison of Edinburgh in 1829 [4]. Hence, the definition of nephrotic as we know it today, with heavy albuminuria, hypoalbuminemia and oedema, as a result of a kidney disease, was finally established in 1830.

In the following century as histological techniques improved, the attention was drawn to the microstructure of the kidneys in health and disease. In 1905 the distinction between "*nephritis*" for "*inflammatory*" lesions of the kidneys and "*nephrosis*" for "*degenerative*" lesions of the kidneys was introduced by the German pathologist Friedrich von Müller [6]. In the 1930s the transition from the clinical and pathological entity "*nephrosis*" to "*nephrotic syndrome*" as a picture not directly related to glomerulonephritis or inflammatory renal disease was gradually established [7]. And in 1950s with introduction of the percutaneous renal needle biopsy by Poul Iverson and Claus Brun in Denmark the understanding that nephrotic syndrome can be associated with a variety of different types of histopathology evolved [6].

Nephrotic syndrome in childhood

Nephrotic syndrome in children is defined as per KDIGO criteria [8]:

- Generalized oedema
- Protein excretion ≥40 mg/m^2/hour, Urine Protein/Creatinine Ratio ≥ 200mg/mmol, or 3+ Protein on urine dipstick
- Hypoalbuminemia ≤ 25g/l (≤ 2.5g/dl)

The incidence reported is 4.7 (range 1.15 - 16.9) per 100,000 children worldwide aged below 16 years with a male to female ratio of 2:1 [9]. The incidence depends on the geographical area and varies between different ethnicities [10]. The highest incidence is reported among South Asian, African American and Arab children [11].

Idiopathic nephrotic syndrome

Idiopathic nephrotic syndrome (INS) refers to nephrotic syndrome where an underlying disease is not identified and is the commonest form of nephrotic syndrome in childhood [12].

If a renal biopsy is performed in children with INS, typically only minimal changes are seen, most commonly foot process effacement on electron microscopy. Hence, the entity is often also referred to as “minimal change disease” (MCD) and the terms MCD and INS are used interchangeably [13]. In a minority of children with INS, in addition to such minimal changes, focal and segmental glomerulosclerosis (FSGS) is seen [14]. It is under ongoing discussion whether MCD and FSGS represent different stages of one disease or if they should be considered as separate entities because of their different likelihood to respond to steroid treatment, associated with different prognosis and outcome [15]. Patients with FSGS are often (but not always) resistant to steroid treatment, are more likely to progress to end stage renal disease and to relapse after renal transplantation. In contrast, patients with MCD typically respond to steroid treatment and may have a milder disease course. Interestingly some children with initial MCD might at a later point develop histological findings of FSGS, further complicating this discussion.

At presentation, most children with MCD can be distinguished clinically from those with other biopsy findings such as FSGS or secondary NS [14]. Patients with MCD are usually younger than six years at presentation, have no macroscopic haematuria, normal complement levels and normal renal function [14]. Therefore, most of the children presenting with nephrotic syndrome do not undergo a biopsy in the first place, but are subjected to a course of steroids first [16]. Children are consequently classified according to their response to steroid treatment, which has great implication for treatment, prognosis and outcome [17]:

Steroid sensitive nephrotic syndrome (SSNS): The majority of children (approximately 90%) do respond to steroid treatment [16], which is defined as complete remission within the initial 4 weeks of corticosteroid treatment [8]. After the initial response to steroids a large number of children with SSNS would experience one or more relapses. It is estimated that approximately half of the patients with SSNS will have a frequently-relapsing and/or steroid-dependent course [17].

Steroid resistant nephrotic syndrome (SRNS): SRNS is defined as failure to achieve complete remission after 8 weeks of corticosteroid treatment [8]. Those patients have an increased risk of development of chronic kidney disease and a less favourable long-term outcome [17]. These children usually necessitate treatment with other immunosuppressive agents.

Secondary nephrotic syndrome

Nephrotic syndrome caused by an underlying process affecting the glomerulus or as part of a systemic disease is referred to as secondary nephrotic syndrome:

- Post-infectious glomerulonephritis
- Nephritis secondary to systemic lupus erythematosus
- Nephritis associated with vasculitis, such as Henoch-Schönlein purpura, or with granulomatosis with polyangiitis and microscopic polyangiitis
- IgA nephropathy
- Other causes include Alport syndrome and haemolytic uremic syndrome as well as sickle cell disease

Congenital and infantile nephrotic syndrome

Congenital nephrotic syndrome (CNS) and infantile nephrotic syndrome (INS) are separate entities where patients present either within the first 3 months or the first year of life with severe proteinuria, hypoalbuminemia and oedema. Both are associated with a high morbidity and mortality [18]. Those patients typically do not respond to steroid treatment and may suffer from severe complications, such as recurrent infection, thrombosis and impaired growth [19]. In many centres, including developed countries, active treatment was not offered until the 1980s. CNS and INS is primarily caused by mutations within the *NPHS1* gene that encodes for Nephrin [20]. Further, mutations in other genes including *NPHS2, PLCE1, WT1, LAMB2, PDSS2* and *COQ2* can cause CNS and are associated with clinically heterogeneous phenotypes [21-24].

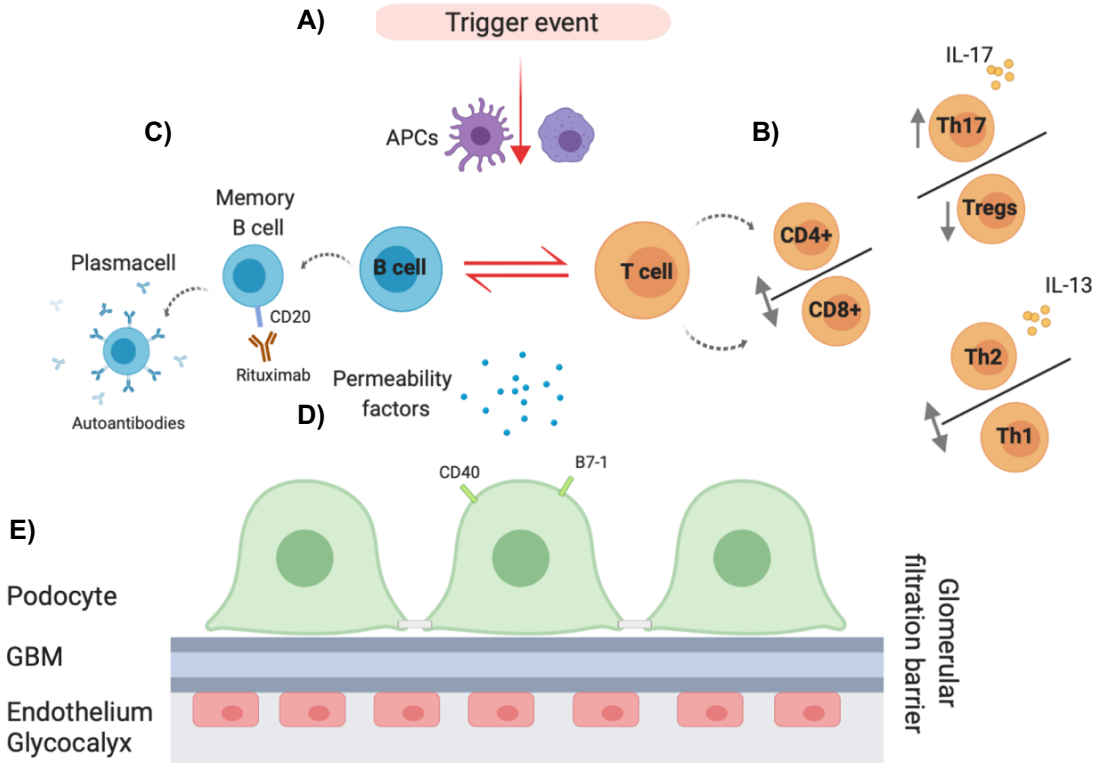
Chapter 2. Pathogenesis of idiopathic nephrotic syndrome

The glomerular filtration barrier

The glomerular filtration barrier consists of three layers 1) the fenestrated endothelial cells covered by a glycocalyx, 2) the glomerular basement membrane (GBM) and 3) podocytes. Podocytes are highly differentiated cells, whose cell bodies have multiple extensions, called foot processes, wrapping around the glomerular capillary loops. The foot processes are arranged in an interdigitating pattern and are interconnected by slit diaphragms which together form the glomerular filter.

INS is associated with changes in podocyte architecture. These changes consist of loss or effacement of the podocyte foot. The exact pathogenesis behind podocyte foot process effacement, its relationship to proteinuria and also its response to steroids and the variable courses are not fully understood and cannot be predicted. Theories around the pathogenesis are detailed in Figure 1.

Figure 1 *Pathogenesis of idiopathic nephrotic syndrome*



Legend *Summary of proposed mechanisms leading to idiopathic nephrotic syndrome. APCs: Antigen-presenting cells; GBM: glomerular basement membrane*

A) *The onset is often related to trigger events, such as viral infections, allergies, vaccinations. Trigger events are thought to stimulate the immune system, antigen presenting cell (APC), T and/or B cells.*
B) *The involvement of T cells in idiopathic nephrotic syndrome has been long time speculated. Studies support a reduction of CD4+ cells with increased CD8+ cells, however, findings are controversial and no clear predominance of either has been proven. An imbalance between Th2 and Th1 cells with an increase in the production of the Th2-specific interleukin-13 (IL-13) has been discussed, but no clear cytokine profile has been identified. Recently, an imbalance between Th17 cells and regulatory T cells (Tregs) towards increased activity of Th17 cells and its Interleukin 17 (IL-17) has been observed.*
C) *The involvement of B cells is supported by the beneficial role of rituximab, a CD20 antibody. A correlation between the recovery of memory B-cells after rituximab treatment and relapse has been observed. The role of autoantibodies, including anti-CD40, has been discussed.*
D) *Since long time the role of circulating permeability factors such as hemopexin, the soluble form of the urokinase-type plasminogen activator receptor and the cardiotrophin-like cytokine factor 1 has been discussed to directly alter the podocyte function, leading to foot process effacement and disruption of the glomerular permeability barrier.*
E) *The podocytes themselves can express specific molecules such as CD40 or CD80 (B7-1) which can interact with the immune system.*

Circulating permeability factor

This model of nephrotic syndrome pathophysiology proposes that in patients with INS, normal kidneys exist in an abnormal environment. The theory that circulating permeability factors play a role in the development of the disease has been supported by several observations and studies primarily on FSGS patients.

Recurrence of NS in the donor kidney of FSGS patients after transplantation of a non FSGS kidney and the efficacy of plasma exchange to induce remission in patients with FSGS support the theory of a causative factor in the plasma [25,26]. Further, the transplantation of a kidney into a recipient with the disease resulted in immediate recurrence, but when the kidney was subsequently transplanted into a patient without the disease it induced remission in the transplant kidney [27,28]. In animal models it has been shown that proteinuria can be induced in healthy rats by transfusing serum of a FSGS patient into the animal [28]. For MCD the evidence of a circulating factor is less striking. However, the evidence is supported by data showing that albuminuria can be induced in rats after infusion of peripheral blood cell products from patients with MCD [29].

Throughout the years, several molecules have been shown to be able to modify the shape and the proprieties of podocytes and to induce proteinuria in experimental *ex vivo* and *in vivo* conditions. We here summarize the most discussed circulating factors that have been hypothesized to be associated with MCD and FSGS [30].

Lagrue *et al* [31] were one of the first studying the effect of plasma factors on vascular permeability in guinea pig skin capillaries. The injection of supernatants derived from lymphocytes of MCD patients resulted in significantly higher vascular permeability compared with controls. The authors concluded that a vascular permeability factor (VPF) is generated by stimulated T lymphocytes from patients with MCD [31]. Several studies followed, however, the exact nature of the VPF remains unknown. Further, VPF seems not to be specific for MCD, but is also common in other glomerular disease and more importantly, direct evidence that it acts on the glomerular capillary wall permeability and causing proteinuria is not provided [32].

The study of VPF in guinea pig skin models has clear limitations because of the discrepancies between vascular and glomerular permeability. Hence, subsequent studies focused on characterizing a glomerular permeability factor using rat models and looking at the glomerular histology and proteinuria as permeability markers [33].

The study of the glomerular permeability factor led to the identification of hemopexin [34]. Hemopexin is a protein mainly produced in the liver and is increased in the acute phase reaction to inflammation or infection. In normal conditions hemopexin is inactive, but it gets activated in certain circumstances [34]. Activated hemopexin has serine protease activity and can alter the glomerular filtration barrier [34]. *In vivo*, activated hemopexin provoked reversible proteinuria in rats together with podocyte foot process effacement. In children with MCD relapses, activated hemopexin was found to be increased [34]. Activated hemopexin is believed to alter the glomerular permeability by causing a nephrin-dependent remodelling process of the podocytes and a degradation of the glycocalyx [35]. However, the crucial process leading to the activation of hemopexin remains unclear [30].

For FSGS, cardiotrophin-like cytokine 1 (CLC-1), a galactose binding factor and member of the interleukin 6 family has been proposed as a permeability factor [36]. It was intensively studied by a group in the US who analysed the plasma of patients with FSGS recurrence after transplant [36,37]. The group used a functional assay of isolated rat glomeruli that showed changes in the glomerular permeability to albumin when incubated with the patients' plasma. With subsequent purification steps the same group concluded that the permeability factor resides in a 30- to 50-kDa plasma fraction and identified CLC-1 as being enriched in FSGS patients plasma [36,37]. They showed that CLC-1 increases the glomerular permeability to albumin and induces proteinuria in rats [36]. The same group hypothesized that galactose administered to patients with FSGS might prevent the development of CKD by preventing the binding of CLC-1 on galactose residues present at the surface of podocytes [38]. However, a clinical trial in paediatric patients with SRNS did not confirm the hypothesis [39].

Recently, the role of urokinase-type plasminogen activator receptor (uPAR) in its soluble form (suPAR) has been the centre of discussion as a proposed circulating permeability factor causing FSGS [40]. uPAR is a membrane bound urokinase

receptor and its stimulation has been linked to foot process effacement and proteinuria [41]. Cleavage of uPAR results in its soluble form suPAR. Studies demonstrated that both overexpression of uPAR on podocytes and the administration of suPAR cause proteinuria in mice by deforming the podocyte cytoskeleton [40]. suPAR was elevated in two-thirds of FSGS patients (significantly higher than in patients with MCD) and in those having recurrence after transplantation [42]. However, subsequent studies have given conflicting results and at present, there is no consensus that suPAR plays a pathophysiological role in FSGS [40].

Despite intensive research into this topic a clearly pathogenic circulating factor has not been convincingly identified yet (Reviewed in [30,37]). The data presented above, however, suggest that a possible circulating factor in INS might be generated by mononuclear cells such as T cell or B cells, pointing to a role of the immune system.

Immune dysregulation

Since decades the involvement of the immune system in the development of SSNS has been suspected. This is driven by the observation that the disease is commonly triggered by infection, such as upper respiratory infections and a good therapeutic response to corticosteroids, as well as to other immunosuppressants. It is also supported by histological findings, where a full reversibility of the foot process effacement with corticosteroid treatment is seen [43].

T cells

The pathophysiological role of T-lymphocytes in the development of the disease was already suggest in the 1970s [44]. Further supporting the idea that T lymphocytes are the source of a possible permeability factor is the observation that remission can be induced by steroids and cyclophosphamide, which both suppress a T cell mediated response. Furthermore, measles infection can induce remission in patients with INS by affecting the cell-mediated immune system and suppressing T cell subsets [45], whereas T cell lymphomas such as Hodgkin's lymphoma, can trigger INS in patients with subsequent chemotherapy inducing remission [46].

Widespread research into the involvement of T cells in the pathophysiology of INS is published. The results can be summarized as follows [47]:

Studies demonstrated that patients with INS show an imbalance of T-cell subpopulations, with a reduction of CD4+ T helper (Th) cells [48,49]. However, other studies could not replicate these findings and did not find any difference in the expression of CD4+ or CD8+ T cells in patients with INS [50].

In a different line of research, the involvement of cytokines in the pathogenesis of INS was investigated. Cytokines are small proteins that function as soluble mediators and are produced by immune and non-immune cells. Several cytokines were discussed to be related to INS, however, no distinct Th1 or Th2 cytokine profile has been identified. Reported results are variable and their interpretation remains difficult as studies show heterogeneity in their methodologies [51,52].

Recently, an imbalance between Th17 and T regulatory (Tregs) cells towards Th17 cells and their cytokine IL-17 has been reported in children with INS [53]. Also in adult patients with MCD, an increase in Th17 cells and IL-17 with a decrease of Tregs was observed and the imbalance returned to normal after effective corticosteroid therapy [54]. Further, it was observed that IPEX syndrome, a genetic disease leading to dysfunction of regulatory T cells, was complicated by MCD [55]. Hence, the involvement of Th17 cells and the imbalance between Th17 cells and Tregs is one of the emerging topics in the pathogenesis of INS, but does need further studies and confirmation.

B cells

The role of B cells in the pathophysiology of the disease is primarily supported by the beneficial effect of rituximab, a monoclonal antibody directed against CD20 [56]. Several clinical trials have confirmed the beneficial role of rituximab on inducing and maintaining remission in patients with steroid dependent or frequent relapsing SSNS. However, CD20 is not expressed on antibody producing plasma cells, hence the action of rituximab on these cells is assumed to be indirect via inhibiting the regeneration of plasma cells from activated CD20+ B cells [57].

The role of antibodies in INS is not yet clarified. The involvement is supported by data showing that the permeability factor may be an immunoglobulin or may bind to an immunoglobulin [58]. Further, reduced IgG levels have been observed in patients with SSNS also during remission [59]. And lately the role of anti-CD40 antibodies in recurrent FSGS patients has been discussed [60]. CD40 is a transmembrane protein that is mainly expressed on antigen presenting cells, including B cells, but was also found on podocytes. Stimulation of B cells via CD40 leads to activation and differentiation into memory B cells and plasma cells. In FSGS patients, CD40 was found to be expressed on cultured podocytes and circulating anti-CD40 IgG were identified in the serum of FSGS patients [60]. This gives rise to speculation that autoantibodies play a role in INS.

Podocytes

Apart from direct dysfunction of B or T cells, changes in the function of the podocytes themselves linked to the immune system are speculated to be involved in the disease pathogenesis [47]. As an example, altered expression of the diseased podocyte of CD80 (B7-1), a T cell co-stimulatory molecule, was speculated to contribute to the pathogenesis of proteinuria by disrupting the glomerular filter and discussed as a novel molecular target for treatment of INS [61]. Further studies are needed to confirm these theories.

In summary, over the years, a large body of evidence has emerged supporting the theory that the immune system plays a pathogenic role in disease development. However, the exact pathophysiological mechanism remains unknown.

Genetics of steroid sensitive nephrotic syndrome

To date, only a limited amount of data on the genetics of SSNS is available, but there is epidemiological evidence supporting that SSNS might be associated with genetic variants in one or multiple genes. This is based on the following observations.
Firstly, there is evidence of familial aggregation in SSNS as children with common backgrounds (*e.g.* siblings) have a higher risk of developing SSNS [62,63]. This can be secondary to genetic factors but may also represent a common environmental

exposure (*e.g.* lifestyle, diet). Secondly, there are differences in the prevalence of SSNS between different ethnical groups. In a small epidemiologic study of children with SSNS in the UK, it was shown that the incidence of MCD among Asians was higher compared to non-Asian children [64]. Differences in the course of SSNS in regards to relapses and steroid dependency has been observed between patients with Caucasian and African American or Hispanic background [65]. However, although different in many genomic characteristics, including allele frequencies, different ethnic groups also differ in environmental factors, lifestyle, and culture. Therefore, we can hypothesise that the susceptibly to develop SSNS is based on specific genetic risk variants, either alone or in combination with environmental triggers.

For (familial) SRNS multiple monogenic mutations have been described as causal for the disease [66]. For SSNS, despite several studies on families with multiple generations affected by SSNS, no single gene has been confirmed to cause the disease exclusively [63,62,67]. Additionally, reported possible monogenic causes of SSNS fail to explain the apparent contribution of immune dysregulation to the disease [68].

The reason why none of these studies has been conclusive could be that SSNS follows a complex inheritance pattern. Possibly, variation in multiple genes and the interaction between those and with the environment is relevant for disease development. There are a handful studies investigating complex inheritance pattern of SSNS (Table 1). These focus on determining the role of variants in Human Leucocyte Antigen genes and *loci* as genetic risk factors for SSNS. An overview of candidate genes identified until now is provided in Table 1.

Table 1 *Overview of risk loci identified for SSNS*

Study population	Number of cases	Gene	References
UK Caucasian	40	*HLA-DR7* *HLA-DQW2*	Clark *et al* 1990 [69]
US Caucasian	32	*HLA-DQW2*	Lagueruela *et al* 1990 [70]
French and German	161	*HLA-DR7* *HLA-DQB* *HLA-DQA*	Konrad *et al* 1995 [71]
Japanese	30	*HLA-DQB1*	Kobayashi *et al* 1995 [72]
Taiwanese	59	*HLA-DQB1* *HLA-DR*	Huang *et al* 2009 [73]
South Asia	76	*HLA-DRB1* *HLA-DQB1*	Ramanathan *et al* 2015 [74]
[1] South Asia USA white	214 100	*HLA-DQA1* *PLCG2*	Gbadegesin *et al* 2015 [75]
[2] European, African, Maghrebian	385	*HLA-DQA1* *HLA-DQB1* *HLA-DRB1* *BTNL2*	Debiec *et al* 2018 [76]
[3]Japanese Replication	224 216	*HLA-DRB1* *HLA-DQB1*	Jia *et al* 2018 [77]

[1]Exome association study
[2]Transethnic meta-analysis
[3]Genome-wide association study

Only recently, an exome association study described an association of the disease with *HLA-DQA1,* a MHC class II gene, and *PLCG2*, a gene involved with the adaptive immunity [75]. During the preparation of this thesis two genome wide association studies on SSNS were published [77,76]. The first included 4 cohorts with European, African and Maghrebian patients followed by a transethnic meta-analysis of the results. Results showed an association of three risk alleles, *HLA-DQA1, HLA-DQB1, HLA-DRB1* and *BTNL2*, with the disease, supporting the idea that autoimmunity plays a role in the development of the disease [76]. The second was performed in a Japanese population likewise revealing an association with the HLA-DR/HLA-DQ region [77].

The common theme in these studies is the identification of risk *loci* in the HLA region, highlighting the importance of the human leukocyte antigens, which are critical for the immune system to distinguish between self and foreign [78]. Yet, it is the identification of risk *loci* outside the HLA *locus* that arguably provide the most informative mechanistic insights. Prominent examples from nephrology include membranous nephropathy (MN) and IgA nephropathy: in MN, a GWAS identified a *locus* over

PLA2R1, acting as an antigen in the kidney, thus suggesting a genetic predisposition to antibody formation against the PLA2R1 receptor as a crucial disease mechanism [79,80]. Similarly, in IgA nephropathy, GWAS have highlighted the important role of the intestinal immune response, as well as IGA1 antibody glycosylation in the pathogenesis of the disease [81-83].

However, for SSNS further insight in the genetic architecture remains elusive.

Chapter 3. Genetics and Genome wide association studies

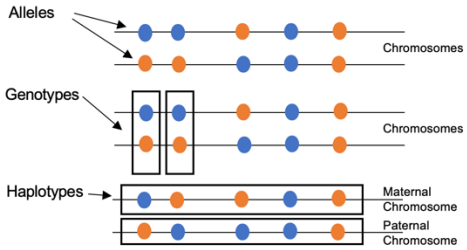
Genetics is the study of heredity and the variation of inherited characteristics. The first written proof of genetics dates back to Hippocrates and Aristotle (460-300 BC) who recognized that characteristics were inherited over generations.

The current understanding of genetics is based on milestones set by Gregor Mendel in 1866 who studied the inheritance of traits in pea plants. The term genetics was introduced by the British biologist William Bateson in 1905 and in 1911 Thomas Morgan proposed that genes are on chromosomes. In 1928, Frederick Griffith discovered that genetic material can transform living bacteria. In 1944, DNA was proven to be the molecule responsible for this transformation and in 1953, James Watson and Francis Crick set a milestone by determining the structure of DNA. Since then technologies have evolved and the molecular understanding of inheritance has increased dramatically. Nowadays, most genetic research aims to better understand the pathophysiology and mechanism of a disease. This led to the development of new treatment strategies and of preventive measurements.

In box 1 a short overview is given of the current understanding of relevant terms in genetics [84,85].

Box 1 *Definitions of commonly used terms [84,85]*

- *Locus*: a specific position on a chromosome
- Variant: a change in the DNA sequence compared to a reference DNA sequence
- Allele: different forms of a variant at one given *locus*
- Genotype: the particular complement(s) of alleles at one or more specific *loci* (one pair per *locus* in the case of diploid organisms)
- Haplotype: a series of alleles that are inherited together from one single parent



- Phenotype: the biochemical, physiological, and physical characteristics resulting from the interaction of genotype with environment
- Polymorphism: variant, which occurs in more than 1% in a certain population
- SNV (single-nucleotide variant): change of a single nucleotide (A, T, G or C) in a DNA sequence
- SNP (single-nucleotide polymorphism): SNV with a frequency of >1% in the population
- INDELS (insertion/deletions): insertion or deletion of small number of base pairs, usually less than one kilobase [86]
- CNV (copy-number variant): a structural variant that results in gain or loss of a one kilobase or larger DNA segments [87]
- MAF (minor allele frequency): the frequency of the less common allele in a certain population
- Sequencing: process of determining the exact DNA sequence of a specific DNA stretch of an individual
- Genotyping: process of determining sites of known genetic variants in the DNA of an individual
- Risk alleles: a genetic variant that, when present, increases a person's risk for a specific condition, but is not sufficient to cause disease
- SNP array: an array in which patient genotypes are determined by hybridizing the patient's DNA to DNA probes corresponding to hundreds of thousands to millions of SNPs

Mutation and variants

The process that introduces genetic variation in eukaryotes is mutations. Mutation describes any change to a given DNA sequence. Mutations can arise from many different sources and can occur either in somatic cells or in gametes. If the mutations are in germline cells, they can be passed on to the next generation and subsequently to further generations. Mutations that are passed on over a generation are referred to as variants. It is estimated that two human genomes differ at approximately 20 million bases (0.6%) and over 320 million variants in humans are described. The most common form of those variants are single nucleotide variants, where a single nucleotide is replaced by another. Importantly, the majority of variants is in non-coding areas of the genome and most do not recognisably alter the phenotype of an individual.

Genetic recombination and linkage disequilibrium

Genetic recombination forms the basis of the controlled reshuffling of genetic variation in humans and all other eukaryotes. Genetic recombination is based on the exchange of genetic material between maternal and paternal chromosomes during the creation of haploid gametes (sperm and egg cells). Haploid gametes are formed during meiosis, a particular type of cell division. Meiosis starts with DNA replication forming two sets of chromosome pairs. Each pair has a maternal and a paternal copy of the chromosome and the two copies in a pair are called homologues. This is followed by crossing over, where the two homologous chromosomes are aligned, the DNA of the chromosomes is cut at random places and DNA fragments are exchanged between the same place in the paired chromosomes. This process is also called homologous recombination. During this process alleles are put together in a new combination on the chromosomes and hence a genetically unique set of chromosomes which are a mix of paternal and maternal DNA is created.

The likelihood that crossover happens between two *loci*, and consequently they are not inherited together, depends on the physical distance between them. A short distance between two *loci* means that the “target” for cross over between them is very small and therefore unlikely. Whereas a large distance between two *loci* will increase the chance that crossover is happening between them. A way to measure the distance

between two *loci* is to calculate the frequency of recombination events between the *loci*. The measure for the genetic distance between two *loci* depending on the recombination frequency between them is expressed as centimorgan (cm).The physical distance between two *loci* is measured in base pairs (bp).

Crossover is not completely random and there are areas where crossover is more frequent compared to others.

The human genome is structured into genetic regions with a historically low recombination rate separated by regions with a high rate of recombination (hotspots of recombination). The regions with a low recombination rate contain alleles which are more likely to be inherited together; therefore, they are also called linked to each other or in linkage disequilibrium (LD). Whereas alleles that are further away from each other are more likely to have a recombination hotspot between them and therefore are less tightly linked. However, two alleles with a close physical distance but in a hotspot of recombination may still have a lower LD than if they are further apart in a region of low recombination.

A group of alleles that are inherited together because they are in LD with each other is called a haplotype. Regions with a low recombination rate will have, at population level, a distinct number of such haplotypes and are also called haplotype blocks, or in short haploblocks [88].

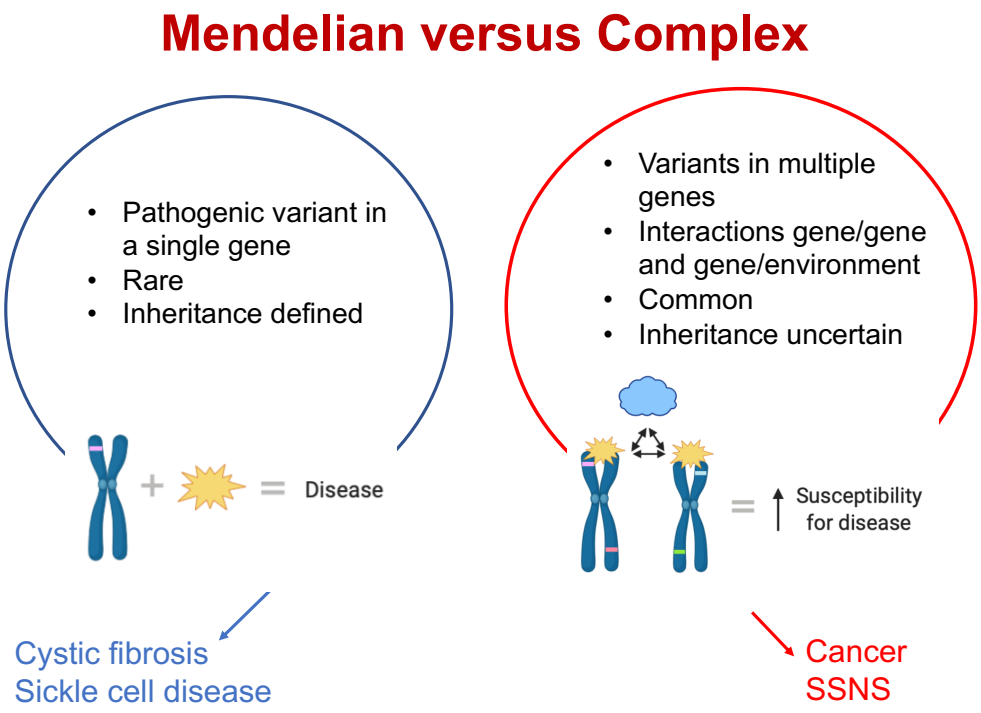
Monogenic versus polygenic disease

Monogenic diseases, also called Mendelian diseases are caused by variants in a single gene. They are generally inherited according to Mendel's Laws and follow either an autosomal recessive, autosomal dominant or X-linked inheritance pattern. Autosomal dominant and X-linked mutations do usually have a strong family history, but mutations can also occur spontaneously in the absence of positive family history (*de novo* mutations). Mendelian traits are generally characterized by a strong genotype-phenotype correlation and they can usually be studied with single family studies.

In contrast to monogenic diseases stand oligogenic diseases, where variants in a few genes are accountable for disease development. Oligogenic diseases represent an intermediate form or a bridge to polygenic and complex diseases where multiple genes are affected.

In complex diseases, additional interactions between genes and between genes and the environment are relevant for disease development (Figure 2). There might be multiple risk alleles with each of them having only a small effect on the patients' risk for developing the disease. They usually show a weak genotype-phenotype correlation. All these factors make the study of complex disease more challenging and single-family studies are often not sufficient to identify relevant variants.

Figure 2 *Mendelian versus complex disease*



Legend *Mendelian diseases are caused by variants in a single gene. They are usually rare and inherited according to Mendel's Laws. In complex diseases, multiple risk alleles, and the interactions between those alleles and between the alleles and the environment are leading to an increased susceptibility for disease.*

The approaches to study Mendelian disease differ from those investigating complex disease. We will shortly describe the following possible techniques: candidate gene testing, linkage analysis and genome wide association studies.

Candidate gene testing

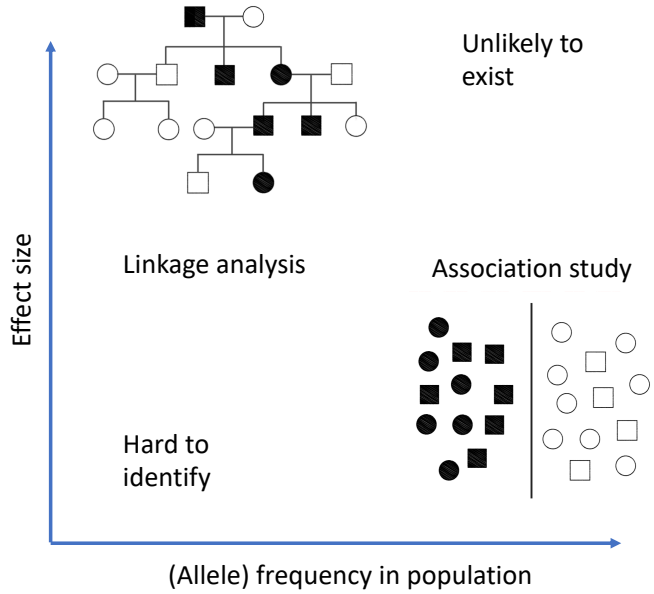
Candidate-gene studies are based on an *a priori* hypothesis of which genes might be involved in the disease development and are usually performed on a population level [89]. Few candidate gene studies have been performed in SSNS, mainly attempting to find variants accounting for the differences in response to steroid or other immunosuppressive treatment [68,90,91]. However, findings of those studies have not been replicated when followed up in subsequent association studies. Because of the uncertainty regarding the number and action of genes involved in SSNS, a more broad and unbiased approach, which uses markers throughout the genome seems more appropriate. The two main approaches investigating the whole genome are linkage studies (using family pedigrees) and association studies (using population data).

Linkage analysis

Linkage analysis is the statistical method for mapping genes responsible for heritable traits to their location on the chromosome. The process involves that markers are genotyped across the genome and tested in pedigrees if they are linked to the trait. This is based on the assumption that some markers of known chromosomal location will be co-inherited with the trait of interest. The markers which have the strongest statistical evidence of linkage to the trait, point towards the location where the gene responsible for the trait is located. In general, many genes will cluster within the linked region and therefore the resolution at the *locus* might be poor. Still, a statistically significant linkage result limits the search for the responsible gene to those in the linked region, thus reducing cost and follow-up time.

Linkage analysis has mainly proven successful in the identification of genes causal for monogenic disease following a Mendelian pattern [92,93]. For complex disease without a clear Mendelian inheritance linkage analysis has been less effective. In complex disease, where each variant has a small effect size, association analyses seem more powerful and a genome-wide association study is the preferred tool (Figure 3) [92,93].

Figure 3 *Linkage analysis versus association studies*



Legend *Linkage analysis is the preferred method to identify rare variants causing Mendelian disease. Linkage analysis is a family-based approach which tests for the co-segregation of alleles within family members. Association analysis is the preferred method to investigate common variants implicated in complex (common) disease and is a population-based approach where allele frequency is compared between unrelated cases and controls. Rare variants of small effect are hard to identify and common variants with a high effect size are unlikely to exist.*

Genome wide association study

A genome wide association study (GWAS) is a hypothesis free approach performed on population level, where SNPs distributed over the whole genome, are used as markers to tag risk *loci* in the genome. In a case-control design every single marker is tested if the frequency of alleles varies between cases and unaffected controls suggesting a relation (positive or negative) of the allele itself or a nearby variant to the susceptibility for the disease (Figure 4).

Figure 4 *Case-control design of association studies*

Legend *A difference in the allele frequencies between cases and controls can be observed. The allele T is more common in cases compared to controls.*

It is important to keep in mind that association does not imply causation of the disease. It may be that the identified SNP is associated with other factors associated with the disease, confounders, but not involved in the causal pathway. Such possible confounders could be *e.g.* the ethnic ancestry. If confounders are ruled out and the association of the SNP with the disease is thought to be true, it is important to keep in mind, that the identified SNP very rarely is the actual causative variant. Most of the identified SNPs are not disease causing themselves but are in LD with a causative variant and therefore an association is found.

The underlying rational why GWAS is currently the favourable approach to study complex disease is the 'common disease, common variant (CDCV)' hypothesis. This hypothesis refers to the idea that common diseases are caused by the combination of common genetic variations with a small effect size each, which can be detected in association studies [94]. Although SSNS represents a rare disease we believe its inheritance pattern is complex. Hence, GWAS using common SNPs to map common risk variants as well as rare risk variants contained in common haplotypes, is expected to identify variants associated with SSNS.

However, most variants found by current GWAS alone or in combination explain only a small proportion of the heritability of complex disease. This gap is referred to as the 'missing heritability' problem [95]. There are different explanations why so little heritability is explained by GWAS findings. These include the theory that an even

larger number of variants with a small effect have yet to be found. Another hypothesis is that multiple rare variants, each with a relatively large effect size, are the major contributors to genetic susceptibility to complex diseases. This model is also referred to as the 'Common Disease, Rare Variant (CDRV)' hypothesis [96]. However, most common genotyping arrays focus on common variants and only poorly detect rare variants. Further, an association with rare variants is difficult to detect via GWAS, unless the effect size is large. Different methods accounting for the overall "mutational load" would be advantageous to detect rare variants. Moreover, structural variants are poorly captured by existing genotyping arrays and might not be represented in most GWAS and GWAS do not account for gene–gene interactions, which may play a substantial role in complex disease. Nonetheless, despite those disadvantages, GWAS is the best established tool to investigate complex diseases (Figure 3) [97].

GWAS is followed by a fine-mapping process which refers to the steps undertaken to analyse the genomic region associated with a disease in order to identify the particular genetic variant responsible for the trait. Fine mapping usually contains the following steps (reviewed in Schaid et al. [98]): Each associated region is investigated in regards to linkage disequilibrium between SNPs using Haploview plots or LocusZoom plots. LocusZoom plots additionally annotate the genes in the investigated region and illustrate the patterns of LD between the lead SNP and the surrounding SNPs. Further fine-mapping can be performed in each region with different statistical methods (e.g. Bayesian methods). Another way to increase fine-mapping resolution is by combining the results of different cohorts, e.g. in a Transethnic Meta-analysis of the results, as performed in our study. The SNPs selected from fine-mapping are then annotated using genomic databases to identify the likely function of the selected SNPs. However, fine-mapping is becoming progressively challenging and variants identified do often explain just a small proportion of the inheritability of the disease. It is becoming increasingly clear that complex traits are highly polygenic, with a large number of variants, regulatory mechanisms and gene–gene or gene–environment interaction involved, that challenges fine-mapping procedures.

Chapter 4. Genotyping and allele encoding

Genotyping for GWAS

In the context of GWAS, genotype describes which two alleles can be found at a specific location. In most cases these are single nucleotide polymorphism (SNP), where one single nucleotide is replaced by another. The two possible alleles at each SNP are usually referred to as minor allele a and major allele A and could be *e.g.* the nucleotide T (allele A) or the nucleotide C (allele a). The genotype of this person at this specific SNP can be either AA, Aa or aa. If the two alleles are identical, it is called a homozygous genotype. If they differ, it is called a heterozygous genotype. Association studies are based on comparing the frequency of the two possible alleles for each given SNP between cases and controls. The process of determining sites of known genetic variants in the DNA is called genotyping. Those variants are in the downstream analysis used to mark areas of association and therefore often referred to as markers. The whole-genome genotyping technology allows the identification of markers catalogued across the whole genome. Depending on the SNP chip used, it can interrogate several million markers per sample [99]. It can detect single nucleotide polymorphism, but also other variants.

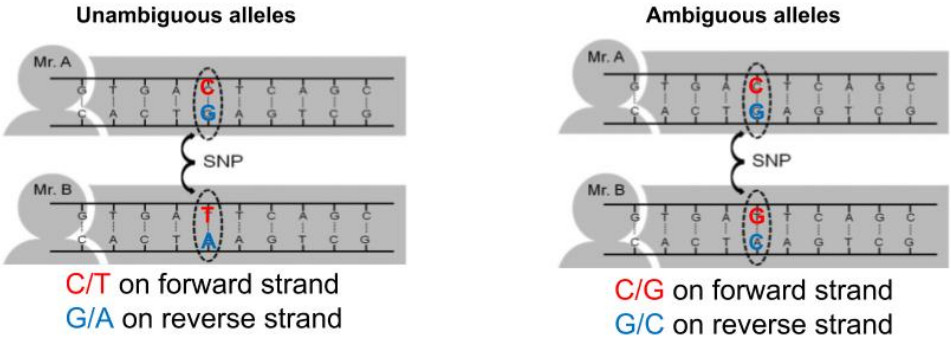
Allele naming and encoding

Association studies are based on comparing the frequency of two possible alleles for each given SNP between cases and controls. In order to get adequate results, the naming and encoding of alleles has to be understood and clearly defined, otherwise it is not possible to compare them. The naming and encoding of alleles is based on the reference genome. The reference genome is an agreed sequence for an organism that generally represents the most common sequences in the global population and is managed by the Genome Reference Consortium (https://www.ncbi.nlm.nih.gov/grc). Different versions of the reference genome have been released, in the form of different genome builds, with the latest at the time of writing this thesis being GRHc38 (March 2017), although for this study GRHc37/hg19 was used (Feb 2009). The human reference genome can be accessed via the National Centre for Biotechnology Information (NCBI) database.

Any documented alterations to the reference genome are called variants. Those variants are recorded in a different database called dbSNP (single nucleotide polymorphism) [100]. DbSNP mainly contains single nucleotide polymorphisms, but can also include other kind of variants. The reference allele for a SNP is defined as the allele found on the reference genome sequence. The alternate alleles are the variants seen and consequently documented and submitted to dbSNP. Very often the reference allele is also the major allele (meaning the more frequent allele) in one population whereas the alternate allele is also the minor allele (meaning the less frequent allele) in the same population. However, because frequencies vary between populations, it could be that the major allele in one population is the minor in another.

The definition of which nucleotide of a certain SNP is the reference and which the alternate allele is depending on the strand (5' to 3' or 3' to 5'). In some SNPs the reference and alternate alleles are a complement to each. A C on the 5' to 3' strand would be a G on its complementary reverse strand. This can lead to ambiguity if the strand is not defined. *E.g.* when referring to the allele C in a SNP where the two options are C/G and the strand assignment is not clear, it is impossible to tell whether the allele C is the reference or the alternate allele (Figure 5).

Figure 5 *Unambiguous versus ambiguous alleles*



Legend *Unambiguous alleles can be uniquely identified without knowledge of strand assignment. For ambiguous alleles the strand assignment is essential in order to uniquely identify allele A and allele a.*

Therefore, it is crucial to know, to which strand the allele name is referring to. In order to deal with that problem, different schemes to encode alleles have been developed.

Encoding schemes

- **Plus/Minus (+/-)**: This encoding scheme is used by the HapMap project (www.hapmap.org). The 5′ end of the + strand is at the tip of the short arm (p-arm) of the chromosome and the 5′ end of the - strand is at the tip of the long arm (q-arm). +/- encoding for a particular SNP may change with the genome build used, hence the genome build must be specified when reporting +/- strands. The alleles are usually named with the GATC nucleotide letters.

- **Forward/Reverse (FWD/REV**): This is the encoding scheme used by dbSNP. Usually dbSNP receives a cluster of submissions per SNP. Those are compared to the reference genome to define their position and orientation and can be either on the forward or reverse strand. The one with the longest flanking sequence is chosen to give the orientation of the RefSNP. Hence, RefSNP are either on the forward or reverse strand in relation to the genome build used, which again can vary between different versions of dbSNP. The alleles are usually named with the GATC nucleotide letters.

- **Illumina TOP/BOT**: Both of the above encoding schemes are dependent on the genome build used and can vary with different versions. Illumina therefore developed its own strand designation scheme which they use internally [101]. This encoding scheme is aimed to solve the genome build problem and is genome build independent**.** Strand designation is defined by taking the flanking probe sequence around the variant into account using TOP and BOT strand. For unambiguous SNPs (A/C or T/G), A and B allele on the TOP strand denote A and T (or C and G, respectively); whereas for BOT strand, A and B allele denote T and A (or G and C, respectively). For ambiguous SNPs (A/T or C/G) the surrounding sequence is taken into account for strand and allele definition. If A or T is on 5' side of the SNP, then it is a TOP strand otherwise it is a BOT strand [101]. The alleles can be encoded with the GATC nucleotide letters or as allele A and B.

- **DESIGN:** In the case of Illumina microarrays, each SNP chip probe is designed against a certain genome build and position. The DESIGN scheme indicates the alleles in relation to the strand the probe was designed against.

Often association studies include data obtained by different groups, which have been generated in different centres and therefore are encoded according to different schemes. That said, for comparing data of different groups it is crucial that the genotyping data is encoded via the same scheme and in relation to the same strand. Therefore, the identification of the encoding of the datasets and the conversion to a common encoding scheme is essential prior to any downstream analysis.

Chapter 5. Imputation

The ideal situation for association studies would be to compare the highest number of genotyped variants between cases and controls. However, the number of genotyped variants is restrained by the SNPchip used and even modern chips with > 1 million variants only cover a fraction of all genetic variants. To overcome missing study data, genotype imputation has been developed as a concept of filling in missing genotypes by putting information into a context.
This can be demonstrated with a “hangman” example:

?at
The dog chases the ?at

Putting missing information in a context, can provide information on the missing data. In genotype imputation, the genotyped data of study samples is compared to a reference panel of haplotypes. This reference panel provides genotype information on a much larger number of markers. Shared haplotype stretches are identified and missing genotypes within a haplotype stretch in the study samples are filled in by the alleles observed in the matching reference haplotype. However, it is not always categorical which haplotype should be used to fill in missing genotypes for a particular sample. For example, in 60% of reference haplotypes genotype A/A was observed at a specific site, whereas in the remaining 40% a different genotype A/a was observed. Imputation output is therefore a probabilistic output. Based on the observed sample data and on the reference data, the genotype with the highest probability is outputted. This can be illustrated when solving our hangman example:

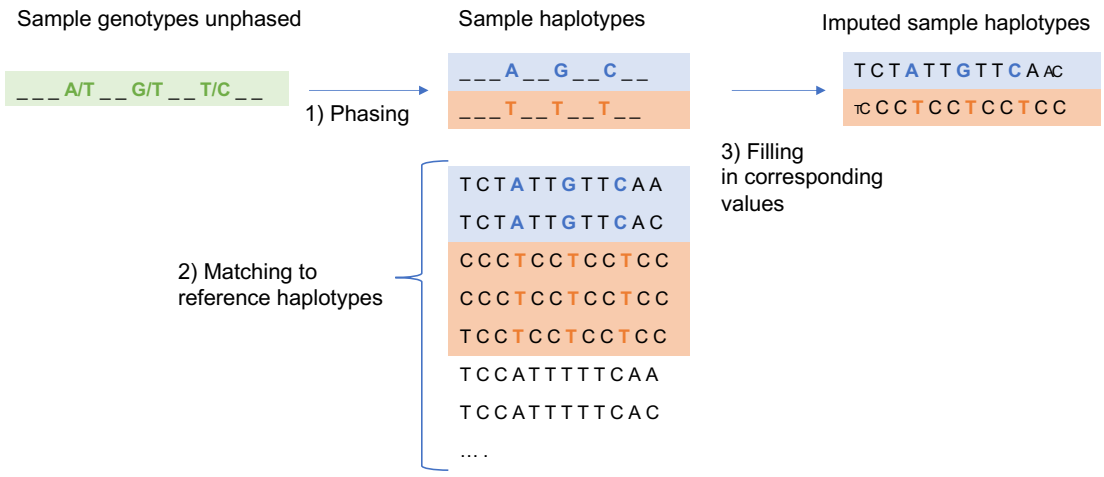
?at
The dog chases the cat – in 97% the ? would stand for a c
The dog chases the rat – in 3% the ? would stand for a r

Models of imputation

Genotype imputation methods were first developed in 2007 by a group from Oxford and the Wellcome Trust Case Control Consortium. The standard model for imputation is based on Hidden Markov Model (HMM) [102,103]:
"A class of statistical model that can be used to relate an observed process across the genome to an underlying, unobserved process of interest." [104]
The general workflow for imputation of unrelated individuals contains the following essential steps [104-106] and is graphically displayed in Figure 6:

- Strand alignment between study dataset and reference panel: The study dataset and the reference panel must be aligned to the same strand in order to provide accurate imputation results.
- Phasing of the study dataset: For each haplotype we have two copies, one on each chromosome (maternal and paternal). Genotyping identifies the two alleles that can be found on the two haplotypes, but doesn't identify which allele is on which copy of the haplotype. Phasing refers to the separation of genotype data into their chromosomal origin to identify which alleles are together on one haplotype. Therefore, genotyped data require computationally pre-phasing before they can be used for imputation. Pre-phasing means to first phase your sample genotypes and then use the estimated sample haplotypes to impute ungenotyped variants from a reference panel.
- After pre-phasing, the haplotypes of each individual in the study dataset are compared to the haplotypes in the reference panel. The reference haplotypes which match the best to each study sample haplotype are selected. The idea is that the haplotypes of the study samples are like a mosaic of haplotypes of the reference panel.
- After identifying the haplotypes with the best match in the reference panel, missing genotypes in the study sample are imputed by copying the genotypes from the matching haplotypes.
- Further analysis of SNPs with imputed genotypes is similar to those genotyped.

Figure 6 *Important steps of imputation process*



Legend *Shown is the process of genotype imputation.*
1) Haplotypes are detected in the unphased genotype sample.
2) Those sample haplotypes are matched to reference haplotypes. The best matching haplotypes are identified (indicated by the same colouring as the sample haplotype).
3) Variants are imputed by filling in the genotypes from the matching reference haplotypes in the sample haplotypes. Note that there is generally more than one matching reference haplotype. Therefore, the corresponding values are filled in probabilistically. E.g., for the first loci of the second imputed sample haplotype this is 1/3 for a T and 2/3 for a C. One can convert this to a 'best guess' genotype, which would be a C.

Imputation accuracy

The whole imputation process is based on probabilities, providing an output with the most probable genotype at each SNP.

For Example:

Allele 1: P(A) = 0.98, P(a) = 0.02

Allele 2: P(A) = 0.14, P(a) = 0.86

This sample is outputted with the genotype A/a for this specific SNP.

The probability with which a genotype is imputed varies, with one genotype being imputed with a very high probability to another genotype imputed with a very low probability. Probability can also be understood as how likely the imputed genotype matches the real (observed) genotype. The probability depends on different factors, *e.g.* the density of surrounding genotyped markers, LD at this certain position, the number of samples in the reference panel. In general, the more context information is

given, and the larger the reference panel is, the higher will be the probability a genotype can be imputed accurately.

Imputation accuracy is the correlation between the real (observed) and the imputed genotype. It is important to take imputation accuracy into account when analysing imputed data [102]. One way to address imputation accuracy is by taking the allele dose into account. Allele dose gives an estimate of how accurately the alternate allele at a specific SNP was imputed. A value of 0.97 means that out of all the haplotypes with an alternate allele at this site, 97% of them are imputed accurately to the alternate allele, if this site was assumed to be not genotyped. The closer the value is to 1.0, the more accurately that site has been imputed.

The accuracy has to be interpreted in conjunction with the minor allele frequency. *E.g* in a data set with 1000 samples, different minor allele frequencies affect the accuracy as follows: If a marker has a minor allele frequency of 40%, and 390 of 1000 are imputed correctly to the minor allele, whereas 10 of 1000 are imputed incorrectly to the major allele, the imputation accuracy is 99%. If the minor allele frequency of this marker is only 0.1%, and no minor allele is imputed at all, the calculated accuracy would even be higher with 99.9%. However, all the information of that marker got lost. In order to use this estimate appropriately it is essential to know the minor allele frequency of each marker.

To overcome this problem most programs use the squared correlation between estimated and true allele dose [102]. The allelic R square (R2) is an imputation quality metric for each imputed marker [102]. Values range between 0 – 1 and larger values of allelic R2 indicate more accurate genotype imputation.

Chapter 6. Human Leucocyte Antigen complex

The MHC (major histocompatibility complex) is a group of genes encoding for proteins involved in the regulation of the immune system as well as other fundamental molecular and cellular processes [107]. Human leukocyte antigen (HLA) is the human version of MHC which historically was identified as a set of antigens involved in transplant rejection, hence the name “major histocompatibility”.

The human leukocyte antigen (HLA) *locus* is a genomic region on the short arm of chromosome 6p21.3 that codes for more than 220 genes. The first complete sequence and gene map of the HLA region dates back to the 1990s [108]. Since then the interest in and research into the HLA region has been extensive, because of its established role in the regulation of the immune system including regulation of inflammation, complement cascade and the innate and acquired immune system. The HLA *locus* is essentially involved in the discrimination between “self” and “non-self” and its role in autoimmunity and development of autoimmune disease has been well established [109,110].

The extended HLA *locus* spans over 7.7 Mb and is divided into subregions named extended class I, classical class I, class III, classical class II and extended class II regions from telomere to centromere. The classical HLA region spans over 3.5 Mb from *ZFP57* to *HLA-DPA3* [107]. The region has a high density of genes, containing more than 200, coding for ligands, receptors, signalling factors and regulatory factors mainly involved in the immune system.

The classical HLA region includes the genes encoding for three basic groups of molecules: HLA class I, HLA class II, and HLA class III. A main characteristic of the HLA genes is that they are highly polymorphic [111]; multiple variants of each gene are known.

Humans have three main HLA class I genes, known as *HLA-A, HLA-B*, and *HLA-C*. Those genes encode a set of structurally related, highly polymorphic proteins historically called antigens as they were identified in transplant rejection. They are present on the surface of all nucleated cells. HLA class I proteins are responsible for

the presentation of antigens from inside the cell to CD8+ cytotoxic T cells. If the immune system recognizes the presented antigen as foreign, a cascade is initiated resulting in cell destruction.

For HLA class II six main genes are known in humans: *HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA,* and *HLA-DRB1*. HLA class II genes encode proteins that are present almost exclusively on the surface of antigen-presenting immune cells, including macrophages, dendritic cells and B cells. MHC II proteins present exogenous antigens that originate from outside the cell from foreign bodies *e.g.* bacteria, to CD4+ helper T cells, leading to the release of lymphokines and initiating the destruction of the antigenic material.

HLA class III genes are less polymorphic and are involved in inflammation and other immune system activities.

The polymorphic nature of the HLA genes allows the immune system to react to a wide range of foreign invaders. By June 2019, the 8 classical HLA genes *(HLA-A,-B,-C,-DRA,-DRB1,-DPA1,-DPB1,-DQA1,* and *-DQB1)* had approximately 24,000 named alleles, encoding for more than 13,000 protein variants [112]. Specific HLA alleles have been linked to the susceptibility of a wide range of disease including different autoimmune diseases, infections and cancer. Also for SSNS, the crucial role of the immune system has been implicated and associations in the HLA region with the disease were identified in previous studies [75,76].

HLA nomenclature

The genes encoding for HLA molecules are highly polymorphic and therefore a systematic nomenclature is necessary. The naming of the HLA genes is defined in a given version of the WHO HLA Nomenclature Report [113]. Each HLA allele name has a unique number corresponding to up to four sets of digits separated by colons (Table 2).

The HLA type is represented by the digits before the first colon (*e.g. HLA-A*02*), also called 1st field. The next set of digits describes the subtype (*HLA-A*02:01*), also called

2nd field. Alleles who differ within these first two sets of digits have at least one nucleotide changed that alters the amino acid sequence of the encoded protein.
Differences in the third set of digits indicates alleles that differ only by synonymous nucleotide substitutions (silent or non-coding substitutions) within the coding sequence (*HLA-A*02:01:01*). Also called 3rd field.
Differences in the fourth set of digits describes alleles that only differ by sequence polymorphisms in the introns, or in the 5' or 3' untranslated regions that flank the exons and introns (*HLA-A*02:01:01:02*). Also called 4th field.

Table 2 *Nomenclature of HLA alleles*

Nomenclature	Meaning
HLA	the HLA region and prefix for an HLA gene
HLA-DRB1	a particular HLA *locus* i.e. DRB1
*HLA-DRB1*13*	a group of alleles that encode the DRB1*13 antigen or sequence homology to other DRB1*13 alleles
*HLA-DRB1*13:01*	a specific HLA allele
*HLA-DRB1*13:01:02*	an allele that differs by a synonymous DNA substitution within the coding region from DRB1*13:01:01
*HLA-DRB1*13:01:01:02*	an allele that differs by DNA substitution outside the coding region from DRB1*13:01:01:01

Legend *adapted from WHO Nomenclature Committee for Factors of the HLA System [113]*

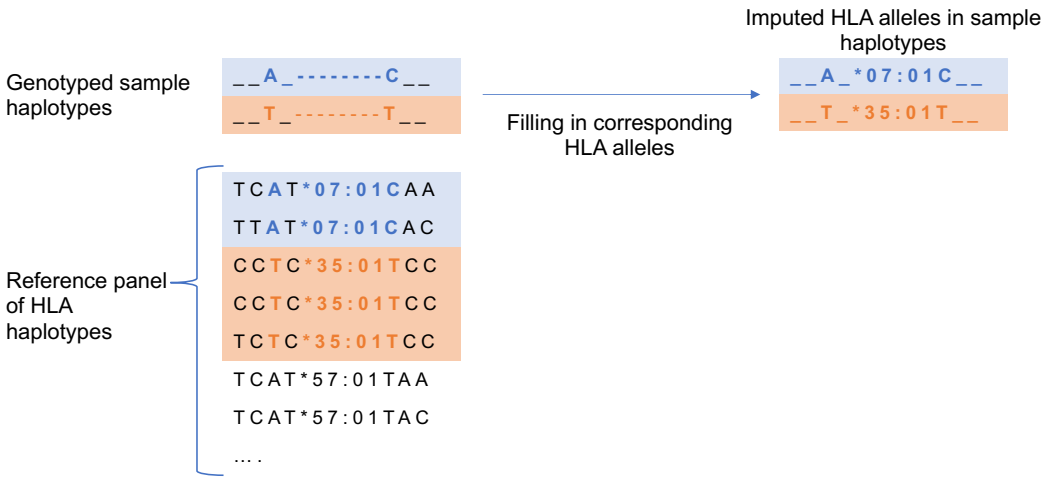
"Low-resolution typing" (antigen or allele family level) is equivalent to serologic typing, and describes resolution at the level of the first two digits or 1st field (*e.g., HLA-DRB1*02, -DRB1*03, -DRB1*04*) [113].
"High-resolution typing" describes the resolution down to the four-digit level or protein level (2nd field) (*e.g. HLA-DRB1*02:01*) [113].
“Allelic resolution” is typing down to a single allele as defined in a given version of the WHO HLA Nomenclature Report [113].

Human Leukocyte Antigen imputation

Direct sequencing of classical HLA alleles is expensive and is not feasible in most of the current association studies. The HLA region itself is characterized by high linkage disequilibrium and earlier studies demonstrated that specific SNPs can be in strong LD to specific HLA alleles [114]. Consequently, a limited number of SNPs can be used

to tag the majority of HLA alleles and imputation of HLA alleles from SNP-level data has become a widespread method [115]. This process is known as HLA imputation and follows the same principles as genotype imputation (Figure 7).

Figure 7 *HLA imputation scheme*



Legend *Shown is the imputation of classical HLA alleles in a genotyped sample set. The reference panel contains genotyped SNPs in the HLA region tagging HLA alleles. The best matching reference haplotypes are identified and HLA alleles are imputed by filling in the HLA alleles from the matching reference haplotypes in the sample haplotypes.*

Different software tools have been developed to perform HLA imputation from SNP-level data [116]. The downstream analysis is identical to that of genotyped SNPs.

Chapter 7. Replication and Meta-analysis

Replication

The gold standard for GWAS findings to be considered as valid is to replicate the findings in an independent cohort (replication cohort). This means to test if an identified allele/variant in one cohort is also associated with the disease in another cohort. This is important to overcome concerns about false positive signals *e.g.* deriving from hidden population stratification effects.

In most of the association studies published to date the discovery cohort is of European ancestry. Reasons for that were the availability of data as well as increased funding opportunities in European countries (larger groups and well-established scientists). Also, technological difficulties in non-European populations, *e.g.* requirement of different SNP chips and markers to account for varying allele frequencies between different population, led to a bias towards European studies.

Using a replication cohort coming from a non-European background can have advantages besides merely replicating the association findings [117]. Association studies detect markers which are in LD with the causative variant and very seldomly they directly detect the causal variant. In a single ethnicity population this association can span over a larger region representing a haploblock. Often it is not clear which of the variant within a haploblock has the strongest association with the phenotype or is possibly causal for the trait. Differences in LD structure across different populations can help narrowing down this region of interest and consequently help in identifying the causal variant. Looking at the overlap of haploblocks associated with the trait in different populations can help to dissect the causal variant form non-causal variants.

On the other hand, using a replication cohort from a different ethnicity, can also lead to difficulties [117]. Differences in allele frequencies between populations can pose a challenge when aiming to replicate findings [118]. A given variant detected in a European ancestry GWAS may be polymorphic or monomorphic in a replication cohort of different ancestry and consequently the risk allele cannot be directly replicated. Also, the prevalence and incidence rates of a disease and trait can vary considerably

between different ethnic populations not only secondary to differences in the genetic architecture, but also because of differences in environmental, lifestyle, and cultural characteristics or the combination of both [118]. Nevertheless, many well-established SNPs have been replicated in transethnic studies [117].

Meta-analysis of GWAS

Transethnic meta-analyses are combining GWAS results of the same trait across genetically diverse populations and can offer a lot of possibilities [76]. Comparisons of GWAS findings across different populations have revealed that the direction of effect of the associated alleles (protective or deleterious) on the trait is often consistent across ethnically different populations. Transethnic meta-analyses are taking advantage of that and are investigating the direction of effect of the risk alleles across different ethnical cohorts [119]. This can be useful in the situation of small sample sizes where the power of the study is limited and the risk allele might not reach the level of significance at $P<5\times10^{-8}$. Rather than replicating the signals at genome wide significance in each cohort separately, combining the results of the separate studies can show the same direction of effect and consequently increases the overall significance across the different ethnic studies [76]. Therefore, in case of a low powered GWAS, a transethnic meta-analysis can detect a common direction of effect for risk alleles across different populations and add more strength to the study.

Chapter 8. Hypothesis and aims

Hypothesis

The hypothesis of this thesis was that there is a genetic susceptibility for steroid sensitive nephrotic syndrome. Previous studies had shown that SSNS is a complex disease with involvement of the immune system. We hypothesised that besides the known association of SSNS and the HLA region, variants outside the HLA region are implicated in the disease development. We further hypothesised that the identification of these variants outside the HLA region, will ultimately increase the understanding of the involvement of the immune system in the disease and probably guide towards an antigen relevant for the disease.

We decided to use a hypothesis free approach with a population-based study. We hence intended to perform a genome wide association study with the aim to first confirm previous findings on the association of SSNS with HLA *loci* and secondly to identify *loci* outside the HLA region associated with SSNS. We aimed to gain fundamental mechanistic insights into the disease aetiology with the identification of such risk *loci*.

Steps

The thesis was built on following steps:

a) To collect and genotype cases with SSNS
b) To find relevant control datasets for the performance of a genome wide association study on SSNS
c) To optimize quality control steps in the process of performing a genome wide association study
d) To perform association testing to identify risk *loci* associated with the disease
e) To understand and perform imputation in order to increase the density of markers and then repeat the association testing on the imputed dataset
f) To understand and perform HLA-imputation and perform an association test on the imputed HLA alleles
g) To replicate the findings in a different ethnical cohort

h) To perform a meta-analysis of the results in order to increase the significance of findings
i) To investigate the regions of association further and interrogate potential candidate genes
j) To develop a hypothesis for the pathophysiological mechanisms implicated with these candidate genes
k) To develop ideas of how to investigate the suggested pathophysiological mechanisms behind the candidate genes as part of future directions

Part 2: GWAS Methods

Chapter 1. Genome wide association study

Genome wide association studies examine if any genetic variant across the whole genome is associated with the disease, hence the word genome wide association study (GWAS). Tests for GWAS are based on using SNPs. The aim is to identify SNPs where one allele is significantly more common in cases than controls. Statistical methods to test for association of those SNPs with the disease are the basic allele test or logistic regression.

Basic allele testing

A basic allele test (BAT) is based on a 2x2 contingency table which represents the counts for each SNP for the minor allele a and major allele A in cases and controls (Table 3). The test for association is performed for each SNP separately. Under the null hypothesis that there is no association between the allele frequency and the disease, we expect the allele frequency to be the same in cases and controls [120].

Table 3 *2x2 Contingency table*

	Allele A	Allele a	Row total
Cases	a	b	a+b
Controls	c	d	c+d
Column total	a+c	b+d	Total n

Legend *The table contains the counts of the two possible alleles per SNP, allele A and allele a, for cases and controls*

A test of association is calculated by a Chi-square test (χ^2 test) for independence of the rows and columns of the contingency table. The test calculates, if the observed frequency of allele A and allele a differs from the expected frequency of allele A and allele a between cases and controls [120]:

$$\chi^2 = \Sigma [(O - E)^2 / E]$$

Where:

O = Observed frequency

E = Expected frequency

∑ = Summation

χ^2 = Chi-square value

The expected frequency is defined as:

$$E = N / n$$

Where:

N = Number of observations

n = Total n

e.g., referring to the Table 3, the expected frequency of cases having allele A is calculated as:

$$E = \text{row total } (a+b) \times \text{column total } (a+c) / \text{table total } (n)$$

Under the null hypothesis the observed and the expected values are close to each other and therefore O − E will be a small value. When the observed values are not close to the expected values, O – E will be a large value. The chi-square value χ^2 is thus small when the null hypothesis is true, and large when the null hypothesis is not true.

The degree of freedom (d.f.) for a chi-square test is calculated by the following formula:

$$d.f. = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

Hence, for a 2x2 contingency table the d.f. is 1. The degree of freedom together with the chosen significance threshold determines the critical value. If the observed chi-square test value χ^2 is greater than the critical value the null hypothesis is rejected.

The same principle applies when testing genotypes between cases and controls, only that a 2x3 contingency table is required (Table 4).

Table 4 *2x3 Contingency table*

	AA	Aa	aa
Cases			
Controls			

Legend *The table contains the counts of the three possible genotypes per SNP, AA, Aa and aa, for cases and controls*

In order to perform a basic allelic test on genotypes, genotypes AA, Aa, and aa are dissolved into pairs of alleles A and A, A and a, or a and a. Both alleles in a pair stemming from the same sample are connected to the same value of the dependent variable (case or control). The associations with these dissolved alleles and the dependent variable are then tested.

Logistic Regression analysis

Logistic regression analysis is used to calculate the relationship between a dependent and one or more independent variables [121]. The dependent variable is the factor which needs to be understood or predicted and the independent variables are factors that are suspected to have an impact on the dependent variable. In our study, the dependent variable is the occurrence of the disease and the independent variables are the genotypes. As the disease status is not a continuous variable (as for example weight would be) but a binary outcome (either disease yes or not), a binary logistic model was used in our study, which estimates the probability of a binary outcome (case or control) based on one or more independent variables (alleles).

The basic logistic regression model looks like:

The expected value of the phenotype $Pi = E(Yi \mid Xi)$

Yi is the phenotype for individual *i,* therefore *Yi* = 0 stands for controls and *Yi* = 1 for cases

Xi is the genotype of individual *i* at a particular SNP, with the 3 possible options

AA referred as *Xi* = 0

Aa referred as *Xi* = 1

aa referred as *Xi* = 2

Extra factors can be added to adjust for potential confounders: *e.g.* ethnicity (Ei), genotypes at other SNPs (Si) *etc.*

These factors can be included in the formula for the expected value of the phenotype.

$$Pi = E(Yi \mid Xi, Ei, Si,..)$$

The main advantage of logistic regression over the basic alleles test is that it can handle more than two independent variables simultaneously, which is important when correcting for covariates. Covariates are variables or the interaction between variables, which could affect the outcome of the analysis without necessarily being related to the disease. (A simple example would be the association between ice cream sales and bicycle accident. A higher sale of ice cream is associated with a higher number of bicycle accidents. This is not because ice cream increases the risk of having a bicycle accident, but both are associated with the covariate good weather or the season of the year. If not correcting for the factor good weather, one could wrongly assume that ice cream sales and bicycle accidents are directly related). Correcting for covariates allows to see the effects of the remaining variables on the outcome. This also enables to correct for stratification (based on Principal Component Analysis).

Recoding of genotypes to numerical variables is required before performing a regression analysis. In order to recode genotypes, the alleles were classified into major allele (A) *vs.* minor allele (a). Major allele was the more frequent one as counted in the dataset. Recoding to numerical variables followed an additive model:

aa=2, Aa=1, AA=0

P- value

Determining the correct P-value threshold for statistical significance of association studies is crucial to control the number of false positive results without greatly sacrificing true positives. Statistical significance of a test leads to rejection of the null hypothesis. The cut-off is usually set at a p-value below 0.05. This means that in 5% the null hypothesis is rejected when the null hypothesis is in fact true and a false-positive result is detected. This probability is in relation to a single statistical test; in the case of a GWAS, where hundreds of thousands of tests are conducted, each one has its own false positive probability leading to numerous false positive results. Correction for multiple testing becomes necessary. A common method to correct the p-value of 0.05 for multiple testing is the Bonferroni correction [122]. The Bonferroni correction simply divides the p-value by the number of independent tests performed. In the case of GWAS this would mean dividing p by the number of markers tested. However, this approach is conservative and would "overcorrect" for variants that are not truly independent. In genetic data many variants are in strong LD and hence not "independent".

The International HapMap Consortium estimated the ‘effective number of independent tests’ when testing all common (MAF ≥ 0.05) variants across the genome. The results showed that in a European sample set the number of independent common variants is approximately 150 for every 500 kilo base pair region [123]. Extrapolating this number to the whole genome of approx. 3.300 mega base pairs indicates 1,000,000 independent markers over the whole genome.

Hence, for genome-wide significance the p-value of 0.05 should be corrected for 1,000,000 independent tests using the Bonferroni correction:

$$p = 0.05 / 1{,}000{,}000 = 5 \times 10^{-8}; -\log_{10} p = 7.3$$

The results suggest a significance threshold of 5×10^{-8}. In general, a genome-wide significance threshold p - value of 5×10^{-8} (-log 7.3) is broadly accepted for common-variant GWAS and has also been intensively investigated and published in different

papers as in international guidelines [124-126]. For our study the same significance level of 5×10^{-8} (-log 7.3) was chosen.

Odds ratio

The odds ratio (OR) is a measurement for the strength of the association between an exposure (presence of an allele) and an outcome variable (case/control status) in a given dataset. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

The genotypic OR represents the odds for a disease in association with a specific allele. This is calculated by comparing the odds of the disease in an individual carrying the specific allele (*e.g.* allele A) to the odds of the disease in an individual not carrying the specific allele (*e.g.* allele a) [127].

Allele count	Allele A	Allele a
Cases	a	b
Controls	c	d

$$\textit{Odds that allele A occurs in cases} = \frac{a}{c}$$

$$\textit{Odds that allele a occurs in cases} = \frac{b}{d}$$

$$\textit{Odds ratio (OR)} = \frac{Odds\ that\ allele\ A\ occurs\ in\ cases}{Odds\ that\ allele\ a\ occurs\ in\ cases} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a \times d}{b \times c}$$

With each allele A being present in cases, the odds ratio increases. Therefore, an OR of 1 indicates no association between genotype and disease, an OR > 1 indicates that the allele A increases the risk of disease and an OR < 1 indicates that allele A decreases the risk of disease.

Haplotype association test

Haplotypes can be used, comparable to single SNPs, to investigate an association between haplotype or haploblock frequencies and case/control status. The advantage is that haplotypes can capture the combined effect of causal variants which are linked [128]. The problem about haplotypes is that they cannot be observed directly but have to be inferred. Haplotypes can also be tested in blocks, haploblocks. Again, where a haploblock starts and ends cannot be directly observed but is inferred indirectly through the use of algorithms. Details on haplotype and haploblock definition can be found in the introduction part on page 38.

In our study, haploblocks were defined algorithmically with standard settings. The Haplotype Block Detection algorithm is described in detail in the defining paper by Gabriel *et al* [88]. The haploblock association test was performed using these precomputed blocks.

For the association testing a Chi-squared test was used, which allows to compare the frequency of haploblocks between cases versus controls. The same correction for multiple testing (Bonferroni method) as used for the BAT was applied. A significance level of $p < 5 \times 10^{-8}$ was chosen.

Programs and software tools

SVS

All analyses, if not stated otherwise, was performed in Golden Helix SNP & Variation Suite version 8.8.1 (SVS, http://goldenhelix.com/products/SNP_Variation/index.html). SVS is an analytic program specially developed to perform multifaceted analyses and visualizations on genomic and phenotypic data. The recommended format to store and utilize data in SVS is in dsf files. The files encode matrices which can be visualized in the program. The columns are the makers and the rows represent the individuals. The data cells contain the genotype for each individual at the specific marker. The two alleles of each genotype are represented in their nucleotide form (A,T,G,C).

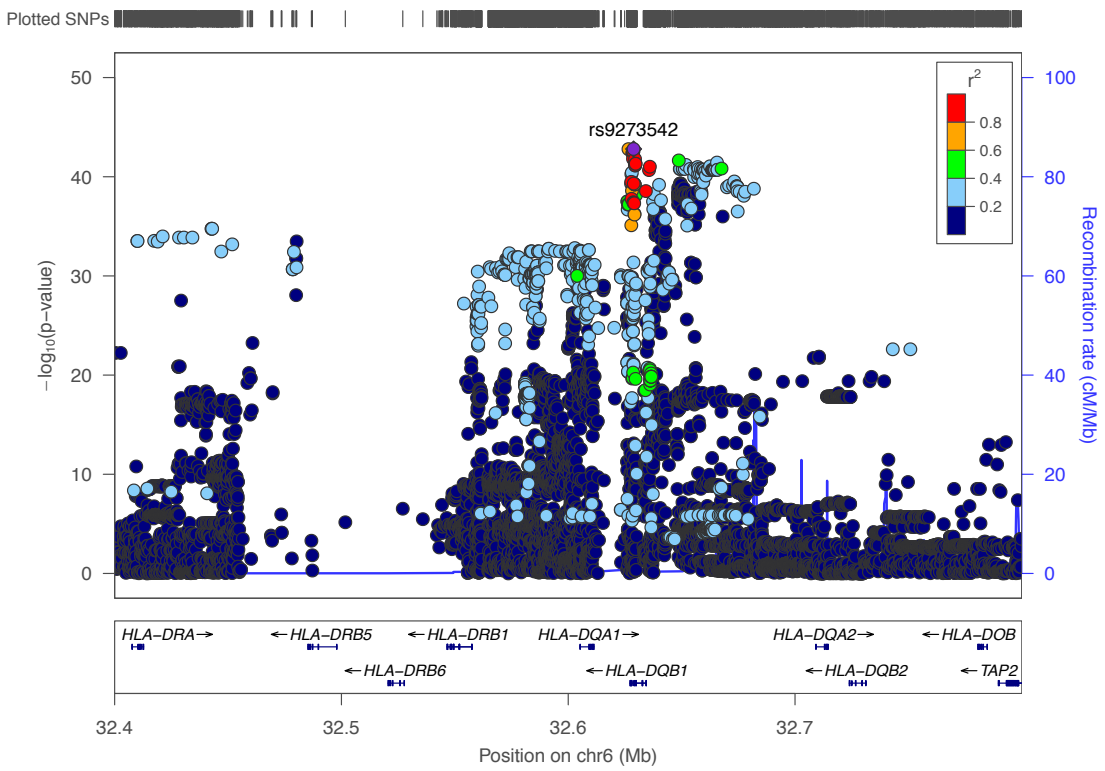
PLINK

For some analyses the software toolset PLINK v1.90 beta was used [129-131]. PLINK is a free, open-source whole genome association analysis toolset. PLINK has no graphical interface, but is a command line program. Input file formats are PED/MAP or BED/BIM/FAM. The PED file is a white-space (space or tab) delimited file and contains information about the individual, the pedigree and the genotypes. The MAP file describes data about each marker. BED/BIM/FAM are the binary version of PED/MAP files. The BED file contains the genotype information. And the BIM file the information about the marker. The FAM file the pedigree and phenotype information.

LocusZoom

LocusZoom is a bioinformatic tool to visualize results of a genome wide association study by generating a regional plot of the area of interest [132].
The software can be downloaded from LocusZoom homepage (http://locuszoom.org). LocusZoom provides useful information about the *locus* including its exact location and genes in that area (GRCh37/hg19 build). An example of a *locus* zoom plot is shown in Figure 8. On the x-axis the chromosomal region and the genes in that region are displayed. On the y-axis the -log of the p-value is indicated. Per standard settings, a purple diamond indicates the SNP with the smallest p-value (index SNP) within the region plotted. SNPs are coloured differently based on their level of LD to the index SNP. The LD is calculated from the 1000 Genomes European reference panel. Recombination hotspots are indicated by blue vertical lines [132].

Figure 8 *Example for locus zoom plot*



Legend *The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log p-value of each marker is on the y-axis. The index SNP (usually the one with the lowest p-value) is annotated with a purple diamond. The colouring of the remaining SNPs indicates the level of LD (r^2) to the index SNP.*

Power calculation

Power calculation was performed using the Michigan Genetic Association Study power calculator (https://csg.sph.umich.edu/abecasis/gas_power_calculator/index.html). As input parameter the number of cases and controls with an assumed disease prevalence of 1:10,000 and a significance level of 5×10^{-8} was used. Calculated was the minimum genetic risk score that can be detected assuming an additive model with a power of 0.8, when the allele frequency in controls is 0.1.

Chapter 2. Cases and Controls

In order to perform a successful GWAS cases and controls need to be carefully identified and it has to be ensured that data is of high quality. Details on case and control collection as well as quality control steps are provided in the following chapters.

Case cohort

Patients with SSNS enrolled in this study were identified by different collaborators (Table 5). All patients were assessed carefully and diagnosed with nephrotic syndrome as per KDIGO 2012 guidelines [8]:

- Oedema
- Protein excretion ≥40 mg/m^2/hour, or Urine Protein/Creatinine Ratio ≥ 200mg/mmol, or 3+ Protein on urine dipstick
- Hypoalbuminaemia ≤ 25g/l (≤ 2.5g/dl)

Steroid sensitive NS was defined according to KDIGO 2012 guidelines [8] if the patient achieved complete remission (urine protein/creatinine ratio < 20 mg/mmol or < 1+ of protein on urine dipstick for 3 consecutive days) within 4 weeks of corticosteroid therapy. No distinction was made between patients who had none, one or multiple relapses or were steroid dependant.

Excluded were all patients who developed nephrotic syndrome secondary to systemic diseases, malignancies, medications, and other conditions. An overview of the case datasets from different collaborators is provided in Table 5.

Table 5 *Overview of case datasets provided by collaborators*

Dataset	Number	Comment
European collaborators		
DB (family) samples	40	UK cohort with a high rate of multi-ethnic migrants
DB samples	89	UK cohort with a high rate of multi-ethnic migrants
NW samples	11	UK cohort with a high rate of multi-ethnic migrants
LEV samples	14	Belgium cohort, exact ethnicity not provided
European samples	109	European cohort, exact ethnicity not provided
Dutch samples	109	Dutch cohort, exact ethnicity not provided
PREDNOS	178	Participants in the PREDNOS study (PMID: 31156083)
PREDNOS2	162	Participants in the PREDNOS2 study (PMID: 24767719)
Total	**712**	
South East Asian collaborators		
JK (family) samples	10	Asian cohort, exact ethnicity not provided
JK 1 samples	38	Asian cohort, exact ethnicity not provided
JK 2 samples	100	Asian cohort, exact ethnicity not provided
Asian samples	203	Asian cohort, exact ethnicity not provided
Sri Lanka samples	162	Sri Lanka cohort, exact ethnicity not provided
Total	**513**	

Legend *DB, NW, LEV, JK: Initials of the individual collaborators. PREDNOS: PREDnisolone in NephrOtic Syndrome; PREDNOS2: Trial of short course daily prednisolone therapy at the time of upper respiratory tract infection in children with relapsing steroid sensitive nephrotic syndrome.*

Sample preparation

After informed consent was obtained, based on the relevant locally approved protocols, DNA was extracted from the patients' whole blood samples and sent to the laboratory at the Royal Free Hospital, London. Each sample was assigned an internal sample ID. The sample ID contained information on the collaborator where it stemmed from together with a unique sample number. All information that could directly trace back to the individual (such as names, date of birth, *etc.*) were removed for data protection purposes. The samples were then plated on 96-well plates with 200ng of DNA in each well. Genotyping itself was performed by a specialised team at the core facility at ICH (Institute for Child Health) UCL (University College London) Genomics.

Genotyping

The SNPchip chosen for our study was the Infinium Multi-Ethnic Global BeadChip v.A1 (MEGA chip) from Illumina, CA, USA. The chip contains more than 1.7 million (1,779,818) markers and was designed to generate a multi-ethnic genotyping chip, that can be used to investigate associations in populations from different ethnicities. A special characteristic of the SNPchip was that it contains a large number of both common and rare variants, which will be relevant in the downstream analysis of our study. The data sheet with detailed information can be downloaded on https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/multi-ethnic-global-data-sheet-370-2016-001.pdf

Illumina (Array type) bead array processing

Processing was carried out at UCL Genomics (UCL Great Ormond Street Institute of Child Health, London) in accordance with the Infinium LCG Assay protocol (Illumina Inc, San Diego, USA). Briefly, in a deep well plate 200ng of high quality genomic DNA is whole genome amplified overnight (37°C, 20-24 hours), then fragmented (37ºC for 1 hour and 15 mins in a hybridisation oven), precipitated and resuspended in hybridisation buffer. Samples are hybridised onto beadchips using a liquid handling robot (Freedom Evo, Tecan Ltd, Switzerland) and incubated at 48°C for 16-24 hours. Unhybridized and non-specifically hybridized DNA is washed away, and the beadchip is prepared for staining and extension. Single-base extension of the oligos on the beadchip, using the captured DNA as a template, incorporates detectable labels on the beadchip and determines the genotype call for the sample. The process of single base extension and staining is carried out using liquid handling robot (Freedom Evo, Tecan Ltd, Switzerland). The staining procedure itself involves signal amplification by multi-layer immunohistochemical staining. Finally, the beadchips are scanned using the iScan scanner with autoloader (Illumina Inc, San Diego, USA). Data is generated in raw intensity files (IDAT) format. The raw IDAT files are processed by Genomestudio software (https://www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html)

Control cohort

We initially used a set of 432 European controls from a previous project which were collected by collaborators at Oxford University (Oxford controls). By searching for publicly available datasets, we subsequently identified two additional control sets, one from the Illumina website and one from the WTCCC project. In summary the control dataset was obtained from 3 independent sources. The details on each dataset are described below.

Oxford controls

This dataset was already available in our laboratory and was previously provided by collaborators from Oxford University. Data is also available through the European Genome Archive (EGAD00010000144 and EGAD00010000520) [133,134]. The dataset contained 432 samples consisting of patients who declared themselves as European. Genotyping was performed at Oxford University on a HumanOmniExpress-12 v1_J (n=144) Chip with 730,525 markers and on a HumanOmniExpress-12v1_A (n=288) Chip with 733,202 markers. The combined dataset, which only included markers present in both datasets, consisted of 730,397 markers.

Illumina ethnicity controls

This dataset was available from the Illumina website (http://www.illumina.com) and contained 270 samples. These samples were based on the Hapmap project dataset and comprised individuals from 4 different populations: CEU (Central European), YRI (African), JPT (Japanese) and CHB (Han-Chinese). The dataset is therefore referred as the Illumina ethnicity controls. 90 of the samples were known to be of European ancestry and were used as controls for the European cohort analysis. Genotyping was performed on a HumanOmniExpress-12v1_C Chip with 731,442 markers.

Wellcome Trust Case Control Consortium controls

We further identified a large control set from the Wellcome Trust Case Control Consortium (WTCCC) which contained 5,604 samples. This is the combined dataset of the 1958 birth cohort and the UK blood service control group controls. The data was made available through the WTCCC website (https://www.wtccc.org.uk) and published previously as a control set for GWAS [135].

The 1958 birth cohort (n=2,867), also known as the National Child Development Study, is a control set of individuals born in England, Wales and Scotland in 1958 [136]. Individuals were initially studied for perinatal mortality and over 17,000 births survivors were followed up to 42 years [136]. DNA was extracted from subjects with self-reported white ethnicity. The second subset was the UK blood service controls (n=2,737), which are provided by the UK blood service and includes DNA from healthy blood donors. The subjects were about equally divided into males and females and both control subgroups were geographically widely distributed across the UK.

Genotyping of both datasets was performed on an Illumina Human 1.2M Duo custom BeadChip with 1,106,184 markers. Details can be found on https://www.wtccc.org.uk and https://www.ebi.ac.uk/ega/studies/EGAS00000000028.

An overview of the case and the 3 control datasets is provided in Table 6.

Table 6 *Overview of case and 3 control datasets*

Dataset	Case/Controls	Sample number	European	Microarray	Variant count
SSNS	Case	1225	?	MEGA	1,779,818
Illumina	Control	270	90	OmniExpress	731,442
Oxford	Control	432	432	OmniExpress	730,397
WTCCC	Control	5,604	5,604	DuoCustom	1,106,184

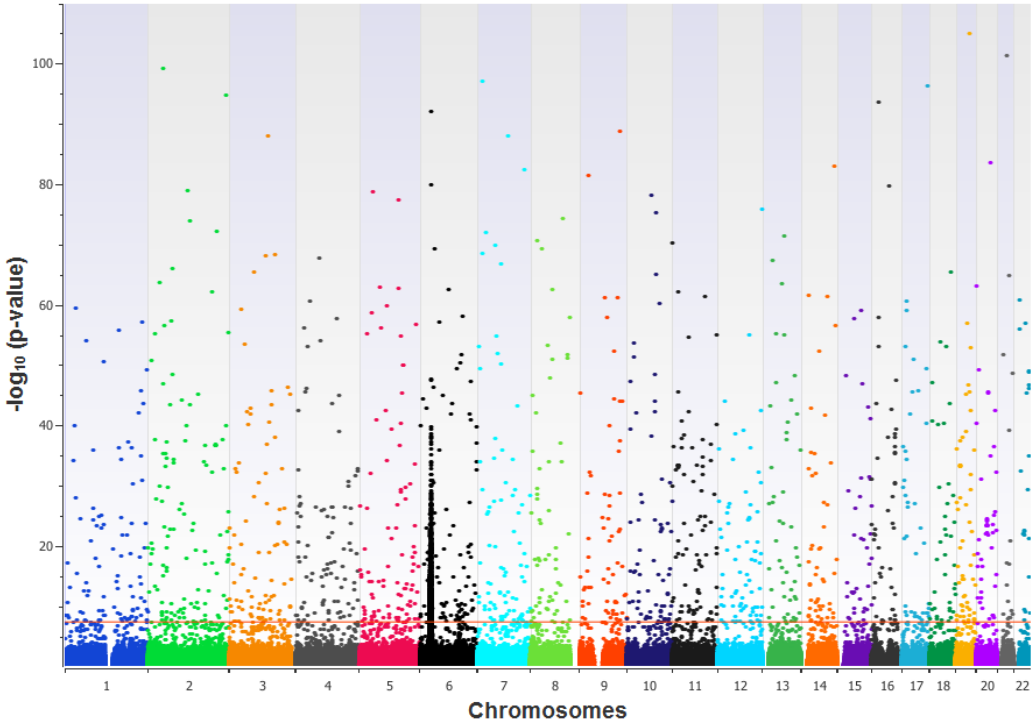
Legend *WTCCC: Welcome Trust Case Control Consortium*

Chapter 3. Data encoding and *REMEDY*

Many research groups outsource their genotyping process or use datasets genotyped in different labs and batches. This can lead to differences not only during the genotyping process but also how the data is exported from Genomestudio. Genomestudio allows the data to be exported in different file formats as well as different encoding schemes. The file format can usually be identified by visual examination of the dataset, whereas the identification of the encoding scheme of a dataset can be challenging. However, the knowledge of the encoding is essential to compare datasets of different sources as well as to identify and correct inconsistencies between the different batches of data prior to further analysis.

In our study we used cases and controls sourced from different centres and genotyped in different laboratories. The initial association analysis of cases versus controls showed a very noisy picture (Figure 9) with thousands of markers above the significance threshold line. This suggested systematic errors and we started to investigate systematic differences between the datasets, suspecting differences in the encoding schemes.

Figure 9 *Manhattan plot before processing with REMEDY*



Legend *Manhattan plot for our initial SSNS GWAS. The red line indicates the genome wide significance threshold. The plot demonstrates high levels of noise referring to thousands of markers above the genome-wide significance threshold line. This suggested systematic differences between the datasets which led to the development of REMEDY.*

Comparing data with different encoding schemes

Raw data generated on Illumina arrays is processed by converting the raw intensity files (IDAT) to a common output format using Genomestudio. The user can choose between different encoding schemes for the output data:

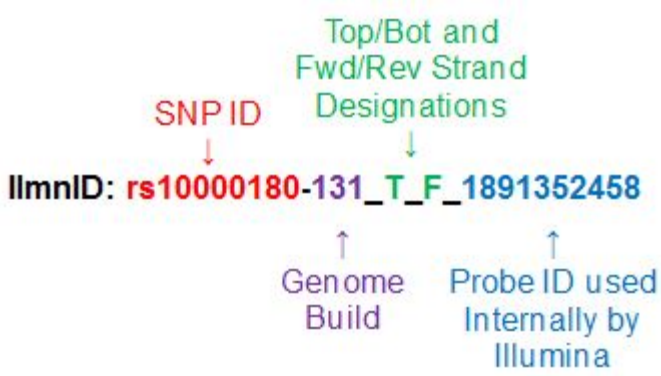
- DESIGN: This output format is the most basic and does not change the encoding scheme of genotyped data from the raw data. This scheme is strand nonspecific and calls the variant according to the strand on which the probe was designed against. This can be a mix of TOP/BOT or +/- or FWD/REV.
- TOP: This output format provides data in the TOP/BOT encoding scheme.
- FWD: This output format provides data in the FWD/REV encoding scheme.

More details on each encoding scheme can be found in the introduction part on page 45. Depending on the genotyping laboratory, personal preferences *etc.* the output

format of datasets can vary. These data cannot be compared directly but the source encoding and strand designation of the different datasets need to be defined first and then converted to a unified encoding format before proceeding with further analysis.

The information how to convert between the encoding schemes can be found in the Illumina manifest file. The Illumina manifest file for each microarray provides information on the internal Illumina ID, SNP rsID, genome build used and the calculated strand designation for the TOP/BOT and FWD/REV strand (Figure 10).

Figure 10 *Example for strand information provided in the Illumina manifest file*



Legend: *Screenshot from the Illumina online web page (http://emea.support.illumina.com/bulletins/2016/05/infinium-genotyping-manifest-column-headings.html). The internal Illumina ID (IlmnID) provides the following information: SNP ID: rs number of the SNP; Genome Build: The NCBI Genome Build the specific probe in this specific manifest is referring to; Strand designation: The calculated strand encoding for TOP/BOT and FWD/REV strand of the Illumina strand (DESIGN strand).*

This file can be used to convert between any of the schemes mentioned. The key is within the Illumina probe ID which provides both its TOP/BOT and FWD/REV strand designation. For example, T_F indicates TOP and FWD, while B_R would indicate BOT_REV. The DESIGN scheme is used as a bridge to transcode between the different schemes. For example, to change from TOP to FWD, the genotypes would be converted to the DESIGN encoding first using the internal Illumina ID information and then encode to the destination scheme again using the internal Illumina ID information. With the manifest file and this knowledge converting one dataset to the other or recoding all to a common encoding scheme is possible.

During the course of investigations, we developed together with Chris Cheshire, a computer scientist PhD student [137], a software based on this knowledge, which

enabled to identify the encoding scheme of the data and to convert between different encoding schemes. The software was named *REMEDY*. A short summary of the characteristics of the program is given below.

REMEDY

In summary, *REMEDY* is an in-house software utilised to re-encode genotypes uniformly to the Genomic Forward encoding scheme based on the genotyping manifest file following internal quality control and matching to dbSNP [137].

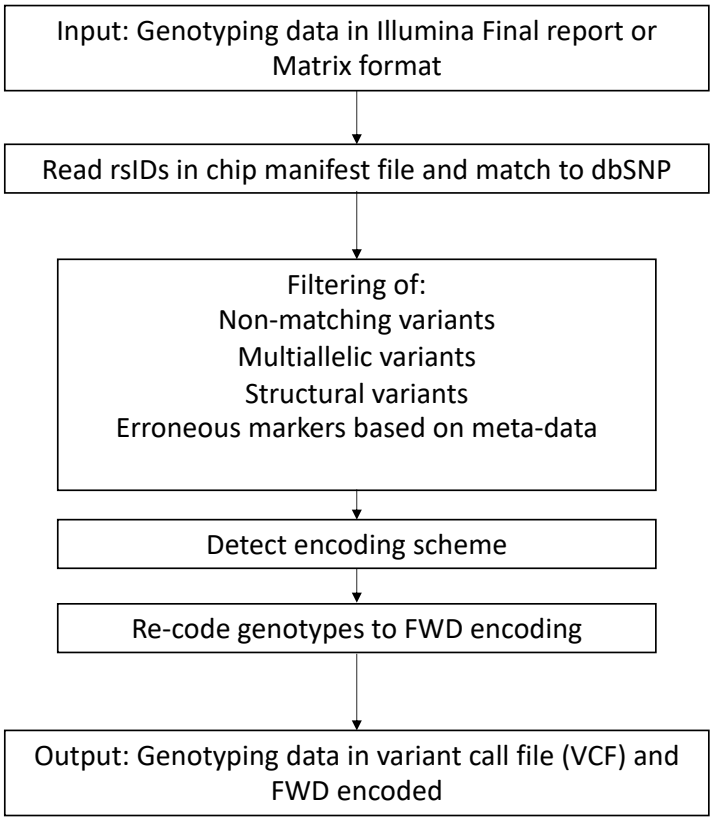
REMEDY accepts two different file formats of genotyped data as input, either the Illumina final report or the Illumina matrix format. The Illumina matrix format is the simpler format, giving one variant per row for one sample per column. The Illumina final report format gives one variant per sample per row and additional information on genotyping, probe information in the following columns.

After loading the dataset into *REMEDY*, the rsID of each variant is matched to dbSNP to obtain more information about the variant. Based on this, quality control steps are performed. This includes filtering for variants that are not matching to dbSNP, are multiallelic in dbSNP, are structural variants or where meta-data suggest problems with the variant. The dbSNP version *REMEDY* matches to is dbSNP version 150.

To detect the encoding scheme of a dataset each SNP is compared to the same SNP genotyped in different schemes. The scheme with the highest proportion of matches over the whole genotyped data is defined as the one used for this dataset. For example, if all genotypes match to the FWD version of a SNP then it can be assumed the dataset itself was encoded to FWD. If the match is roughly 50/50 split between TOP and FWD it can be assumed that the dataset was DESIGN encoded, as this scheme is strand unspecific.

The workflow of *REMEDY* is summarised in Figure 11.

Figure 11 *REMEDY pipeline*



Legend *The in-house software REMEDY pipeline, demonstrating conversion from raw data files to FWD encoded data for use in subsequent analysis.*

All case and control datasets were run through *REMEDY*. The detected encoding schemes for each dataset are shown in Table 7.

Table 7 *Overview of encoding schemes of case and control datasets*

Dataset	Status	Microarray	Encoding
SSNS	Cases	MEGA	DESIGN
Illumina	Controls	OmniExpress	FWD
Oxford	Controls	OmniExpress	FWD
WTCCC	Controls	DuoCustom	TOP

Legend *SSNS: Steroid sensitive nephrotic syndrome; WTCCC: Wellcome Trust Case Control Consortium; FWD: FWD/REV encoding scheme; TOP: Illumina TOP/BOT encoding scheme; DESIGN: Illumina DESIGN encoding scheme*

All datasets were converted to FWD encoding of the FWD/REV encoding scheme and imported into SVS (SNP & Variation Suite v8.6.0).

Chapter 4. Quality control

The genotyping process is not perfect and bears many possible sources of errors leading to poor quality data. Poor quality of the DNA samples, contamination or mix-up of the samples, poor DNA hybridization to the array or poorly performing genotyping probes are some possible sources of errors. Thus, samples and markers have to undergo strict quality assessment and control (QC) before conducting an association analysis.

Our quality control procedures are based on protocols published in Nature [138] and on suggestions of the Golden Helix SNP & Variation Suite (SVS) manual. QC steps were adapted throughout the thesis to optimize the findings of our study.

The QC procedure consists of steps applied per sample, per marker and to address population stratification. An overview is given in box 2.

Box 2 Summary of QC steps

QC step	Function
Per sample	
Call rate	To exclude individuals who have a high rate of missing genotypes.
Heterozygosity rate	To remove individuals with a high or low rate of heterozygous genotypes as this could indicate poor quality sample or inbreeding.
Identity by descent	Measurement for relatedness of samples. To exclude duplicated or related individuals.
Per marker	
X/Y chromosomes	To account for statistical and methodical challenges in analysing sex chromosomes.
Call rate	To exclude markers that are missing in a large proportion of individuals.
Allele count	To exclude markers with more than 2 alleles.
Minor allele frequency	To exclude markers where the minor allele has a frequency below a certain threshold.
Hardy-Weinberg Equilibrium	To exclude markers that deviate from the Hardy-Weinberg Equilibrium as this could indicate genotyping error or population selection.
Population stratification correction	
Principal component analysis	To select a homogenous ethnical group as allele frequencies can differ between ethnicities and therefore population stratification can lead to false association results.

QC steps per samples

First, QC steps per sample were carried out. As genotyping of cases and controls was performed on different SNP chips and in different centres, quality control steps per sample were first done independently in the case and each control set and subsequently on the combined case-control set.

Call rate

The call rate (CR) describes the proportion of genotypes per sample with non-missing data. Samples with a high number of missing genotypes will present with a low call rate. This can be secondary to variation in DNA quality and concentration and would affect the overall results of an association study. Therefore, samples with missing data of more than 10% (equal a call rate below <90%) were removed.

Heterozygosity rate

The heterozygosity rate describes the proportion of heterozygous genotypes for each individual. Samples with an unexpectedly high number of heterozygous genotypes could reflect low sample quality whereas samples with an unexpectedly low number of heterozygous genotypes can be a sign of inbreeding.
To detect samples with deviating heterozygosity rates we first calculated the mean heterozygosity rate for all samples for each dataset and then removed samples deviating with more than 3 standard deviation (SD) +/- from the mean.

Identity by descent

A GWAS is based on single, unrelated individuals. However, duplicated samples could be missed or individuals could be related more closely to another than assumed. Inclusion of those samples can bias the results.
Identity by descent (IBD) is a measure of how many alleles at any marker in each pair of two individuals are shared because of a common ancestor [139]. IBD reflects the grade of relatedness of those two individuals.

- Duplicate samples or identical twins should have 100% of alleles coming from the same ancestral chromosome.

- Siblings should have 50% of the alleles coming from the same ancestral chromosome
- Half-siblings should have 25% of the alleles coming from the same ancestral chromosome
- Unrelated individuals should theoretically have 0% of the alleles coming from the same ancestral chromosome.

Based on this, the thresholds should be IBD =1 for duplicates or monozygotic twins, IBD =0.5 for first-degree relatives, IBD =0.25 for second-degree relatives and IBD =0.125 for third-degree relatives. Due to genotyping error, linkage disequilibrium and population structure there is often some variation around these theoretical values and it is recommended to remove one individual from each pair with an IBD >0.1875, which is halfway between the expected IBD for third- and second-degree relatives [138]. Therefore, only samples with an IBD ≤0.1875 were included in the study.

For the analysis of relatedness, it is recommended to use only independent markers. This can be achieved by linkage disequilibrium (LD) pruning. LD pruning is a method to identify and deactivate markers that are in LD with other markers that are left active. This will reduce the overall number for markers to a subset which is independent of each other. LD pruning was performed in SVS with the default settings. Only LD pruned markers were used for the analysis of relatedness.

To identify a maximum of unrelated samples in our case and controls dataset we used the software PRIMUS_v1.9.0 [140], which can be downloaded at https://primus.gs.washington.edu/primusweb/res/documentation.html

PRIMUS is an open source program for pedigree reconstruction (PR) and Identification of the Maximum Unrelated Set (IMUS) [140]. Only the IMUS mode was used for our study and is an algorithm that identifies the maximum set of unrelated individuals with a defined threshold of relatedness in any dataset [140]. PRIMUS is run in command line option and uses PLINK. A relatedness threshold of second degree was chosen for the case and control datasets. This was operated by adding in PRIMUS the command line option ‘‘--rel_threshold 0.1875.’’

Following command was used for analysis:

```
#Convert PED/MAP to BIM/BED/FAM
plink --file ssnsibd

#Calculate relatedness
plink --bfile plink --genome

#Run PRIMUS with higher threshold
PRIMUS_v1.9.0/bin/run_PRIMUS.pl -p plink.genome --no_PR -t
0.1875 -o PRIMUS_0.1875
```

The output file indicating the maximum of unrelated samples was imported into SVS for further analysis.

QC steps per markers

In a large GWAS, where hundreds of thousands of markers are tested, even a small percentage of erroneous markers, can lead to a thousand false signals. In contrast, with every marker removed from the study a potentially important association may be lost. Therefore, it is essential to find the optimal balance between removing markers that can cause false positive results and omitting essential information.

In order to optimize QC per markers, two tests were performed:

a) different thresholds for each QC step were tested and examined with respect to the results in the association study and
b) the difference between applying the QC steps on the combined case and control set or on each dataset separately was examined.
Results are displayed in the results part on page 106.

Following QC steps were applied on markers.

Exclusion of X and Y chromosome

X and Y chromosome were excluded from subsequent analysis. This has been a common decision for many published GWAS because of the unique analytical challenges the sex chromosomes present [141]. The significance of variants on the X and Y chromosomes is harder to asses, first simply because there are two copies of X in women and only one in men, so the signals obtained for variants when genotyping are lower in men than in women. Further, one of the two X chromosomes gets randomly inactivated in female cells (X inactivation). It is not yet possible to tell which variant is on the active and which is on the silent version of the X chromosome.

Allele count

The allele count (AC) describes how many alleles can be found at a specific marker. One marker can be either biallelic or multiallelic. *E.g.* Looking at a single nucleotide polymorphism, some individuals may carry the nucleotide A and others the nucleotide C for this specific SNP. This SNP would be called a biallelic marker with one allele being nucleotide A and the other allele nucleotide C. If further individuals carry the nucleotide T at this specific SNP, then the SNP would have 3 possible alleles, nucleotide A, C and T and would be called multiallelic (>2 alleles). In association studies it is difficult to account for multiallelic SNPs, simply because of the test statistics, and therefore most SNPs are chosen to be at sites which are biallelic. If nevertheless markers with AC more than 2 were found in the sample set, those were excluded in this study.

Call rate

The call rate (CR) of a specific marker describes the percentage of individuals in which it is genotyped. A low CR of a marker can reflect problems during genotyping or with the sequencing method for this specific marker. Classically, markers with a call rate less than 95% are removed from further study, though thresholds can vary depending on the quality of data [138]. We tested different scenarios for cut-off levels shown in the results part on page 107.

Minor allele frequency

The minor allele frequency (MAF) indicates the frequency of the least often occurring allele at a specific *locus*. Generally, SNP arrays include SNPs with a wide distribution of MAF from nearly monomorphic (MAF <0.5%) to very common (MAF ≈50%) SNPs. Historically, many GWAS focused on SNPs with a MAF >10% [142]. This was because of concerns about genotyping accuracy of low frequency alleles and because larger sample sizes are required to detect association at those markers. Research suggests that the MAF threshold should depend on the sample size of the study, with larger sample sizes allowing for lower MAF thresholds [143,144].

The SNP chip used for genotyping of our case dataset contained a high number of low frequency alleles. Different MAF thresholds, 5%, 1% and 0.1%, were explored to find the optimal cut-off for our study. Results are shown in the results part on page 109.

Hardy-Weinberg Equilibrium

The Hardy-Weinberg Equilibrium (HWE) theorem assumes that allele frequencies in a population remain stable over generations. This is under the assumptions that a) mating is random, b) the population is not under selection, c) it is infinitely large, d) no mutations occur and e) no emigration or immigration happens [145].

Based on the assumptions that the allele frequency is stable, the genotype frequencies can be calculated according to following equation:

$$p^2 + 2pq + q^2 = 1$$

p stands for the frequency of allele A and q for the frequency of allele a

According to the equation:

p^2 gives the frequency of homozygous individuals carrying the genotype “AA”
2pq gives the frequency of heterozygous individuals carrying the genotype “Aa”
q^2 gives the frequency of homozygous individuals carrying the genotype “aa”

In genetic association studies, where the number of individuals with each genotype is known, the allele frequencies can be calculated and assessed for deviation from the HWE. If there is a significant deviation from the HWE, with a chosen p-value as cut-off at a certain marker this could be a sign of violation of the rules above or, more likely, of genotyping errors. Therefore, markers not following the HWE are excluded from further analysis.

Different cut-off levels were tested to find the optimal level for this study. The results are shown in the results part on page 113.

Population stratification

Population stratification describes the presence of different subpopulations (*e.g.* individuals with different ethnic backgrounds) in a study [146].
Different subpopulations can have differences in their allele frequencies not secondary to the disease but due to different ancestry, which can lead to false positive associations and/or mask true association [146]. In general, a careful selection of cases to an ancestry matched control set can minimize the effect of population stratification. However, in our study the case cohort consisted of people from different ancestries, with no well-defined ethnical background, whereas the control cohorts were self-declared Europeans. We had to overcome the problem of different ethnicities in our case cohort and find a method to select for Europeans only. The method of principal component analysis was chosen and performed in SVS.

Principal component analysis and inflation factor lambda

A principal component analysis is a statistical technique of finding patterns in a high dimensional dataset. It aims to reduce multidimensionality in a dataset with multiple correlated data points to a smaller, interpretable format [147]. With mathematical methods correlated variables are converted into a linear set of uncorrelated variables, the principal components. The first principal component represents the largest possible variance in the dataset and accounts for as much of the variability as possible. The second principal component represents the second largest variance, and so on. The principal components can be plotted in a two dimensional manner and visualized as a scatter plot [148].

The computational requirement for PCA is depending on the number of samples and markers. Consequently, the process of principal component analysis can last up to more than 12 hours if performed on a large dataset with many markers and samples, even on a high-performance modern computer. A way to reduce the computational requirement and consequently the time needed, is to reduce the number of markers on which the principal component analysis is performed on. Hence, before performing PCA on our dataset, we tested if the reduction of markers affects the results of the principal component analysis. We reduced the number of markers in 5 steps and compared the first 5 principal components. A reduced number of 20,000 markers did not affect the results and consequently 20,000 randomly selected markers were used to perform principal component analysis in our dataset.

The first 10 principal components were calculated for all cases and European controls using EIGENSTRAT and an additive model. EIGENSTRAT is the name for the program, developed by the Broad Institute, which implements the PCA correction technique in SVS [147]. Different cut off levels for removal of outliers were tested. The analysis was performed six times: with no removal of outlier, and with removal of outlier with a standard deviation of more than 4, 3.5, 3, 2.5 and 2 respectively. The results are shown on page 102.

Further, the inflation factor lambda was calculated for each scenario. The genomic inflation factor lambda is a way to measure stratification in a population [149]. It is calculated as the ratio of the median of the empirically observed distribution of alleles to the expected distribution of alleles in the situation of a basic allele test [149]. A moderate to large lambda (1.1-1.2 and >1.2) reflects a higher grade of stratification (less homogeneous dataset), whereas a small lambda (1.0-1.1) reflects a lower grade of stratification (more homogenous dataset). A large lambda can cause false positive associations. To achieve a small lambda usually many outlier samples must be excluded, thereby reducing the number of samples in a dataset. This on the other hand can be problematic in the situation of a limited number of cases/controls or a rare variant [150].

Analysis was once performed without previous quality control steps applied on markers and then with quality control steps applied on markers, in order to see if this influences the PCA results. PCA without previous QC steps on markers revealed a so-called batch effect (stratification secondary to the inclusion of different batches of datasets in the analysis). Thus, all PCAs were done after QC steps.

Chapter 5. Imputation

Imputation is a standard method in GWAS for increasing the marker density and thus the resolution of association testing (see introduction page 49).

Beagle

For imputation of our datasets Beagle version 5.0 was used [151]. It was developed by Browning *at el* and first published in September 2006. Since then it has undergone several updates and new versions have been released. The latest version at the date of this study was version 5.0, which was released in 2018 [151].

Reference panel

In 2007, the main reference panel was provided by the HapMap Project consisting of 210 individuals with 420 different haplotypes for 3.1 million SNPs [152]. In 2015, the phase 3 reference panel of the 1000 Genomes Project was released consisting of 2,504 individuals from 26 worldwide populations with 5,008 different haplotypes for each of 88 million SNPs [153]. This reference panel (1000 Genomes Project Phase 3 data version 5a) was used for this study and was downloaded from the Beagle homepage (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/).

Imputation was performed on the combined case and control dataset, after stringent QC for samples and markers. The detailed approach to imputation consisted of following steps.

Pre-imputation filtering

Only cases and controls of the final European dataset were processed for imputation. Markers were removed because of CR <0.99, AC >2, MAF <0.01 in all datasets and outside the HWE p<0.001 in the control dataset. The filtered and combined dataset was processed in variant call format (vcf).

Split vcf file into 22 chromosomes

In order to speed up the imputation process via parallelization the vcf file was split into individual chromosomes using vcf tools. The splitvcf utility splits a single vcf file into

multiple vcf files corresponding to the respective chromosome. The following command was used:

```
vcftools --vcf dataset.vcf --chr 1 --out chr1 –recode
```

Displayed is the command for chromosome 1. The command was repeated 22 times for each of the autosomal chromosomes.

Imputation via Beagle v5

After splitting the vcf file into each chromosome imputation was performed with Beagle v5 using following command:

```
java -Xmx26g -jar /usr/lib/jvm/java-8-openjdk-
amd64/jre/lib/ext/beagle5.jar
map=/mnt/data/projects/imputation/beagle/map/plink.chr1.GRCh37
.map
gt=/mnt/data/projects/imputation/input/dataset_chr1.recode.vcf
ref=/mnt/data/projects/imputation/beagle/1kgo_ref_panel/chr1.1
kg.phase3.v5a.b37.bref3
out=/mnt/data/projects/imputation/imputed/dataset_chr1_output
window=10.0
```

As previously, the command was repeated 22 times for each of the autosomal chromosomes.

Output files

Beagle generates two output files. The log file gives a summary of the analysis that includes the Beagle version, the command line arguments and the run time. The vcf.gz file is a bgzip-compressed vcf file that contains phased, non-missing genotypes for all non-reference samples. Further, following information for each marker can be found in the vcf file. A “DR2” field which gives the estimated squared correlation between the estimated allele dose and the true allele dose. An “AF” field which gives the estimated alternate allele frequencies in the imputed samples. The “IMP” mark if the marker is imputed.

Post imputation filtering

The principles of quality control per marker post imputation is similar to those for genotyped markers. The aim is to remove all markers, that could cause false positive associations. Additionally, imputation accuracy has to be addressed during the quality control process.

Filtering on allelic R square

The first step was to account for imputation accuracy. Imputation is based on probabilities. The genotype outputted is the most probable one for each sample at each marker according to a complex calculation algorithm [102]. One way to address imputation accuracy is by using the allelic R square. The newest version of Beagle 5.0 has replaced the allelic R square by the dosage R square. DR2 is the estimated squared correlation between estimated allele dose and true allele dose. More details on imputation accuracy and the allelic R square can be found in the introduction part on page 51.

The idea is to remove all markers which are imputed below a certain threshold of DR2. Beagle itself does not give a suggestion for a cut-off value. Other imputation programs, *e.g.* IMPUTE2, mention a cut-off of 0.3-0.5 in their manual [154] however they point out that there is no universal cut-off value for post-imputation SNP filtering, but the value depends on the specific analysis. In the manual of MACH, another imputation program, a minimal cut-off of 0.3 is recommended for filtering out poorly imputed markers of bad quality [155]. A recently published GWAS suggested a more stringent cut-off such as >0.8 [156]. In accordance with this study we also decided to use a cut-off of 0.8.

The first filtering step post imputation was to retain only those markers with a DR2 >0.8 using bcftools. Following command line instructions were used for filtering for DR2 score:

```
for chr in {1..22};
do echo ${chr}
bcftools filter -i 'DR2>=0.8' -Oz
imputed/chr${chr}_imputed.vcf.gz -o
filtered/chr${chr}_DR2_0.8.vcf.gz
```

Copy number variants

Imputation not only infers genotypes where SNPs are present, but also those with copy number variants or insertion and deletions. Any kind of variant, in which more than 1 nucleotide is altered, is complex in the downstream analysis and therefore can cause false positive results. We therefore only kept SNPs for further analysis.

```
#!/bin/bash
for chr in {1..22};
do echo ${chr}
plink2 --vcf filtered/chr${chr}_DR2_0.8.vcf.gz --vcf-idspace-
to _ --const-fid --out filtered/chr${chr}
plink2 --bfile filtered/chr${chr} --make-bed --snps-only --out
filtered/chr${chr}_snpsonly
```

MAF, CR and HWE on controls

As outlined before, markers with a low minor allele frequency, a low call rate or outside the HWE, can introduce false positive results. We therefore, repeated these QC steps on the imputed dataset using PLINK. Markers with a MAF <1% (--maf 0.01) and CR <99% (--geno 0.01) were removed.

```
#!/bin/bash
for chr in {1..22};
do echo ${chr}
plink2 --bfile filtered/chr${chr}_snpsonly --make-bed --geno
0.01 --maf 0.01 --out bimbam/chr${chr}_filt
```

Markers showing significant deviation from HWE (P<0.001) in the control individuals were removed using SVS.

Import to SVS

The output data were imported to SVS for further analysis. Chromosomes 1-22 were merged and the appropriate marker map was applied. Downstream analysis was the same as with genotyped markers only.

Chapter 6. HLA imputation

We aimed to impute human leukocyte antigen alleles from genotyped SNPs within the major histocompatibility complex (MHC) region on chromosome 6. Different programme tools have been developed to perform HLA imputation from SNP- level data, including SNP2HLA [116]. SNP2HLA is a freely available tool set and has been used in multiple previous studies for HLA imputation. SNP2HLA is using the software package Beagle for HLA imputation.

Reference panel

As a reference panel the HapMap CEU reference dataset was used. The dataset contains 124 samples with 248 haplotypes.

Analysis

As input the final dataset for the European GWAS with a subset of 1,189 genotyped SNPs overlapping with the HapMap CEU reference dataset were used. HLA imputation was performed using SNP2HLA v1.0.3 (http://software.broadinstitute.org/mpg/snp2hla/) with default parameters.

Only HLA alleles were used for analysis. Post imputation quality control included removal of imputed HLA alleles with a quality score $R2<0.8$.

Logistic regression with adjustment for the first ten PCs of ancestry was used to test for association of each HLA allele with SSNS. Conditional analysis of the lead HLA alleles was performed using a logistic regression model.

Chapter 7. Post association analysis

The association analysis detected several variants in association with SSNS. In the post association analysis we investigated if the lead variants outside the HLA *locus* alter the expression of any neighbouring genes.

Multitissue eQTL

Quantitative traits are genetic variants that are highly correlated with gene expression. Hence, expression quantitative trait loci (eQTLs) refers to genetic variants that alter the expression of one or more genes.

GTEx

Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation (http://gtexportal.org). The data available on the GTEx portal is derived from 54 non-diseased tissue samples within nearly 1000 individuals. Correlations between genotype and tissue-specific gene expression levels were obtained and are available on the platform.

We queried the GTEx portal (http://gtexportal.org; information obtained October 2019) to investigate if any of the lead variants associated with SSNS were known to alter expression of genes in any of the 54 available tissues included in this database, including immune cells.

Chapter 8. Meta-analysis

METAL

A commonly used tool to perform a meta-analysis is METAL [119,76].
In short, METAL uses the direction of the effect of a risk allele combined with the p-value for this allele observed in each single study to calculate an overall effect size and p-value for the specific allele.
The direction of effect is determined using a reference panel for the relevant population. A positive effect with respect to the reference allele A represents the situation where an increased number of allele A copies is associated with an increased risk of disease. A negative effect with respect to the reference allele A, would represent a situation where an increased number of allele A copies is associated with a decreased risk of the disease.

The sample size of each study is taken into account when weighing the effect size of each study result [119]. Therefore, this method can combine the evidence of association from individual studies by using appropriate weights.

SVS method for meta-analysis

The method available in SVS are based on METAL concept. In our study we used a sample-size based approach and fixed-effect algorithm output as provided by SVS [119]. This means, that for every study and marker, the p-value, effect direction, and sample size (for weighing purposes) are taken into account. From these, a Z-score and an overall p-value are calculated.

Part 3: Results

Chapter 1. Pre-analysis: Cases and controls explorations

Case cohort

The cases selected for this study were 1225 children diagnosed with SSNS according to the KIDIGO guidelines. The samples were provided from different collaborators from Europe (n = 712) and South East Asia (n = 513) and therefore expected to be from different ethnicities. The samples were genotyped together and retrieved as a combined dataset.

Genotyping and data processing

DNA was extracted from whole blood samples and genotyping was performed uniformly at UCL Genomics (UCL Great Ormond Street Institute of Child Health, London). The chip used for genotyping was the Infinium Multi-Ethnic Global BeadChip. Data on 1,779,818 markers was collected in raw IDAT format from UCL Genomics.

The dataset was uploaded in *REMEDY* and the results showed that thousands of variants on the microarray were unsuitable for GWAS; Approximately 10,000 markers had rsIDs not matching with dbSNP, 150,000 markers were multi-allelic in dbSNP, another 20,000 were structural variants and 70,000 showed inconsistence in the strand designation. These variants all have the potential to cause noise in subsequent analysis and were therefore removed by *REMEDY*. We also noted that the data were DESIGN encoded. The genotypes were thus re-encoded uniformly to the Genomic FWD encoding scheme and 1,565,259 SNPs were outputted. 51,276 markers were on X and Y chromosomes and excluded from further analysis. The number of autosomal markers was 1,513,983.

The definition of a successful sample included a call rate $\geq$ 90%. According to this definition, 12 samples were removed from the dataset and the remaining 1,213 were considered as successful.

Control cohort

The control cohort consisted of 6,306 individuals sourcing from 3 different control datasets. The Oxford controls, Illumina ethnic controls and WTCCC controls.

Oxford controls

The Oxford control dataset consisted of 432 samples. All samples were of self-declared European ethnicity. Genotyping had been performed at Oxford University on a HumanOmniExpress-12v1_J (n=144) Chip and on a HumanOmniExpress-12v1_A (n=288) Chip. The combined dataset consisted of 730,397 markers.
The dataset was uploaded in *REMEDY* and the results showed that approximately 1,200 markers had rsIDs not matching with dbSNP, 45,000 markers were multi-allelic in dbSNP, 115,000 showed incoherence in the strand designation and another 1,000 had multiple mappings. These variants were removed by *REMEDY*. The dataset was identified to be FWD encoded. After *REMEDY*, 672,361 markers were outputted. 18,402 markers were on X and Y chromosomes and excluded from further analysis. The number of autosomal markers was 653,959. All samples had a call rate ≥90% and were considered as successful.

Illumina ethnicity controls

The Illumina ethnicity control dataset consisted of 270 samples of which 90 were of European ethnicity. We initially processed all 270 samples, in order to use the representatives of the different populations as a reference panel for the exploration of ethnicity of our datasets. As a control cohort for the GWAS, only the 90 European samples were used. Genotyping had been performed on a HumanOmniExpress-12v1_C Chip. The dataset contained 731,442 markers and was also processed by *REMEDY*. As the SNP chip and its manifest for this dataset is the same as for the Oxford controls, those datasets were processed together via *REMEDY* and the outputted markers were identical. The dataset was also FWD encoded and 672,361 markers were used for analysis. 18,402 markers were on X and Y chromosomes and excluded from further analysis. The number of autosomal markers was 653,959. All samples had a call rate ≥90% and were thus considered successful.

Wellcome Trust Case Control Consortium controls

The WTCCC control dataset consisted of 5,604 samples. All samples were of self-declared European ethnicity. Genotyping was performed on a Illumina Human 1.2M Duo custom BeadChip v1.

The dataset was uploaded in *REMEDY* and the results showed that approximately 100,000 markers had rsIDs not matching with dbSNP, 75,000 markers were multi-allelic in dbSNP and 500 structural variants, 180,000 showed incoherence in the strand designation. These variants were removed by *REMEDY*. The dataset was identified to be TOP encoded and re-coded to FWD encoding. After *REMEDY* processing, 1,065,696 markers were outputted. 40,259 markers were on X and Y chromosomes and excluded from further analysis. The number of autosomal markers was 1,025,437. 55 samples were removed from the dataset as the call rate was less than 90%, whereas 5,549 were considered as successful.

Ethnicity of case-control cohort

As previously noted, a prerequisite for association studies is to compare cases and controls which are ancestry matched. Differences in allele frequencies between populations can otherwise lead to false positive associations and/or mask true association. To understand the variety of ethnicities in our dataset we first wanted to illustrate the ethnicity distribution of all cases and controls, and then concentrate on investigating European and South East Asian cohorts separately.

For this purpose, a common dataset of all cases and controls was created by combining the data matrices, including only overlapping markers. The combined dataset consisted of 1,213 cases and 6,071 controls with 216,015 overlapping markers.

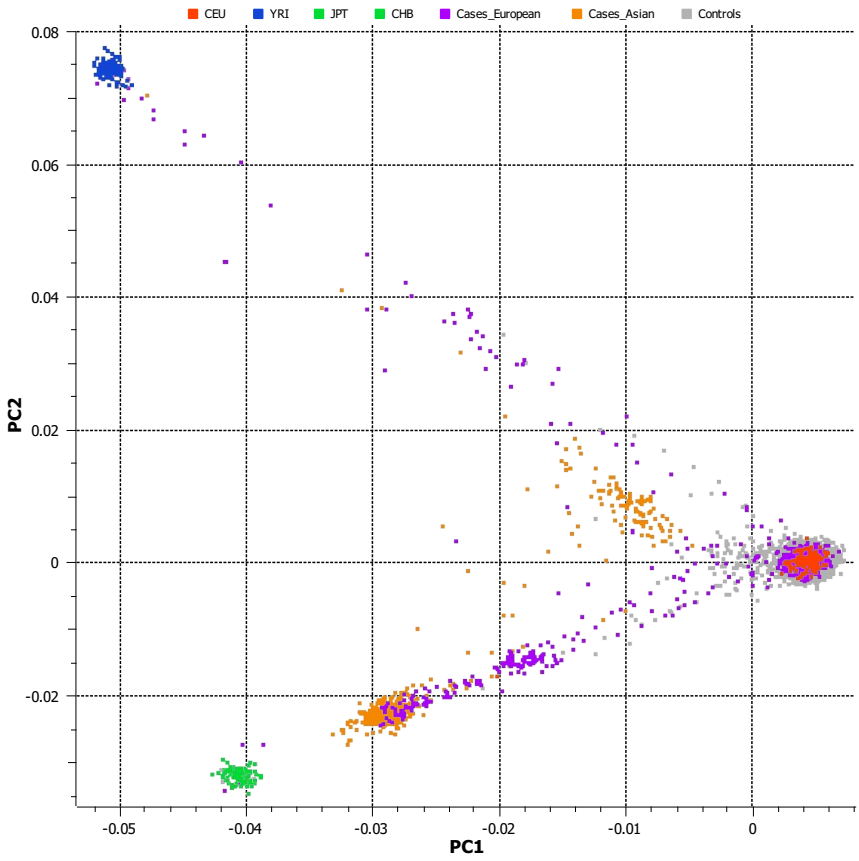
Principal component analysis all cases

For exploration of ethnicities the means of principal component analysis were used (PCA; EIGENSTRAT implemented in SVS) [147]. The principal components for each individual in the case and control dataset were calculated and compared to the Illumina ethnicity control dataset.

The Illumina ethnicity control dataset is based on the HapMap project and includes samples from four populations (YRI, JPT, CHB and CEU) on three continents. These samples are considered representative of the genetic diversity of human populations and widely used as a reference panel to explore and illustrate ethnical diversity in a dataset [157]. YRI stands for samples from the Yoruba in Ibadan, Nigeria, Africa. The Yoruba participants have identified themselves as having four Yoruba grandparents. JPT stands for Japanese from Tokyo and samples were recruited in the area of Tokyo from individuals declaring themselves and their grandparents coming from Japan. CHB stands for samples from Han Chinese from Beijing, China, Asia and were required to have at least three Han grandparents. CEPH stands for samples from the Centre d'Etude du Polymorphisme Humain (CEPH) collection of Utah residents of Northern and Western European ancestry and are labelled as CEU. CEU samples are considered as representatives of European. However, at the point of collection there was no specification of the different subpopulations residing in Europe. Therefore, it is not clear how the pattern of genetic variation *e.g.* between north and south Europe is reflected in this sample set [157].

We performed a PCA for all cases and controls and calculated up to 10 principal components for each sample. The simplest way to understand and interpret the outputted data is by visualizing the results in a scatter plot. The largest dimension of variability is represented by the first two components and therefore the first principal component was plotted on the x-axis against the second principal component y-axis. We illustrated the results in comparison to the Illumina ethnicity controls, which aids to interpret the ethnicity distribution of our samples in relation to the reference populations (CEU, JPT, CHB and YRI). (Figure 12)

Figure 12 *Scatterplot for PCA of all cases and controls*



Legend *Principal component scatter plot. Distribution of all cases (n=1,213) from European and Asian collaborators together with the combined control cohort (n=6,071) along the top two principal components (PC1 and PC2) identified by principal component analysis. The results are visualized in comparison to the Illumina ethnicity controls (CEU, YRI and CHB-JPT). Note that the European cases aggregate around the CEU controls, but a substantial number scatter towards the CHB-JPT controls and AFR controls. The Asian cases aggregate between the CHB-JPT and CEU controls, with a small number scattering closer to the CEU controls. The control dataset mainly clusters around the CEU controls.*

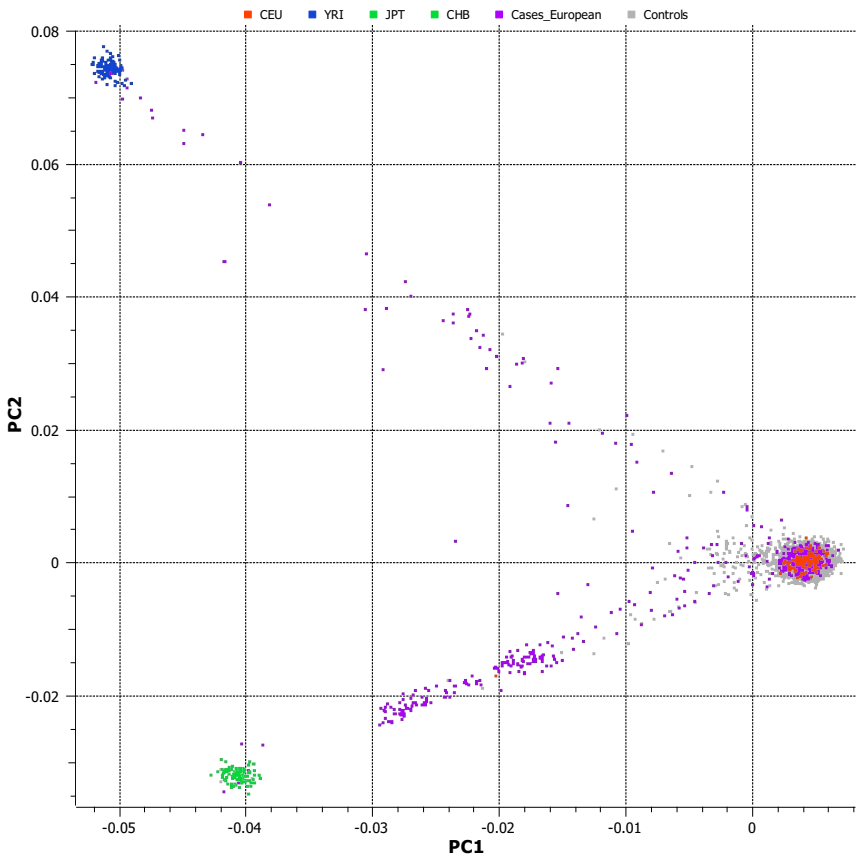
The distance between the clouds of aggregation reflect the amount of stratification between the groups. We could observe the aggregation of the European case dataset around the CEU ethnicity control samples, but a substantial number of samples scattered towards the CHB-JPT or YRI control groups. This revealed that the samples collected from European collaborators contained individuals from different ancestries and subpopulations beside Europe. Case samples collected from the Asian collaborators cluster between the CEU and CHB-JPT ethnicity control groups, an area which is known to represent South Asia (*e.g.* Sri Lanka). This group appeared less stratified. Samples from the control dataset (all self-declared Europeans) clustered

around the CEU control group, genetically confirming the self-declared European ancestry.

Selection of Europeans

In this part of the thesis we focus on the European cohort. This included the 709 cases from European collaborators and 6,071 European controls. The PCA was repeated, only including these datasets to assess the differences in ethnicity between European cases and controls in more detail. (Figure 13)

Figure 13 *Scatterplot for PCA of European cases and controls before removal of outliers*



Legend *Scatter of principal components. Distribution of the presumably European cohort (709 cases and 6,071 controls) along the top two principal components (PC1 and PC2) identified by principal component analysis. The results are visualized in comparison to the Illumina ethnicity controls (CEU, YRI and CHB-JPT). This is done before the outliers, scattering towards YRI and CHB-JPT control group, are removed.*

As observed above, the majority of the 709 samples clustered around the CEU ethnicity control group, but a relevant number of samples revealed not to be of

European ancestries. A way to remove these non-European samples, is to perform a PCA with removal of outliers above a certain standard deviation (SD) threshold. This can be done in steps, with stepwise lowering of the threshold of accepted SD. The aim was to establish an ethnical homogenous group containing as many samples as possible.

Stepwise PCA of cases and controls

The analysis was performed on the 709 cases and 6,071 controls in comparison to the Illumina ethnicity controls. Outliers were removed in 4 steps, with removal of outliers with a standard deviation of more than 5, 4, 3, and 2 respectively (Table 8). The first two principal components for each step were plotted in a scatter plot and inspected (Figure 14).

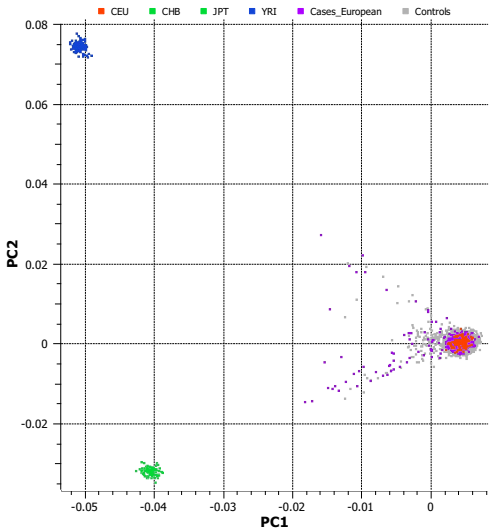
Table 8 *Summary of the influence of removal of outliers with different standard deviations on the number of remaining cases and controls*

	Cases remaining	Controls remaining	IF Lambda
No removal of outlier	709	6,071	9.445
A Removal of outlier SD >5	491	6,051	1.385
B Removal of outlier SD >4	476	6,040	1.288
C Removal of outlier SD >3	459	5,791	1.273
D Removal of outlier SD >2	177	2,383	1.057

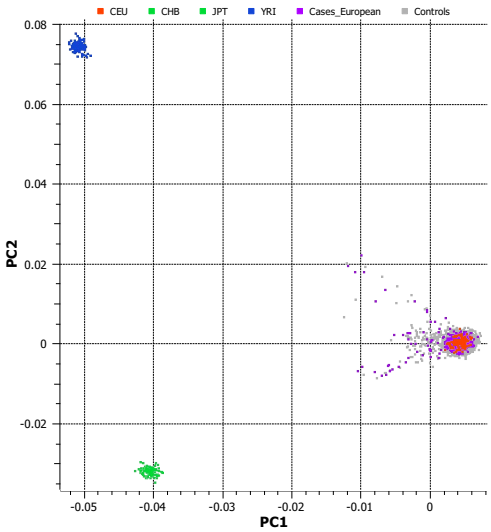
Legend: SD: Standard deviation of the mean; IF: Inflation factor

Figure 14 *Scatterplot for PCA of European cases and controls with stepwise removal of outliers*

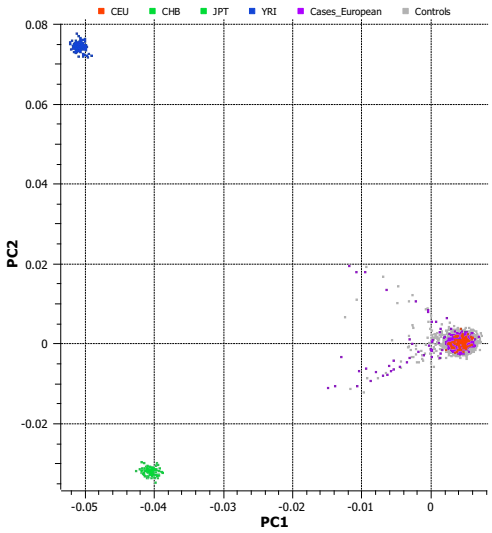
A)



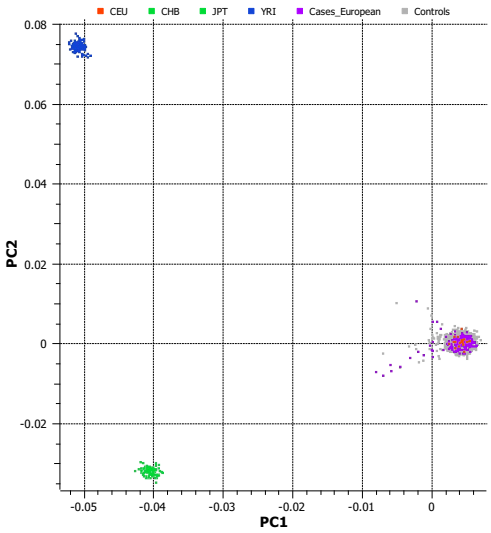
B)



C)



D)



Legend *Distribution of cases and controls along the top two principal components (PC1 and PC2) identified by principal component analysis. The results are visualized in comparison to the Illumina ethnicity controls (CEU, YRI and CHB-JPT). 4 different scenarios are displayed, with increasing number of outliers removed. The number of cases and controls remaining in each scenario is displayed in Table 8.*
A) *Outliers with >5 SD from the mean of the sample set are removed. A large number of samples is still scattering towards the CHB-JPT and YRI controls.*
B) *Outliers with >4 SD from the mean of the sample set are removed. A slight reduction of samples scattering towards the CHB-JPT and YRI controls can be observed.*
C) *Outliers with >3 SD from the mean of the sample set are removed. No further reduction of samples scattering towards CHB-JPY and YRI can be seen.*
D) *Outliers with >2 SD from the mean of the sample set are removed. Less samples are scattering towards CHP-JPT and YRI, but the number of samples remaining has decreased substantially.*
Scenario B appeared to be best compromise between finding the balance of a homogenous group and number of samples remaining.

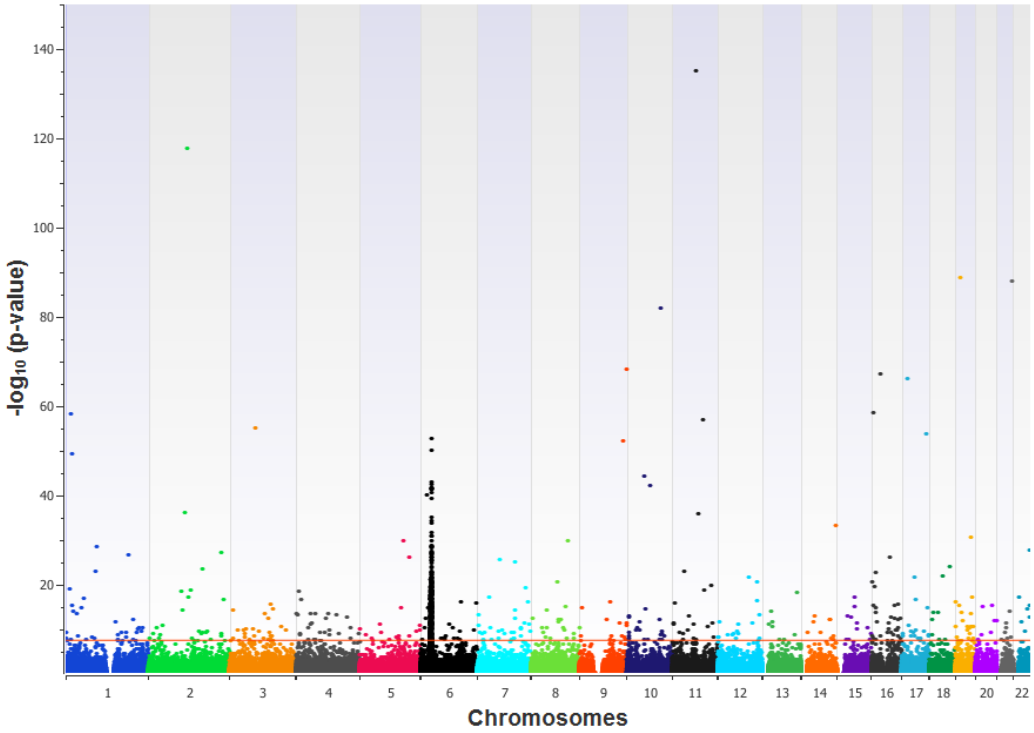
The scenario with removal of outliers > 4 SD and >3 SD deemed very similar in respect to the inflation factor lambda and visual inspection of the graph, whereas the removal of outliers with >2 SD reduced the number of cases and controls drastically with only slight improvement in the homogeneity of the group. As this was the first analysis and we wanted to keep as many cases and controls as possible in the cohort, we decided to use 4 SD as a cut-off. The following analysis was continued with a dataset after removal of European ancestry outliers with SD >4.

Results

We performed a basic allele test for association testing on the now ancestry matched 476 cases and 6,040 controls. All 216,015 markers were tested to assess the results before applying further quality control steps on markers or samples. The significance level was set at 5×10^{-8}. The results were graphically presented in form of a “Manhattan plot”.

Manhattan plot is named after the New York skyline and visualizes the distribution of markers according to their physical position on the genome (x-axis) and their significance value (y-axis). In general, when examining a Manhattan plot we are looking for “skyscrapers”, collections of markers that build-up into the “sky” of significant p-values. Thus, we are not looking for isolated markers with significant p-values, as these likely represent artefacts, but for conglomerates of markers with p-values decreasing, the tighter they are in LD with the leading SNP (the “top floor”), *i.e.* the marker with the lowest p-value. The reason why the surrounding markers have also decreased p-values is that they are in LD with the causal variant, and therefore are found in association with the trait. Markers with low p-values, but without build-up are generally considered as noise. Too much noise seen in a Manhattan plot, can mask build-ups and make it difficult to distinguish true associations. Hence, when applying QC steps the aim is that with every step the Manhattan plot becomes cleaner, meaning the number of false positive associations decreases and build-ups become more visible.

Figure 15 *Manhattan plot for BAT before applying QC steps*



Legend *Manhattan plot for association analysis before applying QC steps. The chromosomal position of the markers is represented on the X-axis corresponding to the human genome GRCh37/hg19. The level of significance is represented on the Y-axis. The red line represents the whole genome wide significance threshold level of* $-\log_{10}$ *of the p-value of* 5×10^{-8}*. A high level of noise can be seen, reflected by a large number of markers above the genome-wide significance level. This image is also referred to as "starting scenario".*

The Manhattan plot for the first association test before applying any QC steps (Figure 15) displayed that many markers reached the level of significance, not necessarily associated with the trait, representing false positive results. Only one build-up on chromosome 6 p-arm can be identified. In order to address false positive results and unmask further possible peaks, quality control steps were implemented on the markers. This first Manhattan plot of the results before applying QC steps is hereafter referred to as the starting scenario.

Chapter 2. Pre-analysis: Quality control optimization

Multiple factors influence the quality and reliability of a GWAS: the genotyping platform used, the data quality of samples, the number of cases and controls, the ethnicities of samples, *etc*.
Consequently, not one standard protocol of quality control procedures can be applied to every dataset, but they must be explored and adapted accordingly. This part of the thesis summarizes the optimization process of the quality control steps.

QC for markers

The first quality control procedures for markers were performed on the cases and controls together under the assumption that genotyping was carried out under similar conditions for all datasets. The dataset containing cases and controls was called combined dataset. This means that the cases and controls were appended (the rows representing the cases were added to the rows representing the controls, keeping only markers which were common in both datasets). After appending the case and control dataset 216,015 markers remained as common markers. We noticed that a significant number of markers got lost as they were not represented in both datasets.
The following quality control steps for markers were addressed:

- Removal of markers with an allele count of more than 2
  (Note that REMEDY removed markers which are multiallelic in dbSNP, whereas this QC step tests if any marker is multiallelic in the actual dataset and removes it)
- Removal of makers with a high missing call rate
- Removal of markers with a low minor allele frequency
- Removal of markers outside the Hard-Weinberg equilibrium

The quality control analysis was performed on 476 cases and 6,040 controls. The total number of overlapping autosomal markers between cases and controls before applying quality control steps was 216,015 markers.

First step: Allele count

We strictly wanted to exclude all multiallelic markers from our study. Any marker which contains more than two alleles has the potential to be misread by the genotyping system and is therefore a possible source of noise. Therefore, we removed any markers with an allele count greater than 2. In our dataset 33 markers had an allele count greater than 2 and were removed from further analysis. The number of remaining markers was 215,982.

Second step: Call rate per marker

The call rate of a marker reflects the quality of the DNA together with the quality of the genotyping array at that specific marker. Some markers might have intrinsic problems during genotyping due to poor hybridization with the probe resulting in a low call rate of this specific marker. To examine the effect of a low call rate on the overall association results we first looked at the overall call rate of all markers and then assessed how different cut-off levels affect the results. The results were graphically presented in form of Manhattan plots. Of 215,982 markers in the combined case-cohort dataset 0 markers had a call rate of 100%, but 201,944 markers had a call rate ≥99%, 210,410 markers ≥97% and 211,758 ≥95%.

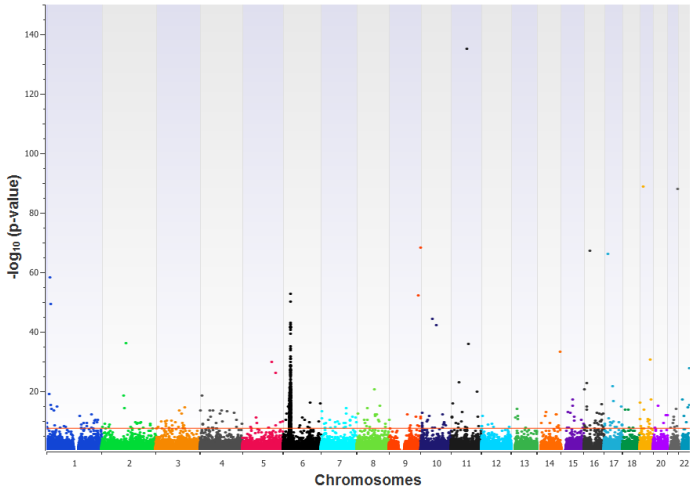
The first scenario was with removal of markers that had a CR < 95%, leaving 211,758 markers for analysis (low stringent) (Figure 16 A). After removing markers with a CR <95%, the Manhattan plot of the association result still displayed a significant amount of noise. This indicated that the cut-off for the CR might be too low. We increased the cut-off to 97% in a next step.

The second scenario was to remove all markers that had a CR <97%, leaving 210,410 markers for analysis (medium stringent) (Figure 16 B). With removal of all markers with a CR <97% the noise was reduced compared to the cut-off of 95% only. We now wanted to test if an even more stringent cut-off for the CR of markers decreases the noise further.

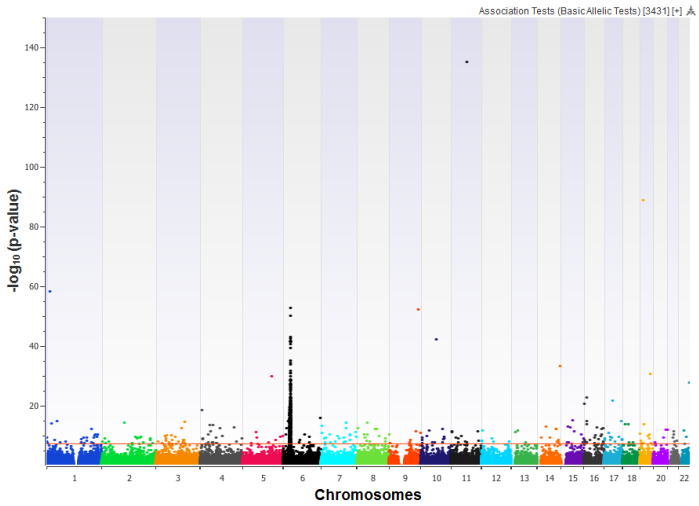
In the last step all markers with a CR <99% were removed, leaving 201,944 markers for analysis (Figure 16 C).

Figure 16 *Manhattan plot after removal of markers with increasing cut-off levels for call rate*

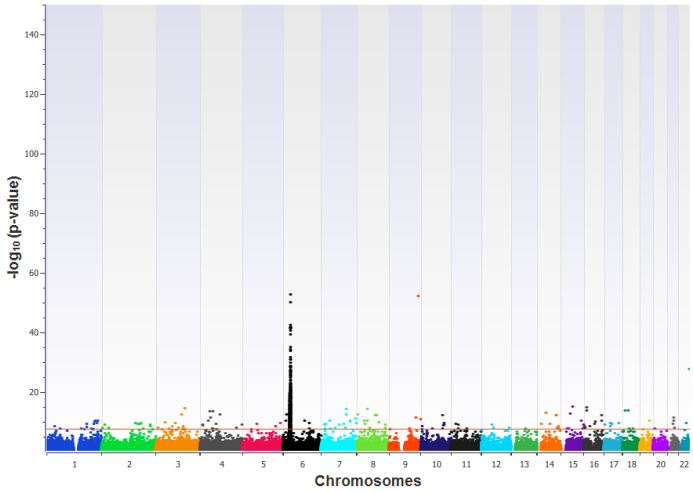
A)



B)



C)



Legend: *Manhattan plots for association analysis with increasing cut-off levels for removal of markers with a low call rate. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log$_{10}$ of the p-value is represented on the y-axis. The level of whole genome wide significance threshold is represented by the red line.*
A) *Manhattan plot of association analysis with removal of markers with a CR <95%. Despite removing markers with a CR <95% the plot continuously showed a high level of noise, indicating that this threshold for CR of markers might be too low.*
B) *Manhattan plot of association analysis with removal of markers with a CR <97%. The amount of noise has reduced from the previous scenario, however, still a large amount of markers reach significance, creating noise and masking possibly true associations.*
C) *Manhattan plot after removal of markers with a CR <99%. The level of noise has reduced further compared to the previous scenario, but the Manhattan plot continues to appear noisy.*

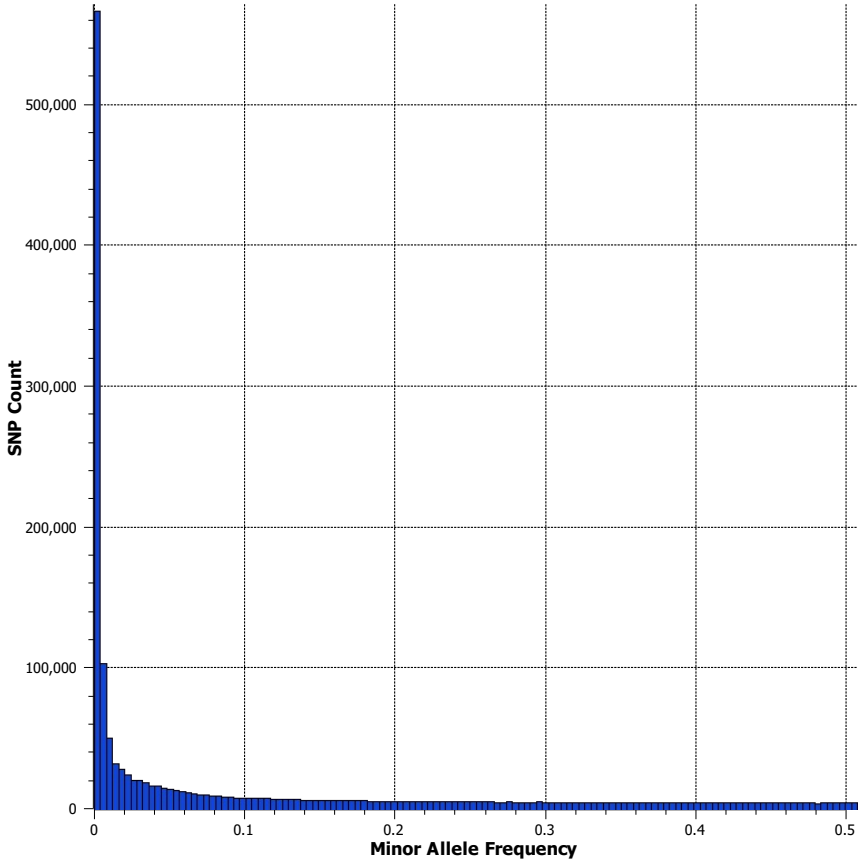
Despite stringent cut-off level for the call rate the Manhattan plot continued to appear noisy, reflecting a large number of false positive associations. This indicated that the removal of markers with a low call rate alone is not sufficient to control for false positive results. We therefore proceeded to implement the next QC step.

Third step: Minor allele frequency

The minor allele is defined as the least common allele at a given *locus* in a defined population. The frequency of this allele is called minor allele frequency. In general, variants with a low MAF can cause false positive associations. This is because the small number of the heterozygote genotypes of these variants can lead to erroneous genotype calling. Further, even when called correctly, association signals caused by these rare SNPs are less robust because they are driven by the genotypes of only a few individuals. Hence, GWAS are mainly based on common markers and therefore most SNPchips, including our control SNPchips, are designed to include mostly common markers (MAF >5%).

The SNPchip (MEGA) used for genotyping of the cases was stated to have a high number of markers with a low MAF. In order to explore the distribution of the MAF in the case dataset, which were genotyped on the MEGA SNPchip, we plotted a histogram for the MAF of the 1,513,983 markers in this dataset.

Figure 17 *Histogram for minor allele frequency of all markers in the case cohort*



Legend *It can be observed that the case SNPchip (MEGA) has a high number of markers with a low MAF. Nearly half of the markers (696,279) have a MAF <0.01.*

Note that the case cohort, had a high number of markers with low MAFs. 259,057 of 1,513,983 markers had a MAF <0.001, 696,279 markers had a MAF <0.01 and 910,234 markers had a MAF of <0.05. This illustrated that nearly half of the markers on the case SNPchip were rare markers, which are commonly defined as MAF <0.01.

A drawback of the high number of markers with a low MAF on our case SNPchip was, that most of the markers with a low MAF were not represented in the control datasets. This is the explanation for the large drop of markers, observed earlier, when combining the case with the control datasets. After combining the datasets only 11,555 markers remained having a MAF ≤0.01.

In order to assess the influence of removing markers with a low MAF further, three scenarios were compared. Removal of markers with a MAF less than 0.1% (less

stringent), less than 1% (medium stringent) and less than 5% (high stringent). The results were graphically presented in form of Manhattan plots.

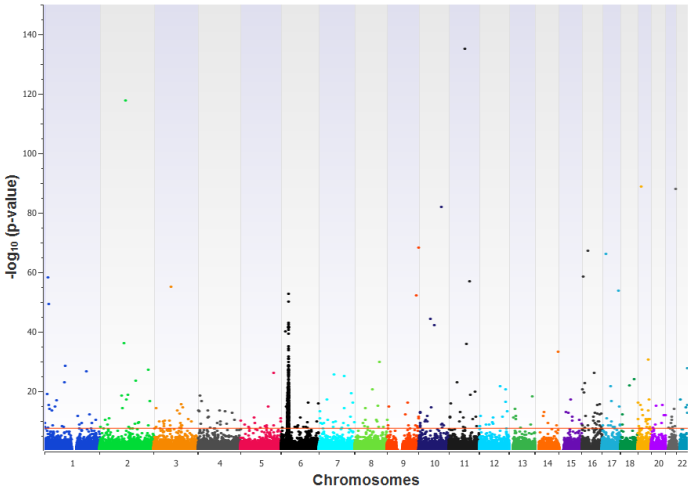
The first scenario was to remove all markers with a MAF <0.001, which was the least stringent cut-off. With that cut-off 5,449 markers were removed, leaving the majority of markers (210,533) for analysis (Figure 18 A).

We repeated the analysis with a more stringent cut-off for the MAF. 11,555 markers with a MAF <0.01 were removed, leaving 204,427 markers for analysis (Figure 18 B).

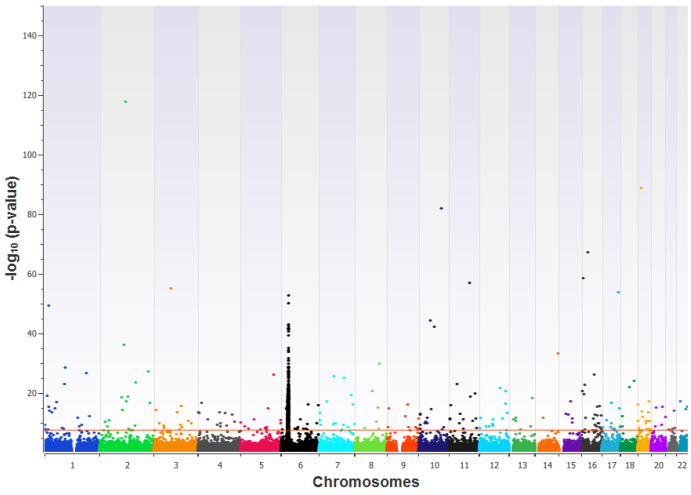
In a last step, markers were removed with a MAF <0.05, which is a common threshold used in GWAS. With that threshold 23,237 markers were removed, leaving 192,745 markers for analysis (Figure 18 C).

Figure 18 *Manhattan plot after removal of markers with increasing cut-off levels for minor allele frequency.*

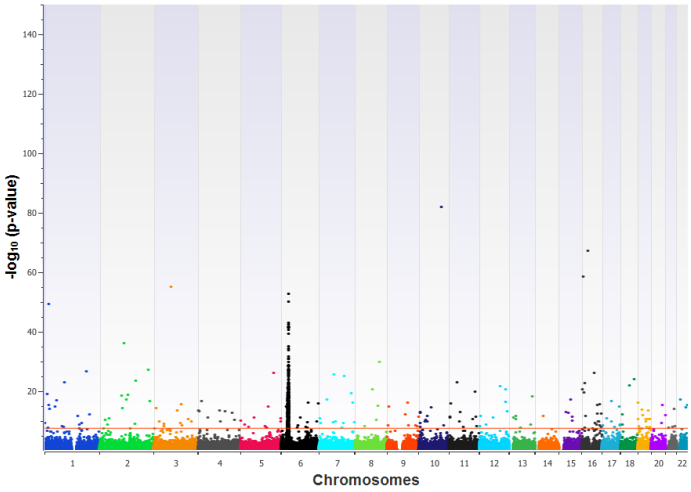
A)



B)



C)



Legend *The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log$_{10}$ of the p-value is represented on the y-axis. The level of whole genome wide significance threshold is presented by the red line.*
A) *Manhattan plot for association analysis with removal of markers with a MAF <0.001. The plot shows a similar level of noise as seen in the starting scenario. This indicates that applying a threshold for MAF of 0.1% alone is not sufficient to improve the level of noise.*
B) *Manhattan plot for association analysis with removal of markers with a MAF <0.01. In comparison to the previous plot with a MAF cut-off of 0.001, the noise has reduced slightly.*
C) *Manhattan plot for association analysis with removal of markers with a MAF <0.05. A minimal further improvement can be observed compared to the previous scenario. However, ongoing noise is masking possible true associations.*

When comparing the Manhattan plots with the three different cut-off levels for MAF, we could see that the noise was the highest when markers with a very low MAF (0.01% - 0.1%) were included in the association test. However, also with including only common markers (>5%), the plot shows a high level of noise, indicating that the MAF alone as a QC is not sufficient to remove spurious markers. Compared to the QC step CR of markers, the removal of markers with a low MAF had a much lower effect to control for false positive associations. We went on to test how much the HWE filtering impacts the outcome of the association test.

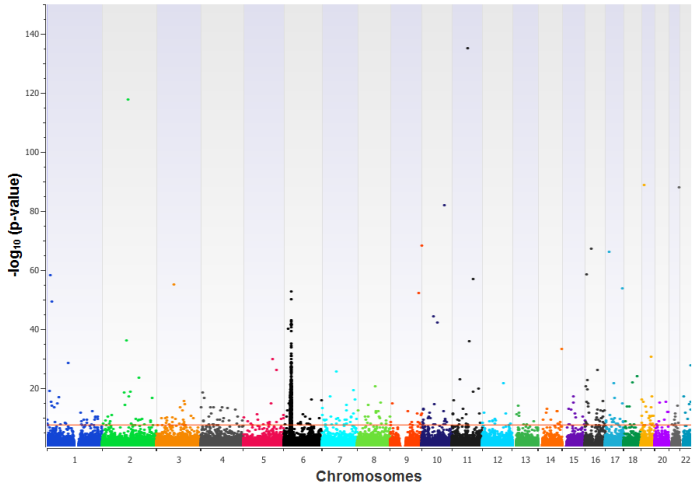
Fourth step: Hardy-Weinberg equilibrium filtering

Extensive deviation from Hardy-Weinberg equilibrium (HWE) can be a sign of a genotyping or genotype calling error and therefore most scientists remove makers that deviate from the HWE above a certain threshold. However, deviations from Hardy-Weinberg equilibrium may also indicate structural changes, and a case sample can show deviations from HWE at a *locus* associated with the disease. To remove this marker would obviously be counter-productive. Therefore, this QC step is only applied on the control dataset.

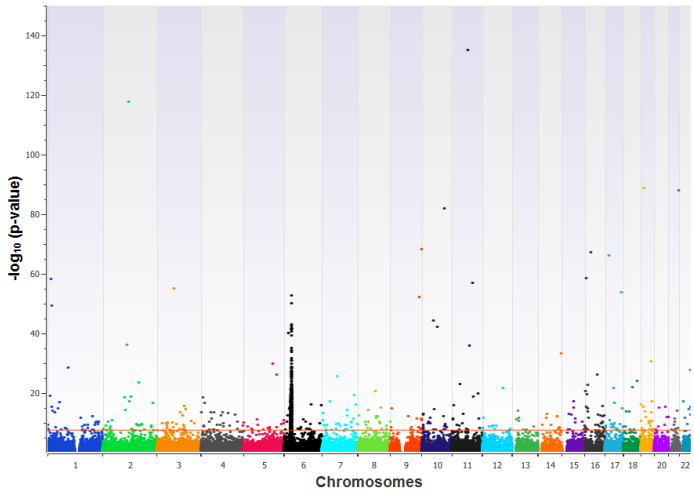
Two thresholds for HWE cut-off were tested. Low HWE (removal of markers outside the HWE with a p-value of less than 0.0001) or high HWE (removal of markers outside the HWE with a p-value of less than 0.001). With a low cut-off for the HWE on the controls ($p < 0.0001$), 962 markers were removed, leaving 215,020 for analysis. With a higher cut-off for the HWE on the controls ($p < 0.001$), 1363 markers were removed, leaving 214,619 for analysis.

Figure 19 *Manhattan plot after removal of markers with increasing cut-off levels for HWE p*

A)



B)



Legend *The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log$_{10}$ of the p-value is represented on the y-axis. The level of whole genome wide significance threshold is presented by the red line.*
The plots with the two different thresholds are displayed next to each other to illustrate that no visible improvement can be seen between the two thresholds.
A) *HWE p <0.0001*
B*) HWE p <0.001*

No visible difference was observed between the two cut-offs.

After testing each of the QC steps separately an overview of the number of markers removed with each step is provided in the table below.

Table 9 *Overview of the number of markers removed by each QC step*

Total SNPs			216,015
QC step	**Threshold**	**Removed**	**Remaining**
Allele count	>2	33	215,982
Call rate	<0.95	4,224	211,758
	<0.97	5,581	210,410
	<0.99	14,038	201,944
Minor allele frequency	<0.001	5,449	210,533
	<0.01	11,555	204,427
	<0.05	23,237	192,745
HWE (ctrls)	p <0.0001	962	215,020
	p <0.001	1,363	214,619

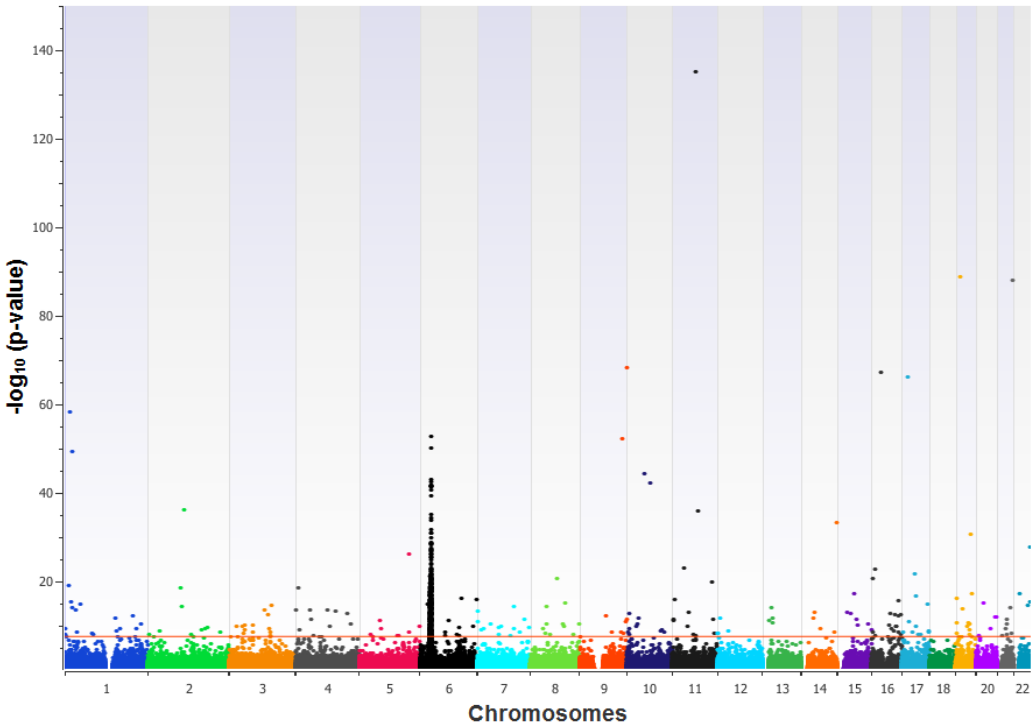
The testing of each QC step separately helped to get a feeling of how each QC step influences the number of markers in the dataset and the association results. Nevertheless, one QC alone did not control sufficiently for false positive associations. We assumed that the combination of the QC is necessary to address false positive results adequately and therefore examined different scenarios of combining the QC steps.

We tested 3 different scenarios of increasing stringency of the quality control steps. For all 3 scenarios a BAT was performed and the inflation factor calculated. The results are demonstrated as a Manhattan plot.

First scenario - Low stringency

In the first scenario the most liberal cut-off levels were used for each QC steps. CR <0.95, MAF <0.001 and HWE on controls p <0.0001. This scenario is referred to as low stringency. Markers remaining for analysis were 205,886. (Figure 20)

Figure 20 *Manhattan plot for scenario 1 – low stringency*



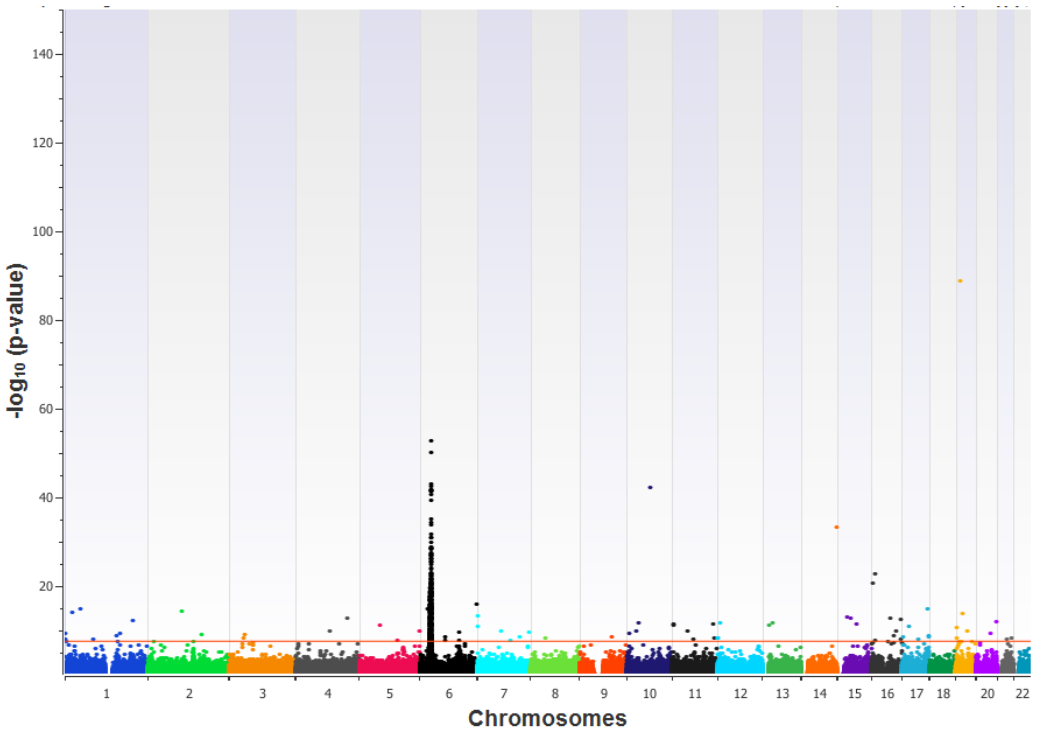
Legend *Manhattan plot for association analysis of 476 cases and 6040 controls after removal of markers with a CR <0.95, MAF <0.001 and HWE on controls p <0.0001. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log$_{10}$ of the p-value is represented on the y-axis. The level of whole genome wide significance threshold is presented by the red line. Inflation factor Lambda: 1.29233. Compared to the starting scenario, where no QC steps were applied, a reduction of noise can be seen.*

We immediately can see an improvement of the results in regards to false positive associations compared to the starting scenario, where no QC steps were applied. However, a significant amount of noise is still visible. Therefore, the quality control steps for this scenario were considered as insufficient to control for false positive results.

Second scenario – medium stringency

In the second scenario we increased the cut-off levels for each QC step to make them more stringent. CR <0.97, MAF <0.01, HWE on controls p <0.001. The number of remaining markers was 199,725. (Figure 21)

Figure 21 *Manhattan plot for scenario 2 – medium stringency*



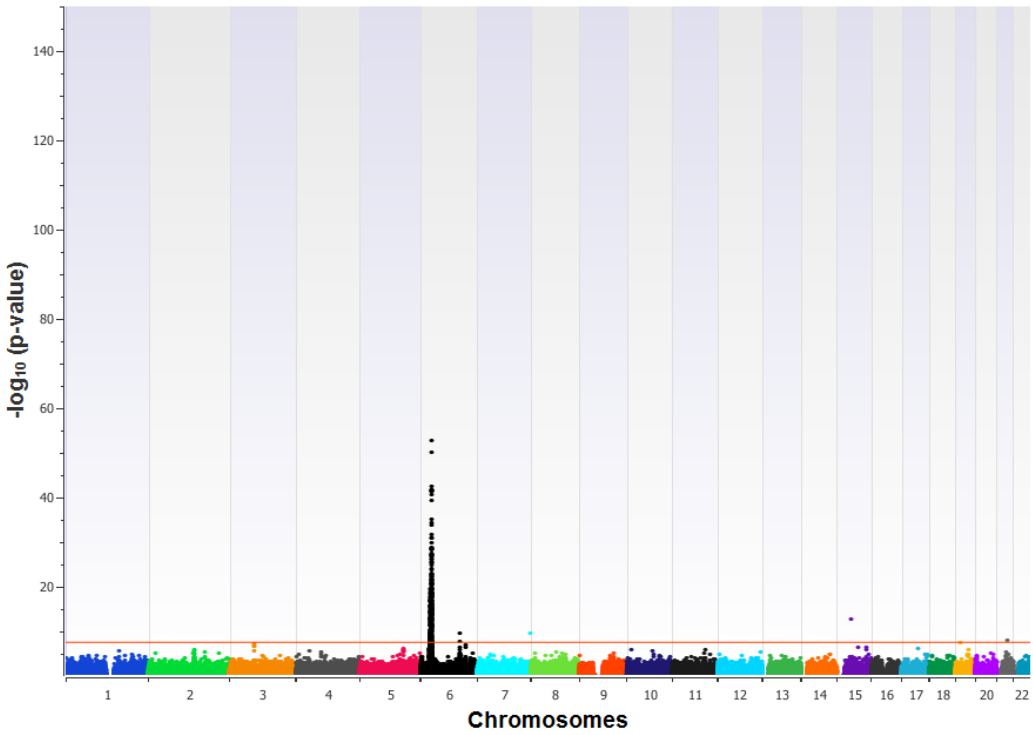
Legend: *Manhattan plot for association analysis of 476 cases and 6040 controls after removal of markers with a CR <0.97, MAF <0.01 and HWE on controls p <0.001. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log₁₀ of the p-value is represented on the y-axis. The level of whole genome wide significance threshold is presented by the red line. Inflation factor Lambda: 1.26453. A significant reduction of noise compared to scenario 1 was observed. Peaks are becoming more visible, however because of ongoing background noise, the demarcation of peaks is still difficult.*

The level of noise has visibly reduced in this scenario. When looking at the Manhattan plot, peaks are starting to unmask. But a clear demarcation of peaks is not possible yet.

Third Scenario – high stringency

In the third scenario we increased the cut-off levels for each QC step to make them more stringent. CR <0.99, MAF <0.05, HWE on controls p <0.001. The number of markers reduced by another approx. 15,000 and 182,331 remained for analysis. (Figure 22)

Figure 22 *Manhattan plot for scenario 3 – high stringency*



Legend: *Manhattan plot for association analysis of 476 cases and 6040 controls after removal of markers with a CR <0.99, MAF <0.05 and HWE on controls p <0.001. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log$_{10}$ of the p-value is represented on the y-axis. The level of whole genome wide significance threshold is presented by the red line. Inflation factor Lambda: 1.2319. Two peaks on chromosome 6, reaching genome wide significance are now clearly demarked. We can see two further markers on chromosome 7 and 15 above the threshold line.*

With every step the Manhattan plot became cleaner and build-ups more visible. When looking at the Manhattan plot of this last scenario, we can see that a second peak was unmasked on chromosome 6 q-arm, which was not visible in the previous scenarios. Therefore, scenario 3 deemed to be the best combination of quality control steps to control for noise.

Until now, QC steps were applied on the combined case control dataset. Considering that genotyping for all 4 datasets was performed on different platforms and therefore the quality of data may vary significantly, we speculated that it might be better to apply the QC steps on each dataset separately before combing them. In order to see if there was a difference between applying QC steps on the combined dataset or on each dataset separately, the scenarios were repeated but with the QC filtering steps applied on each dataset separately.

QC filtering on separate datasets

The following quality control analysis was performed on the case dataset and each control dataset separately. Cases and controls were combined after performing the QC steps. Again, we tested 3 different scenarios of increasing stringency of the quality control steps. Based on the previous results we knew that the cut-off for the MAF of markers included did not seemed to influence the level of noise very drastically and we hence decided to keep markers down to a MAF of 1%. The main impact on the noise was seen from markers with a low genotyping quality, therefore we focused at the high stringency scenario only with respect to CR cut-off. Table 10 provides and overview of how many markers were removed from each dataset with the different scenarios.

Table 10 Overview of the number of markers removed with different stringency scenarios

Stringency level	Total	Low	Medium	High
Cases	1,513,982	867,512	662,700	614,712
Oxford	653,959	608,107	579,926	571,573
Illumina	653,959	587,523	577,781	517,212
WTCCC	1,025,437	872,212	824,265	789,061
Combined datasets	216,015	194,591	186,230	156,794
Final markers	**216,015**	**194,507**	**186,121**	**156,712**

Legend Three levels of stringency were tested: low (AC >2, CR <0.95, MAF <0.001, HWE ctrls p <0.0001), medium (AC >2, CR <0.97, MAF <0.01, HWE ctrls p <0.001) and high (AC>2, CR <0.99, MAF <0.01, HWE ctrls p <0.001). After combing the datasets the QC steps were repeated on the combined dataset, leaving the final number of markers.

For all 3 scenarios a BAT was performed and the inflation factor calculated. The results are demonstrated in Figure 23.

First scenario - Low stringency

In the first scenario markers were removed from each of the 4 datasets if they had a AC >2, CR <0.95 and MAF <0.001. In the control set additionally markers with HWE p <0.0001 were removed. Thereafter the 4 datasets were combined to a case-cohort dataset. The number of markers in the combined dataset was 194,591. The above

mentioned QC steps were repeated on the combined datasets with HWE testing on controls only. The remaining number of markers was 194,507. Association testing was performed on 476 cases and 6040 controls. (Figure 23 A)

We visually compared the Manhattan plot of this scenario to the same scenario with QC steps applied only on the combined cases and controls (Figure 20). It was clearly recognisable that the Manhattan plot for this scenario displayed less noise and build-ups were more visibly identifiable. This confirmed our assumption that it is better to apply the QC steps on each dataset separately before combing them.

Second scenario – medium stringency

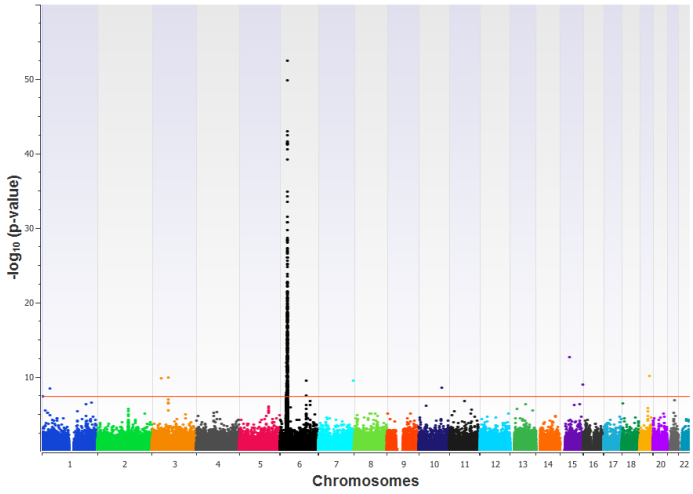
In the second scenario the cut-off levels for all QC steps were set slightly more stringent. Markers were removed from each of the 4 datasets if they had AC >2, CR <0.97 and MAF <0.01. In the control set additionally markers with HWE p <0.001 were removed. Thereafter the datasets were combined leaving 186,230 markers. After repeated QC on the combined dataset, 186,121 markers were brought forward for association analysis. (Figure 23 B)

Third scenario – high stringency

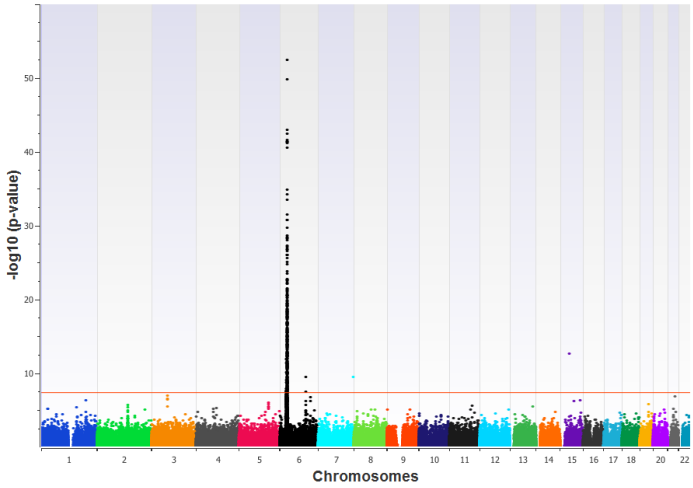
In the third scenario only the cut-off for the call rate was increased further to 99%. After performing the QC on each dataset separately and then combining the dataset, 156,794 markers were left. After repeated QC on the combined dataset 156,712 markers were brought forward for association analysis. (Figure 23 C)

Figure 23 *Manhattan plot after applying QC steps on datasets separately. Scenarios 1 – 3.*

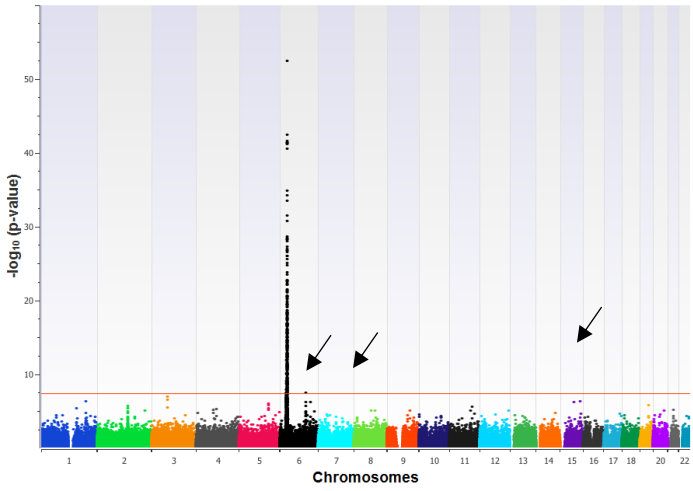
A)



B)



C)



Legend *Manhattan plot of association analysis of 476 cases and 6040 controls with the QC steps per marker applied on each dataset separately before combining them. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log p-value of each marker on the y-axis. The level of whole genome wide significance threshold is presented by the red line. Note, the y-axis range has been reduced to a maximum of 60 to inspect the lower part of the graph better and no markers were reaching -log of the p-value above 60.*
A) *Scenario 1 – low stringency. Manhattan plot after removal of markers with a CR <0.95, MAF <0.001 and HWE on controls p <0.0001 in each dataset separately. Inflation factor Lambda: 1.2358. The level of noise is much lower compared to scenario 1 (Figure 20) with QC steps applied on the combined dataset only.*
B) *Scenario 2 – medium stringency. Manhattan plot after removal of markers with a CR <0.97, MAF <0.01 and HWE on controls p <0.001 in each dataset separately. Inflation Factor Lambda: 1.2238. A further reduction of noise can be seen compared to the previous scenario. Build-ups are clearly distinguishable from the background.*
C) *Scenario 3 – high stringency. Manhattan plot after removal of markers with a CR <0.99, MAF <0.01 and HWE on controls p <0.001 in each dataset separately. Inflation factor lambda: 1.2167. The results yield very similar to the previous scenario. 3 markers on chromosome 6, chromosome 7 and chromosome 15 each, reaching significance in the previous scenario have disappeared now (arrows).*

The results between the last two scenarios are very comparable when looking at the Manhattan plot. For the main peak on Chromosome 6 (p-arm) no differences can be seen by visual inspection of the Manhattan plot, where the density of markers reaching genome wide significance is high. However, 3 markers, which reached significance in the previous scenario are now removed. This is on the Chromosome 6 q-arm the marker rs479536, on Chromosome 7 the marker rs2302443 and on Chromosome 15 the marker rs1898882. These might be true associations and therefore we will look at both scenarios in the final dataset.

QC for samples

After we explored the different QC steps for markers, we went on to examine the samples in more detail. We already removed individuals with a call rate < 90%, as this is recommended by guidelines to do before applying quality control steps on markers. The rationale is, that in a large dataset the removal of a small number of individuals should have little effect on overall power. In contrast, every marker removed from a study is potentially an overlooked disease association and thus the impact of removing one marker is potentially greater than the removal of one individual. Removing individuals with a low call rate first prevents markers being erroneously removed due to a subset of poorly genotyped individuals
In accordance with published guidelines and most papers on GWAS we decided on following quality control steps for samples:

- Removal of duplicates or related individuals
- Removal of individuals with outlying heterozygosity rates of more than 3 standard deviations below or above the mean for the overall samples

Step one: Removal of duplicates or related individuals

Association studies are designed to investigate genetic differences in an independent, not related case and control cohort. This is in contrast to other study settings, *e.g.* linkage analysis, where genetic differences among affected and unaffected family members are studied. In association analysis, any related samples or duplicates could bias the results and should therefore be removed. Of course, already during patient recruitment an individual should not be recruited twice, as well as it should be clear if there are related individuals in the cohort. However, as relation status is not always clear or documented, as well as the combining of samples from different collaborators could also lead to potential errors in respect to this matter it is important to evaluate with an objective method if there are duplicates or related samples in a cohort.

A way to identify duplicates or related samples is by examining how much of the genetic material is shared between two samples. This is reflected in the identity by descent value (IBD). A calculated IBD >0.98 represents duplicate samples, IBD >0.50 would represent 1st degree relatives (parent offspring situation or siblings), IBD >0.25 would represent 2nd degree relatives (1st cousins). As those cut-offs are theoretical and 1st degree relatives not always share exactly 50% of their genetic information, but can also share *e.g.* 45% or 55% the cut-offs used in reality are IBD >0.375 to remove 1st degree and IBD >0.1875 to remove 2nd degrees relatives.

Based on this idea, we first tested the influence of different cut-offs for the IBD calculation using PRIMUS. Results are displayed in the following table.

Table 11 *Overview of duplicate and related samples*

	Total (CR≥0.90)	**Duplicates**	**1st degree relatives**	**2nd degree relatives**	**Remaining**
IBD cut-off		>0.98	>0.375	>0.1875	
SSNS	709	38	18	2	**651**
OXF	432	4	1	1	**426**
ILUM CEU	90	0	30	0	**60**
WTCCC	5549	10	44	13	**5482**

Legend: *IBD >0.98 represents duplicate samples, IBD >0.375 represents 1st degree relatives with taking into account variations around the theoretical cut-off of 0.5. IBD >0.1875 represents 2nd degree relatives, taking in to account the variation around the theoretical cut-off level of 0.25.*

In GWAS a general accepted cut-off level above which samples should be removed is 0.1875 representing 2nd degree relatives. We also used this cut-off level in the downstream analysis.

Step two: Removal of samples with deviating heterozygosity rates

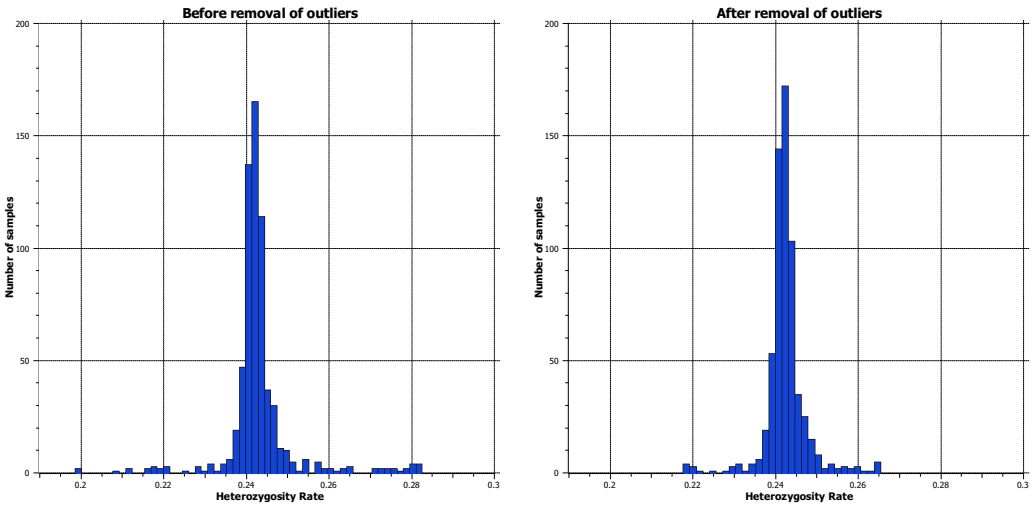
Usually the rate of heterozygous genotypes of samples in a dataset follows a normal distribution. Samples with an either very high or very low rate of heterozygous genotypes can indicate poor sample quality and should be removed from further analysis. Therefore, samples with a heterozygosity rate more than 3 standard deviations below or above the mean for the overall samples were removed.

For the case dataset, 26 samples were deviating more than 3 SDs from the mean heterozygosity rate and removed from the dataset. For the Oxford controls, 6 samples showed a high or low rate of heterozygous genotypes and were removed. For the Illumina CEU controls 1 sample was removed because of extreme heterozygous rate and for the WTCCC controls 82 samples were removed.

The heterozygous rate of all cases before and after removal of outliers is plotted as a histogram in Figure 24 to better understand the impact of removal of outliers.

Figure 24 *Histograms for heterozygosity rate distribution of cases before and after removal of samples with >3 SD from the mean*

A) ***B)***



Legend: *Histogram of the heterozygosity rate of the cases before and after removal of outliers. The heterozygosity rate follows a normal distribution. Outliers with a low or high heterozygosity rate can be seen in graph* ***A)****. In graph* ***B)*** *those outliers were removed.*

Overview QC steps per sample

After applying the above-mentioned QC steps per sample on each dataset, 625 cases were remaining, and a total of 5,879 controls. An overview of how many samples were removed with each QC steps is given in Table 12.

Table 12 *Overview of number of samples removed with each QC step*

Dataset	Total	CR <0.90	IBD >0.1875	Het rate </>3SD	Final
Cases	712	3	58	26	**625**
Oxford	432	0	6	6	**420**
Illumina CEU	90	0	30	1	**59**
WTCCC	5604	55	67	82	**5400**

Legend *CR: Call rate; IBD: Identity by descent; Het rate: Heterozygosity rate, SD: Standard deviation*

With testing the different QC steps for markers and samples we got a better understanding of how each of them affects our dataset. With that knowledge we went back to the original dataset and applied each of the QC steps systematically on samples and markers.

Chapter 3. GWAS European cohort

After assessing the effect of each QC step on our dataset and identifying the optimal combination, we repeated the analysis for the European cohort. The quality control steps were applied to each dataset separately.

Cases

The dataset consisted of 712 cases from collaborators from Europe. The number of markers after *REMEDY* processing was 1,565,259 of which 1,513,983 were autosomal. Previously described quality control steps were applied. First, samples were removed with a CR <0.90 (n=3) leaving 709 cases.

Then, quality control steps on markers were performed before proceeding with further QC steps for samples. Markers were removed because of CR <99% (n=222,889), MAF <1% (n=700,397) and multiallelic (n=0). The remaining number of markers were 669,943.

In a next step quality control per samples was performed removing duplicated and related samples with an IBD >0.1875 (n=58) and samples with a heterozygosity rate >3SD from the mean (n=26). The remaining number of cases was 625.

Controls

Oxford controls

This control dataset consisted of 432 controls of European ethnicities. The number of markers after *REMEDY* processing was 672,361 of which 653,959 were autosomal. No samples had a CR <90.

Markers were removed because of CR <0.99 (n= 13,310), MAF <0.01 (n= 67,724) and HWE p <0.001 (2,889). The remaining number of markers after QC was 571,616.

Six samples had an IBD >0.1875 and were removed from further analysis, leaving 426 samples. Six samples had heterozygosity rate with >3SD away from the mean and was removed, leaving 420 samples in the dataset.

Illumina ethnicity controls

This control dataset consisted of 90 European samples. The number of markers after *REMEDY* processing was 672,361 of which 653,959 were autosomal. No sample had a CR<90.

Markers were removed because of CR<0.99 (n= 79,777) and MAF<0.01 (n= 68,185). Markers in the control datasets were removed if they were outside the HWE with p<0.001 (n=1,795). The remaining number of markers after QC was 516,372**.**

Thirty samples had an IBD >0.1875 and were removed from further analysis, leaving 60 samples. One sample had a heterozygosity rate >3SD away from the mean and was removed, leaving 59 samples in the dataset.

WTCCC controls

This control dataset consisted of 5604 controls of European ethnicities. 55 samples had a CR<0.90 and were removed leaving 5,549 samples for further analysis. The initial number of markers was 1,065,696, with 1,025,437 autosomal markers.

Ten samples had an IBD >0.1875 and were removed from further analysis, leaving 5482 samples. 82 samples had heterozygosity rate with >3SD from of the mean and were removed, leaving 5400 samples in the dataset.

Markers were removed because of CR <0.99 (n=133,778), AC >2 (n=0), MAF <0.01 (n=135,992) and HWE p <0.001 (16,941). The remaining number of markers after QC was 788,849.

Combining datasets

After performing the quality control steps on each control set separately, the datasets were combined to a common control set. The combined number of controls was 5,879 with 372,137 overlapping markers.

Those were then combined with the case dataset. The dataset for further analysis consisted of 625 cases and 5,879 controls with 158,314 overlapping markers.

The QC steps for markers were repeated with HWE testing on controls only. The remaining markers were 158,217.

Ethnicity selection

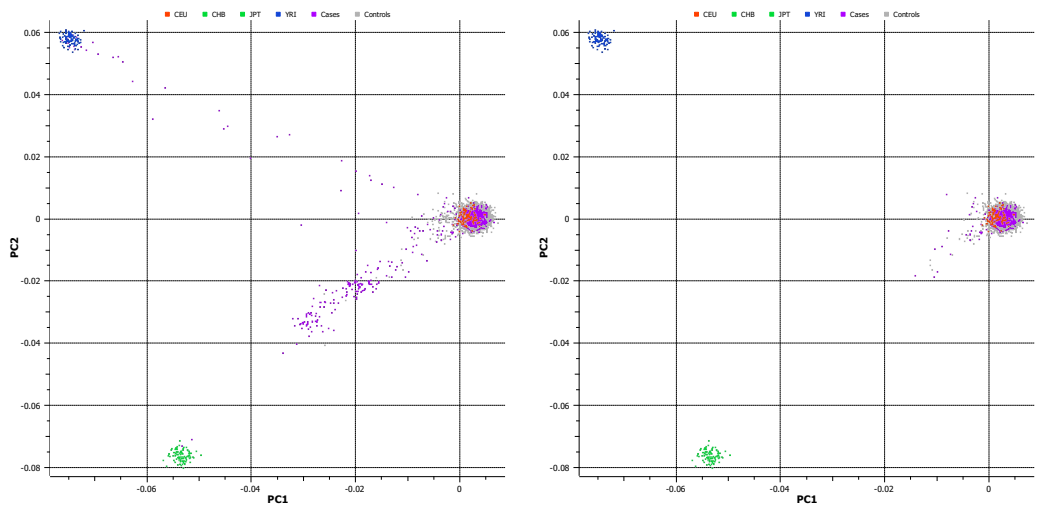
We repeated the PCA on this dataset with the aim to select a homogenous group of cases and controls. The results were visualised as a scatter plot of the first two principal components and were plotted together with the Illumina ethnicity controls (CEU, YRI and CHB-JPT). European ancestry outliers were identified and stepwise removed. As a cut-off for removal of outliers, the standard deviations 4, 3 and 2.5 were tested.

With a SD of 4, 192 cases and 62 controls were removed. With a SD of 3, 203 cases and 237 controls were removed and with a SD of 2.5, 255 cases and 1,028 controls were removed. (Figure 25)

The scenario with removal of outliers with >3 SD deemed to be the best compromise between getting an ancestry matched group and loosing as little cases and controls as possible. The results are visualized in comparison to the Illumina ethnicity controls (CEU, YRI and CHB-JPT) (Figure 25).

Figure 25 *Scatterplot for PCA of European dataset after optimizing QC steps*

A) ***B)***



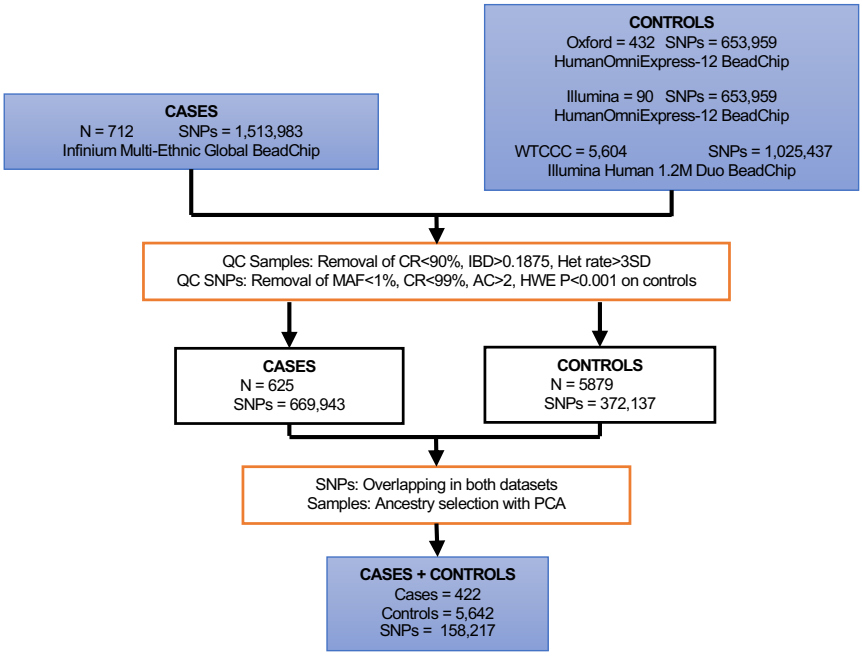
Legend *Scatter plot for PCA for European ancestry selection*
A) *prior to the exclusion of non-European individuals including all cases (n=625) and controls (n=5,879). The scatter plot shows the distribution of cases and controls along the top two principal components (PC1 and PC2). The results are visualised in comparison to the Illumina ethnicity controls (CEU, YRI and CHB-JPT)*
B) *after ancestry selection for Europeans by removal of cases and controls with principal components of >3 SD from the mean. Number of remaining cases (n=422) and controls (n=5,642).*

After PCA with removal of outliers with SD >3 422 cases and 5,642 ancestry matched controls remained in the dataset.

Summary QC steps

An overview of all QC steps and filtering process is provided in Figure 26.

Figure 26 *Flow chart of quality control steps leading to final dataset of pre-imputation GWAS*



Legend *Flowchart providing information on data input and processing. QC: quality control; CR: call rate; IBD: identity by descent; Het rate: heterozygosity rate; MAF: minor allele frequency; HWE: Hardy-Weinberg equilibrium.*

GWAS power calculation

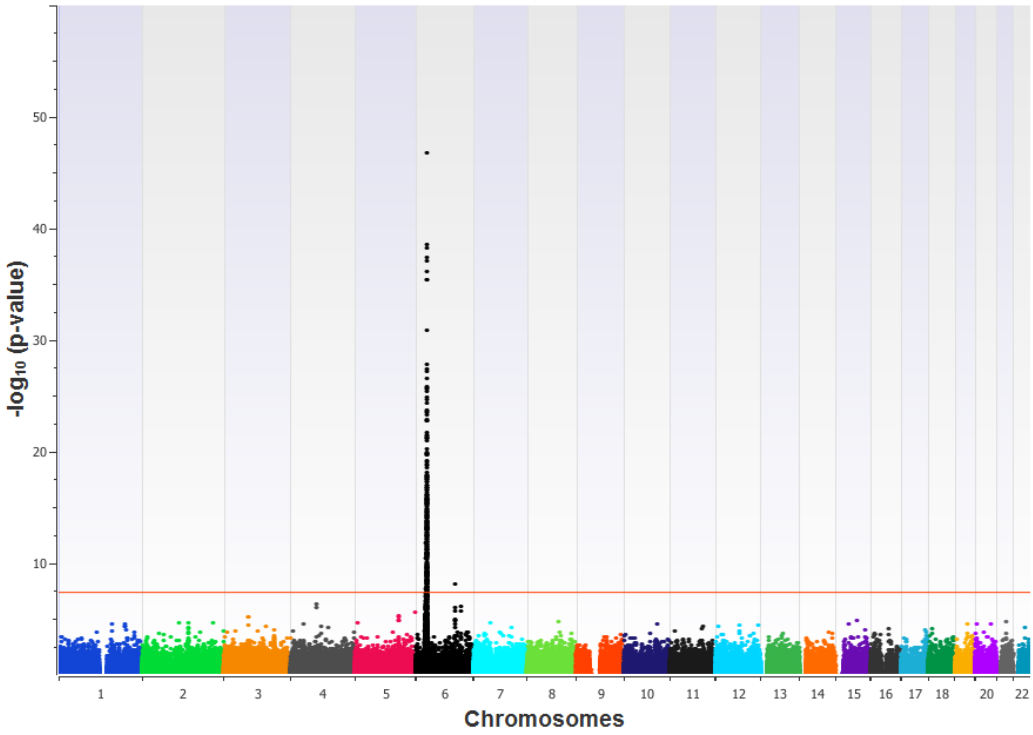
Comparing 422 cases with 5,642 controls using alpha = 5×10^{-8} under an additive model, the power to detect association of an allele with a frequency of 0.1 in controls exceeds 0.8 at a genotype relative risk (GRR) of 2.19 and power to detect the effect of more common alleles exceeds 0.8 at smaller GRRs.

Results

A basic allele test was performed on 422 cases versus 5,642 controls with 158,217 markers. No significant inflation was observed (inflation factor lambda 1.10045). 246

markers were associated with the trait at genome-wide significance level. When inspecting the results of the BAT graphically in form of a Manhattan plot two peaks on chromosome 6 were detected (Figure 27).

Figure 27 Manhattan plot for BAT of European GWAS



Legend *Displayed is the Manhattan plot for the analysis of 422 cases and 5,642 controls with the 158,217 genotyped SNPs. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log p-value of each marker on the y-axis. The level of whole genome wide significance threshold is presented by the red line. Inflation factor lambda: 1.10045 Two peaks are reaching genome wide significance, on Chromosome 6 p-arm and q-arm.*

All of 246 markers reaching genome wide significance were on Chromosome 6, and all but one on the p-arm. One marker (rs648210) was on the q-arm of Chromosome 6.

We noticed that 3 markers which reached significance in the results of the pre-analysis basic allele test were not represented in this dataset (on Chromosome 6 q-arm, on Chromosome 7 and on Chromosome 15). This could be secondary to the stringent cut-off of 99% for CR of markers. Before analysing the association results in more

detail we wanted to repeat the analysis with a cut-off of 97% for CR of markers. The idea was not to miss possibly true positive associations with the disease.

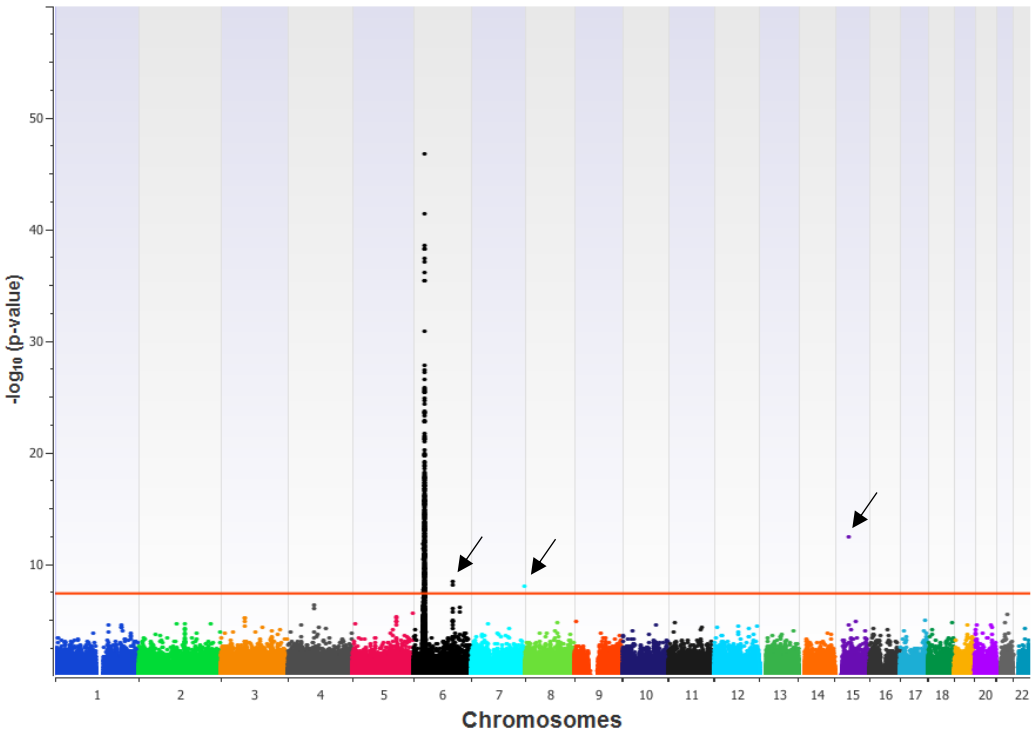
Repeated European GWAS with CR cut-off 97%

As explained above, the selection of the optimal combination of QC steps is a balance between reducing false positive results and retaining possibly true associations. When we expected the results of association analysis during the QC improvement process as well as the results of the association analysis after imputation (explained in the following chapter), we could identify 3 more markers in association with the disease. However, these markers got filtered out during the QC steps with a cut-off of 99% for the call rate of markers. This suggested that the cut-off chosen for the CR caused the removal of a possibly true associations and we wanted to investigate the results with a slight less stringent cut-off for call rate.

In the repeated analysis markers were removed because of CR <0.97, MAF <0.01, and for controls HWE p <0.001. The remaining number of markers after QC were 430,242 markers for the case dataset, 578,036 markers for the Illumina controls, 580,024 markers for the Oxford controls and 824,485 markers for the WTCCC controls. Duplicated and related samples were identified by using PRISMA with a cut-off of 0.187 and samples with a heterozygosity rate with >3SD of the mean were removed. The number of remaining samples was identical to the previous analysis, with a total of 625 cases and 5879 controls. After combining the datasets and selection for Europeans, 187,163 markers and 6,064 samples (422 cases, 5,642 controls) were brought forward to association analysis.

A basic allele test was performed on the dataset showing 291 markers associated with the trait at genome-wide significance threshold. The results are displayed graphically as a Manhattan plot (Figure 28).

Figure 28 *Manhattan plot for BAT of European GWAS with CR for markers 97%*



Legend *Shown is the Manhattan plot for the analysis of 422 cases and 5,642 controls with the 187,163 genotyped SNPs. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log p-value of each marker on the y-axis. Same QC criteria were used as in the previous analysis, except for a cut-off for the CR of markers of <0.97. Note that on Chromosome 6 q-arm additionally the marker rs549262 reaches level of significance, as well as one marker on Chromosome 7 and one on Chromosome 15.*

The associated markers are distributed on 3 chromosomes as follows: 287 markers were on chromosome 6 p-arm, two markers on chromosome 6 q-arm, one marker on chromosome 7 (rs2302443) and one on chromosome 15 (rs1898882). An overview for the top markers at each *locus* reaching genome wide significance is given in Table 13. Each *locus* was explored in more detail in the downstream analysis.

Table 13 Lead SNPs of the *loci* reaching genome wide significance

Locus	Gene	SNP	Minor allele	MAF cases	MAF controls	OR	p-value
6p21.3	*Upstream NOTCH4*	rs479536	T	0.22	0.07	3.47	1.87×10^{-47}
6p21.3	*Upstream HLA-DQB1*	rs4947342	A	0.45	0.24	2.63	4.90×10^{-42}
6p21.3	*Upstream HLA-DRA*	rs9501626	A	0.27	0.11	2.85	3.70×10^{-39}
6q22.1	*DSE*	rs549262	A	0.30	0.40	0.63	4.30×10^{-9}
6q22.1	*FAM162B*	rs648210	G	0.33	0.43	0.65	8.73×10^{-9}
7q36.3	*NOM1*	rs2302443	C	0.55	0.44	1.50	1.22×10^{-8}
15q15.1	*DISP2*	rs1898882	G	0.57	0.44	1.68	4.39×10^{-13}

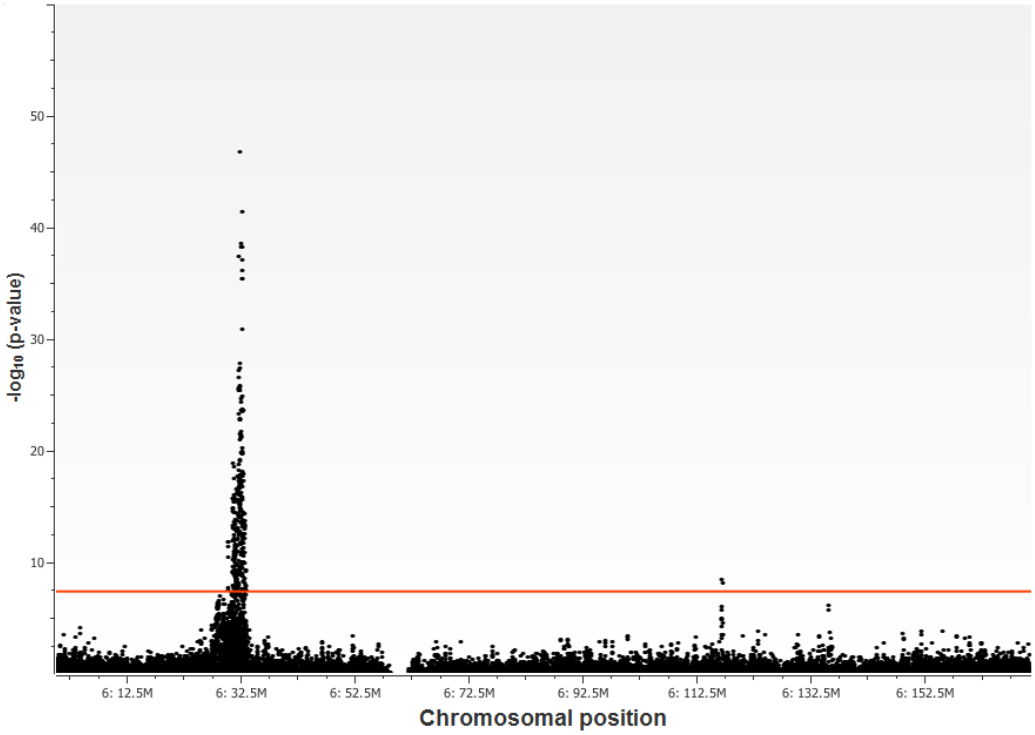
Legend Minor allele frequencies (MAF) and p-values of the lead SNPs for each *locus* reaching significance. Standard filtering steps with a call rate cut-off of 97% for markers was used for quality control. Basic allele test was used for association testing.

Description of regions of association

Chromosome 6

The association results on chromosome 6 were divided in two peaks. One on the p-arm and one on the q-arm of the chromosome (Figure 29).

Figure 29 *Manhattan plot for chromosome 6*



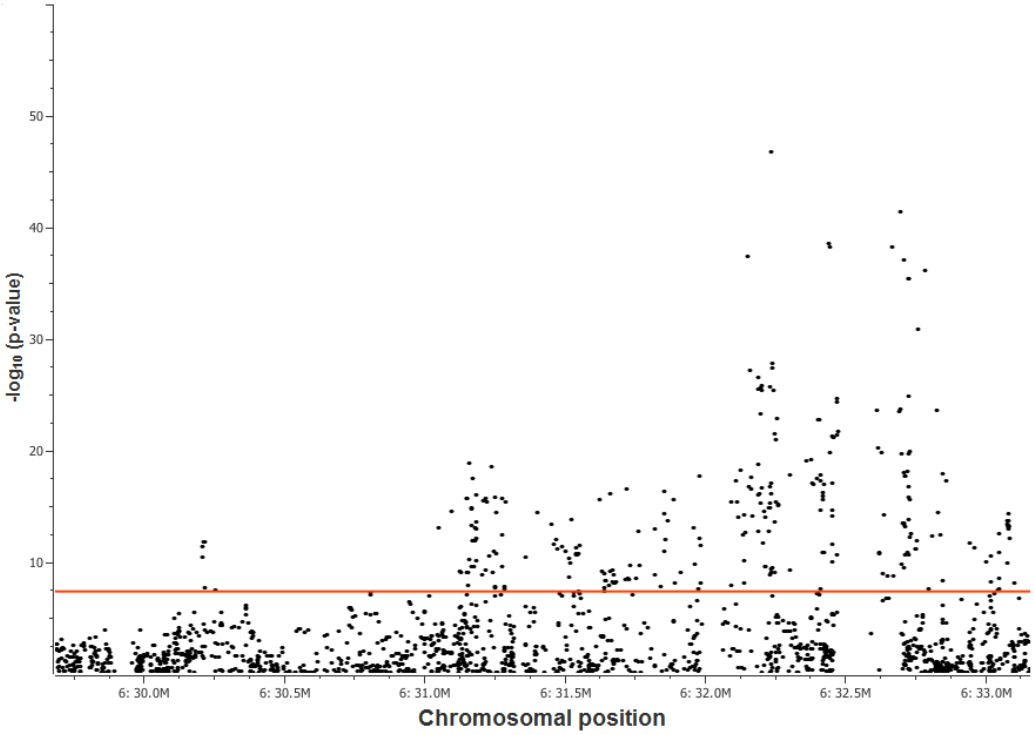
Legend *Manhattan plot for BAT of 422 cases and 5,642 controls with 187,326 genotyped SNPs zoomed in on Chromosome 6. The p and the q-arm are separated by the centromere - the area where no markers are present. Two peaks reaching genome wide significance (indicated as the red line) are visible, one on the p-arm of the chromosome and one on the q-arm.*

The two observed peaks are on each side of the centromere. We therefore investigated them separately.

Chromosome 6 p-arm

The 287 SNPs clustered on chromosome 6 in the 6p30.0 to 6p33.5 region which corresponds to the HLA-MHC *locus* (Figure 30). For the purpose of this study we refer to the classical region spanning from 6:29,640,169 – 6:33,099,120 for the GRCh37/hg19 assembly. The delimiters are the *ZFP57* gene on the telomeric side and the *HLA-DPA3* on the centromeric side [107].

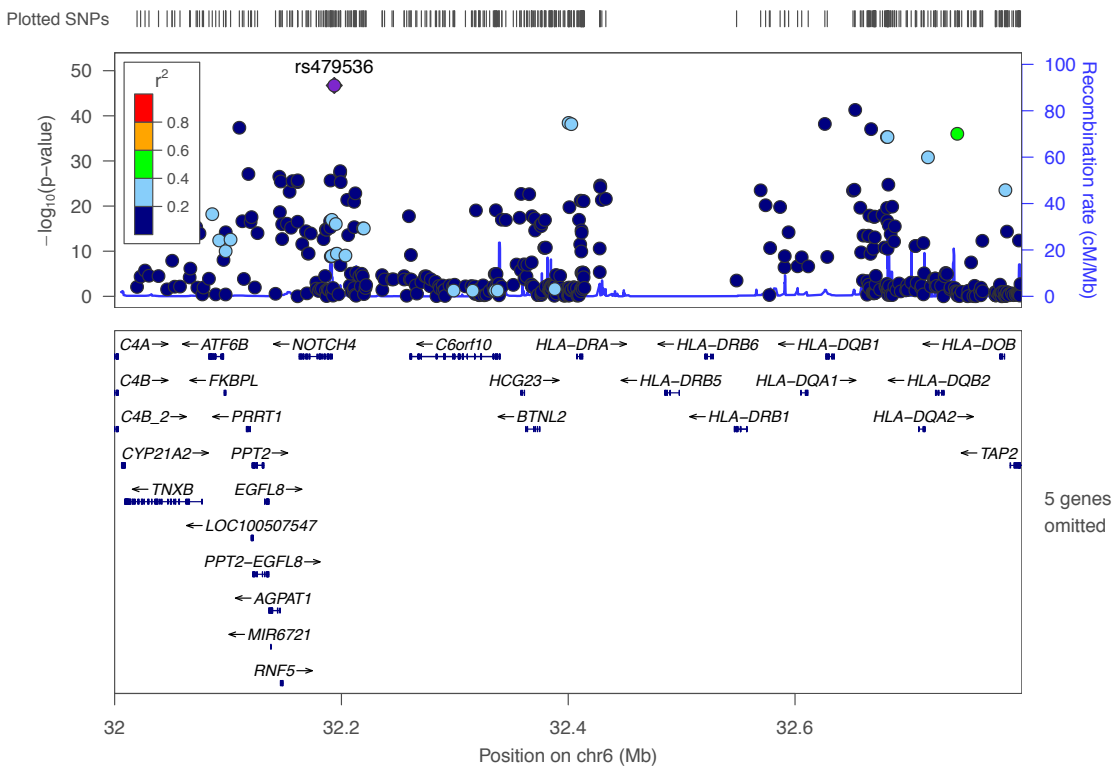
Figure 30 *Manhattan plot for classical HLA region on chromosome 6*



Legend Manhattan plot for BAT of 422 cases and 5,642 controls with 187,326 genotyped SNPs zoomed in on the classical HLA region on chromosome 6p. Multiple SNPs are reaching genome-wide significance.

We investigated the 3 markers reaching the lowest p-values in relation to their surrounding genes (Figure 31). The SNPs were plotted using the software tool *Locus* Zoom, which created a regional plot of the area of interest. The regional plot provides information about the *locus* including the genes in the area. The SNP with the lowest p-value is indicated with a purple diamond and the colouring of the other SNPs indicates their level of LD to this index SNP.

Figure 31 *Locus zoom plot for HLA region*



Legend *Shown is the locus zoom plot for the HLA region on chr6:32.10 – chr6:32.75 (GRCh37/hg19). The SNP with the lowest p-value is rs479536 (annotated with a purple diamond) and is upstream of NOTCH4. The SNP with the second lowest p-Value (rs4947342, not annotated) is upstream of HLA-DQB1. The colouring of the SNP at the approximate position of chr6:32.7 in green indicates that this SNP is in LD with the index SNP at position chr6:32.3, revealing LD expanding over this whole region of 0.4 Mb.*

The SNP with the lowest p-value (rs479536, $p=1.87\times10^{-47}$) was located 2kb upstream of *Notch receptor 4 (NOTCH4)*. The SNP with the second lowest p-value (rs4947342, $p=4.90\times10^{-42}$) was located approximately 18kb upstream of *HLA-DQB1*. And the SNP with the third lowest p-value (rs9501626, $p=3.70\times10^{-39}$) was located 7kb upstream of *HLA-DRA*. Linkage disequilibrium extended for approximately 400 kb between these genes (Figure 31).

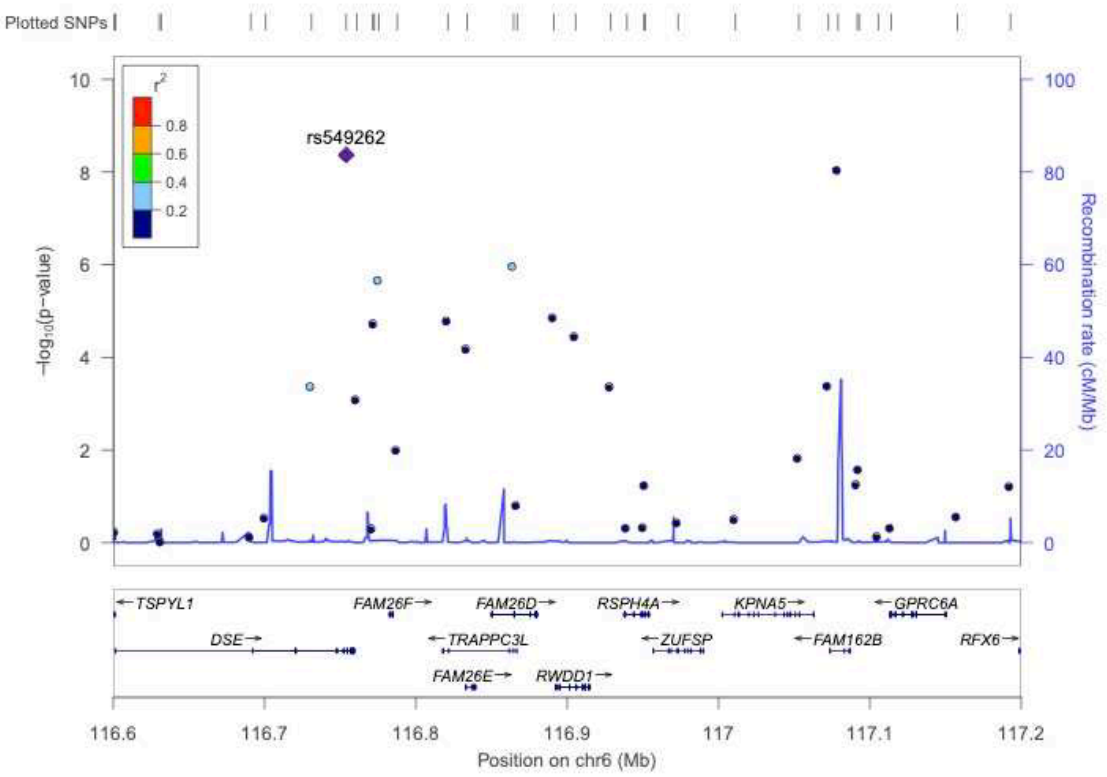
The genes *HLA-DQB1/HLA-DRA* encode for classical HLA alleles. HLA imputation to define the exact alleles was performed in the downstream analysis.

Chromosome 6 q-arm

The 2 markers reaching genome wide significance on the q-arm of chromosome 6 were rs549262 ($p=4.30\times10^{-9}$) and rs648210 ($p=8.73\times10^{-9}$). The SNPs were plotted in

Locus Zoom. Rs549262 was located intronic in the gene *dermatan sulfate epimerase (DSE)* and rs648210 in the intronic region of the gene *Homo sapiens family with sequence similarity 162, member B (FAM162B)*. Linkage disequilibrium extends for approximately 100kb between these genes (Figure 32).

Figure 32 *Locus zoom plot for rs549262*



Legend *Shown is the locus zoom plot for rs549262 on chromosome 6q22.1. The SNP with the lowest p-value is annotated with a purple diamond and is in the gene DSE. The SNP with the second lowest p-value (not annotated) is over the gene FAM162B.*

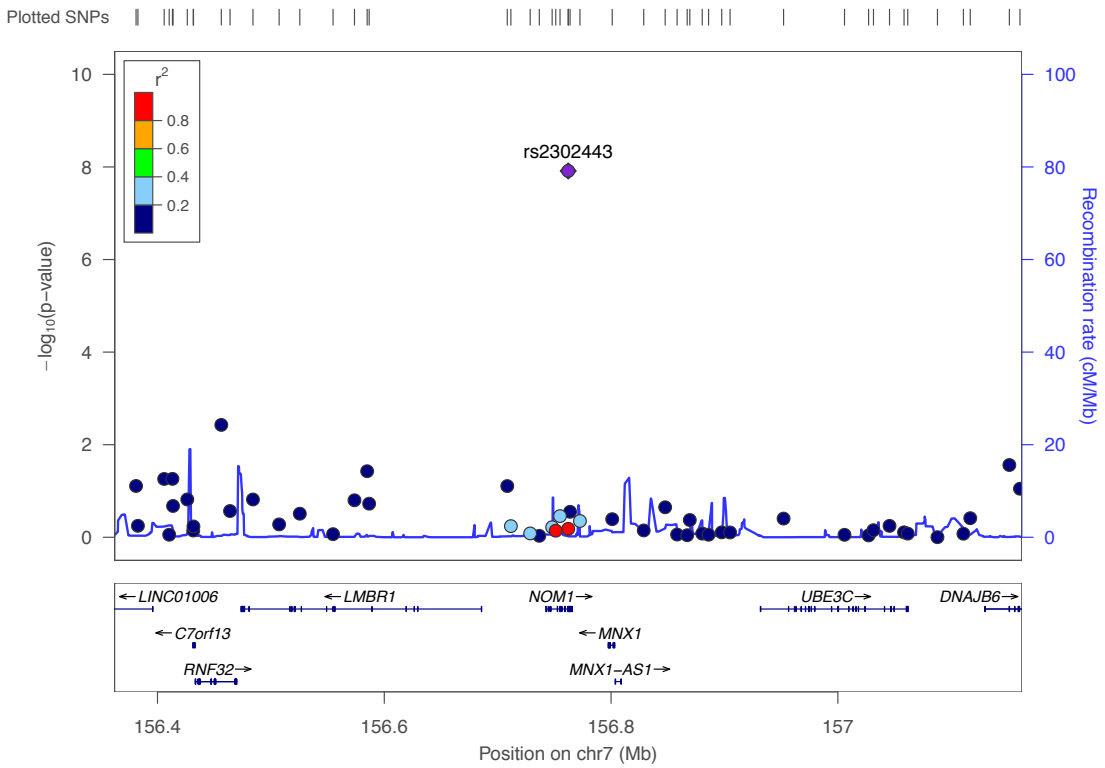
Chromosome 7 and 15

Further, two isolated markers reached significance, rs2302443 ($p=1.22\times10^{-8}$) on Chromosome 7 at the position q36.3 and rs1898882 ($p=4.39\times10^{-13}$) on chromosome 15q15.1.

The SNP rs2302443 on chromosomes 7 is a missense variant in the gene *NOM1*. In the most updated dbSNP version (at time of writing this thesis), build 153, the SNP is annotated as a multiallelic marker with three allele options G/C/T. The marker therefore should have been filtered out already by *REMEDY*, however, as *REMEDY*

refers to dbSNP version 151, the marker survived the QC steps. The marker was plotted with *Locus* Zoom and none of the SNPs in LD with that marker were associated with the disease (Figure 33). The marker was therefore considered as false positive.

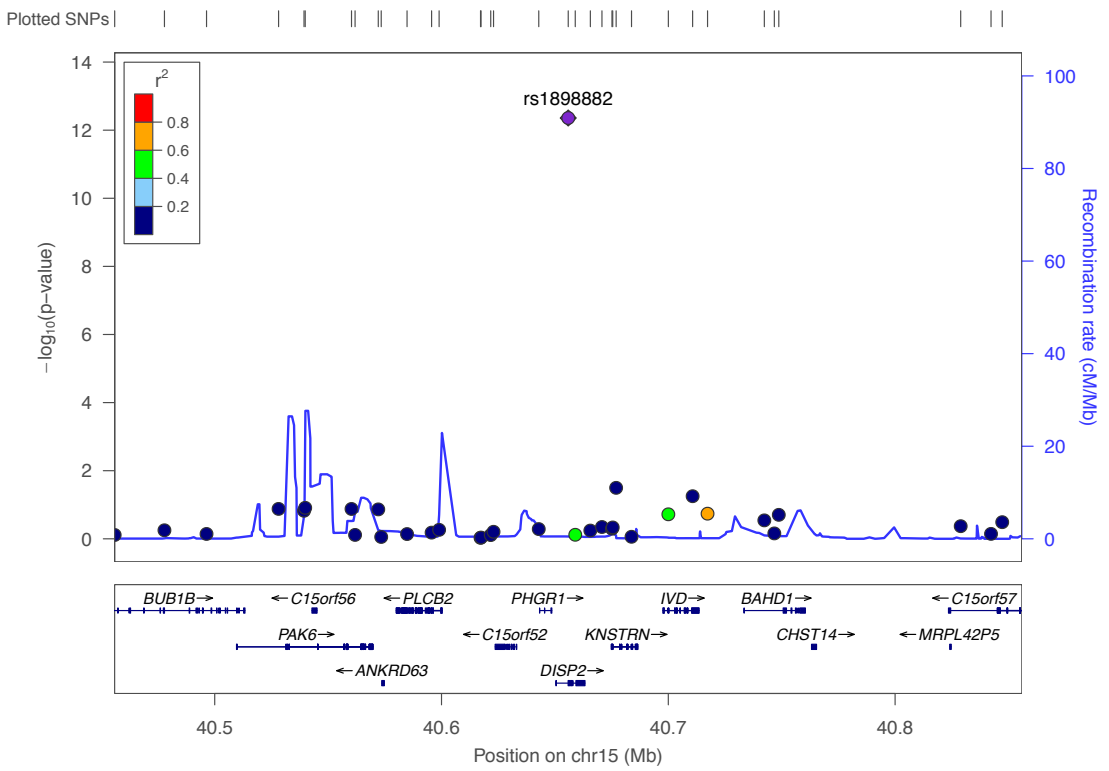
Figure 33 *Locus zoom plot for rs2302443*



Legend *Shown is the locus zoom plot for the marker rs2302443 on chromosome 7. The SNP is located in the gene NOM1. None of the markers in LD with the index SNP have a decreased p-value, suggesting a false positive result.*

The SNP rs1898882 on chromosome 15 is a missense variant in the gene *DISP2 (dispatched RND transporter family member 2).* The marker is biallelic. G is the major allele and C the minor allele with a frequency of 0.46. No other SNPs in LD with this marker were associated with the disease (Figure 34). This SNP was therefore considered as false positive.

Figure 34 *Locus zoom plot for rs1898882*



Legend *Shown is the locus zoom plot for the marker rs1898882 on chromosome 15. The SNP is located in the gene DISP2. No other SNPs in LD with this marker are associated with the disease, suggesting a false positive result.*

In conclusion, the two SNPs rs2302443 on chromosomes 7 and rs1898882 on chromosomes 15 were considered as false positive results. For further investigations, we focused on the two regions on chromosome 6.

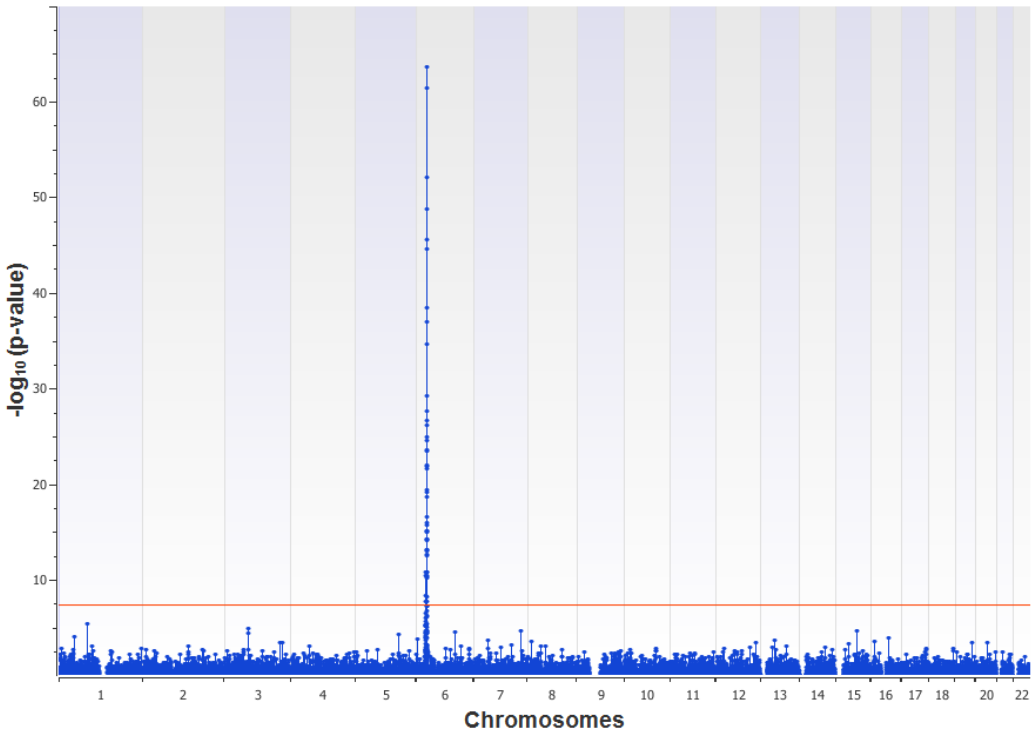
Haplotype association test

A haplotype describes a set of alleles which are inherited together. Haploblocks are stretches of DNA, in which recombination rate is low and therefore haplotypes within a haploblock are inherited together. These haploblocks are bordered by areas of high recombination, marking the start of a new haploblock. Haplotypes and haploblocks can be calculated according to algorithms, and then used for association testing with the disease.

We performed a haplotype and then a haploblock association test on our dataset of 422 cases and 5,642 controls. Using the algorithm provided in SVS, 29,529 haplotypes on 10,399 haploblocks were detected. The statistical test used for the analysis was Chi-squared test, which allows to compare the frequency of haplotypes/haploblocks between cases versus controls.

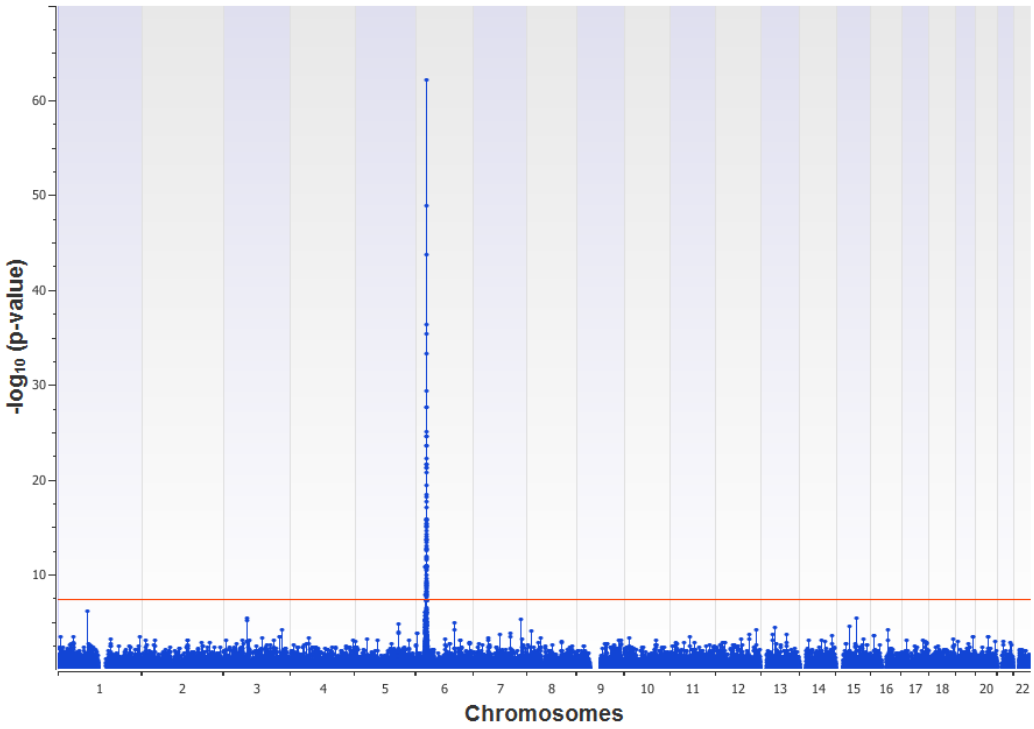
48 haploblocks were associated with the disease at genome wide significance threshold of $p<5\times10^{-8}$ (Figure 35). 82 haplotypes were associated with the disease at genome wide significance threshold (Figure 36). All of them were on Chromosome 6 in the HLA region.

Figure 35 *Manhattan plot for HAT per haploblock*



Legend *Manhattan plot for haplotype association test for comparison of haploblock frequency between cases versus controls. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log p-value of each marker on the y-axis. Each blue dot represents the first marker of a haploblock. One locus reaches genome wide significance. This locus corresponds to the MHC/HLA region, overlapping with the results seen for the basic allele test.*

Figure 36 *Manhattan plot for HAT per haplotype*



Legend *Manhattan plot for haplotype association test for comparison of haplotype frequency between cases versus controls. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log p-value of each marker on the y-axis. Each blue dot represents the first marker of a haplotype. As seen in the haploblock association test, one locus reaches genome wide significance corresponding to the MHC/HLA region.*

The results of the haplotype association test showed that on chromosomes 6 in the HLA region not only single markers are associated with the disease, but also whole haploblocks. The level of significance is higher than in the basic allele test representing the combined effect of markers in a haplotype and haploblock on the association results.

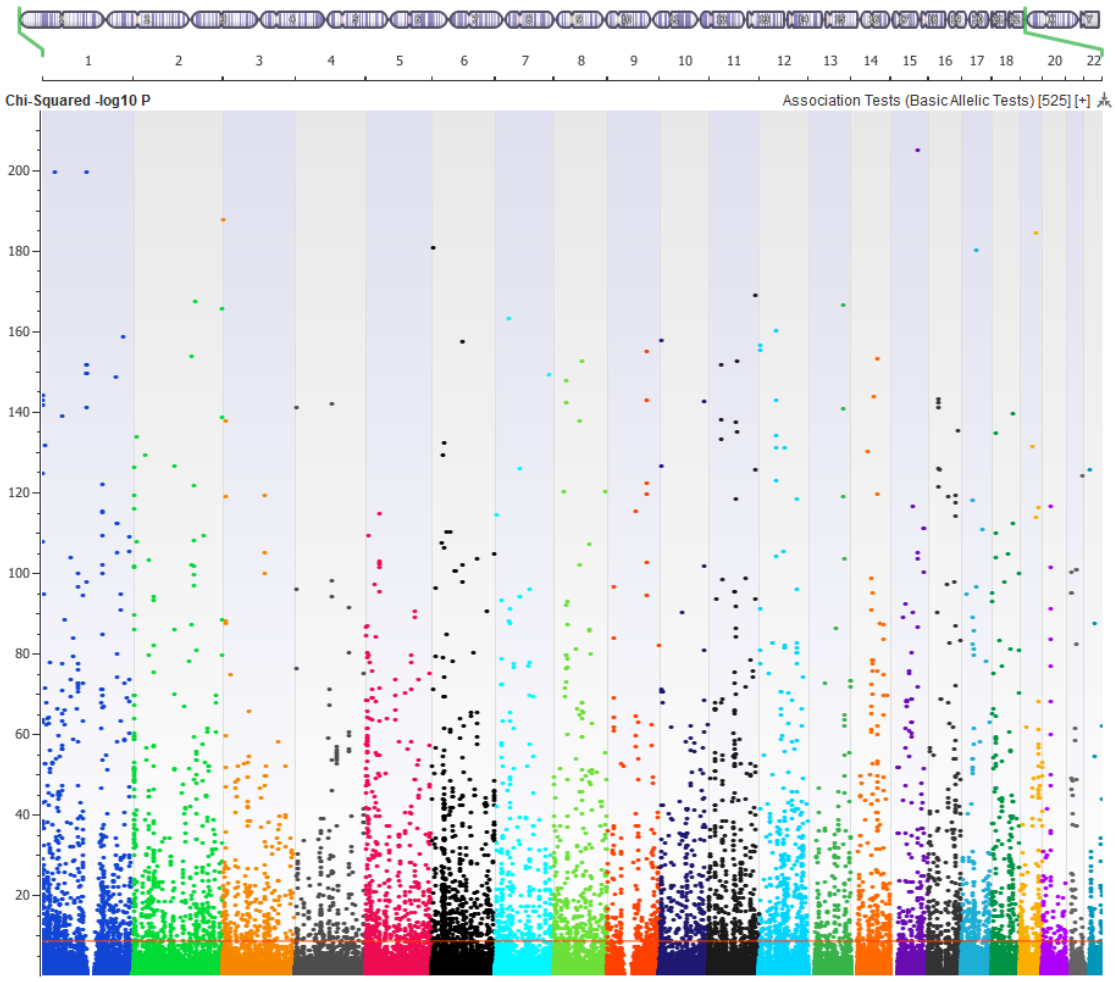
Chapter 4. Imputation European cohort

The ideal situation would be to investigate the highest number of variants possible when comparing cases and controls. However, in the final dataset of our study the density of genotyped markers was relatively low. Many markers dropped when overlapping the case and control datasets and the filtering steps removed additional markers. A low density of markers can cause disease *loci* to be missed as gaps between the markers might be large and whole areas might not be covered or signals are lower as fewer markers are in LD with the causal variant. A way to increase the density of markers is via imputation. Imputation is a statistical method to infer missing genotype data by placing information into a context. It takes the available genotype data and fills in the gaps between with most probable genotypes based on probabilities arising from the surrounding data and reference panels (details can be found in the introduction part on page 49).

Imputation was performed on the final case and control dataset from the GWAS. This dataset comprised 422 cases and 5,642 controls genotyped on 158,217 markers passing the quality control steps. Imputation analysis itself was performed using Beagle 5.0 with the 1000 Genomes Project Phase 3 data (version 5a) as a reference panel. The reference panel included 2,504 samples including representatives from the 5 super populations East Asian, South Asian, African, European, American (detailed methods are described in material and methods part page 89).

Cases and controls were imputed together to a total of 30,761,499 markers. The first attempts of performing an association test on the imputed dataset revealed very noisy results (Figure 37). We consequently applied strict quality control steps on the imputed dataset before further analysis.

Figure 37 *Manhattan plot for first attempt of association test after imputation*



Legend *Shown is the Manhattan plot for the results of the first association test post imputation. No filtering was done on the imputed markers. One can see that a large number of markers are above the genome wide significance threshold causing a noisy picture. Quality control filtering is needed to detect and remove markers causing possible false positive associations.*

Post imputation quality control

Imputation accuracy

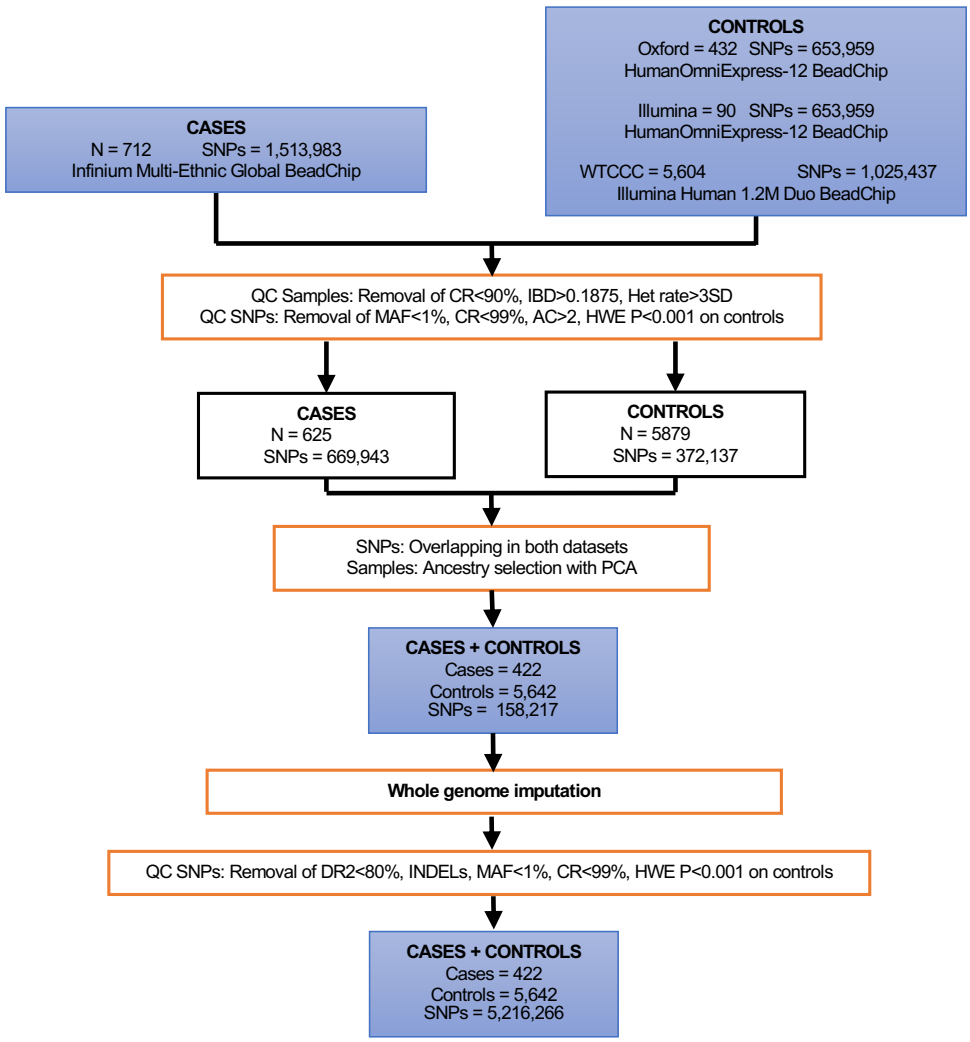
The genotypes outputted are the most probable ones for each marker of each individual. Therefore, each genotype is imputed with a certain probability. The probability refers to how likely the imputed genotype matches the real (observed) genotype. The first quality control step was to keep only genotypes which are imputed above a certain probability threshold.

The parameter representing imputation quality is the dosage R-squared (DR2), discussed in the methods section on page 91. 6,435,046 markers were imputed with a high probability given by a DR ≥80%. All markers with a DR2 <80% were removed Further, 587,586 markers were structural variants and removed from the analysis. Only SNPs were processed for further analysis (n=5,847,460). All markers had a CR ≥99% and 624,229 markers had a MAF <1% and were removed. Nine markers had no assigned rsID and were removed. The cleaned dataset containing 5,223,222 markers and was imported in SVS for further analysis. Additionally, 6,956 markers showed significant deviation from HWE ($p<0.001$) in the control samples and were removed.

Final dataset

The final dataset included 5,216,266 markers. Compared to 158,217 genotyped markers of the genotyped dataset, the density has increased 50-fold (Figure 38).

Figure 38 *Flow chart of quality control steps leading to final dataset of imputed GWAS*



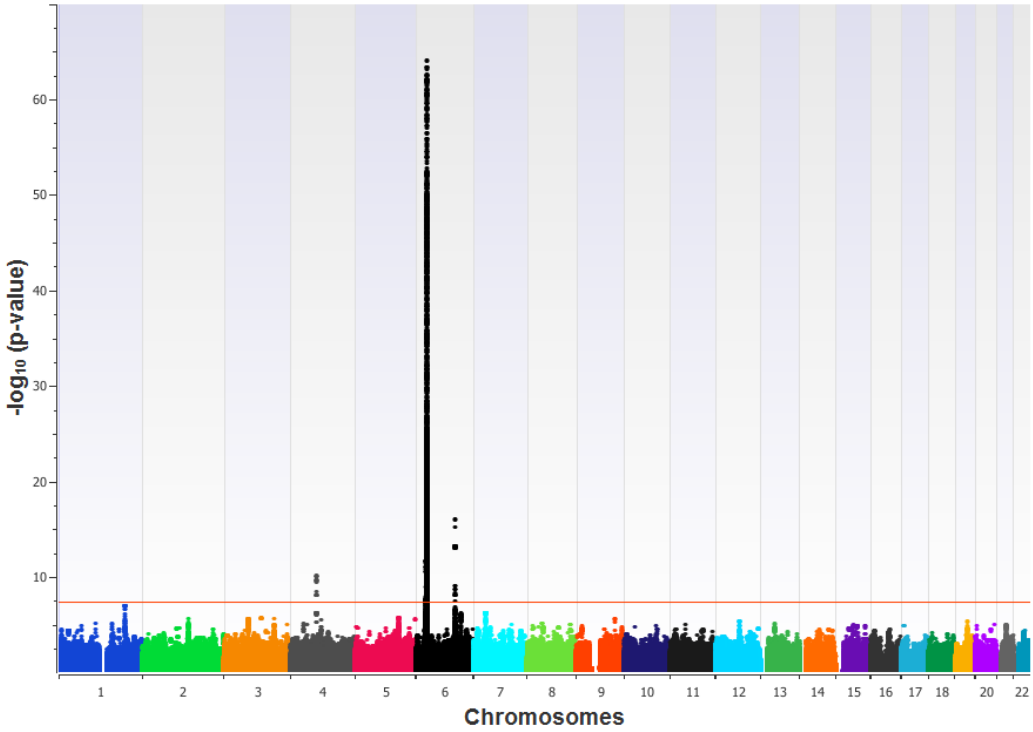
Legend *Flowchart providing information on data input and processing. QC: quality control; CR: call rate; IBD: identity by descent; Het rate: heterozygosity rate; MAF: minor allele frequency; HWE: Hardy-Weinberg equilibrium.*

Association testing on imputed dataset

Basic allele test

A basic allele test was performed on 422 cases versus 5,642 controls with 5,216,266 markers. The inflation factor lambda was 1.0794. 7,321 markers were associated with the trait at genome-wide significance level (Figure 39).

Figure 39 *Manhattan plot for BAT of imputed dataset*



Legend *Shown is the Manhattan plot for the analysis of 422 cases and 5,642 controls with the 5,216,266 imputed SNPs. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log p-value of each marker on the y-axis. The level of whole genome wide significance threshold is presented by the red line. The locus on chromosome 6q has increased in significance and more markers in the same build-up are now above the threshold line. The peak on chromosome 4 was not seen in the Manhattan plot with only genotyped markers, possibly indicating a further association finding.*

The markers reaching genome wide significance were distributed on 2 chromosomes. 17 markers were on chromosome 4 and 7,304 on chromosome 6. A peak on chromosome 4 appeared, which was not seen in the Manhattan plot with only genotyped markers. The *locus* on chromosome 6q has increased in significance and more markers in the same build-up than seen with the genotyped GWAS are above the significance threshold.

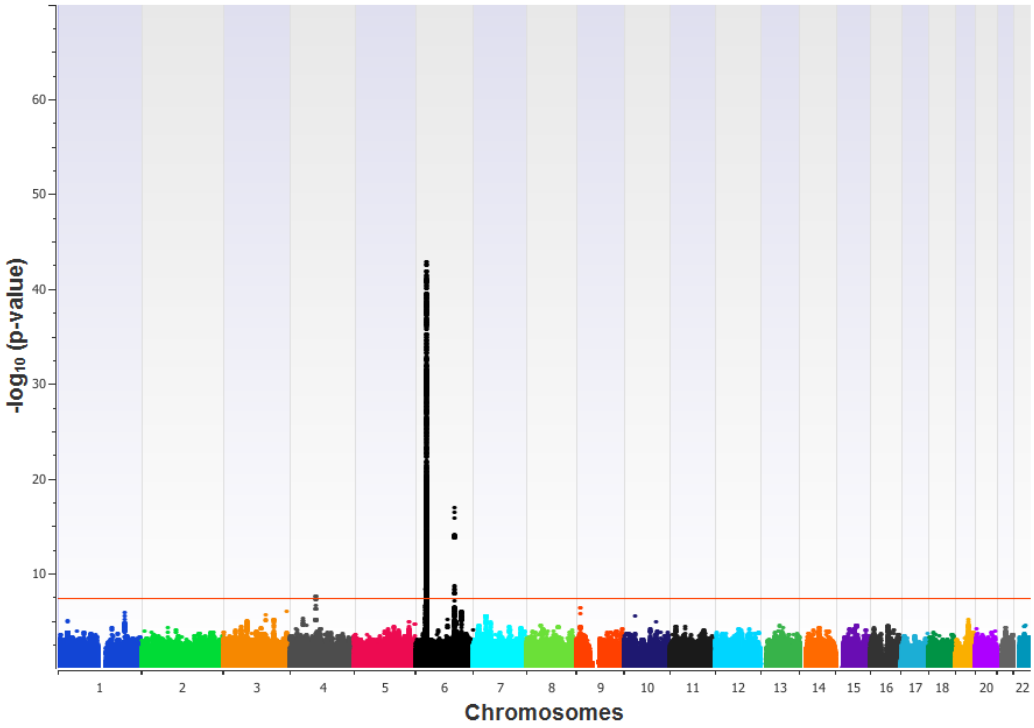
To address possible substratification and to correct for covariates, we decided to repeat the association test using a logistic regression model.

Regression analysis

The main advantage of logistic regression over basic alleles test is that it can handle more than two independent variables simultaneously, which is important in order to correct for covariates. Covariates are variables or the interaction between variables, which could affect the analysis without necessarily being related to the disease. Correcting for covariates, allows to see the effects of the remaining variables on the outcome. This enables to correct for stratification taking the principal components into account.

We performed a regression analysis with correction for the first 10 principal components calculated for this dataset in order to address possible substratification. The genomic inflation factor lambda was 1.027, hence no substantial inflation was noted. The analysis revealed 4,019 markers reaching genome-wide significance. The markers were distributed as previously seen over three *loci*, one on chromosome 4 and two on chromosome 6 (Figure 40).

Figure 40 *Manhattan plot for regression analysis of imputed dataset*



Legend *Shown is the Manhattan plot for the analysis of 422 cases and 5,642 controls with the 5,216,266 imputed SNPs corrected for the first 10 principal components. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log p-value of each marker on the y-axis. The level of whole genome wide significance threshold is presented by the red line. Three loci achieve genome wide significance. Compared to the results of the basic allele test, the overall level of significance has decreased, but the same loci are reaching genome wide significance.*

Details for the top marker of each *locus* reaching genome wide significance are displayed in Table 14.

Table 14 *Lead SNPs of the three loci reaching genome wide significance in the imputed dataset.*

Locus	Gene	SNP	DR2	Minor allele	MAF cases	MAF controls	OR	95% CI	p-value
6p21.3	*HLA-DQB1*	rs9273542	0.89	T	0.51	0.24	3.39	2.86-4.03	1.59×10^{-43}
6q22.1	*CALHM6*	rs2637678	0.96	C	0.26	0.40	0.51	0.44-0.60	1.27×10^{-17}
4q13.3	*PARM1*	rs10518133	0.93	A	0.12	0.06	1.96	1.57-2.45	2.50×10^{-8}

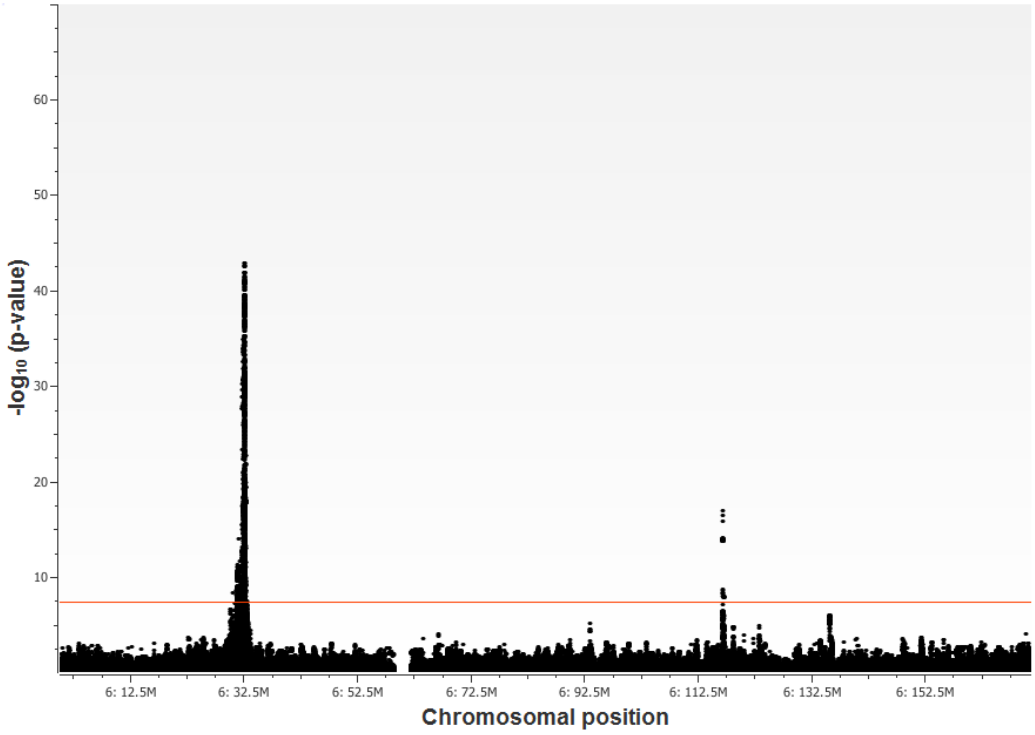
Legend *Minor allele frequencies (MAF) and odds ratio (OR) with 95% confidence intervals (95% CI) for each of the minor alleles of the lead SNPs from the three loci achieving genome-wide significance. DR2 (Dosage R-Squared) indicates the Beagle imputation quality score.*

For easier understanding, we broke down the analysis of the results into the three *loci*.

Chromosome 6 p-arm

The strongest signal seen in the Manhattan plot corresponded to a broad peak on chromosome 6p21.32 in the classical HLA region (Figure 41).

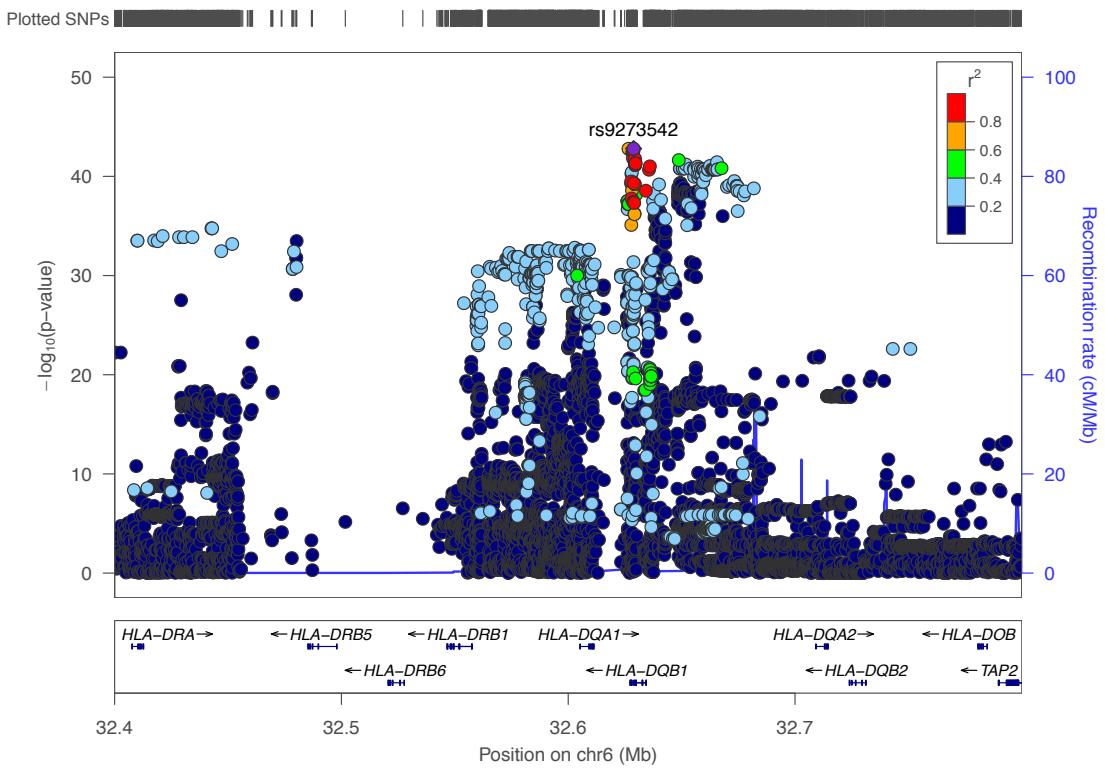
Figure 41 *Manhattan plot for chromosome 6*



Legend *Manhattan plot for regression analysis zoomed in on chromosome 6. Two loci are reaching genome wide significance. The broad peak on chromosome 6 p-arm corresponds to the HLA region.*

The lead markers in this region was rs9273542 ($p=1.59\times10^{-43}$, OR=3.39, 95%CI 2.86-4.03). The marker was imputed with a DR2 score of 0.89 indicating high imputation accuracy. The marker is located in the intronic region of the gene *HLA-DQB1*. The OR of 3.39 indicates that the minor allele is associated with a higher risk for development of the disease. The SNP with the second lowest p-value was 100bp apart, rs9273529 ($p=2.87\times10^{-43}$, OR=3.39, 95%CI=2.85-4.03), and was also located in the intronic region of the gene *HLA-DQB1*. The SNP with the third lowest p-value was rs9273371 ($p=1.64\times10^{-43}$, OR=3.29, 95%CI=2.78-3.89) located intergenic between *HLA-DQA1* and *HLA-DQB1.* The ORs of all three SNPs are very similar, demonstrating the same direction of their effect on disease development. We created a *locus* zoom plot for the region of interest (Figure 42).

Figure 42 *Locus zoom for region on chromosome 6p*



Legend *Shown is the locus zoom for the SNP rs9273542 on chromosome 6. The index SNP is annotated with a purple diamond and is in the gene HLA-DQB1. The SNP with the second lowest p-value (rs9273529, not annotated) is in close proximity (100bp) also in the gene HLA-DQB1. The SNP with the third lowest p-value (rs9273371, not annotated) was approx. 2.5kb apart and intergenic between HLA-DQB1 and HLA-DQA1. The red colouring of the surrounding SNPs indicates that they are in strong LD with the index SNP.*

The *locus* zoom plot indicated that the two SNPs with the second lowest p-values are in strong LD with the lead SNP (Figure 42). We performed a conditional analysis correcting for the lead SNP to test if the association between the other two SNPs and the disease is independent from the lead SNP. A significant association of the 2nd and/or 3rd SNP with the disease after adjusting for the lead SNP would indicate that those SNPs have an independent effect on the risk for disease development. If the SNPs are not further associated with the disease, they are not independently affecting the disease risk.
Additionally, the result can reveal SNPs which are independently associated with the disease, but are masked by the strong signal of the lead SNP.

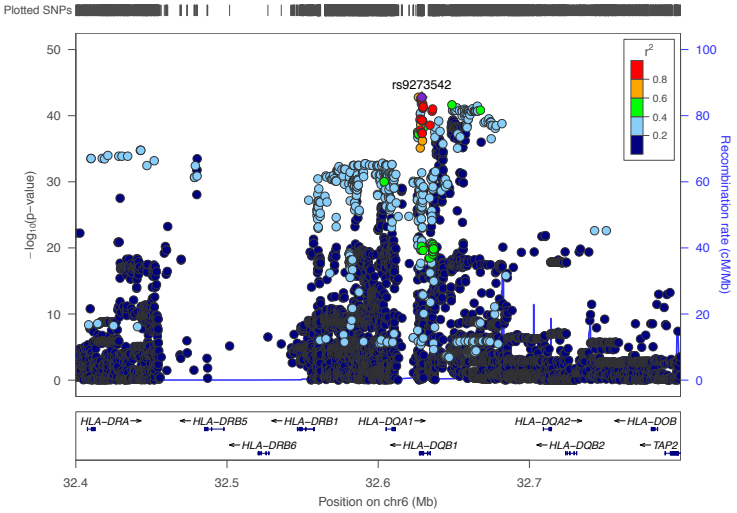
Conditional analysis

We performed a conditional analysis on the lead SNP, rs9273542. The strength of the association in this region decreased. As expected from the LD displayed in the *locus* zoom plot (Figure 43 A), the two SNPs rs9273529 and rs9273371 did no longer reach significance level. The SNP with the lowest p-value changed to rs2858317 ($p=4.29\times10^{-31}$) upstream of *HLA-DQB1* (Figure 43 B).

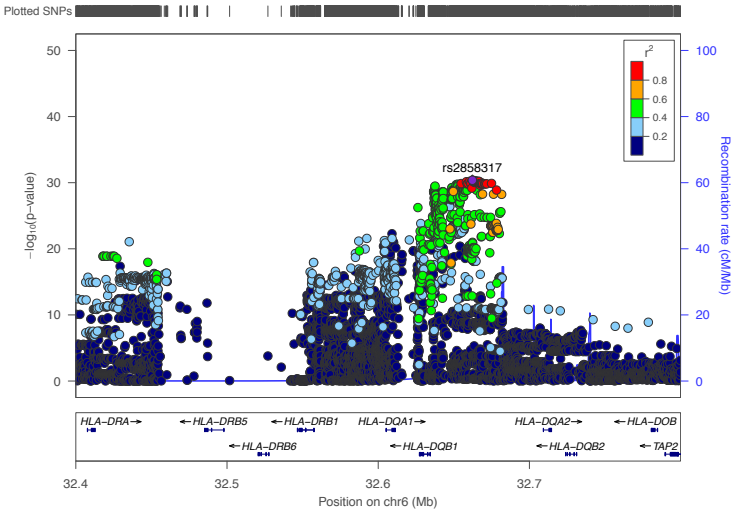
We subsequently conditioned on both markers, rs9273542 and rs2858317. Joint conditioning on the markers reduced the strength of the association in this region further. The marker with the lowest p-value was rs3828799 ($p=2.40\times10^{-8}$) just above the significance threshold (Figure 43 C).

Figure 43 *Locus zoom for region on chromosome 6p before and after conditioning*

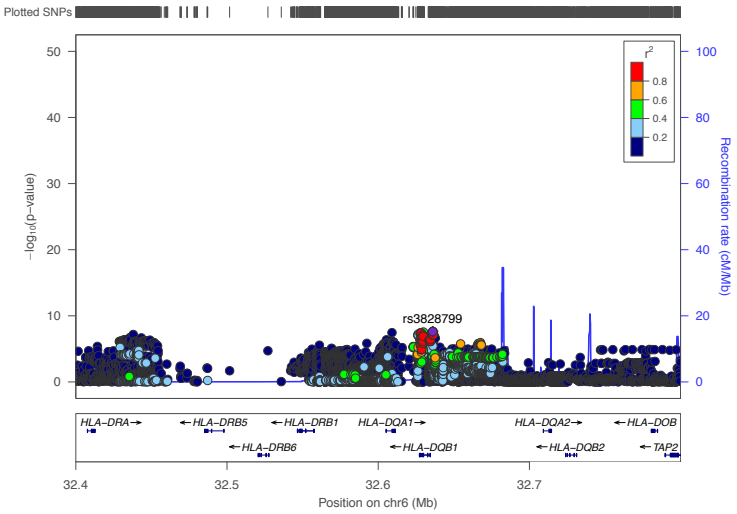
A)



B)



C)



Legend
A) *Locus zoom for rs9273542 before conditioning.*
b) *Locus zoom after conditioning on rs9273542. The SNP rs2858317 on chromosome 6 is the SNP with the lowest p-value after conditioning on rs9273542. The SNP is marked with a purple diamond and is upstream of HLA-DQB1. The red colouring of the surrounding SNPs indicates that they are in strong LD with the index SNP.*
C) *Locus zoom after conditioning on rs9273542 and rs2858317. After conditioning on rs9273542 and rs2858317 the significance level in this region decreased further, indicating that the association at this locus is driven by the two independent markers rs9273542 and rs2858317.*

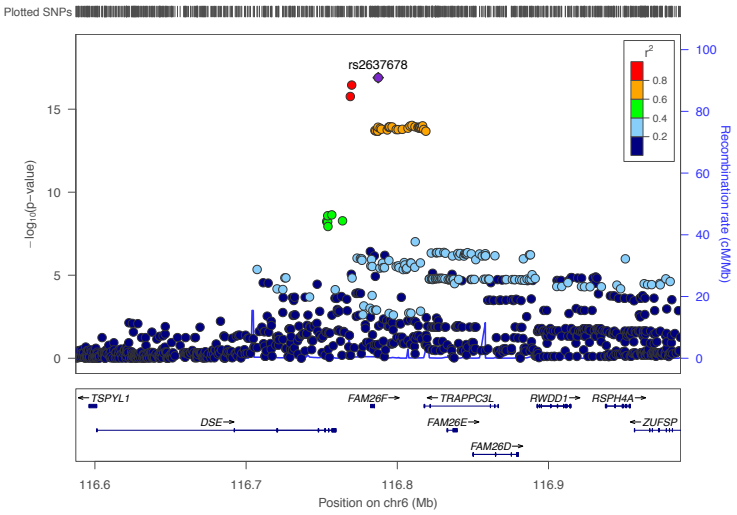
These results indicated that the association at this *locus* is driven by two independent signals, one around rs9273542 and another around rs2858317.

Chromosome 6 q-arm

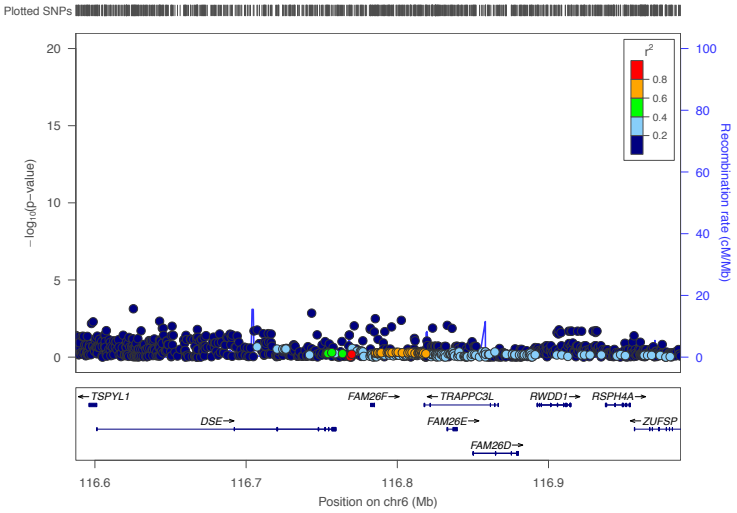
The second *locus* associated with the disease was on chromosome 6q22.1. The lead marker was rs2637678 ($p=1.27\times10^{-17}$, OR=0.51, 95%CI=0.44-0.60) and had an imputation quality score DR2 of 0.96 indicating a high imputation accuracy. The marker was located downstream of the gene *CALHM6* (previously called *FAM26F*) (Figure 44 A) [158]. The OR of 0.51 indicated a protective effect of the minor allele. The SNP with the second lowest p-value was rs2637681 ($p=3.53\times10^{-17}$, OR=0.52, 95%CI=0.44-0.61) approximately 18kb apart from rs2637678, upstream of *CALHM6*. And the SNP with the third lowest p-value was rs2858829 ($p=1.72\times10^{-16}$, OR=0.53, 95%CI=0.45-0.62) another 1kb further upstream of *CALHM6*. Both of them showed a similar OR to the lead SNP with the minor allele being protective.

Figure 44 *Locus zoom for region on chromosome 6q22.1 before and after conditioning*

A)



B)



Legend
A) *Shown is the locus zoom for the SNP rs2637678 on chromosome 6 q-arm. The index SNP is marked with a purple diamond and is downstream of the gene FAM26F (CALHM6). The SNPs with the second and third lowest p-value (rs2637681 and rs2858829) are annotated in red. Both of them are upstream of FAM26F (CALHM6). The red colouring indicates that they are in strong LD with the index SNP.*
B) *Shown is the locus zoom for the peak on chromosome 6 q-arm after conditioning on rs2637678. No further markers reach genome wide significance, indicating that the association at this locus is driven by one single signal.*

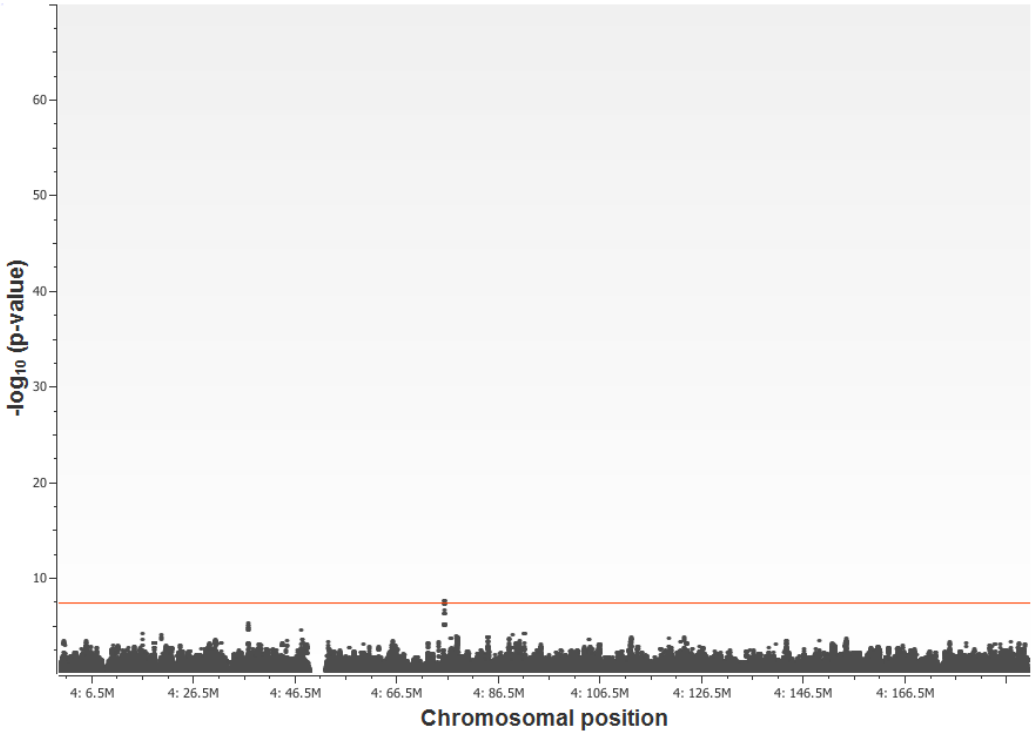
As aforementioned, we wanted to test if the markers with the second and third lowest p-values are independently associated with the disease or if their association is driven by LD with the lead SNP. We performed a conditional analysis with conditioning on rs2637678. After conditioning on rs2637678 no further marker reached the level of

genome wide significance, indicating that a single signal is responsible for driving this association (Figure 44 B).

Chromosome 4

The third *locus* reaching genome-wide significance was on chromosome 4q13.3 (Figure 45).

Figure 45 *Manhattan plot for chromosome 4*

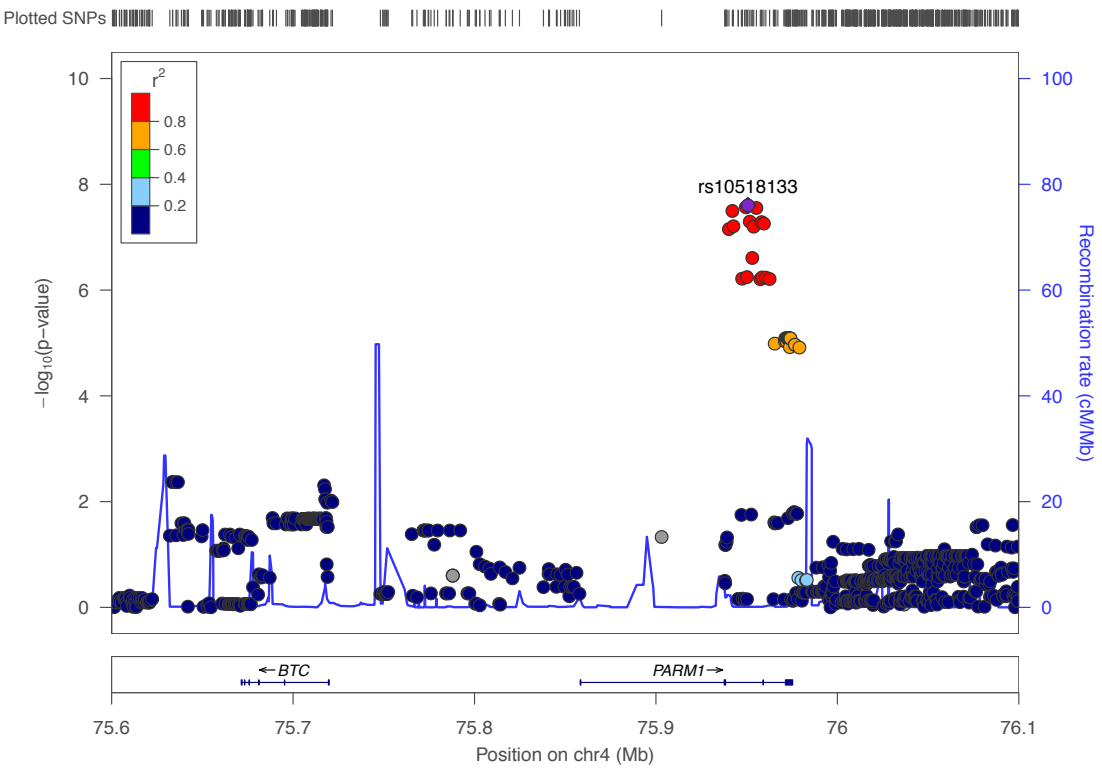


Legend: *Manhattan plot for regression analysis zoomed in on chromosome 4. One locus is reaching genome wide significance.*

The lead marker was rs10518133 ($p=2.50\times10^{-8}$, OR 1.96, 95% CI=1.57-2.45). The marker was imputed with a DR2 score of 0.93, reflecting a high imputation accuracy. The marker was located in the intronic region of the gene *PARM1 (*Figure 46). The OR of 1.96 indicated an increased risk for disease with the minor allele. The SNPs with the second and third lowest p-values, rs72660383 ($p=2.73\times10^{-8}$, OR 1.96, 95% CI=1.57-2.45) and rs17000108 ($p=2.79\times10^{-8}$, OR 1.96, 95% CI=1.57-2.45) were in close proximity to the lead SNP with comparable ORs. After conditioning on the lead

SNP, no marker reached genome-wide significance indicating that the association is driven by a single signal.

Figure 46 *Locus zoom for region on chromosome 4q13.3*



Legend *Shown is the locus zoom for the SNP rs10518133 on chromosome 4. The index SNP is marked with a purple diamond and is in the intergenic region of PARM1. The SNPs with the second and third lowest p-values are among those coloured in red. The red colouring indicates that they are in strong LD with the index SNP.*

eQTL analysis

All three lead SNPs (rs2637678, rs2637681, rs2858829) in the chromosome 6q22.1 *locus* showed strong cis-eQTL effects according to the GTEx database. Rs2637678 showed the highest effect size (NES 0.56, $p=3.0\times10^{-8}$) in EBV-transformed lymphocytes on *CALHM6*. Similar results were obtained for rs2637681 (NES 0.56, $p=1.1\times10^{-7}$) and rs2858829 (NES 0.54, $p=2.8\times10^{-7}$) with the largest effect size in EBV-transformed lymphocytes altering the expression of *CALHM6*. For all three variants the minor allele increased the expression of *CALHM6*, indicating that in cases (where the minor allele was less common) the expression of *CALHM6* is downregulated.

We also investigated how the lead variants affect the gene *DSE*. All three lead SNPs (rs2637678, rs2637681, rs2858829) had the lowest effect size on *DSE* in whole blood (rs2637678: NES -0.14, $p=3.4\times10^{-14}$, rs2637681: NES -0.17, $p=1.3\times10^{-20}$, rs2858829: NES -0.16, $p=1.3\times10^{-18}$). Here, the minor allele decreased the expression of *DSE*, indication that in cases (where the minor allele was less common) the gene *DSE* is upregulated.

No significant eQTLs were found for rs10518133 (tagging the *locus* on Chromosome 4q13.3) in any tissue in the GTEx database.

Further, we queried data from the ENCODE project and found evidence that the SNP rs2858829 lies in a site of strong regulatory activity:

We found an enrichment of H3K27Ac histone marks (high acetylation of histone H3 at lysine 27) at this site, which indicates transcription enhancer activity. Further, we found a clustering of DNAse hypersensitive areas overlapping this region, which also is an indication for regulatory regions in general, and promoter sites in particular. Moreover, chromatin immunoprecipitation followed by sequencing (ChIP-seq) data showed strong evidence of CCAAT/enhancer binding protein beta (CEBPB) transcription factor related activity in this region.

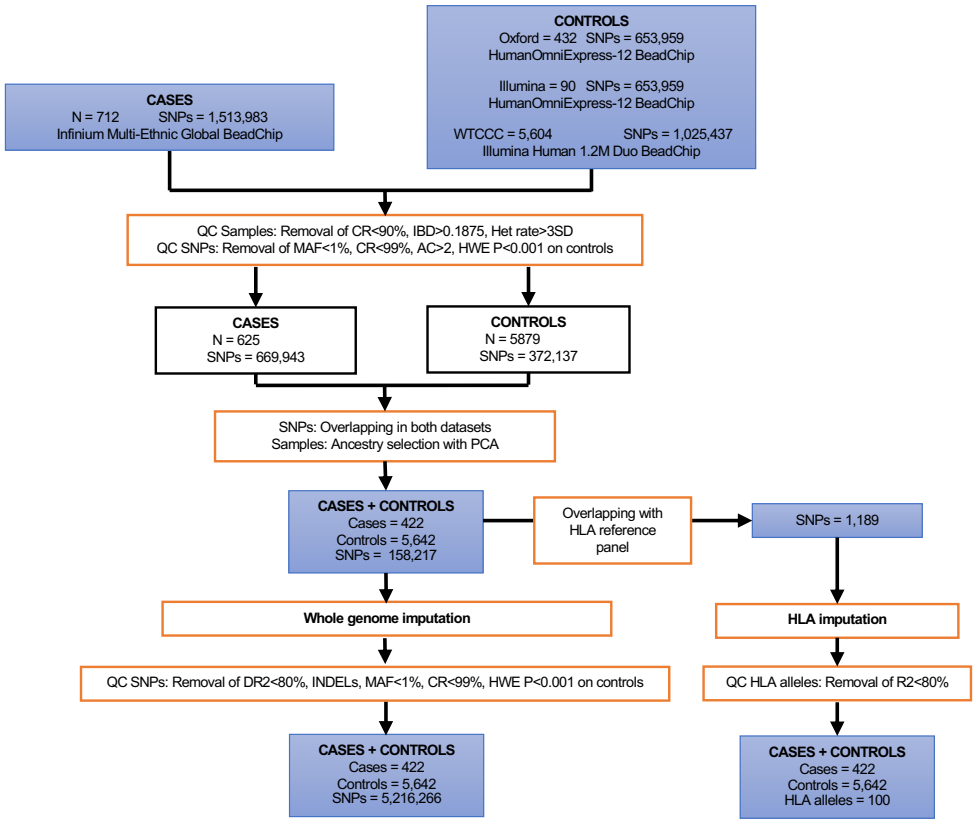
Chapter 5. Human Leukocyte Antigen imputation European cohort

The strongest signal of the entire association study is coming from markers in the HLA region. The lead SNP, rs9273542 is located within the *HLA-DR/DQ* region. We wanted to investigate in more detail which HLA alleles are associated with the disease and which are protective. Direct sequencing of HLA alleles is expensive and was not feasible for this study. The HLA region itself is characterized by a high linkage disequilibrium and earlier studies demonstrated that specific SNPs are in strong LD to specific HLA alleles [114]. Therefore, a limited number of genotyped SNPs can be used to impute the majority of HLA alleles [115]. We used SNP2HLA, which is a freely available software tool, to perform HLA imputation from SNP- level data [116].

Post imputation processing

From the final dataset 1,189 genotyped SNPs were overlapping with the SNPs in the HapMap European reference dataset for HLA imputation. After imputation only HLA alleles with a quality score R2 above 80% were used for downstream analysis. A total of 100 imputed HLA alleles were carried forward for association analysis (Figure 47).

Figure 47 *Flow chart of quality control steps leading to final dataset of HLA imputation*



Legend *Flowchart providing information on data input and processing. QC: quality control; CR: call rate; IBD: identity by descent; Het rate: heterozygosity rate; MAF: minor allele frequency; HWE: Hardy-Weinberg equilibrium.*

HLA association test

HLA association testing was performed using logistic regression with adjustment for the first ten PCs of ancestry. Twelve HLA alleles were significantly associated with the disease. Three represented a lower typing resolution of the same allele. Typing resolution refers to what factor the HLA allele is defined. Details about HLA nomenclature and HLA typing can be found in the introduction part on page 54. As the information of the lower resolution allele is redundant to the higher resolution allele, only the higher resolution ones are displayed and kept for downstream analysis (Table 15).

Table 15 *Risk of classical HLA alleles associated with SSNS*

HLA allele	MAF cases	MAF controls	OR	95%CI	p-value
HLA_DQA1*02:01	0.35	0.15	3.42	2.80-4.16	1.06×10^{-32}
HLA_DQA1*01	0.13	0.38	0.36	0.30-0.43	1.90×10^{-31}
HLA_DRB1*07:01	0.35	0.15	3.26	2.68-3.97	5.62×10^{-31}
HLA_DQB1*02	0.4	0.21	2.43	2.04-2.91	9.77×10^{-22}
HLA_DQA1*01:03	0.02	0.09	0.24	0.15-0.38	1.79×10^{-14}
HLA_DRB1*13	0.04	0.11	0.31	0.22-0.44	2.41×10^{-14}
HLA_DRB1*13:01	0.02	0.08	0.23	0.15-0.37	3.18×10^{-14}
HLA_DQA1*01:01	0.08	0.15	0.46	0.35-0.59	1.53×10^{-10}
HLA_B*08:01	0.2	0.13	2.95	2.05-4.23	9.17×10^{-09}

Legend *Minor allele frequencies (MAFs) for patients and controls and odds ratios (ORs) with 95% confidence intervals (95%CIs) for each of the HLA alleles achieving genome-wide significance are shown.*

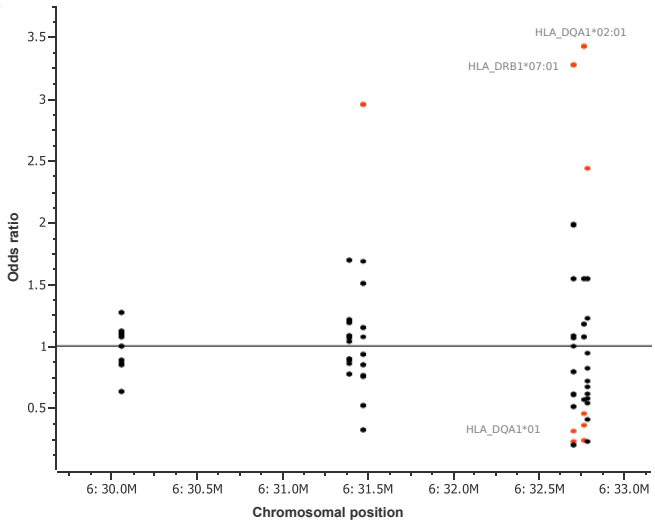
The strongest association was observed with *HLA-DQA1*02:01*. The OR of 3.42 indicated that carriers of this allele have an increased risk for the disease. Also associated with risk to develop the disease was *HLA-DRB1*07:01* with an OR of 3.26. Whereas the HLA allele *HLA-DQA1*01* had an OR of <1, indicating that carrying this allele is protective for disease development (Figure 48 A).

To test if these alleles are independently associated we performed a conditional analysis. After conditioning on *HLA-DQA1*02:01* the strongest signal came from *HLA-DQA1*01* ($p=1.24\times10^{-31}$, OR=0.31, 95% CI=0.25-0.38). This revealed that the allele from *HLA-DQA1*01* is independently protective for disease development (Figure 48 B).

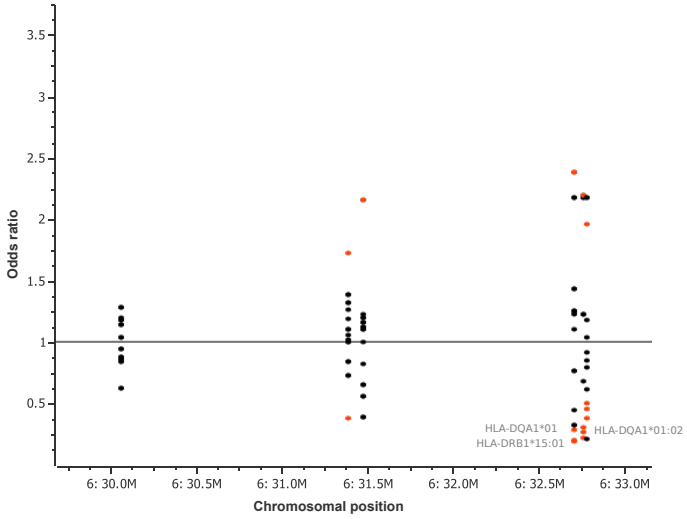
After conditioning on both, *HLADQA1*02:01* and *HLA-DQA1*01,* only two alleles remained independently significant. *HLA-DQB1*03:03* ($p=1.22\times10^{-8}$, OR=0.38, 95%CI=0.26-0.54) and *HLA-DQB1*03* ($p=1.69\times10^{-8}$, OR=0.64, 95% CI=0.55-0.75).

Figure 48 *Results of HLA type association analysis*

A)



B)



Legend *Results for HLA allele imputation for HLA class I (HLA-A, -C and -B) and class II (HLA-DRB1, -DQB1) genes. Shown are the odds ratios (OR) for classical HLA alleles on y-axis and chromosomal position on x-axis. HLA alleles with an OR > 1 are deemed to be disease causing versus HLA alleles with an OR < 1 deemed to be protective. Significantly associated HLA alleles are coloured in red. Indicated are the three HLA alleles with the highest significance level.*
A) *HLA-DQA1*02:01 and HLA-DRB1*07:01 are disease causing, whereas HLA-DQA1*01 seems to have a protective effect.*
B) *After conditioning on HLA-DQA1*02:01. Note that the protective allele HLA-DQA1*01 is essentially unchanged, indicating that its effect is independent of HLA-DQA1*02:01.*

Part 4. Asian cohort GWAS

Chapter 1. Replication cohort

The main findings of the European GWAS was, that a) specific HLA alleles are significantly associated with the disease and b) that there are two *loci* outside the HLA region on chromosome 4q13.3 and chromosome 6q22.1 that are significantly associated with the disease. The gold standard for GWAS findings to be considered as valid is to replicate the findings in an independent cohort (replication cohort). Therefore, we aimed to replicate the findings of the European GWAS in a separate cohort. We had cases and controls provided from Asian collaborators, which we utilised as replication cohort.

Cases

The dataset consisted of 513 cases from collaborators from South East Asia. Those samples were genotyped at ICH (Institute for Child Health) UCL (University College London) Genomics on the Infinium Multi-Ethnic Global BeadChip. After processing with *REMEDY* the dataset contained 1,565,259 of which 1,513,983 were autosomal.

Controls

Samples from 223 healthy controls were provided from the same collaborators. Those 223 were genotyped together with the cases on the Infinium Multi-Ethnic Global BeadChip and processed via *REMEDY*.

Quality control steps

The same quality control steps as for the European study were applied to the Asian cohort. The only difference was that cases and controls were analysed together. This was reasonable as cases and controls were genotyped together in the same batch on the same SNPchip and therefore treated as a single cohort.
First, samples were removed with a CR<0.90 (n=10) leaving 504 cases and 222 controls.

Duplicated and related samples were identified by using PRISMA with a cut-off for IBD of 0.1875. 104 cases and 20 controls were identified to be either identical or related and were removed from the further analysis, leaving 400 cases and 202 controls.
15 samples had a heterozygosity rate with more than 3 standard deviation +/- of the mean and were removed, leaving 387 cases and 200 controls in the dataset. The remaining dataset consisted of 587 samples (387 cases and 200 controls).
The initial number of markers in the dataset was 1,565,259 of which 1,513,983 were autosomal markers. Markers were removed because of CR <0.97 (n=63,727) and MAF <0.01 (n= 739,938). Markers in the control dataset were removed if they were outside the HWE with p <0.001 (13,722). The remaining number of markers after QC was 712,190.

Ethnicity selection

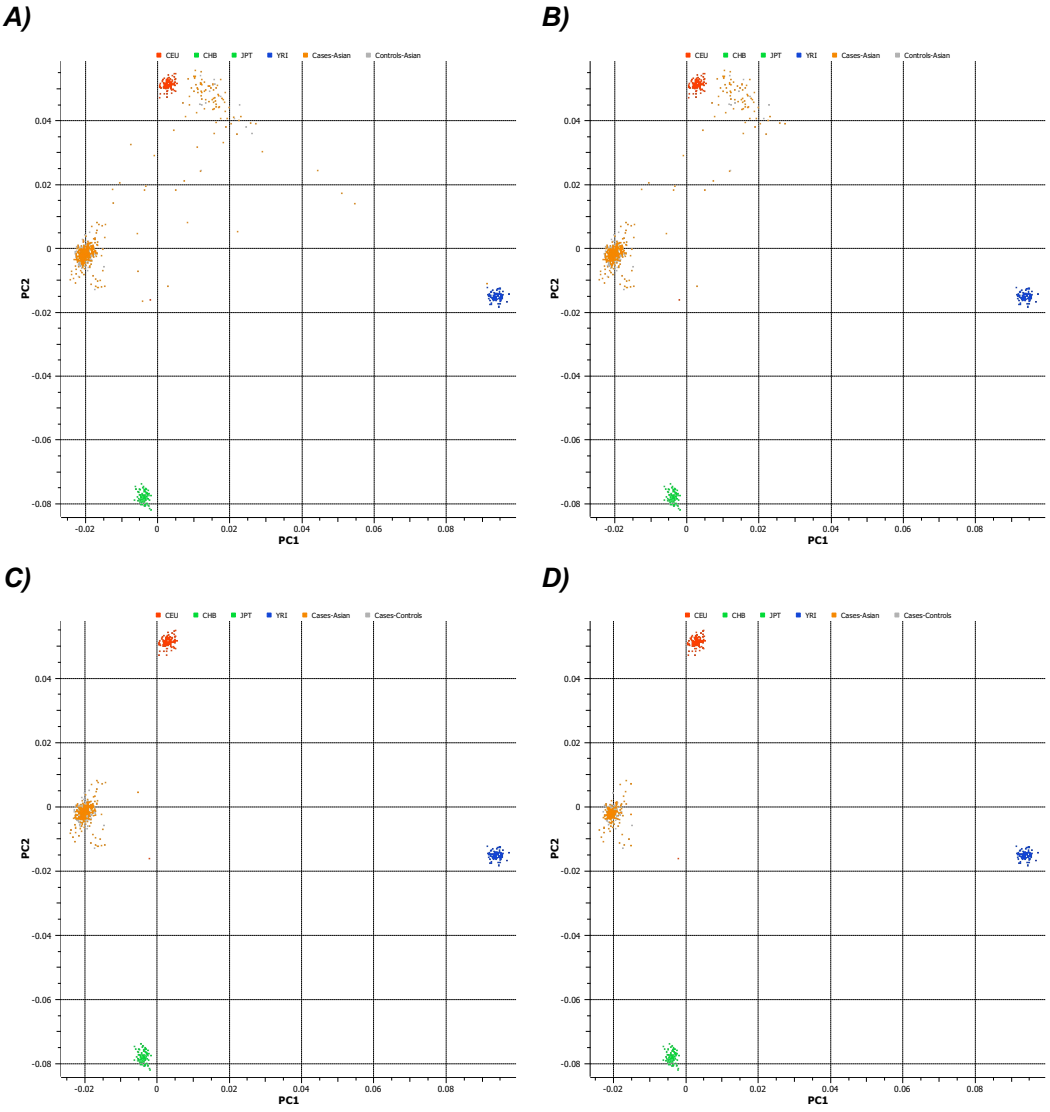
In order to get a homogenous group a PCA was used to identify population stratification. Outliers were visualized in a scatter plot and excluded from the further analysis. As a cut-off for removal of outliers, the standard deviations 3, 2.5 and 2 were tested (Table 16). The results are visualized in comparison to the Illumina ethnicity controls (CEU, YRI and CHB-JPT) (Figure 49). The distribution of the cases and controls in relation to the Illumina ethnicity controls showed that they were clustering between Europeans (CEU) and East Asians (CHB-JPT), confirming their South/West Asian ancestry. However, for simplicity reasons we refer to this cohort as Asian cohort.

Table 16 *Summary of the influence of removal of outliers with different standard deviations on the number of remaining cases and controls*

	Cases remaining	Controls remaining	IF Lambda
A No removal of outlier	387	200	1.709
B Removal of outlier SD >3	358	176	1.522
C Removal of outlier SD >2.5	245	136	1.083
D Removal of outlier SD >2	136	61	1.074

Legend SD: Standard deviation of the mean; IF: Inflation factor

Figure 49 *Scatterplot for PCA of Asian cohort with stepwise removal of outliers*



Legend *Scatter plot of the distribution of cases and controls along the top two principal components (PC1 and PC2). The results are visualised in comparison to the Illumina ethnicity controls (CEU, YRI and CHB-JPT). Number of remaining cases and controls for each scenario can be found in Table 16.*
A) *prior to the exclusion of non-Asian individuals. There is a clustering of samples towards the CEU reference group visible.*
B) *after ancestry selection for Asian by removal of cases and controls with principal components of >3 SDs from the mean. There is still a clustering of samples towards the CEU reference group visible.*
C) *after ancestry selection for Asian by removal of cases and controls with principal components of >2.5 SDs from the mean. The clustering of samples towards the CEU control group is filtered out.*
D) *after ancestry selection for Asian by removal of cases and controls with principal components of >2 SDs from the mean. The number of remaining samples has decreased substantially without generating a visibly more homogenous group.*

The scenario with removal of outliers with >2.5 SD was deemed to be the best compromise between generating a homogenous group and loss of samples.

GWAS power calculation

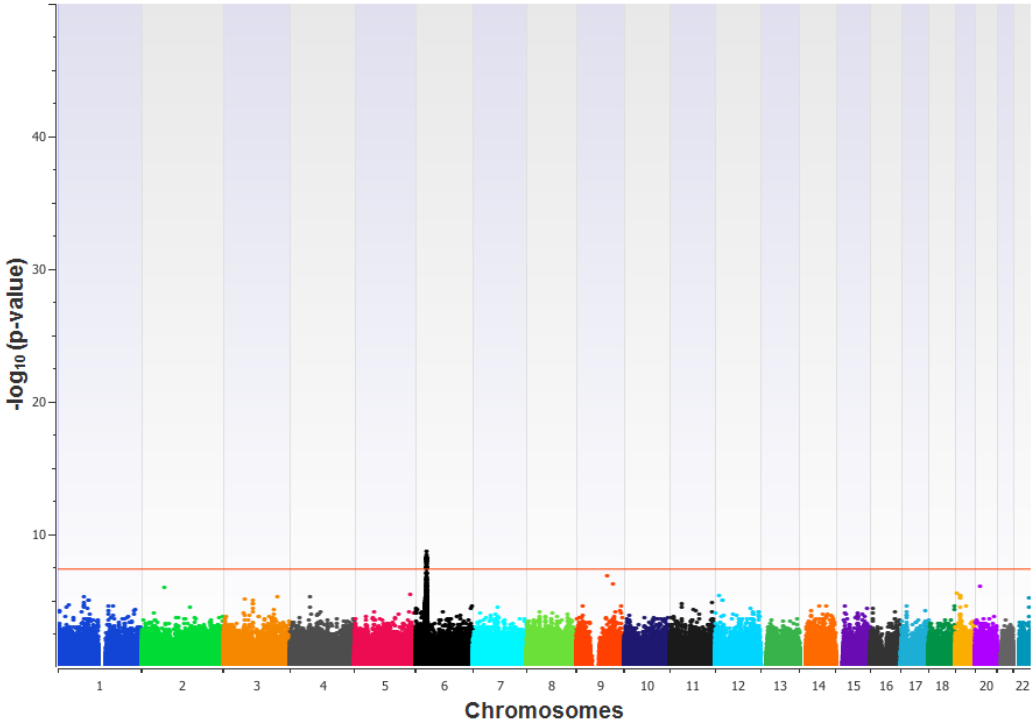
Comparing 245 cases with 136 controls using alpha = 5×10^{-8} under an additive model, the power to detect association of an allele with a frequency of 0.1 in controls at a genotype relative risk (GRR) set at the same level as for the European cohort (2.19) is only 0.023. In order to exceed the power of 0.8 a genotype relative risk of 3.99 would be necessary. This indicates that the Asian cohort is only powered to detect risk alleles with an effect size > 4 at a power of 80%.

Results

A basic allele test was performed on 245 cases versus 136 controls with 712,190 markers. The genomic inflation factor lambda was 1.0837, not indicating stratification. The statistic method used was the basic allele test and the genome wide significance threshold was set at $p=5\times10^{-8}$.

49 markers were associated with the trait at genome-wide significance threshold. All of them were within one peak on chromosome 6p in the HLA region. The results are displayed graphically as a Manhattan plot (Figure 50).

Figure 50 *Manhattan plot for BAT of Asian cohort*



Legend *Shown is the Manhattan plot for the analysis of 245 cases and 136 controls with the 712,190 genotyped SNPs. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log p-value of each marker on the y-axis. Same QC criteria were used as in the discovery cohort. In contrast to the discovery cohort only one locus reaches genome-wide significance on chromosome 6. This locus corresponds to the HLA region. No locus outside the HLA region reaches significance.*

Details for the top markers reaching genome wide significance are displayed in Table 17.

Table 17 *Lead SNPs reaching genome wide significance in the Asian cohort*

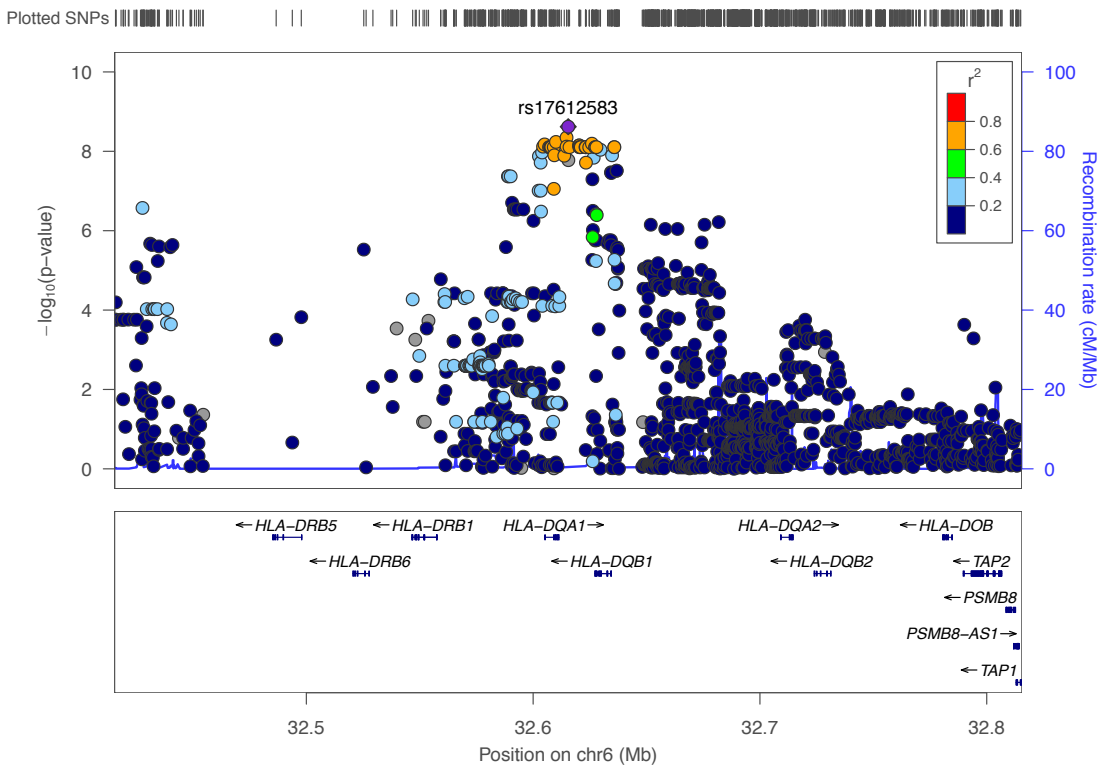
Locus	**Gene**	**SNP**	**Minor allele**	**MAF cases**	**MAF controls**	**OR**	**p-value**
6p21.3	*Upstream HLA-DQA1*	rs17612583	A	0.68	0.46	2.5	2.40×10^{-9}
6p21.3	*HLA-DQA1*	rs17612482	T	0.69	0.47	2.5	4.50×10^{-9}
6p21.3	*HLA-DQA1*	rs34843907	G	0.69	0.47	2.4	5.80×10^{-9}

Legend *Minor allele frequencies and p-values of the lead SNPs associated in the Asian cohort with the disease. Minor allele is defined as calculated in the control cohort.*

Equal to the discovery cohort the main peak on chromosome 6 corresponded to the HLA region. The lead marker was rs17612583 ($p=2.40\times10^{-9}$) approximately 700bp

upstream of *HLA-DQA1* (Figure 51). This *locus* is the same as in the imputed dataset of the European cohort tagged by the SNP rs9273542.

Figure 51 *Locus zoom for rs17612583*



Legend *Shown is the locus zoom for the SNP rs17612583 on chromosome 6p. The index SNP is marked with a purple diamond and is intergenic between HLA-DQA1 and HLA-DQB1. The SNPs with the second and third lowest p-value (not annotated) are in close proximity and are in LD with the index SNP. This is indicated by their orange colouring. Note, that this locus is the same as tagged by the lead SNP rs9273542 of the European cohort.*

The SNP with the second lowest p-value was rs17612482 ($p=4.50\times10^{-9}$) in the exonic region of the gene *HLA-DQA1.* The marker with the third lowest p-value was rs34843907 ($p=5.80\times10^{-9}$) in the intronic region of *HLA-DQA1*. Both of them were in strong LD with the index SNP. This was indicated by their orange colouring in the *locus* zoom plot (Figure 51).

We tested if any marker is independently associated with the disease. After conditioning on the lead SNP rs17612583 no further markers reached genome wide significance indicating that the HLA peak is driven by a single signal.

In contrast to the discovery cohort, no additional *loci* outside the HLA region reached genome–wide significance.

Next, we specifically looked up the lead SNPs (rs2637678 and rs10518133) tagging the *loci* reaching genome wide significance outside the HLA region in the European association study. As the Asian cohort was not imputed, none of the lead SNPs from the European imputed dataset were represented in the Asian cohort. We hence looked up the markers from the pre-imputed European dataset reaching genome wide significance in the 6q22.1 *locus* (rs549262 and rs648210) (Table 18). As this is a single test, a significance level of $p<0.05$ was chosen. The markers rs549262 and rs648210, which tagged the 6q22.1 *locus* in the pre-imputed European GWAS did not reach significance (p-values of 0.117 and 0.375, respectively) in the Asian cohort, but showed the same direction of effect (the minor allele was less common in cases than in controls) (Table 18).

Table 18 *Results in the Asian cohort for the European lead SNPs on chromosome 6q22.1*

Locus	**Gene**	**SNP**	**Minor allele**	**MAF cases**	**MAF controls**	**OR**	**p-value**
6q22.1	*DSE*	rs549262	A	0.17	0.21	0.74	0.12
6q22.1	*FAM162B*	rs648210	G	0.21	0.24	0.85	0.38

Legend *Minor allele frequencies and p-values of the lead SNPs* of the pre-imputed European cohort *in the Asian cohort*

A power calculation to assess if the study was powered to detect an association with a p-value <0.05 at this *locus* was performed. Using the parameters from the European cohort for the marker rs549262 (MAF in controls=0.40, OR=0.63) the Asian cohort would have the power of 0.88 to detect an association at that *locus*. However, when taking the different allele frequencies in the different ethnicities into account, hence using the MAF and the OR of the Asian study (MAF in controls=0.21, OR=0.74) the power to detect an association at a p-value of <0.05 was only 0.39.

In summary, the lead SNP in the Asian cohort (rs17612583) is tagging the same region (*HLA-DQA1/HLA-DQB1*) as the lead SNP (rs9273542) in the European discovery cohort. However, the associations outside the HLA region could not be replicated at this point.

Chapter 2. Meta-analysis European and Asian cohort

The replication cohort confirmed the association in the HLA region. However, the peaks on Chromosome 6 q-arm and Chromosome 4 p-arm could not be replicated in the Asian cohort.

In the situation of small sample sizes where the power of the study is limited a meta-analysis of the results can be useful. The focus is on investigating the direction of effect of the risk alleles, rather than replicating the signals at genome wide significance level.

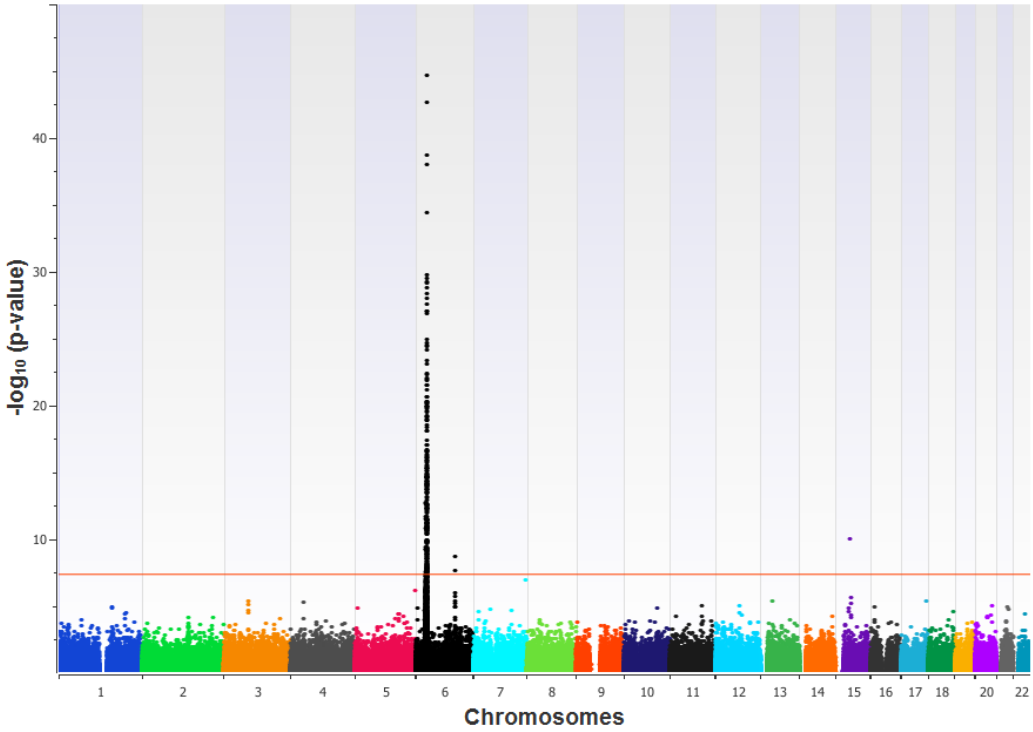
A commonly used tool to perform a meta-analysis is METAL [119,76]. In short, METAL uses the direction of the effect of a risk allele combined with the p-value for this allele observed in each single study to calculate an overall effect size and p-value for the specific allele. The sample size of each study is taken into account when weighing the effect size of each study result [119]. Therefore, this method can combine the evidence of association findings from individual studies by using appropriate weights. More details can be found in the introduction part on page 58.

Results

Pre-imputation European dataset and Asian dataset

For the European dataset the pre-imputation GWAS results of 422 cases and 5651 controls with 187,163 genotyped markers were included. For the Asian dataset 245 cases and 136 controls with the 712,190 genotyped markers were included. The number of overlapping markers and therefore included in the Meta-analysis was 185,384. Two *loci* on chromosomes 6 and one isolated marker on chromosome 15 reached genome wide significance (Figure 52).

Figure 52 *Manhattan plot for the trans-ethnic meta-analysis*



Legend *Shown is the Manhattan plot for the trans-ethnic metanalysis of the pre-imputed European discovery and the Asian replication cohort. The chromosomal position of the markers is represented on the x-axis corresponding to the human genome GRCh37/hg19. The -log p-value of each marker on the y-axis. Two loci on chromosome 6 reach genome wide significance and a single marker on chromosome 15.*

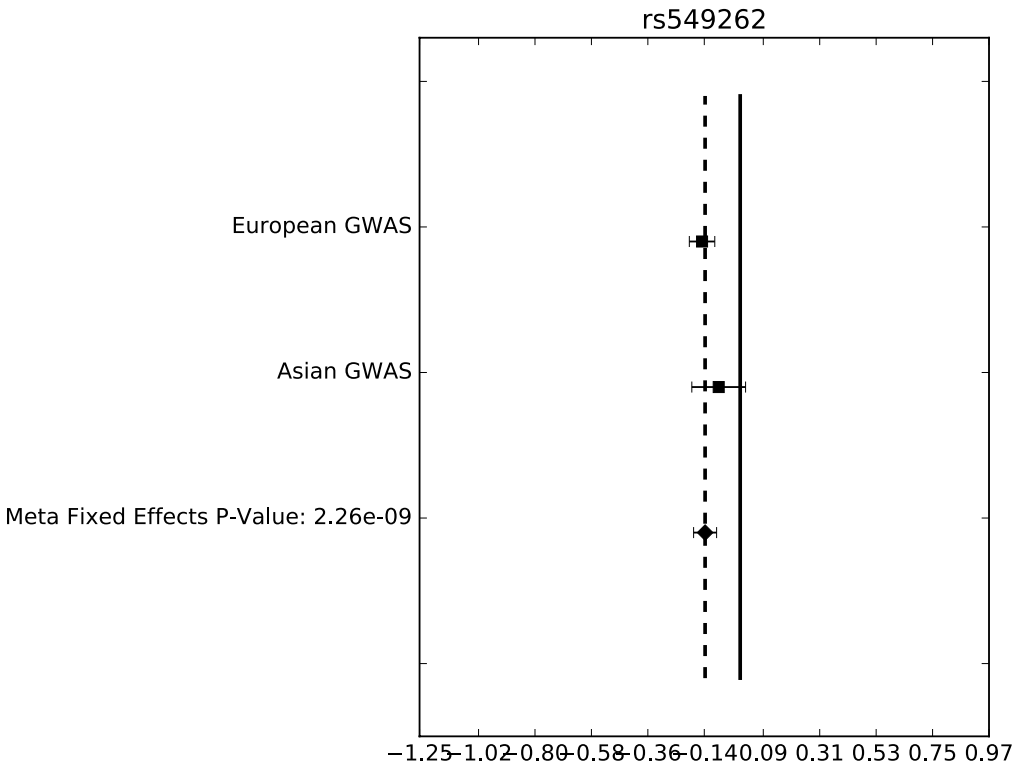
In the Meta-Analysis across the European and the Asian the strongest signal corresponded as previously seen to the HLA region. The marker with the lowest p-value was rs4947342 ($p=2.42\times10^{-45}$, OR=1.39, 95%CI=1.32-1.45) upstream of *HLA-DQB1*. The markers with the second lowest p-value was rs479536 ($p=2.57\times10^{-43}$, OR=1.37, 95%CI=1.31-1.43) upstream of *NOTCH4*.

The second *locus* on chromosome 6 reaching genome wide significance corresponded to the 6q22.1 region. The marker with the lowest p-value was rs549262 ($p= 2.26\times10^{-9}$, OR=0.87, 95%CI=0.83-0.91) in the intronic region of the gene *DSE.* The marker with the second lowest p-value in the same *locus* was rs648210 ($p= 2.36\times10^{-8}$, OR=0.88, 95%Ci=0.84-0.92) in the gene *FAM162B*.

The results for the lead SNP in the 6q22.1 *locus* were plotted in a Forest plot revealing that in both studies the minor allele had a negative effect size. This demonstrated that

the direction of effect of the minor allele was the same in the European and Asian cohort (Figure 53).

Figure 53 *Forest plot for meta-analysis results for rs549262 on chromosome 6q22.1*



Legend: Forest plot for the marker rs549262. The vertical dashed line represents the overall measure of effect for the meta-analysis. The vertical solid line is the line of no effect. If the confidence interval for an individual study or the meta-analysis overlaps with the line of no effect, it demonstrates, that the individual study's effect size does not significantly differ from "no effect". Both GWAS, the European and Asian demonstrate a protective effect of the test (minor) allele.

The one SNP on chromosome 15, rs1898882, reaching genome wide significance, was the same as already considered false positive in the European discovery cohort.

Post-imputation European dataset and Asian dataset

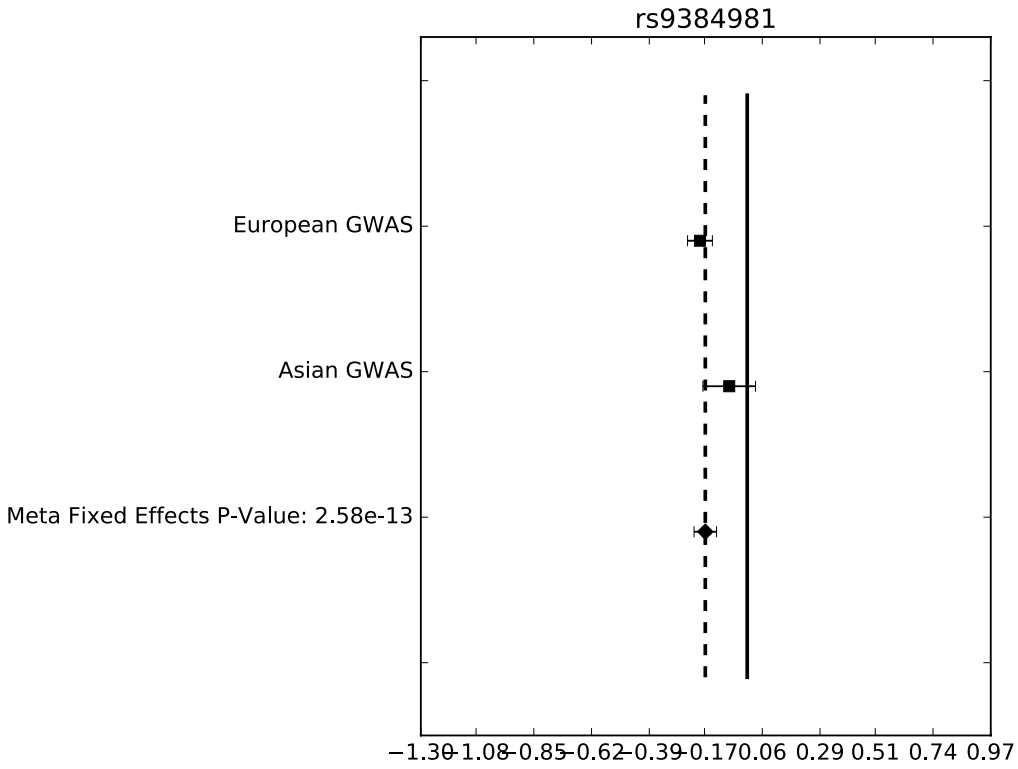
The density of markers was limited in this study because of the low number of markers in the pre-imputed European dataset. We repeated the analysis using the post-imputed European dataset. The number of overlapping markers increased to 386,378. Again, two *loci* reached genome wide significance, one on chromosome 6 p-arm

corresponding to the HLA region and one on chromosome 6 q-arm corresponding to the 6q22.1 *locus*.

The marker with the lowest p-value was rs2856696 ($p=7.79\times10^{-67}$, OR=1.48, 95%CI=1.42-1.55) downstream of *HLA-DQB1*, followed by rs9273471 ($p=2.34\times10^{-66}$, OR=1.48, 95%CI=1.42-1.55) intronic in *HLA-DQB1*, tagging the HLADR/DQ region.

On the chromosome 6 q-arm additionally the marker rs9384981 ($p=2.58\times10^{-13}$, OR=0.85, 95%CI=0.81-0.88) downstream of *CALHM6* was detected, tagging the chromosome 6q22.1 region (Figure 54).

Figure 54 *Forest plot for meta-analysis results for rs9384981 on chromosome 6q22.1*



Legend: *Forest plot for the marker rs9284981. The vertical dashed line represents the overall measure of effect for the meta-analysis. The vertical solid line is the line of no effect. If the confidence interval for an individual study or the meta-analysis overlaps with the line of no effect, it demonstrates, that the individual study's effect size does not significantly differ from "no effect". Both GWAS, the European and Asian demonstrate the same direction of effect for this marker.*

No additional *loci* reached genome wide significance.

Part 5. Discussion

In this thesis, we present our findings on *loci* outside the HLA region associated with SSNS as well as confirm previous association findings of SSNS within the HLA region. We performed a GWAS on the so far largest cohort of SSNS patients providing new insights into possible pathomechanisms of the disease. The identified *loci* strongly suggest an immunological component in the risk for development of SSNS.

Loci identified

HLA *locus* and SSNS

By far the strongest association found in our study is located in the HLA region. In our European discovery cohort, all three lead SNPs identified, rs9273542, rs9273529 and rs9273371, are located within the HLA-DR/DQ region, specifically in and around the genes *HLA-DQB1* and *HLA-DQA1*. These findings were replicated in the Asian cohort, where the lead SNP rs17612583 also tagged the region around the genes *HLA-DQB1* and *HLA-DQA1*, and in the transethnic meta-analysis of the European and Asian cohort. This is in accordance with previous GWAS published on SSNS. In Gbadegesin *et al*, four SNPS (rs1129740, rs9273349, rs1071630, and rs1140343) reached exome-wide significant threshold and all four of them were within or around the genes *HLA-DQA1* and *HLA-DQB1* [75]. In Debiec *et al*, the lead SNP in the European discovery cohort was located between the genes *HLA-DQB1* and *HLA-DQA2* and in the Maghrebian replication cohort between the genes *HLA-DRB1* and *HLA-DQA1* [76]. In both cohorts LD is extending over the whole HLA-DR/DQ region. Debiec *et al* further performed a transethnic meta-analysis across four cohorts and the one SNP reaching genome-wide significance was again around *HLA-DQB1*. The cohort ethnically most unrelated to ours was studied by Jia *et al,* who published a GWAS on Japanese SSNS patients. The top peak in this Japanese GWAS was also around the gene *HLA-DQB1*, indicating that the association with HLA-DR/DQ is conserved across ethnic boundaries [77].

Imputation of HLA alleles was performed in the European cohort. We identified that the composite haplotype *HLA-DQA1*02:01; HLA-DRB1*07:01; HLA-DQB1*02* was

associated with the strongest risk of disease development. The same haplotype was identified by Debiec *et al* to be associated with the disease [76]. Interestingly, South Asian children with SSNS carry the same risk haplotype [159]. The fact that the haplotype is preserved over different ethnicities strongly supports its relevance in the disease pathogenesis. Conversely, *HLA-DQA1*01, HLA-DQA1*01:03* and *HLA-DRB1*13* appear to be protective. *HLA-DQA1*01* was also identified as protective in the South Asian cohort [159].

Remarkably, in the Japanese cohort, the common risk haplotype of the European and South Asian population (*HLA-DQA1*02:01; HLA-DRB1*07:01; HLA-DQB1*02)* was not replicated [77]. The haplotype associated with the strongest risk for disease development was *HLA-DRB1*08:02, DQB1*03:02* and *HLA-DRB1*13:02, DQB1*06:04* appeared to be protective [77]. These findings may indicate that the associated HLA alleles differ depending on geographic and ethnic origin [77].

Table 19 *Overview of HLA alleles significantly associated with SSNS*

HLA allele	**Discovery cohort**	**Debiec et al [76]**	**Adeyemo et al [159]**	**Jia et al [77]**
	European	European	South Asian	Japanese
Deleterious				
*HLA-DQA1*02:01*	X	X	X	
*HLA-DRB1*07:01*	X	X	X	
*HLA-DQB1*02*	X	X	X	
*HLA-DRB1*08:02*				X
*HLA-DQB1*03:02*				X
Protective				
*HLA-DQA1*01*	X		X	
*HLA-DQA1*01:03*	X			
*HLA-DRB1*13*	X			
*HLA-DRB1*13:02*				X
*HLA-DQB1*06:04*				X

Legend *Classical HLA alleles identified in SSNS cohorts to be associated with the disease.*

However, the underlying mechanisms how the different HLA class II alleles impact the susceptibility for immune-mediated disorders including SSNS, autoimmune or inflammatory diseases is not clearly understood. The highly polymorphic nature of HLA genes and the extensive LD across the MHC region has made the identification of the causal variants in the MHC region and the disease a challenging task. Further, the majority (if not all) MHC-associated diseases are following a complex inheritance

pattern where the combination of particular HLA alleles with other genetic variants is crucial for increasing the susceptibility for the disease. In addition, non-inherited factors including environmental influences and epigenetic factors may also alter the risk for disease development. Additionally, immune-mediated diseases are commonly heterogeneous in their clinical presentation. All these factors complicate and hamper the detection of the underlying mechanisms how MHC molecules influence disease susceptibility.

However, for the better interpretation of the role of MHC molecules in immune mediated disease I will shortly summarize the principles of autoimmunity and immune dysregulation.

Principles of Autoimmunity

Autoimmunity describes the failure of the adaptive immune system to distinguish between pathogens and self-antigens leading to erroneous immune response that causes tissue damage [160]. The two major representatives of the adaptive immune system are T cells and B cells.

T cells derive from immature precursors cells in the bone marrow and migrate to the thymus, where they develop into two effector cell lines, CD8+ cytotoxic T cells or CD4+ T helper cells. Activation of T cells requires antigen-presentation via specialised antigen presenting cells (APC), such as macrophages or dendritic cells.
An intracellular antigen is presented via HLA class I molecules to the CD8+ cells. This leads to the production of specific cytotoxins and consequently to apoptosis of the target cell. An extracellular antigen is presented via HLA class II molecules to CD4+ T helper cells. This leads to the differentiation of T helper cells into several subsets of T helper and regulatory cells. Different cytokine profiles released by APC cells promote the differentiation into the subtypes of T cells, which then mediate different immune responses.

The second arm of the adaptive immune system are B cells. B cells are antibody producing cells. The immature precursor cells are stored in the bone marrow. In the periphery, mature B cells express the antigen-sensing immunoglobulin IgM on their

cell surface that can bind antigens, also referred to as B cell receptor (BCR). The presentation of an antigen in presence of T helper cells activates the mature B cells to develop into either antibody-producing plasma cells or memory B cells. Five isotypes, or classes, of antibodies (IgM, IgD, IgG, IgA, and IgE) exist. Initially activated B cells produce low-affinity IgM, but subsequently they experience an isotype switch, producing high-affinity IgG (in a smaller proportion IgA, or IgE) type antibodies [161]. Plasma cells can be short or long-living. The short-lived plasma cells die after a few days, whereas long lived plasma cells return to the bone marrow and continue producing antibodies independently of antigen exposure. [162]

Antibodies have three main roles in eliminating pathogens. First, antibodies can neutralize the recognised pathogen via binding it (*e.g.* the virus) and preventing it from entering a cell. Secondly, antibodies can activate other immune cells, such as macrophages, which then attack the pathogen and thirdly, antibodies are activators of the classic pathway of the complement system by binding to C1q. [162]

In addition to the role of B cells of producing antibodies, B cells also interact with T cells [162]. B cells are involved in the antigen presentation to both, CD8+ and CD4+ T lymphocytes [163]. Secondly, B cells are involved in the co-stimulation of T cells leading to T cell activation and memory. Further, B cells produce inflammatory cytokines (IL-6 and TNF-alpha), which lead to T cell activation and differentiation [163,162]. On the other hand, B cells are also assumed to have a regulatory function on T cell activation via the production of inhibitory cytokines (IL-10 and IL-35) [163,162].

Autoimmunity derives if these mechanisms turn against an antigen of the own body (self-antigen), rather than a pathogen. There are several immune mechanisms to induce self-tolerance and hence protect against B and T cells reacting to a self-antigen. In general, those are divided in central and peripheral tolerance mechanisms [160]. Central tolerance is established in the thymus and bone marrow during the maturation of lymphocytes, where B and T cells are eliminated if they react against a self-antigen [160]. However, this process is not perfect and not all possible self-antigens are presented in the thymus and bone marrow. Further, tolerance has to

address the change occurring in the body with time. Hence, additional peripheral self-tolerance mechanism ensure that B and T cells do not react against a self-antigen. For example, if a T cell detects a self-antigen, but no additional signals, such as cytokines, are present, the T cell will be inactivated [160].

If any of those mechanisms fail, autoimmunity can develop against any antigen in the body. For example, in type 1 diabetes the immune system reacts against the pancreatic ß-cells. In membranous nephropathy, PLA2R is a recognized antigen [80]. However, how self-tolerance is disturbed and how autoimmunity is triggered remains to be understood.

Hundreds of associations between immune-mediated disease and the HLA locus, which encode the receptors that are expressed by APC to trigger the immune response, have been reported. In general, different HLA class II alleles in this locus could possibly influence the HLA - peptide - T cell receptor (TCR) interactions and promote the activation of autoreactive effector CD4+ T cells [164,165]. HLA gene expression levels and differential HLA protein stability could further influence the activation of the immune system [164,165]. This could subsequently lead to increased TCR-mediated activation of autoreactive effector T cells, as well as to a reduction of regulatory T cells [164,165]. The threshold for T cell activation could be enhanced by impaired regulatory mechanisms of the immune system [164,165]. Arguably, it is the detection of associations of non-HLA genes that point towards impacted regulatory mechanism and hence substantially aid to understand immune-mediated disease, including in SSNS.

Non-HLA genes in the HLA *locus*

The majority of associations identified by GWAS on autoimmune disease to date involve the HLA region. Interestingly, in previous studies on autoimmune disease, conditional analysis on the top signals in the HLA region demonstrated multiple independent association signals within the region. For example, the international multiple histocompatibility complex and autoimmunity genetics network performed conditional regression analyses on seven autoimmune disorders (systemic lupus

erythematosus, Crohn's disease, ulcerative colitis, rheumatoid arthritis, myasthenia gravis, selective IgA deficiency, multiple sclerosis) in the HLA region [110]. For each of the diseases, after conditioning on the top marker, multiple further association signals were identified, revealing that the HLA region displays a complex effect, involving multiple *loci*, on autoimmune disorders.

Similarly, previous studies of SSNS identified independent association signals within the HLA region. In the study by Debiec *et al,* serial conditional analysis of the HLA association revealed an independently associated SNP in the HLA region located in close proximity of *BTNL2* (rs9348883) [76]. However, in our GWAS we were unable to confirm an independent association with this gene. This could reflect the different ethnicities analysed, as the *BTNL2* signal was primarily driven by the African cohort in that study [76].

Associations outside the HLA region

The adaptive immune response is triggered by antigen presentation via HLA molecules to T cells and is modulated by regulatory mechanisms, elements of which are encoded in non-HLA genes [78]. B cells as part of the humoral immune response are capable of producing antibodies directed against a specific antigen. Consequently, the identification of immune regulatory genes or/and a possible antigen could provide the most informative insights into the complex architecture of the dysregulated immune response in specific diseases.

In the European cohort we found two *loci* outside the HLA region achieving genome-wide significant associations with SSNS. For the further discussion, it is important to note that our GWAS cannot pinpoint the causal allele or gene and therefore we are using a “closest gene” approach to discuss the genes in these *loci* tagged in our GWAS.

6q22.1 *locus*

The strongest signal outside the HLA region in the European cohort was on chromosome 6q22.1 with the lead SNP rs2637678 ($p=1.27\times10^{-17}$, OR=0.51). This marker was imputed, however a high imputation score of DR2 of 0.96 makes us confident about its accuracy.

Interestingly, Debiec *et al* also described two SNPs reaching suggestive levels of significance rs59882675 and rs2858829 outside the HLA *locus*. The second SNP reached a p-value of 6.8×10^{-8} and is also on chromosome 6q22.1 in the intergenic region between *CALHM6* and *DSE*. The SNP rs2858829 reported by Debiec *et al* is identical to one of the lead SNPs in this *locus* identified in our study where it reached a p-value of 1.72×10^{-16}. We consider the existence of this published suggestive association at rs2858829 as an independent confirmatory evidence for our genome-wide significant finding. However, one important limitation of using this study as a replication cohort has to be taken into account. Debiec *et al* neither provide information on the minor allele frequency in cases and controls, nor is the OR given, hence we do not know the direction of effect in their study. As the majority of cases in the Debiec *et al* study were of European origin we can assume a comparable minor allele frequency. Nevertheless, information on the OR would be needed to confirm the same direction of effect.

CALHM6 as candidate gene

The lead SNP rs2637678 is located downstream the gene *CALHM6 (Calcium Homeostasis Modulator Family Member 6)*. *CALHM6* was previously also annotated as *FAM26F* or *INAM (IRF-3-dependent NK-activating molecule)*, before the accepted consensus name became *CALHM6* [166,158]. *CALHM6* is located on chromosome 6q22.1 and consists of three exons. It is neighboured by the genes *KRT18P22* and *TRAPPC3L* (GRCH37/hg19). Sequence alignment showed that *CALHM6* has remained evolutionary conserved [166].

Malik *et al* investigated CALHM6 using an *in silico* approach and provided essential insight in the structure and function of the protein [166]. CALHM6 is expected to be a membrane protein, based on multiple predicted transmembrane helices. It is predicted to contain a single, well conserved calcium homeostasis modulator domain, rendering it member of calcium homeostasis modulator family [166]. Members of the calcium homeostasis modulator family (*e.g.* CALHM1) were shown to control cytosolic calcium concentration and speculated to be pore-forming ion channels [167]. Based on

function prediction tools CALHM6 is assumed to be a cation channel, however, the exact function and even the type of channel remains vague [166].

Further, the protein structure of CALHM6 is predicted to have an immunoglobulin-like fold, which could facilitate interactions and potential synapses between immune cells [166]. Indeed, CALHM6 was originally identified as a membrane protein that is involved in dendritic cell mediated activation of natural killer cells [168]. Thereafter, numerous studies reported the protein to be involved in infection, stress and immune response and hence CALHM6 is thought to have an important role in the regulation of the immune system [169].

CALHM6 is highly expressed in the spleen and lymphocytes. A study investigating *CALHM6* expression with real time PCR in monkey (*Rhesus macaques*) lymphocytes, showed that *calhm6* is highest expressed in CD4+ T cells, followed by CD8+ T cells and CD20+ B cells [170]. When interrogating expression platforms such as The Human Blood Atlas, *CALHM6* was found to be highly expressed on non-classical and intermediate monocytes as well as naïve and memory B cells [171].

The same study on monkey (*Rhesus macaques*) lymphocytes further demonstrated that INF-γ is the strongest stimulator for *calhm6* expression, followed by INF-α and TNF alpha. IFN-γ is a pro-inflammatory cytokine produced by NK cells, CD4+ helper cells, CD8+ T and B cells. This may indicate that *CALHM6* plays a role in IFN-γ response in both the innate and adaptive part of the immune system [172]. The review by Malik *et al* even proposes that *CALHM6* expression level is a hallmark for IFN-γ-led immune responses (reflected in its previous name *INAM* - IRF-3-dependent NK-activating molecule) [166]. This has been confirmed by other studies, reporting that *CALHM6* has an increased expression as part of inflammatory responses, such as during different viral and bacterial infections, and is differentially expressed as part of immune responses [173,174].

Little is known about the mechanism of action of *CALHM6*, however, *CALHM1*, another member of the CALHM superfamily, has been shown to be an ATP release

channel [175]. As ATP is a reported trigger of apoptosis also in immune cells [176,177], one can speculate that *CALHM6* is involved in lymphocyte apoptotic mechanisms.

In relation to our disease, SSNS, these findings are consistent with the concept that altered immune regulation is a key risk factor for the development of SSNS. Hence, the role of *CALHM6* as one of our candidate genes and constituting an immunological component of the disease seems conclusive.

To assess whether our lead variants affect the expression of *CALHM6* we performed an expression quantitative trait *locus* (eQTL) analysis [178,179]. All three lead variants (rs2637678, rs2637681, rs2858829) are strong eQTLs associated with *CALHM6* expression. It is important to notice that for the lead variant (rs2637678) the minor allele was more common in controls than cases indicating that the major allele is the risk allele. Hence, the major (risk) allele was associated with decreased *CALHM6* expression. This leads to the hypothesis that the presence of the risk allele may impair an important immune regulatory role of *CALHM6* in lymphocytes.
The variant rs2858829 lies additionally in a site of strong regulatory activity according to the ENCODE data. This is indicated by increased histone acetylation, a pattern typical of enhancers, by DNaseI hypersensitivity clusters, which is associated with gene expression regulation, and by ChIP-seq data showing transcription factor binding sites in this region.

Interestingly, in a GWAS on European patients with ulcerative colitis (UC) the same *locus* (rs2858829, $p= 8.97\times10^{-9}$, OR=1.12, 95%CI=1.08–1.16) was identified [180]. The eQTL analysis revealed the same *cis* regulatory evidence on the expression of *CALHM6* gene. However, the risk allele showed the opposite direction, increasing the expression of *CALHM6*. The same group further performed microarray studies and showed that *CALHM6* was consistently upregulated in inflamed colon mucosa tissue form UC patients compared to healthy controls [180]. Although this study did not explain the mechanism of action of *CALHM6*, it is suggesting that differential expression of *CALHM6* is crucial in determining the immune response of an individual in a disease condition [180].

DSE as candidate gene

It is important to note that GWAS identify variants that represent a haploblock rather than a specific gene associated with the disease. Thus, it is possible that another gene than *CALHM6* in the same haploblock may actually be causally associated with the disease. Interestingly, the lead variants in the imputed dataset also alter the expression of the gene *Dermatan sulfate epimerase-1 (DSE)*. Hence, it could be that this gene (or even both *CALHM6* and *DSE*) is the key to the pathogenesis of SSNS.

DSE encodes an enzyme, named dermatan sulfate epimerase-1, that converts chondroitin D-glucuronic acid to dermatan L-iduronic acid (IdoA) during the biosynthesis of dermatan sulfate (DS) [181]. The presence of L-iduronic acid distinguishes dermatan sulfate from chondroitin sulfate (CS). DS and CS are members of a large family of polysaccharides called glycosaminoglycans (GAGs). Covalently bound to a core protein they form proteoglycans (CS/DS-PGs), such as decorin and biglycan [182].

GAGs are an essential part of the glycocalyx surrounding the endothelial cells of the glomerular filtration barrier [183]. Several studies are supporting that GAGs are involved in the charge-selective and/or size-selective permeability properties of the glomerular basement membrane. Specially heparan sulfate (HS) has been considered an important part for maintaining the glomerular filtration barrier [184,183]. However, findings are controversial and a recent study on a heparan sulfate gene knock out zebrafish model showed that the glomerular permeability was unaffected [185].

The gene *DSE* is ubiquitously expressed, with a high expression in the kidneys. An experiment with *dse-1* knock out mice showed that the mutant kidneys only contained 4% of the activity of the wild-type samples, indicating that *DSE* is the predominant enzyme in the kidneys [182]. The antibodies LKN1 and GD3A12 are tagging specific DS domains. The GD3A12 tagged DS domain was found in rat kidneys around the large blood vessels and less in the peritubular space and near the Bowman's capsule [186]. The LKN1 tagged DS domain was found in the tubular interstitial space, but not in the healthy glomerulus [187].

Changes in the CS/DS content and modifications have been found in different animal models for renal disease. Oversulfated CS/DS with low-iduronic acid content was found in a rat model for tubulointerstitial nephritis [188]. In diabetic rats the production of DS in the mesangial cells was much increased compared to healthy controls [189]. Recently, a study in humans looked at the expression of different DS domains in kidney biopsies of patients with glomerular disease (FSGS, MN and SLE) [190]. The two antibodies tagging DS, LKN1 and GD3A12, were used to investigate the expression of the different DS domains. In healthy control kidneys, neither DS domains are expressed in the glomerulus. However, in patients with glomerular diseases the DS domain tagged by LKN1 was highly expressed in the glomerulus. This suggests a role of this DS domain in glomerular disease including FSGS [190].

All our three lead variants are eQTLs for *DSE*. The minor allele leads to decreased expression of the gene. Considering that in our SSNS cohort the minor allele was protective, the risk allele (major allele) increases the expression of *DSE*. An increased expression of *DSE* could lead to an overproduction of DS. In the assumption that FSGS reflects just a more progressed stage of SSNS, overexpression of DS in the glomerulus could be implicated also in the pathogenesis of SSNS.

4q13.3 *locus*

PARM1

We found a genome-wide significant association with a *locus* on chromosome 4q13.3 (lead SNP rs10518133, $p=2.50\times10^{-8}$, OR=1.96), which is located within the intronic region of the gene *PARM1 (Prostate androgen-regulated mucin-like protein 1). PARM1* was mainly linked and investigated in relation to prostate cancer [191]. An increased expression of the gene has been seen in androgen dependent cell lines, and decreased expression upon castration, indicating that the gene expression is regulated by androgen [191]. A recent study proposed *PARM1* as a potential prognostic tumour biomarker for colorectal carcinoma [192]. However, it has a wide tissue distribution with especially high expression in heart, kidney and placenta [191].

The protein itself is member of the mucin-family. It has been further implicated to have an oncogenic role, especially in CD8+ T cell leukemia [193]. Further, in cardiac

research, *PARM1* was found to be expressed in the endoplasmatic reticulum of cardiac myocytes and is speculated to be involved in cardiac remodelling in hypertensive heart disease [194]. There is currently no obvious association of *PARM1* with renal or autoimmune disease. Yet, the fact, that *PARM1* is regulated by androgens may be of interest: a significantly higher proportion of boys suffer from SSNS than girls. Moreover, androgen levels increase in puberty, a typical time when the disease spontaneously resolves. This, of course, invites speculation that a sex-steroid regulated protein is involved in the disease mechanism and justifies further investigation of this locus, assuming it can be replicated in other cohorts.

Nevertheless, the significance of this finding in our GWAS remains unclear, because we were unable to replicate it in the Asian cohort or the transethnic meta-analysis. Interestingly, one of the two *loci* reaching suggestive significance in Debiec *et al*, tagged by the SNP rs59882675 with a p-value of 5.9×10^{-8} is also on chromosome 4q13 approximately 250kb away from our lead marker. Rs59882675 is within the intronic region of *Betacellulin (BTC*), the next gene upstream of *PARM1* on chromosome 4q [76]. The lead SNP in our study and the SNP identified by Debiec *et al* are however separated by a strong recombination hotspot (> 50cM/Mb). This hotspot is at least equally strong in African populations, so that the different ethnicities cannot explain the separation of this *locus* between the previous and our study.

BTC

Betacellulin belongs to the epidermal growth factor (EGF) family that signals through the EGF receptor [195]. The best reported biological role of *BTC* is its action on pancreatic ß-cells. *BTC* stimulates ß-cell proliferation in animal models but also in humans [196]. It can convert a number of non-ß-cells to insulin producing cells. Studies further demonstrated that *BTC* is a ß-cell growth factor and its injection into diabetic rats or mice can improve glucose tolerance and ß-cell volume [197].
Another function of *BTC* is its role in reproduction. Together with other growth factor receptor ligands, *BTC* plays a central role in the transmission of LH signals in the ovarian follicles and oocyte maturation [198].

Additionally, *BTC* has been implicated in cancer pathophysiology. Overexpression of *BTC* was observed in different forms of cancer including endometrial adenocarcinoma, hepatocellular carcinoma and multiple pancreatic cancer cell lines [195]. Further, *BTC* plays a role in angiogenesis by stimulating the proliferation of different types of vascular smooth muscle cells [199]. No function directly related to the podocytes, proteinuria or the immune system has been reported.

Further insight into the function of *BTC* was gained via *btc* knockout mice and a mouse overexpressing *btc* [195]. The knockout mice showed no kidney-relevant phenotype [195]. The mice overexpressing *btc* revealed multiple phenotypic changes, including increased bone mass and gastric epithelium hyperplasia. Again, no phenotype involving the kidneys or the immune system was described [195].

Comparison of findings European – Asian GWAS

HLA *locus*

The strongest association in both studies, the European and the Asian cohort, was located in the HLA region. Hence, we can state that the association with the HLA region was replicated in the Asian cohort.
In the European discovery cohort, the three lead SNPs identified were located within the HLA-DR/DQ region, specifically in and around the genes *HLA-DQB1* and *HLA-DQA1*. All three lead SNPs were associated with an increased risk of the disease. These findings were replicated in the Asian cohort, where all three lead SNPs (rs17612583, rs17612482 and rs34843907) also tagged the region around the genes *HLA-DQB1* and *HLA-DQA1* and were associated with an increased risk of the disease. However, in order to know if the same HLA alleles are associated in both ethnicities with the disease, HLA imputation of the Asian cohort is necessary, which was outside the scope of this thesis. Hence, we are currently able to say that the association with the HLA *locus* was replicated in the Asian cohort; if the same HLA alleles are associated with the disease in both ethnical cohorts remains a question for future studies.

6q22.1 *locus*

In the European cohort, the strongest association outside the HLA region was within the 6q22.1 *locus*. In the Asian cohort this *locus* was not significantly associated with the disease. Again, it has to be taken into account that the Asian cohort was not imputed and therefore the lead SNPs were not represented in this cohort. However, the SNP rs549262, which is in the same *locus* and was significantly associated in the pre-imputation European SSNS cohort, was also represented in the Asian cohort. When testing if the association at this single SNP can be replicated in the Asian cohort, the p value did not reach significance level of <0.05, but the OR showed the same direction of effect. A power calculation revealed that the number of cases and controls in the Asian cohort would have been sufficient to detect an association with a significance threshold of p-value < 0.05 at the marker rs549262. This calculation was based on the assumption that the allele frequency and the odds ratio are the same in the Asian and in the European cohort. However, in the Asian cohort the allele frequency of this SNP is lower than in the European cohort and the required number of cases and controls needed to detect an association, is hence higher. Arguably, the Asian study was underpowered to detect this region as significant [200]. Also differences in the LD pattern can result in different results of association testing [200]. Either, the genetic architecture of the European and Asian SSNS patients is simply different and the causal variant is a different one or, if it is the same, the effect size varies between these two ethnicities. Indeed, it has been shown that a relevant number of variants identified by GWAS in one ethnical group are not showing a comparable genetic effect in other ethnicities [201]. In addition, the same allele can increase risk for one disease but be protective for another, as has been shown for eight *loci* in a study analysing ten autoimmune diseases with paediatric age of onset [202].

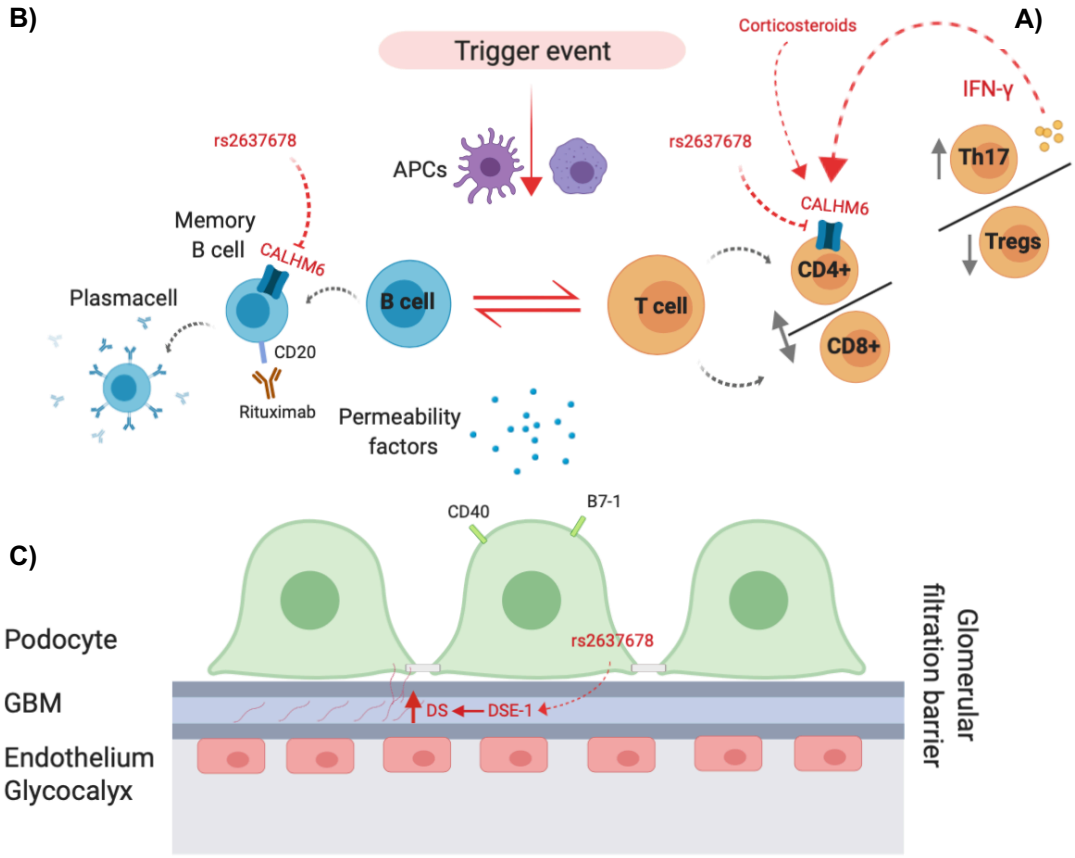
The transethnic meta-analysis of the European and the Asian cohort did not reveal any further *loci* of significance. The same *loci* as in the European GWAS did reach genome wide significance also in the meta-analysis, reflecting that the results are mainly driven by the European cohort. Importantly, the lead SNP in the chromosome 6q22.1 *locus*, showed in both cohorts the same direction of effect, supporting the relevance of this finding.

Disease mechanism

At present, the pathomechanisms leading to podocyte injury and altered glomerular filtration barrier in nephrotic syndrome are still unknown. A detailed overview of the current understanding of the pathogenesis of idiopathic nephrotic syndrome is outlined in the introduction section on page 26. The main theory of disease mechanism proposes that minimal change disease involves lymphocytes that produce autoantibodies and/or a yet unknown circulating permeability factor, which induces podocyte injury increasing the permeability of the glomerular basement membrane.

We subsequently want to outline how our findings fit into and impact this current hypothesis of pathomechanisms underlying SSNS (Figure 55).

Figure 55 *Proposed impact of our findings on disease mechanisms*



Legend *Summary of proposed mechanisms leading to SSNS.*
APCs: Antigen-presenting cells; GBM: glomerular basement membrane; DSE-1: Dermatan sulfate epimerase; DS: Dermatan sulfate
A) *CALHM6 is expressed on CD4+ lymphocytes and is assumed to have an important immune regulatory function, possibly via mediating apoptosis of lymphocytes. IFN-γ is a strong inducer of CALHM6 expression. IFN-γ is also produced by Th17 cells, and hence could be the mediator of an autoregulatory mechanism of CD4+/Th17/Tregs cells. The identified lead variant rs2637678 downregulates the expression of CALHM6, which could lead to an alteration of this immune regulatory mechanism. In contrast, corticosteroids increase the expression of CALHM6, hence the treatment with corticosteroids could restore the immune regulatory function of CALHM6. This then leads to remission of nephrotic syndrome.*
B) *CALHM6 is also expressed on memory B cells a subpopulation of B cells which also carry CD20. The recovery of memory B cells after rituximab has been linked to relapse of nephrotic syndrome. Hence, altered immune regulatory mechanism because of downregulated CALHM6 in memory B cells could be involved in the pathomechanisms of SSNS. Treatment with Rituximab addresses those B cells, which also show altered CALHM6 expression, and hence the removal of these cells can induce remission.*
C) *Another pathway could be that the identified lead variant rs2637678 upregulates the expression of DSE, which consequently leads to an increased production of DS in the glomerulus. Increased expression of DS has been associated previously with FSGS. However, how this could play a role in SSNS pathophysiology remains to be elucidated.*

Permeability factors

Over the years, several molecules have been shown to be able to modify the permeability of podocytes including hemopexin, cardiotrophin-like cytokine 1 (CLC-1) and the soluble form of urokinase-type plasminogen activator receptor (suPAR). A detailed summary of these most relevant permeability factors associated with MCD and FSGS is given in the introduction section on page 28. Our findings do not favour any of those factors, but rather contribute to the hypothesis that a permeability factor stems from T- or B cell dysregulation, or both.

T cell involvement

Since decades, SSNS was considered a T cell mediated disease. This was originally based on following observations. The treatment with steroids and cyclophosphamide, knowing to alter cell-mediated immune responses, leads to remission of steroid sensitive nephrotic syndrome. The injection of lymphocytes supernatant of patients with nephrotic syndrome in healthy controls led to an alteration of vascular permeability [31]. In paediatric patients with nephrotic syndrome remission was induced during measles infections, which led to prolonged depression of cell mediated immunity including T-cell subset reduction [45]. And the observation that nephrotic syndrome in patients with T cell lymphomas occurred and the treatment with chemotherapy resulted in spontaneous remission of nephrotic syndrome is supporting T cell involvement [46]. Since then, numerous studies investigating T cell involvement in the disease pathogenesis are published.

Considering that *CALHM6* has highest expression on CD4+ T cells, recent findings deserve closer attention: CD4+ T helper (Th) cells differentiate upon activation into different types of effector cells, which can be identified by their distinctive cytokine production profiles. The best known are INF-γ producing Th1 cells and IL-4 and IL-13 producing Th2 cells. Additionally, CD4+ Th cells can differentiate into so-called Th17 cells and regulatory T cells [203,204]. Th17 cells produce the signature IL-17. Th17 cells have been reported to participate in inflammation and autoimmunity processes. Conversely, CD4+ cells which differentiate into regulatory T cells (Tregs) downregulate the immune response.

An imbalance between Th17/Treg cells towards Th17 cells as well as increased levels of the Th17-related cytokine (IL-17) have been observed in adult minimal change nephrotic syndrome patients and with effective steroid therapy a normal ratio could be restored [54]. This was confirmed by Matsumoto *et al.* [205] who found that urinary IL-17 excretion is increased in minimal change nephrotic syndrome patients in relapse and restored to baseline with remission. Also, in children with idiopathic nephrotic syndrome an increased number of Th17 cells and decreased number of Tregs have been found [53]. IL-17 has been associated with podocyte damage and recent research showed that blockage of IL-17 led to improvement of albuminuria in a diabetic mouse model [206]. Interestingly, Th17 cells have been shown to produce not only their signature interleukin IL-17, but are also an important source of IFN-γ. The central role of these IFN-γ expressing Th17 cells is especially confirmed in other autoimmune disease [207].

In our study we were not able to identify any genes involved in the Th17/IL-17 pathway itself, however, based on all those observations and the convincing evidence that *CALHM6* plays an important role in immune regulatory process, we suggest the following hypothesis (Figure 55A).

CALHM6 is highly expressed on CD4+ cells and probably involved in sustaining the balance between Th17 and Tregs cells. In patients with SSNS the expression of *CALHM6* is downregulated, hence the immune regulatory role of *CALHM6* is altered possibly leading to an imbalance between Th17 and Tregs cells. In healthy state, the release of IFN-γ by Th17 cells may be part of an autoregulatory process, leading to an upregulation of *CALHM6* and hence controls the immune response. In patients with SSNS the expression of *CALHM6* in response to IFN-γ is insufficient to regulate the Th17 led immune response.
A study on glucocorticoid response in rheumatoid arthritis patients showed that *CALHM6* was significantly upregulated in CD4+T cells of steroid responders in comparison to non-responders [208]. Hence, we speculate that the treatment with corticosteroids leads to an upregulation of *CALHM6*, which leads to reinstatement of immune regulatory processes and therefore an improvement of the disease is observed in patients with SSNS. The differential expression of *CALHM6* could be the

key to different responses to steroid treatment, as well as changes in steroid sensitivity over time.

A key effect of glucocorticoids on the immune system is also the induction of lymphocyte apoptosis [209]. Together with the observation that another member of the CALHM superfamily is an ATP release channel and ATP is a known trigger for cell toxicity and apoptosis, one can speculate that *CALHM6* is mediating its effect on lymphocytes via inducing apoptosis. Impaired *CALHM6* function may enhance an inappropriate immune response, leading to SSNS, which can be suppressed by administrating glucocorticoids.

In summary, we speculate that the identified *locus*, rather than explaining the mechanism why these children get proteinuria, may explain important immune regulatory processes and why patients with SSNS respond to steroid treatment.

How Th17 cells act on the podocytes and subsequently their permeability is not yet understood but has been lately discussed by Saleem *et al*, who speculate that Th17 cells are the source of a possible permeability factor [210]. Saleem *et al* proposes that Th17 cells release a serine protease, which has yet to be identified, which cleaves PAR-1 on the podocytes [210]. This further leads to activation of the JNK and p38 MAPK pathways which have been shown to be involved in podocyte injury. In the downstream pathways JNK and p38 MAPK converge at paxillin leading to phosphorylation, which induces a state of podocyte hypermotility, causing the observed podocyte damage. [210] Further, Th17 cells have been implicated in different diseases as the mediators of different steroid responsiveness [211,212]. Saleem *et al* even postulates that Th17 cells are involved in mechanism leading to steroid resistance in nephrotic syndrome [210].

B cell involvement

On the other hand, there is supportive evidence that SSNS is a B cell mediated disease rather a T cell mediated disease. The strongest argument for the involvement of B cells in the disease pathogenesis is that sustained remission can be achieved after administration of Rituximab. Rituximab is a monoclonal antibody that targets

CD20, which is specifically expressed on the cell surface of B cells, except for precursor and plasma cells. The beneficial effect of Rituximab and subsequent B cell depletion has been shown in several retrospective studies and randomized controlled trials (reviewed in [213]). Additionally, calcineurin inhibitors, which are also successfully used in treatment of SSNS are also acting on B cells. Further, EBV is known to cause resistance of B cells to apoptosis and has been shown to increase the risk for developing SSNS [214,215].

However, if, and more importantly how B cells are contributing to SSNS pathogenesis is still being debated. One theory proposes that antibodies produced by plasma cells, are the mediators of the disease. However, the target of Rituximab CD20 is not expressed on plasma cells, which of course complicates explaining this pathway of action. Vivarelli *et al* recently demonstrated that the effect of Rituximab is related to a prolonged depletion of memory B cells, rather than of total CD19+ B cells, implicating the importance of memory B cells in the disease development. Delayed reconstitution of memory B cells was related to a longer time until relapse [216].

Based on this observation, we propose another theory how our candidate gene, *CALHM6,* could be involved in the disease pathogenesis (Figure 55B). *CALHM6* is not only expressed on T cells but also on B cells. Interestingly, the highest expression of *CALHM6* within B cells is in naïve B cells and memory B cells, whereas the expression is low in plasma cells. Naïve B cells and memory B cells are the subtypes of B cells which also carry the surface antigen CD20 and have been implicated to SSNS relapse. Hence, we hypothesize that *CALHM6* plays an important role in regulatory processes of memory B cells, possibly via inducing apoptosis. In patients with SSNS the expression of *CALHM6* is downregulated, and hence those immune regulatory processes in memory B cells are disturbed. Treatment of SSNS patients with Rituximab is beneficial, as this leads to depletion of those insufficiently regulated B cells.

Alternatively, the effectiveness of rituximab in INS could be mediated through other or additional pathways. Rituximab interacts with sphingomyelin phosphodiesterase acid-like 3b protein (SMPDL), which is expressed by podocytes and Th17 cells [213]. *In*

vitro experiments showed that cross-reaction of rituximab with SMPDL-3b on podocytes prevents the downregulation of SMPDL-3b and downstream deregulation of the actin cytoskeleton [217]. Also, a study in rheumatoid arthritis demonstrated that rituximab reduces Th17 cell response, thus making a logical connection among rituximab, Th17, and nephrotic syndrome [213,218,219]. Further, targeting B cells with rituximab may affect costimulatory pathways involved in T cell activation [220].

Hence, whether altered immune regulatory mechanism because of reduced *CALHM6* expression affects only B cells or T cells or perhaps both lymphocyte cell lines in parallel has to be further investigated. Both cell lines could be affected at the same time by administering Rituximab. This could explain the major beneficial role of this drug in the disease.

Role of DS in SSNS

We further speculate that our second candidate gene, *DSE,* could be involved in the pathophysiology of SSNS. Based on the findings that our lead variant increases the expression of *DSE*, we propose that in SSNS patients *DSE* is overexpressed, which could lead to an increased production of DS in the kidneys. In patients with FSGS, which arguably represents a different spectrum of the same disease, DS was found to be overexpressed in the glomerulus of patients' kidney biopsies [190]. Hence, we speculate that in the healthy state, DS is not expressed in the glomerulus of the kidneys, but in the presence of our variant, DS is expressed in the kidneys, where it possibly serves as an antigen (Figure 55C). However, the lack of immunoglobulins in kidney biopsies of MCD patients does not support an antibody-antigen mechanism in the pathogenesis of SSNS. Therefore, it remains to be elucidated what role DS plays in the pathophysiology of SSNS and how it possibly leads to structural changes in the podocyte's cytoskeleton.

Conclusion

This thesis demonstrates the complexity of SSNS and the ongoing lack of understanding of multiple components possibly involved in the disease mechanism. However, our findings support the involvement of the MHC *locus* and *CALHM6* as an immune regulator gene in the disease development. This is a basis on which future

research can be built on. Additionally, the role of *DSE* as a possible antigen opens new insights and warrants further investigation.

Limitations

Our study has several limitations.

First, only limited clinical information was available for our patients. It would have been of interest to know the age at disease onset in order to perform regression analysis of age of onset and number of risk alleles. Further, information on the course of the disease in terms of relapse rates, steroid dependency, duration of the disease would have been valuable. This would have allowed to stratify our patients into different subgroups. And also, the prescription of 2nd or 3rd line agents and response to them would have been of great interest. Hence, it would be recommended as part of the future directions of this study to collect these clinical information and perform further analysis.

Secondly, the control datasets were sourced from different collaborators or publicly available databases. Hence, all three control sets as well as the case dataset were genotyped on different platforms and needed to be combined thereafter. This initially led to systematic errors caused by strand inconsistency, however, we managed to overcome this limitation by developing the program *REMEDY*. Another downside of combining data from different genotyping platforms and SNPs is the limited number of overlapping markers. We overcame this problem by imputing the datasets to a common reference panel and therefore increased the density of markers substantially. Generally, imputation has become an accepted tool in GWAS and we could show that the same *locus* (6q22.1) was identified with the imputed, but also with the genotyped markers only dataset. This provides evidence that the significant *loci* are not secondary to imputation inaccuracy, but are true associations identified in the GWAS.

Another important limitation is that we were not able to replicate the 6q22.1 and 4q13.3 *loci* in our Asian replication cohort. This is likely due to the low power of the replication study due to a low number of cases and controls. Still, the independent identification of exactly the same *locus* at 6q22.1 by Debiec *et al* provides strong confirmatory evidence for our results at this *locus*. However, Debiec *et al* did not provide detailed information on the risk allele at 6q22.1 and we are thus unable to assess whether the

direction of effect of the risk allele is the same for both studies. Future independent replication is needed to strengthen the findings at this *locus*.

Future directions

The first step could be to collect more detailed clinical information on the patients, including the age of onset of disease, the course of the disease (no relapse versus frequent relapsing or steroid depended patients), information if the disease continued into adulthood. The age of onset of disease could be correlated to the number of risk alleles to test if there is a relationship between lower age of onset and more risk alleles. Regarding the course of the disease, the patients could be divided into subgroups (no relapse, frequently relapsing (FR) SSNS and steroid-dependent (SD) SSNS) and a repeated GWAS performed for each separately. Further, the groups could be compared against each other to investigate differences in their genomes. Another parameter for severity of disease is the prescription of 2^{nd} or 3^{rd} line agents and response to them. This information could also be used to stratify the patients in different groups. The aim would be the identification of markers of disease severity and very importantly markers that help to predict the patients' course of disease. If we were able to predict the course of the disease (no relapse, frequent relapses or steroid dependent), we could use this to guide the treatment of SSNS patients. Hence, the identification of differences in the genetic architecture between those groups would be of high clinical relevance.

Another important future direction is the replication of our findings in a larger cohort, which is powered to detect further associations in the *loci* outside the HLA-region. This could be achieved by increasing the number of cases and controls for the Asian cohort, as well as performing whole genome imputation and HLA imputation on the Asian cohort. Findings from such a study could not only provide an independent replication of our findings but could also provide more insight in the genetic similarities and differences of SSNS patients across different ethnical groups.

Several *loci* in our study are just below the significance threshold line. We speculate that by increasing the number of cases, these suggestive *loci* would possibly reach genome wide significance. Hence, repeating the European GWAS with a larger number of cases could also lead to the identification of more significant *loci*. This could be achieved by including adult MCD patients in the study. In adults minimal change disease causes 10% to 15% of primary nephrotic syndrome, after FSGS and membranous nephropathy [221]. An early kidney biopsy is essential for diagnosis and drives the therapeutic approach. That said, the histologic picture of MCD is identical in adults and children. Hence, in future studies, adults with a biopsy proven MCD, could be included.

Further, it would be very interesting to see if the genetic architecture is the same for steroid sensitive and steroid resistant (SRNS) patients. This could be addressed by performing a separate GWAS on steroid resistant patients. Another option would be to develop a genetic risk score for SSNS based on our identified lead variants and assess if these variants also increase the risk for SRNS. In case *CALHM6* is involved in steroid responsiveness, as we speculated above, we would expect that in steroid resistant patients the risk score from variants including the *CALHM6 locus* is different to the risk score not including the variants in the *CALHM6 locus*.

We have to keep in mind that genome-wide association studies signal highlight regions of associations, which often contain multiple genes and the lead variants may not be causal to the disease phenotype. To further investigate specifically the association on the 6q22.1 region several steps would be possible. First, sequencing of this *locus* in a subset of cases and controls would capture all variants in this area. Subsequently an association study for this *locus* with the sequenced data could give more information on which variant has the strongest association with the disease.

Another way forward would be to test the hypothesis that the lead variant identified in this study is affecting the expression of *CALHM6* and *DSE*. This could be done by quantifying the expression of those genes in patient samples. The expression of *CALHM6* in whole blood samples of patients with SSNS in relapse could be compared to controls. Or for *DSE*, as we think this might be overexpressed in the kidneys leading

to an altered production of DS in the glomerulus, it would be a reasonable step to measure the expression of *DSE* in cells from biopsy samples of patients with SSNS in relapse, compared to controls. In addition, the anti-DS antibodies GD3A12, LKN1 and 2A12 directed against DS protein could be used to investigate the tissue distribution of DS in patients' kidney samples compared to control samples.

Altogether, there are multiple ways forward from our current findings and the results from this thesis give ground to various new possible studies into the disease mechanism of SSNS.

References

1. Pal A, Kaskel F (2016) History of Nephrotic Syndrome and Evolution of its Treatment. Front Pediatr 4:56. doi:10.3389/fped.2016.00056

2. Hippocrates, Chadwick J, Mann WN (1950) The Medical works of Hippocrates : a new translation from the original Greek made especially for English readers. Blackwell, Oxford

3. Cameron JS (1985) Five hundred years of the nephrotic syndrome: 1484-1984. Ulster Med J 54 Suppl:S5-19

4. Christison R (1829) Observations on the Variety of Dropsy Which Depends on Diseased Kidney. Edinb Med Surg J 32 (101):262-291

5. Johnson G (1846) On the minute anatomy and pathology of Bright's disease of the kidney, and on the relation of the renal disease to those diseases of the liver, heart, and arteries, with which it is commonly associated. Med Chir Trans 29:1-24

6. Cameron JS, Hicks J (2002) The origins and development of the concept of a "nephrotic syndrome". Am J Nephrol 22 (2-3):240-247. doi:10.1159/000063768

7. Leiter L (1931) Nephrosis. Medicine 10 (2):135-242

8. Group KDIGOGW (2012) KDIGO clinical practice guideline for glomerulonephritis. Kidney international Supplement 2 (2):139-274

9. El Bakkali L, Rodrigues Pereira R, Kuik DJ, Ket JC, van Wijk JA (2011) Nephrotic syndrome in The Netherlands: a population-based cohort study and a review of the literature. Pediatr Nephrol 26 (8):1241-1246. doi:10.1007/s00467-011-1851-8

10. Banh TH, Hussain-Shamsy N, Patel V, Vasilevska-Ristovska J, Borges K, Sibbald C, . . . Parekh RS (2016) Ethnic Differences in Incidence and Outcomes of Childhood Nephrotic Syndrome. Clin J Am Soc Nephrol 11 (10):1760-1768. doi:10.2215/CJN.00380116

11. McKinney PA, Feltbower RG, Brocklebank JT, Fitzpatrick MM (2001) Time trends and ethnic patterns of childhood nephrotic syndrome in Yorkshire, UK. Pediatr Nephrol 16 (12):1040-1044. doi:10.1007/s004670100021

12. Eddy AA, Symons JM (2003) Nephrotic syndrome in childhood. Lancet 362 (9384):629-639. doi:10.1016/S0140-6736(03)14184-0

13. The primary nephrotic syndrome in children. Identification of patients with minimal change nephrotic syndrome from initial response to prednisone. A report of the International Study of Kidney Disease in Children (1981). J Pediatr 98 (4):561-564

14. Barnett H, Edelman C, Greifer I, Spitzer A, Freeman K, Arneil G (1978) Nephrotic syndrome in children: prediction of histopathology from clinical and laboratory

characteristics at time of diagnosis. A report of the International Study of Kidney Disease in Children. Kidney Int 13:159-165

15. Maas RJ, Deegens JK, Smeets B, Moeller MJ, Wetzels JF (2016) Minimal change disease and idiopathic FSGS: manifestations of the same disease. Nature Reviews Nephrology 12 (12):768

16. Primary nephrotic syndrome in children: clinical significance of histopathologic variants of minimal change and of diffuse mesangial hypercellularity. A Report of the International Study of Kidney Disease in Children (1981). Kidney Int 20 (6):765-771. doi:10.1038/ki.1981.209

17. Tarshish P, Tobin JN, Bernstein J, Edelmann C (1997) Prognostic significance of the early course of minimal change nephrotic syndrome: report of the International Study of Kidney Disease in Children. Journal of the American Society of Nephrology 8 (5):769-776

18. Holtta T, Bonthuis M, Van Stralen KJ, Bjerre A, Topaloglu R, Ozaltin F, . . . Groothoff JW (2016) Timing of renal replacement therapy does not influence survival and growth in children with congenital nephrotic syndrome caused by mutations in NPHS1: data from the ESPN/ERA-EDTA Registry. Pediatr Nephrol 31 (12):2317-2325. doi:10.1007/s00467-016-3517-z

19. Jalanko H (2009) Congenital nephrotic syndrome. Pediatr Nephrol 24 (11):2121-2128. doi:10.1007/s00467-007-0633-9

20. Kestila M, Lenkkeri U, Mannikko M, Lamerdin J, McCready P, Putaala H, . . . Tryggvason K (1998) Positionally cloned gene for a novel glomerular protein--nephrin--is mutated in congenital nephrotic syndrome. Mol Cell 1 (4):575-582

21. Hinkes BG, Mucha B, Vlangos CN, Gbadegesin R, Liu J, Hasselbacher K, . . . Arbeitsgemeinschaft fur Paediatrische Nephrologie Study G (2007) Nephrotic syndrome in the first year of life: two thirds of cases are caused by mutations in 4 genes (NPHS1, NPHS2, WT1, and LAMB2). Pediatrics 119 (4):e907-919. doi:10.1542/peds.2006-2164

22. Gbadegesin R, Hinkes BG, Hoskins BE, Vlangos CN, Heeringa SF, Liu J, . . . Hildebrandt F (2008) Mutations in PLCE1 are a major cause of isolated diffuse mesangial sclerosis (IDMS). Nephrol Dial Transplant 23 (4):1291-1297. doi:10.1093/ndt/gfm759

23. Zenker M, Aigner T, Wendler O, Tralau T, Muntefering H, Fenski R, . . . Reis A (2004) Human laminin beta2 deficiency causes congenital nephrosis with mesangial sclerosis and distinct eye abnormalities. Hum Mol Genet 13 (21):2625-2632. doi:10.1093/hmg/ddh284

24. Diomedi-Camassei F, Di Giandomenico S, Santorelli FM, Caridi G, Piemonte F, Montini G, . . . Emma F (2007) COQ2 nephropathy: a newly described inherited mitochondriopathy with primary renal involvement. J Am Soc Nephrol 18 (10):2773-2780. doi:10.1681/ASN.2006080833

25. Senggutuvan P, Cameron JS, Hartley RB, Rigden S, Chantler C, Haycock G, . . . Koffman G (1990) Recurrence of focal segmental glomerulosclerosis in transplanted kidneys: analysis of incidence and risk factors in 59 allografts. Pediatric Nephrology 4 (1):21-28

26. Andresdottir MB, Ajubi N, Croockewit S, Assmann KJ, Hibrands LB, Wetzels JF (1999) Recurrent focal glomerulosclerosis: natural course and treatment with plasma exchange. Nephrology Dialysis Transplantation 14 (11):2650-2656

27. Gallon L, Leventhal J, Skaro A, Kanwar Y, Alvarado A (2012) Resolution of recurrent focal segmental glomerulosclerosis after retransplantation. New England Journal of Medicine 366 (17):1648-1649

28. Zimmerman S (1984) Increased urinary protein excretion in the rat produced by serum from a patient with recurrent focal glomerular sclerosis after renal transplantation. Clinical nephrology 22 (1):32-38

29. Tanaka R, Yoshikawa N, Nakamura H, Ito H (1992) Infusion of peripheral blood mononuclear cell products from nephrotic children increases albuminuria in rats. Nephron 60 (1):35-41

30. Davin J-C (2016) The glomerular permeability factors in idiopathic nephrotic syndrome. Pediatric nephrology 31 (2):207-215

31. Lagrue G, Xheneumont S, Branellec A, Hirbec G, Weil B (1975) A vascular permeability factor elaborated from lymphocytes. I. Demonstration in patients with nephrotic syndrome. Biomedicine/[publiee pour l'AAICIG] 23 (1):37-40

32. Maas RJ, Deegens JK, Wetzels JF (2014) Permeability factors in idiopathic nephrotic syndrome: historical perspectives and lessons for the future. Nephrology Dialysis Transplantation 29 (12):2207-2216. doi:10.1093/ndt/gfu355

33. Yoshizawa N, Kusumi Y, Matsumoto K, Oshima S, Takeuchi A, Kawamura O, . . . Niwa H (1989) Studies of a glomerular permeability factor in patients with minimal-change nephrotic syndrome. Nephron 51 (3):370-376

34. Bakker WW, van Dael CM, Pierik LJ, van Wijk JA, Nauta J, Borghuis T, Kapojos JJ (2005) Altered activity of plasma hemopexin in patients with minimal change disease in relapse. Pediatric Nephrology 20 (10):1410-1415

35. Lennon R, Singh A, Welsh GI, Coward RJ, Satchell S, Ni L, . . . Saleem MA (2008) Hemopexin induces nephrin-dependent reorganization of the actin cytoskeleton in podocytes. Journal of the American Society of Nephrology 19 (11):2140-2149

36. Savin VJ, Sharma M, McCarthy ET, Sharma R, Reddy S, Dong J, . . . Kopp J (2008) Cardiotrophin like cytokine-1: Candidate for the focal glomerular sclerosis permeability factor. J Am Soc Nephrol 19:59A

37. McCarthy ET, Sharma M, Savin VJ (2010) Circulating permeability factors in idiopathic nephrotic syndrome and focal segmental glomerulosclerosis. Clinical Journal of the American Society of Nephrology 5 (11):2115-2121

38. Savin VJ, McCarthy ET, Sharma R, Charba D, Sharma M (2008) Galactose binds to focal segmental glomerulosclerosis permeability factor and inhibits its activity. Translational Research 151 (6):288-292

39. Sgambat K, Banks M, Moudgil A (2013) Effect of galactose on glomerular permeability and proteinuria in steroid-resistant nephrotic syndrome. Pediatric Nephrology 28 (11):2131-2135

40. Maas RJ, Deegens JK, Wetzels JF (2013) Serum suPAR in patients with FSGS: trash or treasure? Pediatric nephrology 28 (7):1041-1048

41. Wei C, Möller CC, Altintas MM, Li J, Schwarz K, Zacchigna S, . . . Reiser J (2008) Modification of kidney barrier function by the urokinase receptor. Nature Medicine 14 (1):55-63. doi:10.1038/nm1696

42. Wei C, El Hindi S, Li J, Fornoni A, Goes N, Sageshima J, . . . Saleem M (2011) Circulating urokinase receptor as a cause of focal segmental glomerulosclerosis. Nature medicine 17 (8):952

43. Vivarelli M, Massella L, Ruggiero B, Emma F (2017) Minimal Change Disease. Clinical journal of the American Society of Nephrology : CJASN 12 (2):332-345. doi:10.2215/CJN.05000516

44. Shalhoub R (1974) Pathogenesis of lipoid nephrosis: a disorder of T-cell function. The Lancet 304 (7880):556-560

45. Lin C-Y, Hsu H-C (1986) Histopathological and immunological studies in spontaneous remission of nephrotic syndrome after intercurrent measles infection. Nephron 42 (2):110-115

46. Audard V, Larousserie F, Grimbert P, Abtahi M, Sotto J, Delmer A, . . . Delarue R (2006) Minimal change nephrotic syndrome and classical Hodgkin's lymphoma: report of 21 cases and review of the literature. Kidney international 69 (12):2251-2260

47. Colucci M, Corpetti G, Emma F, Vivarelli M (2018) Immunology of idiopathic nephrotic syndrome. Pediatric Nephrology 33 (4):573-584

48. Kemper MJ, Zepf K, Klaassen I, Link A, Müller-Wiefel DE (2005) Changes of lymphocyte populations in pediatric steroid-sensitive nephrotic syndrome are more pronounced in remission than in relapse. American journal of nephrology 25 (2):132-137

49. Lama G, Luongo I, Tirino G, Borriello A, Carangio C, Salsano ME (2002) T-lymphocyte populations and cytokines in childhood nephrotic syndrome. American Journal of Kidney Diseases 39 (5):958-965

50. Lapillonne H, Leclerc A, Ulinski T, Balu L, Garnier A, Dereuddre-Bosquet N, . . . Deschênes G (2008) Stem cell mobilization in idiopathic steroid-sensitive nephrotic syndrome. Pediatric Nephrology 23 (8):1251-1256

51. Weening JJ (2004) Role of the immune system in the pathogenesis of idiopathic nephrotic syndrome. Clinical Science 107 (2):125-136

52. Araya CE, Wasserfall CH, Brusko TM, Mu W, Segal MS, Johnson RJ, Garin EH (2006) A case of unfulfilled expectations. Cytokines in idiopathic minimal lesion nephrotic syndrome. Pediatric Nephrology 21 (5):603. doi:10.1007/s00467-006-0026-5

53. Shao XS, Yang XQ, Zhao XD, Li Q, Xie YY, Wang XG, . . . Zhang W (2009) The prevalence of Th17 cells and FOXP3 regulate T cells (Treg) in children with primary nephrotic syndrome. Pediatric Nephrology 24 (9):1683-1690

54. Liu L-l, Qin Y, Cai J-f, Wang H-y, Tao J-l, Li H, . . . Li X-w (2011) Th17/Treg imbalance in adult patients with minimal change nephrotic syndrome. Clinical Immunology 139 (3):314-320. doi:https://doi.org/10.1016/j.clim.2011.02.018

55. Hashimura Y, Nozu K, Kanegane H, Miyawaki T, Hayakawa A, Yoshikawa N, . . . Matsuo M (2009) Minimal change nephrotic syndrome associated with immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome. Pediatric Nephrology 24 (6):1181-1186. doi:10.1007/s00467-009-1119-8

56. Iijima K, Sako M, Nozu K, Mori R, Tuchida N, Kamei K, . . . Ohtomo Y (2014) Rituximab for childhood-onset, complicated, frequently relapsing nephrotic syndrome or steroid-dependent nephrotic syndrome: a multicentre, double-blind, randomised, placebo-controlled trial. The Lancet 384 (9950):1273-1281

57. Leandro MJ (2013) B-cell subpopulations in humans and their differential susceptibility to depletionwith anti-CD20 monoclonal antibodies. Arthritis research & therapy 15 (1):S3

58. Dantal J, Godfrin Y, Koll R, Perretto S, Naulet J, Bouhours J-F, Soulillou J-P (1998) Antihuman immunoglobulin affinity immunoadsorption strongly decreases proteinuria in patients with relapsing nephrotic syndrome. Journal of the American Society of Nephrology 9 (9):1709-1715

59. Kemper MJ, Altrogge H, Ganschow R, Müller-Wiefel DE (2002) Serum levels of immunoglobulins and IgG subclasses in steroid sensitive nephrotic syndrome. Pediatric nephrology 17 (6):413-417

60. Delville M, Sigdel TK, Wei C, Li J, Hsieh S-C, Fornoni A, . . . Jackson A (2014) A circulating antibody panel for pretransplant prediction of FSGS recurrence after kidney transplantation. Science translational medicine 6 (256):256ra136-256ra136

61. Reiser J, Von Gersdorff G, Loos M, Oh J, Asanuma K, Giardino L, . . . Schwarz K (2004) Induction of B7-1 in podocytes is associated with nephrotic syndrome. The Journal of clinical investigation 113 (10):1390-1397

62. Xia Y, Mao J, Jin X, Wang W, Du L, Liu A (2013) Familial steroid-sensitive idiopathic nephrotic syndrome: seven cases from three families in China. Clinics 68 (5):628-631

63. Motoyama O, Sugawara H, Hatano M, Fujisawa T, Iitaka K (2009) Steroid-sensitive nephrotic syndrome in two families. Clinical and experimental nephrology 13 (2):170-173

64. Feehally J, Kendell N, Swift P, Walls J (1985) High incidence of minimal change nephrotic syndrome in Asians. Archives of disease in childhood 60 (11):1018-1020

65. Kim JS, Bellew CA, Silverstein DM, Aviles DH, Boineau FG, Vehaskari VM (2005) High incidence of initial and late steroid resistance in childhood nephrotic syndrome. Kidney international 68 (3):1275-1281

66. Saleem MA (2013) New developments in steroid-resistant nephrotic syndrome. Pediatric Nephrology 28 (5):699-709

67. Ruf RG, Fuchshuber A, Karle SM, Lemainque A, Huck K, Wienker T, . . . Hildebrandt F (2003) Identification of the first gene locus (SSNS1) for steroid-sensitive nephrotic syndrome on chromosome 2p. Journal of the American Society of Nephrology 14 (7):1897-1900

68. Lane BM, Cason R, Esezobor CI, Gbadegesin R (2019) Genetics of Childhood Steroid Sensitive Nephrotic Syndrome: An Update. Frontiers in pediatrics 7:8

69. Clark AGB, Vaughan RW, Stephens HA, Chantler C, Williams DG, Welsh KI (1990) Genes encoding the β-chains of HLA-DR7 and HLA-DQw2 define major susceptibility determinants for idiopathic nephrotic syndrome. Clinical Science 78 (4):391-397

70. Lagueruela CC, Buettner TL, Cole BR, Kissane JM, Robson AM (1990) HLA extended haplotypes in steroid-sensitive nephrotic syndrome of childhood. Kidney international 38 (1):145-150

71. Konrad M, Mytilineos J, Bouissou F, Scherer S, Gulli MP, Meissner I, . . . Schärer K (1994) HLA class II associations with idiopathic nephrotic syndrome in children. Tissue antigens 43 (5):275-280

72. KOBAYASHI T, OGAWA A, TAKAHASHI K, UCHIYAMA M (1995) HLA-DQB1 allele associates with idiopathic nephrotic syndrome in Japanese children. Pediatrics International 37 (3):293-296

73. Huang Y-Y, Lin F-J, Fu L-S, Lan J-L (2009) HLA-DR,-DQB typing of steroid-sensitive idiopathic nephrotic syndrome children in Taiwan. Nephron Clinical Practice 112 (2):c57-c64

74. Ramanathan ASK, Senguttuvan P, Chinniah R, Vijayan M, Thirunavukkarasu M, Raju K, . . . Krishnan JI (2016) Association of HLA-DR/DQ alleles and haplotypes with nephrotic syndrome. Nephrology 21 (9):745-752

75. Gbadegesin RA, Adeyemo A, Webb NJ, Greenbaum LA, Abeyagunawardena A, Thalgahagoda S, . . . Mid-West Pediatric Nephrology C (2015) HLA-DQA1 and PLCG2 Are Candidate Risk Loci for Childhood-Onset Steroid-Sensitive Nephrotic Syndrome. J Am Soc Nephrol 26 (7):1701-1710. doi:10.1681/ASN.2014030247

76. Debiec H, Dossier C, Letouze E, Gillies CE, Vivarelli M, Putler RK, . . . Ronco P (2018) Transethnic, Genome-Wide Analysis Reveals Immune-Related Risk Alleles and Phenotypic Correlates in Pediatric Steroid-Sensitive Nephrotic Syndrome. J Am Soc Nephrol 29 (7):2000-2013. doi:10.1681/ASN.2017111185

77. Jia X, Horinouchi T, Hitomi Y, Shono A, Khor SS, Omae Y, . . . and for the Research Consortium on Genetics of Childhood Idiopathic Nephrotic Syndrome in J (2018) Strong Association of the HLA-DR/DQ Locus with Childhood Steroid-Sensitive Nephrotic Syndrome in the Japanese Population. J Am Soc Nephrol. doi:10.1681/ASN.2017080859

78. Hu X, Daly M (2012) What have we learned from six years of GWAS in autoimmune diseases, and what is next? Curr Opin Immunol 24 (5):571-575. doi:10.1016/j.coi.2012.09.001

79. Sekula P, Li Y, Stanescu HC, Wuttke M, Ekici AB, Bockenhauer D, . . . Kottgen A (2016) Genetic risk variants for membranous nephropathy: extension of and association with other chronic kidney disease aetiologies. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association. doi:10.1093/ndt/gfw001

80. Stanescu HC, Arcos-Burgos M, Medlar A, Bockenhauer D, Kottgen A, Dragomirescu L, . . . Kleta R (2011) Risk HLA-DQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy. N Engl J Med 364 (7):616-626. doi:10.1056/NEJMoa1009742

81. Wuttke M, Kottgen A (2016) Insights into kidney diseases from genome-wide association studies. Nature reviews Nephrology 12 (9):549-562. doi:10.1038/nrneph.2016.107

82. Kiryluk K, Li YF, Scolari F, Sanna-Cherchi S, Choi M, Verbitsky M, . . . Gharavi AG (2014) Discovery of new risk loci for IgA nephropathy implicates genes involved in immunity against intestinal pathogens. Nature genetics 46 (11):1187-1196. doi:10.1038/ng.3118

83. Gale DP, Molyneux K, Wimbury D, Higgins P, Levine AP, Caplin B, . . . Barratt J (2017) Galactosylation of IgA1 Is Associated with Common Variation in C1GALT1. Journal of the American Society of Nephrology : JASN 28 (7):2158-2166. doi:10.1681/Asn.2016091043

84. Education N. https://www.nature.com/scitable/definition/genotype-234/.

85. Institute NHGR. https://www.genome.gov/genetics-glossary.

86. Sehn JK (2015) Chapter 9 - Insertions and Deletions (Indels). In: Kulkarni S, Pfeifer J (eds) Clinical Genomics. Academic Press, Boston, pp 129-150. doi:https://doi.org/10.1016/B978-0-12-404748-8.00009-5

87. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, . . . Chen W (2006) Global variation in copy number in the human genome. nature 444 (7118):444

88. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, . . . Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296 (5576):2225-2229. doi:10.1126/science.1069424

89. Tabor HK, Risch NJ, Myers RM (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3 (5):391-397. doi:10.1038/nrg796

90. Safan MA, Elhelbawy NG, Midan DA, Khader HF (2017) ABCB1 polymorphisms and steroid treatment in children with idiopathic nephrotic syndrome. British journal of biomedical science 74 (1):36-41

91. Moussa A, Mabrouk S, Hamdouni H, Ajmi M, Tfifha M, Omezzine A, . . . Bouslama A (2017) MDR-1 and CYP3A5 Polymorphisms in Pediatric Idiopathic Nephrotic Syndrome: Impact on Susceptibility and Response to Steroids (Preliminary Results). Clinical laboratory 63 (7):1233-1242

92. Altshuler D, Daly MJ, Lander ES (2008) Genetic Mapping in Human Disease. Science 322 (5903):881-888. doi:10.1126/science.1156409

93. Ott J, Wang J, Leal SM (2015) Genetic linkage analysis in the age of whole-genome sequencing. Nature Reviews Genetics 16:275. doi:10.1038/nrg3908

94. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17 (9):502-510

95. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, . . . Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461 (7265):747-753. doi:10.1038/nature08494

96. Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19 (3):212-219. doi:10.1016/j.gde.2009.04.010

97. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. The American Journal of Human Genetics 90 (1):7-24

98. Schaid DJ, Chen W, Larson NB (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics 19 (8):491-504

99. Illumina. Whole-Genome Genotyping Technology. https://www.illumina.com/techniques/popular-applications/genotyping/whole-genome-genotyping.html?langsel=/us/.

100. Sherry ST, Ward M, Sirotkin K (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9 (8):677-679

101. Illumina I (2006) 'TOP/BOT' strand and 'A/B' allele (Technical Note). https://www.illumina.com/documents/products/technotes/technote_topbot.pdf.

102. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84 (2):210-223. doi:10.1016/j.ajhg.2009.01.005

103. Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165 (4):2213-2233

104. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11 (7):499-511. doi:10.1038/nrg2796

105. Browning BL, Browning SR (2016) Genotype Imputation with Millions of Reference Samples. Am J Hum Genet 98 (1):116-126. doi:10.1016/j.ajhg.2015.11.020

106. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5 (6):e1000529. doi:10.1371/journal.pgen.1000529

107. Shiina T, Hosomichi K, Inoko H, Kulski JK (2009) The HLA genomic loci map: expression, interaction, diversity and disease. Journal of human genetics 54 (1):15

108. The M (1999) Complete sequence and gene map of a human major histocompatibility complex. Nature 401 (6756):921

109. Choo SY (2007) The HLA system: genetics, immunology, clinical testing, and clinical implications. Yonsei medical journal 48 (1):11-23

110. Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, . . . Rioux JD (2008) Defining the role of the MHC in autoimmunity: a review and pooled analysis. PLoS genetics 4 (4):e1000024

111. Janeway Jr CA, Travers P, Walport M, Shlomchik MJ (2001) The major histocompatibility complex and its functions. In: Immunobiology: The Immune System in Health and Disease. 5th edition. Garland Science,

112. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG (2015) The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43 (Database issue):D423-431. doi:10.1093/nar/gku1161

113. Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, . . . Trowsdale J (2010) Nomenclature for factors of the HLA system, 2010. Tissue Antigens 75 (4):291-455. doi:10.1111/j.1399-0039.2010.01466.x

114. de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, . . . Rioux JD (2006) A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 38 (10):1166-1172. doi:10.1038/ng1885

115. Leslie S, Donnelly P, McVean G (2008) A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet 82 (1):48-56. doi:10.1016/j.ajhg.2007.09.001

116. Jia X, Han B, Onengut-Gumuscu S, Chen WM, Concannon PJ, Rich SS, . . . de Bakker PI (2013) Imputing amino acid polymorphisms in human leukocyte antigens. PLoS One 8 (6):e64683. doi:10.1371/journal.pone.0064683

117. Li YR, Keating BJ (2014) Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med 6 (10):91. doi:10.1186/s13073-014-0091-5

118. Adeyemo A, Rotimi C (2010) Genetic variants associated with complex human diseases show wide variation across multiple populations. Public health genomics 13 (2):72-79

119. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26 (17):2190-2191. doi:10.1093/bioinformatics/btq340

120. An Introduction to Genetic Analysis. (2000) W. H. Freeman; . https://www.ncbi.nlm.nih.gov/books/NBK21907/.

121. Walker SH, Duncan DB (1967) Estimation of the probability of an event as a function of several independent variables. Biometrika 54 (1):167-179

122. Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. PLoS Comput Biol 8 (12):e1002822. doi:10.1371/journal.pcbi.1002822

123. Consortium IH (2005) A haplotype map of the human genome. Nature 437 (7063):1299

124. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9 (5):356-369. doi:10.1038/nrg2344

125. Fadista J, Manning AK, Florez JC, Groop L (2016) The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet 24 (8):1202-1205. doi:10.1038/ejhg.2015.269

126. Barsh GS, Copenhaver GP, Gibson G, Williams SM (2012) Guidelines for genome-wide association studies. PLoS Genet 8 (7):e1002812. doi:10.1371/journal.pgen.1002812

127. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT (2011) Basic statistical analysis in genetic case-control studies. Nat Protoc 6 (2):121-133. doi:10.1038/nprot.2010.182

128. Morris RW, Kaplan NL (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 23 (3):221-233

129. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, . . . Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81 (3):559-575. doi:10.1086/519795

130. Purcell S. v1.9 edn.,

131. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7. doi:10.1186/s13742-015-0047-8

132. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, . . . Willer CJ (2010) LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26 (18):2336-2337. doi:10.1093/bioinformatics/btq419

133. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, . . . McGee C (2014) Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science 343 (6175):1246949

134. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, . . . Knight JC (2012) Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. Nature genetics 44 (5):502

135. Wellcome Trust Case Control C (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 (7145):661-678. doi:10.1038/nature05911

136. Power C, Elliott J (2006) Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol 35 (1):34-41. doi:10.1093/ije/dyi183

137. Cheshire C (2019) Bioinformatic Investigations Into the Genetic Architecture of Renal Disorders. University College London,

138. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT (2010) Data quality control in genetic case-control association studies. Nat Protoc 5 (9):1564-1573. doi:10.1038/nprot.2010.116

139. Stevens EL, Heckenberg G, Roberson ED, Baugher JD, Downey TJ, Pevsner J (2011) Inference of relationships in population data using identity-by-descent and identity-by-state. PLoS Genet 7 (9):e1002287. doi:10.1371/journal.pgen.1002287

140. Staples J, Nickerson DA, Below JE (2013) Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. Genet Epidemiol 37 (2):136-141. doi:10.1002/gepi.21684

141. Accounting for sex in the genome (2017). Nature Medicine 23:1243. doi:10.1038/nm.4445

142. Cupples LA, Arruda HT, Benjamin EJ, D'Agostino RB, Demissie S, DeStefano AL, . . . Gottlieb DJ (2007) The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. BioMed Central,

143. Ardlie KG, Lunetta KL, Seielstad M (2002) Testing for Population Subdivision and Association in Four Case-Control Studies. The American Journal of Human Genetics 71 (2):304-311. doi:https://doi.org/10.1086/341719

144. Hong EP, Park JW (2012) Sample size and statistical power calculation in genetic association studies. Genomics Inform 10 (2):117-122. doi:10.5808/GI.2012.10.2.117

145. Andrews C (2010) The Hardy-Weinberg Principle. Nature Education Knowledge 3(10):65

146. Population Stratification (2006). In: Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1444-1445. doi:10.1007/3-540-29623-9_8284

147. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38 (8):904-909. doi:10.1038/ng1847

148. Jolliffe I (2014) Principal Component Analysis. Wiley StatsRef: Statistics Reference Online

149. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55 (4):997-1004

150. Georgiopoulos G, Evangelou E (2016) Power considerations for lambda inflation factor in meta-analyses of genome-wide association studies. Genet Res (Camb) 98:e9. doi:10.1017/S0016672316000069

151. Browning BL, Zhou Y, Browning SR (2018) A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet 103 (3):338-348. doi:10.1016/j.ajhg.2018.07.015

152. International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, . . . Stewart J (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449 (7164):851-861. doi:10.1038/nature06258

153. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, . . . Abecasis GR (2015) A global reference for human genetic variation. Nature 526 (7571):68-74. doi:10.1038/nature15393

154. IMPUTE2. https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#info_metric_details.

155. MACH. http://csg.sph.umich.edu/abecasis/mach/tour/imputation.html.

156. Sekula P, Li Y, Stanescu HC, Wuttke M, Ekici AB, Bockenhauer D, . . . Kottgen A (2017) Genetic risk variants for membranous nephropathy: extension of and association with other chronic kidney disease aetiologies. Nephrol Dial Transplant 32 (2):325-332. doi:10.1093/ndt/gfw001

157. International HapMap C (2004) Integrating ethics and science in the International HapMap Project. Nature reviews Genetics 5 (6):467-475. doi:10.1038/nrg1351

158. Wain HM, Lush M, Ducluzeau F, Povey S (2002) Genew: the human gene nomenclature database. Nucleic Acids Research 30 (1):169-171

159. Adeyemo A, Esezobor C, Solarin A, Abeyagunawardena A, Kari JA, El Desoky S, . . . Gbadegesin R (2018) HLA-DQA1 and APOL1 as Risk Loci for Childhood-Onset Steroid-Sensitive and Steroid-Resistant Nephrotic Syndrome. American journal of kidney diseases : the official journal of the National Kidney Foundation 71 (3):399-406. doi:10.1053/j.ajkd.2017.10.013

160. Gutierrez-Arcelus M, Rich SS, Raychaudhuri S (2016) Autoimmune diseases—connecting risk alleles with molecular traits of the immune system. Nature Reviews Genetics 17 (3):160

161. Carsetti R, Rosado MM, Wardmann H (2004) Peripheral development of B cells in mouse and man. Immunological reviews 197 (1):179-191

162. Hoffman W, Lakkis FG, Chalasani G (2016) B Cells, Antibodies, and More. Clinical Journal of the American Society of Nephrology 11 (1):137-154. doi:10.2215/cjn.09430915

163. Lund FE, Randall TD (2010) Effector and regulatory B cells: modulators of CD4+ T cell immunity. Nature Reviews Immunology 10 (4):236

164. Dendrou CA, Petersen J, Rossjohn J, Fugger L (2018) HLA variation and disease. Nature Reviews Immunology 18 (5):325

165. Caillat-Zucman S (2017) New insights into the understanding of MHC associations with immune-mediated disorders. Hla 89 (1):3-13

166. Malik U, Javed A, Ali A, Asghar K (2017) Structural and functional annotation of human FAM26F: A multifaceted protein having a critical role in the immune system. Gene 597:66-75. doi:10.1016/j.gene.2016.10.029

167. Dreses-Werringloer U, Lambert J-C, Vingtdeux V, Zhao H, Vais H, Siebert A, . . . Hannequin D (2008) A polymorphism in CALHM1 influences Ca2+ homeostasis, Aβ levels, and Alzheimer's disease risk. Cell 133 (7):1149-1161

168. Ebihara T, Azuma M, Oshiumi H, Kasamatsu J, Iwabuchi K, Matsumoto K, . . . Seya T (2010) Identification of a polyI:C-inducible membrane protein that participates in dendritic cell–mediated natural killer cell activation. The Journal of Experimental Medicine 207 (12):2675-2687. doi:10.1084/jem.20091573

169. Malik U, Javed A (2016) FAM26F: An Enigmatic Protein Having a Complex Role in the Immune System. Int Rev Immunol:1-11. doi:10.1080/08830185.2016.1206098

170. Javed A (2012) Gene expression pattern and functional analysis of CD8+ T cells from individuals with or without anti HIV/SIV noncytolytic activity. Niedersächsische Staats-und Universitätsbibliothek Göttingen,

171. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, . . . Pontén F (2015) Tissue-based map of the human proteome. Science 347 (6220):1260419. doi:10.1126/science.1260419

172. Chmielewski S, Olejnik A, Sikorski K, Pelisek J, Blaszczyk K, Aoqui C, . . . Bluyssen HA (2014) STAT1-dependent signal integration between IFNgamma and TLR4 in vascular cells reflect pro-atherogenic responses in human atherosclerosis. PloS one 9 (12):e113318. doi:10.1371/journal.pone.0113318

173. Grumann D, Scharf SS, Holtfreter S, Kohler C, Steil L, Engelmann S, . . . Broeker BM (2008) Immune cell activation by enterotoxin gene cluster (egc)-encoded and non-egc superantigens from Staphylococcus aureus. The Journal of Immunology 181 (7):5054-5061

174. Kim MJ, Romero R, Kim CJ, Tarca AL, Chhauy S, LaJeunesse C, . . . Kusanovic JP (2009) Villitis of unknown etiology is associated with a distinct pattern of chemokine up-regulation in the feto-maternal and placental compartments: implications for conjoint maternal allograft rejection and maternal anti-fetal graft-versus-host disease. The Journal of Immunology 182 (6):3919-3927

175. Taruno A, Vingtdeux V, Ohmoto M, Ma Z, Dvoryanchikov G, Li A, . . . Foskett JK (2013) CALHM1 ion channel mediates purinergic neurotransmission of sweet, bitter and umami tastes. Nature 495 (7440):223-226. doi:10.1038/nature11906

176. Nagy PV, Fehér T, Morga S, Matkó J (2000) Apoptosis of murine thymocytes induced by extracellular ATP is dose-and cytosolic pH-dependent. Immunology letters 72 (1):23-30

177. Trautmann A (2009) Extracellular ATP in the immune system: more than just a "danger signal". Sci Signal 2 (56):pe6. doi:10.1126/scisignal.256pe6

178. Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nature Reviews Genetics 7 (11):862-872. doi:10.1038/nrg1964

179. Battle A, Montgomery SB (2014) Determining causality and consequence of expression quantitative trait loci. Human genetics 133 (6):727-735. doi:10.1007/s00439-014-1446-0

180. Julia A, Domenech E, Chaparro M, Garcia-Sanchez V, Gomollon F, Panes J, . . . Marsal S (2014) A genome-wide association study identifies a novel locus at 6q22.1 associated with ulcerative colitis. Human molecular genetics 23 (25):6927-6934. doi:10.1093/hmg/ddu398

181. Maccarana M, Kalamajski S, Kongsgaard M, Magnusson SP, Oldberg Å, Malmström A (2009) Dermatan Sulfate Epimerase 1-Deficient Mice Have Reduced Content and Changed Distribution of Iduronic Acids in Dermatan Sulfate and an Altered Collagen Structure in Skin. Molecular and Cellular Biology 29 (20):5517-5528. doi:10.1128/mcb.00430-09

182. Malavaki C, Mizumoto S, Karamanos N, Sugahara K (2008) Recent advances in the structural study of functional chondroitin sulfate and dermatan sulfate in health and disease. Connective tissue research 49 (3-4):133-139

183. Jeansson M, Haraldsson B (2006) Morphological and functional evidence for an important role of the endothelial cell glycocalyx in the glomerular barrier. American Journal of Physiology-Renal Physiology 290 (1):F111-F116. doi:10.1152/ajprenal.00173.2005

184. Kanwar YS, Linker A, Farquhar MG (1980) Increased permeability of the glomerular basement membrane to ferritin after removal of glycosaminoglycans (heparan sulfate) by enzyme digestion. The Journal of cell biology 86 (2):688-693

185. Khalil R, Lalai RA, Wiweger MI, Avramut CM, Koster AJ, Spaink HP, . . . Baelde HJ (2019) Glomerular permeability is not affected by heparan sulfate glycosaminoglycan deficiency in zebrafish embryos. American Journal of Physiology-Renal Physiology 317 (5):F1211-F1216. doi:10.1152/ajprenal.00126.2019

186. Gerdy B, Yamada S, Kobayashi F, Purushothaman A, van de Westerlo EM, Bulten J, . . . van Kuppevelt TH (2009) Dermatan sulfate domains defined by the novel antibody GD3A12, in normal tissues and ovarian adenocarcinomas. Histochemistry and cell biology 132 (1):117-127

187. Lensen JFM, Wijnhoven TJM, Kuik LH, Versteeg EMM, Hafmans T, Rops ALWMM, . . . van Kuppevelt TH (2006) Selection and characterization of a unique phage display-derived antibody against dermatan sulfate. Matrix Biology 25 (7):457-461. doi:https://doi.org/10.1016/j.matbio.2006.06.003

188. Koshiishi I, Hasegawa T, Imanari T (2002) Quantitative and qualitative alterations of chondroitin/dermatan sulfates accompanied with development of tubulointerstitial nephritis. Archives of biochemistry and biophysics 401 (1):38-43

189. Hadad SJ, Michelacci YM, Schor N (1996) Proteoglycans and glycosaminoglycans synthesized in vitro by mesangial cells from normal and diabetic rats. Biochimica et Biophysica Acta (BBA) - General Subjects 1290 (1):18-28. doi:https://doi.org/10.1016/0304-4165(95)00183-2

190. Lensen JFM, van der Vlag J, Versteeg EMM, Wetzels JFM, van den Heuvel LPWJ, Berden JHM, . . . Rops ALWMM (2015) Differential Expression of Specific Dermatan Sulfate Domains in Renal Pathology. PloS one 10 (9):e0134946-e0134946. doi:10.1371/journal.pone.0134946

191. Fladeby C, Gupta SN, Barois N, Lorenzo PI, Simpson JC, Saatcioglu F, Bakke O (2008) Human PARM-1 is a novel mucin-like, androgen-regulated gene exhibiting proliferative effects in prostate cancer cells. International Journal of Cancer 122 (6):1229-1235. doi:10.1002/ijc.23185

192. Liu Y-R, Hu Y, Zeng Y, Li Z-X, Zhang H-B, Deng J-L, Wang G (2019) Neurexophilin and PC-esterase domain family member 4 (NXPE4) and prostate androgen-regulated mucin-like protein 1 (PARM1) as prognostic biomarkers for colorectal cancer. Journal of Cellular Biochemistry 120 (10):18041-18052. doi:10.1002/jcb.29107

193. Charfi C, Levros L-C, Edouard E, Rassart E (2013) Characterization and identification of PARM-1 as a new potential oncogene. Molecular Cancer 12 (1):84. doi:10.1186/1476-4598-12-84

194. Isodono K, Takahashi T, Imoto H, Nakanishi N, Ogata T, Asada S, . . . Matsubara H (2010) PARM-1 is an endoplasmic reticulum molecule involved in endoplasmic reticulum stress-induced apoptosis in rat cardiac myocytes. PloS one 5 (3):e9746-e9746. doi:10.1371/journal.pone.0009746

195. Dahlhoff M, Wolf E, Schneider MR (2014) The ABC of BTC: Structural properties and biological roles of betacellulin. Seminars in Cell & Developmental Biology 28:42-48. doi:https://doi.org/10.1016/j.semcdb.2014.01.002

196. Demeterco C, Beattie GM, Dib SA, Lopez AD, Hayek A (2000) A role for activin A and betacellulin in human fetal pancreatic cell differentiation and growth. The Journal of Clinical Endocrinology & Metabolism 85 (10):3892-3897

197. Yamamoto K, Miyagawa J-i, Waguri M, Sasada R, Igarashi K, Li M, . . . Yamagata K (2000) Recombinant human betacellulin promotes the neogenesis of beta-cells and ameliorates glucose intolerance in mice with diabetes induced by selective alloxan perfusion. Diabetes 49 (12):2021-2027

198. Park J-Y, Su Y-Q, Ariga M, Law E, Jin S-LC, Conti M (2004) EGF-like growth factors as mediators of LH action in the ovulatory follicle. Science 303 (5658):682-684

199. Mifune M, Ohtsu H, Suzuki H, Frank GD, Inagami T, Utsunomiya H, . . . Eguchi S (2004) Signal transduction of betacellulin in growth and migration of vascular smooth muscle cells. American Journal of Physiology-Cell Physiology 287 (3):C807-C813

200. Zanetti D, Weale ME (2018) Transethnic differences in GWAS signals: A simulation study. Annals of Human Genetics 82 (5):280-286. doi:10.1111/ahg.12251

201. Ntzani EE, Liberopoulos G, Manolio TA, Ioannidis JP (2012) Consistency of genome-wide associations across major ancestral groups. Human genetics 131 (7):1057-1071

202. Li YR, Li J, Zhao SD, Bradfield JP, Mentch FD, Maggadottir SM, . . . Hakonarson H (2015) Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. Nat Med 21 (9):1018-1027. doi:10.1038/nm.3933

203. Martinez GJ, Nurieva RI, Yang XO, Dong C (2008) Regulation and function of proinflammatory TH17 cells. Ann N Y Acad Sci 1143:188-211. doi:10.1196/annals.1443.021

204. Dong C (2008) T H 17 cells in development: an updated view of their molecular identity and genetic programming. Nature Reviews Immunology 8 (5):337

205. Matsumoto K, Kanmatsuse K (2002) Increased urinary excretion of interleukin-17 in nephrotic patients. Nephron 91 (2):243-249

206. Lavoz C, Matus YS, Orejudo M, Carpio JD, Droguett A, Egido J, . . . Ruiz-Ortega M (2019) Interleukin-17A blockade reduces albuminuria and kidney injury in an accelerated model of diabetic nephropathy. Kidney international 95 (6):1418-1432

207. Chen Y, Chauhan SK, Shao C, Omoto M, Inomata T, Dana R (2017) IFN-γ–Expressing Th17 Cells Are Required for Development of Severe Ocular Surface Autoimmunity. The Journal of Immunology 199 (3):1163-1169. doi:10.4049/jimmunol.1602144

208. Fritsch-Stork R, Silva-Cardoso S, Groot MK, Broen J, Lafeber F, Bijlsma J (2016) Expression of ERAP2 and LST1 is increased before start of therapy in rheumatoid arthritis patients with good clinical response to glucocorticoids. Clinical and experimental rheumatology 34 (4):685-689

209. Banuelos J, Lu NZ (2016) A gradient of glucocorticoid sensitivity among helper T cell cytokines. Cytokine Growth Factor Rev 31:27-35. doi:10.1016/j.cytogfr.2016.05.002

210. May CJ, Welsh GI, Chesor M, Lait PJ, Schewitz-Bowers LP, Lee RWJ, Saleem MA (2019) Human Th17 cells produce a soluble mediator that increases podocyte motility via signaling pathways that mimic PAR-1 activation. American Journal of Physiology-Renal Physiology 317 (4):F913-F921. doi:10.1152/ajprenal.00093.2019

211. McKinley L, Alcorn JF, Peterson A, Dupont RB, Kapadia S, Logar A, . . . Kolls JK (2008) TH17 cells mediate steroid-resistant airway inflammation and airway hyperresponsiveness in mice. J Immunol 181 (6):4089-4097. doi:10.4049/jimmunol.181.6.4089

212. Banuelos J, Cao Y, Shin SC, Lu NZ (2017) Immunopathology alters Th17 cell glucocorticoid sensitivity. Allergy 72 (3):331-341. doi:10.1111/all.13051

213. Ravani P, Bonanni A, Rossi R, Caridi G, Ghiggeri GM (2016) Anti-CD20 antibodies for idiopathic nephrotic syndrome in children. Clinical Journal of the American Society of Nephrology 11 (4):710-720

214. Clybouw C, Mchichi B, Mouhamad S, Auffredou MT, Bourgeade MF, Sharma S, . . . Vazquez A (2005) EBV Infection of Human B Lymphocytes Leads to Down-Regulation of Bim Expression: Relationship to Resistance to Apoptosis. The Journal of Immunology 175 (5):2968-2973. doi:10.4049/jimmunol.175.5.2968

215. Dossier C, Sellier-Leclerc A-L, Rousseau A, Michel Y, Gautheret-Dejean A, Englender M, . . . Simon T (2014) Prevalence of herpesviruses at onset of idiopathic nephrotic syndrome. Pediatric nephrology 29 (12):2325-2331

216. Colucci M, Carsetti R, Cascioli S, Casiraghi F, Perna A, Ravà L, . . . Vivarelli M (2016) B Cell Reconstitution after Rituximab Treatment in Idiopathic Nephrotic Syndrome. Journal of the American Society of Nephrology 27 (6):1811-1822. doi:10.1681/asn.2015050523

217. Fornoni A, Sageshima J, Wei C, Merscher-Gomez S, Aguillon-Prada R, Jauregui AN, . . . Chen L (2011) Rituximab targets podocytes in recurrent focal segmental glomerulosclerosis. Science translational medicine 3 (85):85ra46-85ra46

218. van de Veerdonk FL, Lauwerys B, Marijnissen RJ, Timmermans K, Di Padova F, Koenders MI, . . . Joosten LAB (2011) The anti-CD20 antibody rituximab reduces the Th17 cell response. Arthritis & Rheumatism 63 (6):1507-1516. doi:10.1002/art.30314

219. Zhu Z (2014) Rituximab Down-Regulate Th17 Cell Differentiation in Diffuse Large b-Cell Lymphoma Patients. Blood 124 (21):5479-5479. doi:10.1182/blood.V124.21.5479.5479

220. Bertelli R, Bonanni A, Di Donato A, Cioni M, Ravani P, Ghiggeri GM (2016) Regulatory T cells and minimal change nephropathy: in the midst of a complex network. Clinical & Experimental Immunology 183 (2):166-174. doi:10.1111/cei.12675

221. Nachman PH JJ, Falk RJ (2008) Primary glomerular disease. . In: BM B (ed) The Kidney, vol 8th. Saunders Elsevier, Philadelphia, pp 987–1066